Performance assessment tasks in the TIMSS study: can we learn from them?

A thesis presented in partial fulfilment of the requirements

for the degree of

Master of Educational Studies

in Mathematics at Massey University

Robyn Vivian Caygill

1998

ABSTRACT

Within the context of the reform of curricula in the education system, assessment methods and activities are also being reformed. There has been little research into the new methods and activities of assessment or of the impact these methods and activities will have on both the learning of students and the assessment of that learning. The International Association for the Evaluation of Educational Achievement (IEA) in its comparative study, the Third International Mathematics and Science Study (TIMSS), included some *hands-on* investigations, called performance assessment tasks, as some of the activities that assessed student learning. The student performances on two of the mathematics performance assessment tasks, *dice* and *packaging* were examined in this thesis, particularly in relation to student performances on some of the multiple-choice tasks also used in the study. In addition, the performances of some subgroups of the 207 standard three and 276 form three students who attempted each task were compared. The subgroupings were based on student responses to questions on gender, ethnicity, language of home, socio-economic status, and value of mathematics.

Many students were found to perform differently when their performances were compared in the multiple-choice and performance assessment questions that had similar content. Students were more likely to give no response to the performance assessment tasks than the multiple-choice tasks, particularly at the standard three level. For some, but not all, of the performance questions there was a smaller difference between the educationally disadvantaged subgroups of students and their peers, when compared with the differences between them on the multiple-choice tasks.

ACKNOWLEDGMENTS

Many people have contributed, in different ways, to the completion of this thesis. Firstly, I wish to thank the staff of the Comparative Education Research Unit within the Ministry of Education, particularly Steve May and Megan Chamberlain, for supplying the data files and answering my questions promptly. I also wish to thank the many friends who kept me sane, listened to me moaning, looked after Paul, and generally kept me going.

Thanks to Glenda Anthony who supervised and guided me through the challenges of the thesis work.

Finally, to Graham, my friend, husband, severest and kindest critic, thank you for all the accommodations and contributions you have made to a thesis which was *harder to give birth to than Paul.*

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

1 INTRODUCTION

Teaching is multi-faceted; it involves learning by the student, and assessment of that learning to aid: the student, the teacher, and the education system. The relationship between learning and assessment is therefore a symbiotic one; both impact on each other. In order to optimise learning, students need to be active participants in the education process (Khattri & Sweet, 1996; Romberg, 1992; Shepard, 1992). As we recognise students as active learners, we need to allow them to create their own responses. The assessments we use to evaluate student learning need to reflect what is important in mathematics (Wiggins, 1992). The skills and knowledge assessed, however, indicate what the **examiners** consider to be important, and this can impact upon subsequent student learning (Cooney, Badger, & Wilson, 1993; Ferrara & McTighe, 1992).

Changes to our current assessment practices are taking place within the wider context of changing theories of learning, changing technologies, and reform of curricula and mathematics education (Niss, 1993). As practitioners try to apply new theories of learning to the classroom, they realise that teaching methods need to alter (Burton, 1992; Elliot & Fuchs, 1997; Romberg, 1992; Wilson, 1992a). Within New Zealand, the focus of the mathematics curriculum has altered, and it is now based on a problem-solving approach to teaching and learning, with an emphasis on catering for individual needs (Ministry of Education, 1992). As the skills of problem-solving are important in the learning process, they should also be part of what we assess (Linn & Dunbar, 1992).

One aspect of the change in assessment practice has been the use of different activities or forms of assessment. The use of performance assessment tasks is promoted as part of an appropriate response to criticisms of traditional assessments (for example, Baker, Linn, & Herman, 1996; Griffin & Nix, 1991; Kulm, 1994; Wang & Lane, 1996). Proponents argue that performance assessment tasks assess a greater range of student learning than written fixed-choice tests and are more appropriate than traditional tests for assessing higher-order thinking (Gipps, 1994; Stiggins, 1991). Performance assessments are touted as being more aligned with: good instruction, the real world, and

outcomes important to society, than traditional assessments (Linn & Burton, 1994). Performance assessments have the ability to enhance teaching, learning, and attitudes towards mathematics (Baggett & Ehrenfeucht, 1996; Kulm, 1994). They are more appropriate for the educationally disadvantaged (Dochy & Moerkerke, 1997), and they allow process as well as product to be evaluated (Bateson, Nicol, & Schroeder, 1991; Gronlund, 1993).

Performance assessments compare less favourably with traditional methods of assessment when the costs associated with their creation, administration, and marking are examined (Harnisch, 1994; Nuttal, 1992). They can also take a lot of time to administer, and questions have been raised as to whether results of performance assessments can be valid, reliable, or generalisable (Adams & Wilson, 1996; Stephens & Sullivan, 1997). Furthermore, a change to using performance assessments, without a change in conceptions of mathematical knowledge and theories of learning, could continue to support an assessment system which tests only a narrow range of skills and facts and which requires only regurgitation rather than exposes thinking processes and student abilities (Baker, 1993; Lamon & Lesh, 1992; Wolf, 1994).

Despite these potential problems with performance assessments, the practice of teaching to the test, as well as the inclusion of high-stakes testing in education programmes, make compelling reasons to include performance assessments in large-scale testing programmes as well as classrooms.

1.1 The Third International Mathematics and Science Study

During this time of reform, the International Association for the Evaluation of Educational Achievement (IEA) began a comprehensive study of mathematics and science education called The Third International Mathematics and Science Study (TIMSS). Around 40 countries, including New Zealand, are involved in TIMSS.

New Zealand is currently participating in this comprehensive study of mathematics and science in primary and secondary schools. This large-scale comparative study is obtaining information about students' knowledge and abilities in mathematics and science as well as attitudes towards these subjects. TIMSS is also gathering information

about the cultural environments, teaching practices, curriculum goals, and institutional arrangements that are associated with student achievement. The measurement of students' knowledge and skills in this study of mathematics and science, utilises a number of types of assessment: multiple-choice questions, free-response questions, and performance tasks.

Three main reasons were given for the inclusion of performance assessment tasks in the design of TIMSS. Firstly, the maths and science curricula of most countries involved in the study have practical skills as an important component of students' learning. Secondly, it would be difficult to assess all the skills, content, and processes expected from these students with only pencil-and-paper questions. Thirdly, there is a growing body of evidence that some students are better able to demonstrate their understanding of maths and science concepts with performance assessment tasks.

This study uses some of the data collected in the TIMSS study in New Zealand. The focus is on the analysis of some of the mathematics performance tasks in relation to some of the multiple-choice questions and free-response questions and also in relation to some of the background information provided by the students.

1.2 What is assessment?

The Task Group on Assessment and Testing in the UK describe assessment as:

> *"a general term encompassing all methods customarily used to appraise the performance of an individual pupil or group. It may refer to a broad appraisal including many sources of evidence and many aspects of a pupil's knowledge, understanding, skills and attitudes; or to a particular occasion or instrument. An assessment instrument may be any method or procedure, formal or informal, for producing information about pupils: e.g. a written test paper, an interview schedule, a measurement task using equipment, a class quiz."*
>
> (cited in Griffin & Nix, 1991, p. 3)

In New Zealand assessment takes place informally and formally at an individual student level, in class groups, and in age cohort groupings. In 1994, the Ministry of Education published a handbook to assist schools to develop their assessment policy. This

document, *Assessment Policy to Practice*, has the improvement of student learning as the central purpose of assessment. In general, the purposes of assessment include formative aspects, that is assessment as a diagnostic tool to inform instruction and learning, as well as summative aspects, that is assessment used for reporting progress, certification, accountability, and monitoring (Ministry of Education, 1994; Resnick & Resnick, 1996). Looking at assessment from the view of informing instruction and learning, Burton (1992) states that the purposes of assessment are to inform the learner and thereby the carer, the teacher, the next educational stage and the employer.

Assessment is a comparative process. Both methods of assessment, norm-referenced or standards-based assessment, are used to compare students. In norm-referenced assessments students are compared with other students; in standards-based assessments they are compared with standards, that is, the assessments determine whether the student can perform some task and to what standard.

This thesis is more concerned with the activities of assessment than the methods. Assessment activities range "*from informal monitoring by observation to the more formal use of standardised tests, and includes observation, oral question and answer, conferencing or interviewing, self-assessment and self-reporting by students, peer assessment, exemplars or benchmarks, tests and checklists prepared by teachers, and standardised tests*" (Ministry of Education, 1994, p. 9). Standardised tests are tests that have been constructed, usually by a testing agency, and administered to a representative sample of students in order to establish normative data. They include instructions on administration and marking so that test conditions are similar to those used to create the normative data. Standardised tests are norm-based assessments but may also contain standards-based information. Peer assessment is assessment undertaken on behalf of a student by other students following some agreed format or process. Exemplars or benchmarks are examples of work that show the levels of achievement for students to aim towards.

Carr and Ritchie (1991) detail assessment activities they have observed being used in primary schools in New Zealand. These activities include the use of norm-based written tests, mastery-based written tests, and interview schedules. In mastery-based written

tests students are required to demonstrate in some way that they have *mastered* the skills, attitudes or understandings that are the objectives of the teaching programme.

In addition to the discussion of assessment, *Assessment Policy to Practice* defines evaluation as "*the process of making a judgment about the effectiveness of a teaching and learning programme, or about an individual's progress, based on assessment information*" (Ministry of Education. 1994, p. 10). Assessment activities are part of a continuous process of evaluation of students, lessons and programmes.

1.3 Assessment terms and activities

The term *Performance Assessment* has been used in TIMSS to refer to *hands-on* tasks requiring sustained, integrated strategies or routine practical procedures. TIMSS recognises that all achievement test items assess student performance, including the multiple-choice and free-response items in the booklets, but have restricted the term performance assessment to hands-on tasks.

Such forms of assessment have been called *performance measurement, alternative assessment* and *practical assessment*. For example, Haertel (1992) uses the term *performance measurement* with the same intention of meaning as the TIMSS use of performance assessment. He states that "*Performance measurement calls for examinees to demonstrate their capabilities directly, by creating some product or engaging in some activity*" (cited in Gipps, 1994, p. 99). Aschbacher (1991) notes that alternative assessments comprise a number of different approaches, such as "*portfolios of student work over time, exhibits or displays of knowledge and skills, open-ended questions with no single right answer, and hands-on experimentation*" (p. 276). TIMSS have chosen the approach of hands-on experimentation as their method of performance assessment, with the view that there is not necessarily a single correct approach or answer to some of the problems and questions posed.

The Educational Testing Service (1995) have defined performance assessment and authentic assessment as interchangeable terms. They use these terms to describe assessments that "*engage students in hands-on activities, often involving the creation of*

a product or the construction of a response" (p. 7). Note however that the term performance assessment has also been used by the Educational Testing Service, and in other IEA studies (Wolf, 1994) to refer to extended-response questions requiring only paper and pencil that are scored on more than just results. For an example of what is meant by an extended-response question consider the problem:

> *Factorise the equation* $x^3 + 2x^2 + 2x + 4 = 0$.
> *Show all your workings.*

These extended-response items together with short answer items have been grouped together in TIMSS under the term *free-response* items. The problem illustrated above has been referred to in TIMSS as a free-response question rather than a performance assessment task, because it can be tackled entirely by pencil-and-paper strategies and does not require sustained, integrated strategies or routine practical procedures. According to Bateson, Nicol, and Schroeder (1991) "*most measurement texts subdivide the types of items ... into ... those which require a limited response or* **short answer** *items, those which require an extended written response or* **essay** *items, and those which require hands-on physical tasks or* **performance** *items*" (p. 7, emphasis in the original). The short-answer items in TIMSS were items that required a one or two word response or a simple calculation. The extended-response items required a series of calculations or a paragraph containing more than one idea. Free-response items were included in TIMSS together with multiple choice-items in a written test. The performance assessment items were administered to students in a separate testing session.

1.4 Aim of the study

This research examines the use of performance assessment tasks in TIMSS as well as their use in mathematics education. Performance assessment tasks are increasingly being touted as the *new way* forward in assessment reform and were included in the design of TIMSS along with more traditional forms of assessment. The inclusion of the performance assessment tasks in TIMSS was justified on the grounds that: firstly, the maths and science curricula of most countries involved in the study have practical skills as an important component of students learning; secondly, it would be difficult to assess

all the skills, content, and processes expected from these students with only pencil-and-paper questions; and thirdly, there is a growing body of evidence that some students are better able to demonstrate their understanding of maths and science concepts with performance assessment tasks. The overall aim of this thesis is to examine the assumptions underlying these reasons, with a particular focus on the third reason - ie do students better demonstrate their understanding of maths concepts during performance assessment tasks than with the pencil-and-paper tasks?

Specifically, the research questions are:

(1) How do the TIMSS tasks examined relate to New Zealand's mathematics curriculum?

(2) Do some students better demonstrate their understanding of maths concepts in the performance assessment tasks than in the multiple-choice or free-response tasks examined?

(3) Which groups of students better demonstrate their understanding of maths concepts with performance assessment tasks?

(4) Are the performance assessment tasks better tasks for analysing student's conceptual understandings than other forms of tasks?

1.5 Relevance of this research

Assessment plays an important part in informing the learning process. Discovering what students know and can do, as well as how they think about mathematics are stated as important goals of assessment in *Mathematics in the New Zealand Curriculum* (Ministry of Education, 1992). Teachers can use the information provided by assessments to appropriately cater for the individual needs of their students and to assess the programme they have provided. Students can determine their areas of strengths and weaknesses, can celebrate their strengths and give greater attention to improving their weaknesses.

However, it is important that any assessment activity used is appropriate to the purposes of the assessment, to the objectives of mathematics, and to those being assessed. The mathematics curriculum asserts that "*Traditional time-constrained pencil and paper tests have proved unreliable indicators of Maori achievement in the past*" (Ministry of Education, 1992, p. 13). Performance assessment tasks, it has been

suggested, may be more appropriate for students who have been disadvantaged by traditional tests. It is important to examine the possible impacts of performance assessment tasks on students, particularly those who have been disadvantaged by traditional tests. For example, do the performance assessment tasks in TIMSS give ethnic minority students more of an opportunity to show what they know and can do?

The curriculum document also states that "*Teachers know that students are capable of solving quite difficult problems when they are free to use concrete apparatus to help them think the problems through*" (Ministry of Education, 1992, p. 13). The logical extension of this statement is that concrete apparatus should be included in assessment activities. Can the teaching and assessment of mathematics be improved by better access to or greater provision of apparatus?

It is also important to consider that performance assessment tasks requiring apparatus are expensive to administer. If we can get all the information we need about students' mathematical abilities, and allow students to demonstrate the full extent of their abilities by administering appropriately developed written tests, then it would be both a pointless exercise and an irresponsible use of taxpayers' money to continue using performance assessment tasks for national and international testing. If, however, these tasks better inform us about the knowledge and skills of students, particularly Maori students and female students, then this may encourage teachers to include more of these types of tasks in both the learning and the assessment of mathematics in the classroom.

New Zealand is participating in TIMSS because international studies in education are an important part of our efforts to improve the teaching and learning processes. The research findings may encourage improvements to the methodology currently used in institutional and national assessment programmes. A far-reaching implication could be that further reform to traditional modes of assessment is required in the whole of the education system, including higher education.

2 ASSESSMENT IN MATHEMATICS EDUCATION

In many countries, curricula, assessment practices, and philosophies underpinning education are currently being reviewed, questioned, and changed. In New Zealand there have been changes in curricula, assessment practices and philosophies, with input from both those involved in education and the wider community. Change is a revolutionary and evolutionary process; and mathematics education and assessment in New Zealand continues to change.

In this literature review the past, present and the future of assessment practices in mathematics are examined. Firstly, a brief overview of past and present assessment in mathematics in New Zealand is presented. Some of the beliefs and ideas that underpin these methods of assessment are examined. The impact of assessment on teaching and learning is discussed in order to highlight concerns with many current assessment practices. The ideas and reasons for the changes in curricula and assessment are presented. Assessment will continue to contain elements of past assessments, and thus the principles of good assessment are examined. Performance assessment is lauded as part of the solution to the problems with past assessment methods. The points in favour of performance assessments are examined along with what is needed for a successful change in assessment practice. The challenges that are faced in change and the technical issues surrounding the move to performance assessment are discussed.

2.1 Historical review of assessment

The philosophies in education world-wide are undergoing change and, along with this, assessment is also undergoing a transformation. "*Traditionally, assessment in mathematics has been focused on a quite narrow range of procedures*" (Ministry of Education, 1992, p. 18). The old order in assessment, in particular multiple-choice and standardised testing, is based on the behaviourist model of education. The behaviourist model views learning as occurring in small increments from simple to complex and thus discrete aspects can be decontextually tested (Elliot & Fuchs, 1997; Khattri & Sweet, 1996; Romberg, 1992). Assessment strategies have also been based on an absolutist

paradigm: that is mathematics is an existing body of unchallengeable truths (Burton, 1992). Under the old order in education, teaching is seen as telling, schools are seen as factories (Romberg, 1992), and mathematics education focuses on teaching techniques (Neyland, 1994).

Assessment and evaluation of student learning in New Zealand mathematics classrooms in the recent past has been based mainly on mastery or standards based pencil-and-paper tests (Forbes, 1996). Some oral presentations such as rote repetition of times-tables and observation by teachers of written and diagrammatic work have also been included in classrooms. Manipulatives and measurement devices have been used in mathematics classrooms during teaching but seldom during formal assessments.

Necessarily, at the early primary level, assessments have been more likely to be observations of students interacting with manipulatives and performing activities or oral question and answer. With the introduction of the Beginning School Mathematics (BSM) resource into primary schools, formal records of observations and results of interviews have been kept based on the objectives in this programme. Carr and Ritchie (1991) report the common use of norm-based written tests, mastery-based written tests, and interview schedules in primary schools. Standardised tests, mainly the Progressive Achievement Tests (PAT) which are multiple-choice tests produced by the New Zealand Council for Educational Research (NZCER), have also been used in classrooms to assess student learning in comparison to a norm group (Ministry of Education, 1994). NZCER also have produced diagnostic tests which are now available on the Worldwide Web. The National Educational Monitoring Project, now in its fourth year of monitoring the outcomes of primary education, uses performance tasks to assess students. Task results are recorded either by the students or on video during interviews and group tasks.

Until recently, the assessments used for certification in New Zealand secondary schools, and subsequently to inform the next educational stage and the employer, have been entirely supervised time-limited pencil-and-paper tests. While at the fifth form level, schools have been able to have their students graded for certification based on

assessments written within their school, school-based results have been moderated by an externally written supervised time-limited multiple-choice test. At sixth form there was a written supervised time-limited test called University Entrance. University Bursary and University Scholarship examinations at the seventh form level were totally externally written supervised time-limited pencil-and-paper assessments. Generally the focus of these summative assessments was on the repetition of taught facts and procedures (Begg, 1991).

More recently, the sixth form has become entirely internally assessed and thus school-based. At the seventh form level, an internally assessed component has been introduced in the University Bursary Mathematics with Statistics paper. Unit standards as a method of reporting student competencies, and thus as objectives for teaching and assessment, are still being introduced into schools at the time of writing. The influence of unit standards upon school-based assessments varies greatly from school to school and teacher to teacher at this stage (Nightingale, 1997).

2.2 The purposes of assessment

As mentioned in chapter one, the Ministry of Education (1994) states that the improvement of student learning is the central purpose of assessment. In general, the purposes of assessment include its formative aspects, as a diagnostic tool to inform instruction and learning, as well as its summative aspects, as in reporting progress, certification, accountability, and monitoring (Ministry of Education, 1994; Resnick & Resnick, 1996). Looking at assessment from the view of informing instruction and learning, Burton (1992) states that the purposes of assessment are to inform the learner and thereby: the carer, the teacher, the next educational stage, and the employer.

The different purposes of assessment are sometimes in conflict with one another and as such, any discussion on the forms of assessment must also look at the purposes for assessment (Gordon & Bonilla-Bowman, 1996). The purposes of assessment may determine the suitability of an assessment instrument. Within a classroom, both formative and summative assessments take place, whereas even if the intended purpose

of a large-scale national assessment is to improve teaching and learning, the possibilities of reporting progress and comparing results have not been ignored.

Two popular philosophies underpin the type of results required from large scale cohort testing:

a) the social Darwinist argument of survival of the most able in the education community, and

b) education is for equipping individuals for employment and economic well-being.

Philosophy (a) requires an assessment system that sorts students so the most able can be found. Philosophy (b) requires an assessment system which embodies criteria that correspond to suitable skills for employees.

Clements and Ellerton (1996, p. 144) criticise the competitive philosophy.

> *"It would be impossible to assess the extent of the psychological damage to individuals, and the collective damage to society, brought about by competitive pencil-and-paper public examinations. Yet, politicians, bureaucrats, and indeed people from all walks of life, seem to believe that a strong competitive examination is the cornerstone of economic progress and equality of educational opportunity."*

At the primary level, there have been no external whole-cohort examinations. However, recent moves calling for greater accountability, and the availability of benchmarks may change this situation. At the secondary level, with the summative assessments for certification of students, classroom instruction can become too focussed towards *the test*. Swan (1993) notes that if the purpose of assessment is confused then the result may be that assessment and record keeping become the focus in classrooms rather than student learning. The focus may also be narrowed to include only isolated techniques rather than the desired ability to use mathematics on non-routine real-world problems.

2.3 The impact of assessment on teaching and learning

It is desirable that assessment has a positive impact on instruction and learning, particularly because this is the main purpose of assessment. However, high stakes measures, that is measures that affect promotion, class placement, report grades and such, are inclined to cause teachers and students to focus on tests rather than learning (Resnick & Resnick, 1996). For example, Barnes, Clarke, and Stephens (1995) found that the teachers they questioned in New South Wales were strongly influenced by the style of assessment and the types of assessment activities assessed in the external examinations and these played a powerful role in determining which aspects of the mathematics curriculum were valued by the teachers.

It is not only high stakes tests that affect the learning process. Any assessment sends messages to students about what knowledge, skills, processes and attitudes are important for them to learn and develop (Cooney, Badger & Wilson, 1993; Ferrara & McTighe, 1992). Students' views of the nature of mathematical thinking are influenced by assessment forms (Clements & Ellerton, 1996). Written tests, in particular multiple-choice tests, have been accused of fostering a one-right answer mentality (Hambleton & Murphy, 1992) which has the potential to limit thinking and creativity rather than foster creative thinking in students.

Many authors (Kulm, 1994; Leder, 1990; Ministry of Education, 1992; Suzuki & Harnisch, 1996) wish to foster creativity and higher-order thinking in students. Leder (1990, p. 26) states that "*Mathematics classrooms need to become places where originality, independent and creative thinking and imagination are valued. ... this implies using an investigative open-ended approach whenever possible.*" If we want this type of approach to be valued in our schools, then we should be including this approach in our assessment programmes.

It is clear that problem solving is an ability that is held in high esteem both in New Zealand and in other countries (Kulm, 1994; Ministry of Education, 1992; Suzuki and Harnisch, 1996). Reasoning and communicating skills are also considered important

(Ministry of Education, 1992; Suzuki & Harnisch, 1996). Stiggins (1991, p. 267) noted that "*higher order thinking and problem-solving processes often are very complex, sometimes requiring many steps, more than one problem solver, the application of knowledge and skills from more than one school subject at a time, and the completion of some of the work outside school.*" Thus our assessment processes should reflect both the importance of these skills and the complexity of them. Performance assessment tasks have the ability to reflect the learning outcomes that are important to our society (Gipps, 1994).

If too much emphasis is placed on factual knowledge in assessments, or if tests are constructed to focus on a narrow range of objectives and isolated techniques, or if getting a final answer is what is rewarded in a marking scheme, then these signal what are held important by teachers, parents, and consequently, students (Griffin & Nix, 1991; Kulm, 1994; Linn, 1988). In contrast, written comments of a substantive nature given as feedback to students after assessments, have been found to be associated with more positive emotions, self-ratings of ability, and learning orientation, than check marks and number of questions correct (Stipek, Salmon, Givvin, Kazemi, Saxe, & MacGyvers, 1998). A narrowing range of content in tests can cause a narrowing of the taught and perceived curriculum (Carr & Ritchie, 1991). Rowntree (1987, p. 156) calls this Macnamara's Fallacy, "*making the measurable important rather than the important measurable.*" The ministerial working party on Assessment for Better Learning (Project ABLE, Ministry of Education, 1990) believed that if assessment tasks in the primary school levels are limited to assessing numeracy and literacy skills then the curriculum provided by schools could be limited.

Class level or national level assessment measures used in any testing programme should be constructed with the potential impact on teaching and learning in mind. Any test that will lead students and teachers to want to improve performances should use appropriate measures. Resnick and Resnick (1996) and Linn, Baker and Dunbar (1991) suggest that appropriate measures are direct measures, that is measures through which students demonstrate their abilities and knowledge by active performance in a way that is valued

by society. Indirect measures are classified as those for which students cannot study such as IQ tests. Indirect measures do not directly sample explicit curriculum content. If indirect measures are used, they may distort instruction by becoming the focus and the once valued content and skills will be neglected.

It is clear that some measures fall between the two categories of direct and indirect measures, so it is more important to focus on direct measures as the ideal rather than indirect measures as those to be avoided. Many writers would not classify multiple-choice questions as direct measures, even though they sample explicit curriculum content, because they do not allow active demonstration of abilities and knowledge. Resnick and Resnick (1996) state that direct measures have the power to stimulate and support a social system that works toward learning improvement and generates an understanding about the act of learning. Direct measures reconcile assessment with instruction giving authenticity to the assessment process and constructively inform the actions of the school community (Clarke, 1992).

Shepard (1992) laments the use of the standardised test for ranking teachers, schools, districts and states which has caused teachers to teach to these tests. Stake (1995 - cited in Clements & Ellerton, 1996) observed that an increased call for accountability in the United States resulted in teachers emphasising the so-called basics to the detriment of deep understanding of concepts. A greater emphasis on pencil-and-paper tests, Stake suggests would lead to "*overstandardisation, oversimplification, over-reliance on statistics, student boredom, increased numbers of dropouts, a sacrifice of personal understanding and, probably, a dimunition of diversity in intellectual development*" (Stake, 1995, p. 213 cited in Clements & Ellerton, 1996). Clement and Ellerton (1996) say that Stake's cry was ignored as administrators and practitioners strove to establish benchmarks and check the quality of their education system. The administrators and practitioners believe that pencil-and-paper tests are still needed to achieve this aim.

It is believed that in many multiple-choice tests students who have acquired test-taking skills have significant advantages (Wiley & Haertel, 1996). Haney and Madaus (1992) go further with their comments, reporting that standardised tests may corrupt the

processes of teaching and learning. These beliefs are backed up with research such as that of Gay and Thomas (1993), and Clements and Ellerton (1995).

2.4 Criticisms of assessments

Much of the literature on assessment includes statements to the effect that the rise in diversity of assessments came from a dissatisfaction with multiple-choice tests in particular, and more generally, commonly used written achievement or standardised tests (for example Haney & Madaus, 1992; Niemi, 1996; Shepard, 1992; Wiley & Haertel, 1996). Assessments are criticised from a number of different but often interrelated aspects. These aspects include: question type such as multiple-choice or fixed-response questions, test type such as standardised tests, adequacy of test construction, the way in which student answers have been analysed, and the uses of results of assessments.

Traditional written tests are criticised as providing only indirect measures of student learning (Niemi, 1996) and are seen only as indicators or correlates of other valued performances (Linn et al, 1991). They give limited information on the conceptions students bring to problems, and lack information on the strategies students use to solve problems; thus it is often difficult to diagnose precisely the difficulties students face (Heuvel-Panhuizen & Gravemeijer, 1993; Niemi, 1996; Webb & Romberg, 1992; Wilson, 1992a). Moreover, many written tests do not give students the opportunity to show all they know, at best requiring prescribed replication of information or methods given in class (Haney & Madaus, 1992; Heuvel-Panhuizen & Gravemeijer, 1993). The method of reporting results such as a single mark result or item score does not indicate the knowledge used in passing tests (Carr & Ritchie, 1991).

Much of the dissatisfaction with multiple-choice tests in particular, but also other indirect assessments, arises from the belief that they focus on, and encourage the use of, easily tested simple skills rather than higher-order thinking skills and creative endeavours (Haney & Madaus, 1992; Resnick & Resnick, 1996; Wiley & Haertel, 1996). Others criticise the focus on a narrow range of mathematical skills and procedures, because they are not considered comprehensive enough for a proper

assessment of ability to use mathematics effectively outside of the classroom (Brown, O'Gorman, & Du, 1996; Gronlund, 1993; Hambleton & Murphy, 1992; Kulm, 1994; Linn, 1988; Linn & Burton, 1994; Ministry of Education, 1992; Stiggins, 1991). Lesh and Lamon (1992, p. 6) acknowledge concerns that when most large-scale, high-stakes standardized tests are evaluated in terms of their alignment with the curriculum "*the understandings and abilities that are assessed tend to represent only narrow, obsolete, and untypical conceptions about (i) the nature of mathematics, (ii) the nature of real-life situations in which mathematics is useful in our modern world, and (iii) the nature of the knowledge and abilities that contribute to success in the preceding kinds of situations.*" Likewise, Webb and Romberg (1992) criticise multiple-choice and fixed-choice tests as being a reflection of the view of mathematics as a set of discrete skills and concepts rather than an integrated body of knowledge. The focus on a narrow range of skills in pencil-and-paper tests can cause this narrow range to be seen as the entirety of mathematics (Carr & Ritchie, 1991).

The goals of assessment have changed so that "*there is now a greater emphasis on building capabilities for work on extended complex tasks often as part of a team*" (Wiley & Haertel, 1996, p. 66). Thus assessment programmes should include the assessment of the full range of skills and procedures included in curricula, not just memorisation and reproduction of formulae and results (Bateson, Nicol, & Schroeder, 1991). Kulm (1994, p. 6) states that "*Traditional tests provide the opportunity to assess only a narrow range of student capability. Students who are visual thinkers, who are kinesthetically able, or who succeed best in a group setting are prevented from exhibiting their best performances in mathematics.*" The marking of assessment tasks often does not favour the divergent thinkers, those who think creatively in a situation rather than replicate the normally accepted methods or answers (Rowntree, 1987). We need to examine whether the assessment tasks we use, give students a chance to show what they know.

Many traditional pencil-and-paper tests have focussed on the product of the student's work (Rowntree, 1987). Rowntree notes that in marking the answer only, we may fail to recognise that a student, though finding an incorrect answer, has correctly reasoned or

set up the appropriate procedures, but made a silly slip in mechanical arithmetic. Rowntree also suggests reasons why a correct answer may be misleading. These include: recall of the answer previously worked out, copying without understanding, guessing wildly or intelligently, meaning something different, and believing something different.

In Gay and Thomas' (1993) study of 199 seventh and eighth grade students, the students were given a multiple-choice test and then were asked to write explanations of how they decided upon their answers. They were subsequently interviewed to gain further insight into their understanding of the mathematical concepts tested. Gay and Thomas found that some students who got the answer correct, were making the judgement using flawed logic and some students who got the answer wrong were able to accurately explain some of the concepts during the interview. Similarly, Clements and Ellerton (1995), in a study of 65 year-eight students in New South Wales, found that one quarter of the responses to the multiple-choice and short-answer questions they analysed inadequately assessed student understanding. That is, students had an incorrect answer to the question but were assessed as having a full understanding of the concepts required to complete the question, or had a correct answer but were assessed as having no understanding. As a result of this sort of misclassification of student response, students themselves may be misclassified when grouping or promoting students (Carr & Ritchie, 1991) particularly where a small number of items have been used on a test.

Lesh and Lamon (1992, p. 9) point out that it is more than just the question type which causes a problem in the assessment of student answers.

> *"Not using multiple-choice items may do little good if we continue to use problems and scoring procedures that impose artificial constraints by allowing only a single type and level of correct answer, because such constraints tend to trivialize the interpretation and model-refinement phases of problems where deeper and higher-order mathematical understandings tend to be emphasized."*

A pass on an examination does not necessarily indicate mastery of the content of the course or abilities in higher-order mathematical thinking. Williams (1993) found that while students in an introductory statistics class were able to obtain a pass grade in the examination, their understanding of the concepts and procedures taught during the year was poor. Just after the mid-semester test, most of the students selected for in-depth observation were not able to complete a task similar to a textbook example and two weeks prior to the final examination, they could not use the skills they had been taught in novel situations.

Many critics (Griffin & Nix, 1991; Hambleton & Murphy, 1992; Haney & Madaus, 1992; Ministry of Education, 1992), say that standardized tests or traditional written tests are biased against some kinds of students, for example: minority students, those with limited proficiency in English, females, and students from low-income families. Clements and Ellerton (1996) ask: "*Is a test **really** fair if some of those asked to do it cannot demonstrate relevant understandings because they cannot read or comprehend the written questions*?" (p. 142 - emphasis in original). They say that research into the reasons for errors in arithmetic word problems on pencil-and-paper tests has consistently indicated that well over half of the errors are associated with difficulties with reading, comprehension, and transformation.

2.5 Review of current philosophies of assessment in mathematics education

During 1992 a new mathematics curriculum, Mathematics in the New Zealand Curriculum (Ministry of Education, 1992), for primary and secondary students was introduced, and at the time of writing it is the current mathematics curriculum. This curriculum reflects current changes in philosophy of education and trends in assessment. The goals of mathematics education now embody general goals for problem-solving, communication and critical attitude rather than solely emphasising traditional skills in arithmetic. The overall goal is now the development of higher-order thinking skills rather than discrete packets of knowledge (Cole, 1990; de Lange, 1992; Ministry of Education, 1992). However, traditional arithmetical skills are still part of the

mathematics curriculum with an emphasis on learning them in the context of real-life problems (Ministry of Education, 1992). Cole (1990) states that higher-order thinking versus basic skills and facts is not a dichotomy of right and wrong but two conceptions with *right* in both. Cole concludes that we must find a framework that integrates basic skills, facts, higher-order skills, and advanced knowledge and this is the aim of Mathematics in the New Zealand Curriculum.

Mathematics is a human activity involving exploring, conjecturing, searching for patterns, communicating; and the objective of school mathematics is to ensure students grow in mathematical power (Romberg, 1992). Teaching is now seen as the creation of discourse communities; and schools are places where students produce their own knowledge, develop an in-depth understanding of a few propositions, and where collaboration and use of tools and resources to solve problems are rewarded (Romberg, 1992).

Constructivist theories about knowledge and learning are espoused as a good basis upon which to build good practice in teaching and assessment. Begg (1996, p. 5) gives the implications in constructivist theories as:

> *"3.1 knowledge is personally constructed from new experiences,*
>
> *3.2 every learner has ideas prior to learning and these affect the way that they make sense of what they are being taught,*
>
> *3.3 learning is not transmitted by linguistic communication but language is a tool to help students construct learning,*
>
> *3.4 the teachers role is to help ensure that individual constructions are modified in line with the accepted views of communities of practice (mathematicians), and*
>
> *3.5 theories about the world are provisional."*

Along with this change in philosophy, curricula worldwide have expanded in content, working forms, and activities. The expansion in content includes:

> *"... aspects of applications and modeling, cooperation with other subjects on topics of common interest, philosophy and history of mathematics, problem-oriented creativity, explorations and experiments aided by computers and informatics."*
>
> (Niss, 1993, p. 4).

These changes are in response to changes in societal needs. Stiggins (1991) noted that currently society needs more than just *information memorizers*, we also need *information managers*. We need people who can solve problems, work together collaboratively, and communicate findings and explanations (Ministry of Education, 1992).

During this time of philosophy and curricula change, ideas on how assessments should be carried out have also changed. Teachers are now encouraged to:

> *"avoid carrying out only tests which focus on a narrow range of skills such as the correct application of standard algorithms. While such skills are important, a consequence of a narrow assessment regime which isolates discrete skills or knowledge is that students tend to learn in that way. Mathematics becomes for them a set of separate skills and concepts with little obvious connection to other aspects of learning or to their world."*
>
> (Ministry of Education, 1992, p. 15).

The new order in assessment implies active student production of evidence of learning (Khattri & Sweet, 1996) and a social constructivist paradigm: that is a socially constructed body of personal and accepted knowledge (Burton, 1992; Elliot & Fuchs, 1997; Romberg, 1992; Wilson, 1992a).

2.6 A good assessment

Educators aim towards a future in which assessments improve teaching and learning. The multiple purposes of assessment provide a framework on which principles of assessment have been developed by a number of authors.

The principles of good assessments are that they should:

a) improve teaching and learning (Burton, 1992; de Lange, 1992; Linn & Baker, 1996; Ministry of Education, 1990, 1994);

b) maximise benefits for students (Ministry of Education, 1990)

c) focus on the individual, in particular upon the changes in the conceptions of an individual learner over time (Wilson, 1992b) and encourage the learner to progress (Begg, 1996; Ministry of Education, 1990, 1994);

d) enable the learner to have success (Burton, 1992);

e) be appropriate to all those being assessed (Burton, 1992; Ministry of Education, 1990, 1994);

f) enable students to demonstrate what they know (Burton, 1992; de Lange, 1992; Wiggins, 1992) and allow teachers to find out what they are thinking (Begg, 1996);

g) be engaging, motivating, and stimulate the best possible performance on the part of the student (Cooney et al, 1993; Linn & Baker, 1996; Ministry of Education, 1990; Wiggins, 1992).

h) recognise that some students give responses that do not match our expectations well, and so allow for greater flexibility in item scoring and in interpretation of the tests results (Shepard, 1992; Wilson, 1992a);

i) be able to be solved in a variety of ways (Cooney, Badger, & Wilson, 1993);

j) require complex mental processes from students (Ministry of Education, 1990; Shepard, 1992);

k) place emphasis on uncoached explanations and real student products (Shepard, 1992);

l) operationalise all goals of mathematics education (de Lange, 1992; Shepard, 1992; Wiggins, 1992);

m) be clearly related to both important objectives and real mathematics, that is mathematics as it is practised in the real world (Burton, 1992; Ferrara & McTighe, 1992; Rowntree, 1987; Wiggins, 1992; Wilson, 1992b);

n) be not unlike the style of learning but should blend with it (Burton, 1992; Ferrara & McTighe, 1992; Shepard, 1992);

o) not be driven by ability to score objectively (the term used as in objective tests) (de Lange, 1992);

p) be clear in their criteria for successful performance (Burton, 1992; Ministry of Education, 1990, 1994);

q) be practical (Ferrara & McTighe, 1992; de Lange, 1992);

r) be appropriate to the audience for the results of the assessment (Ferrara & McTighe, 1992).

These principles are consistent with the movement that calls for performance assessments to dominate, although some principles have more prominence for different sectors of the movement than for others. These principles require much education amongst those responsible for test construction before they become reflected in all assessments. Many of the principles arise out of a philosophy that students construct their own interpretation and meanings when learning (Begg, 1996; Wilson, 1992a, 1992b), that testing should not be seen as external to the student and separate from the learning process (Brown, 1992; Burton, 1992; Ferrara & McTighe, 1992; Ministry of Education, 1994; Shepard, 1992), and that tests should be constructed to reflect what is important (Wiggins, 1992).

Further to these principles, Linn and Dunbar (1992) feel that it is important that higher-order reasoning and problem-solving skills be properly represented in the definition of achievement that is embodied in all forms of assessment. Determining the students' understandings and revealing thought processes are considered integral to good assessment. Students should not only know how to complete a problem but also how

to explain it (Szetela, 1993). Dager Wilson and Chavarria (1993) suggest using problems that reveal the students' ability to record their reasoning process as well as their mathematical abilities or inabilities. Wilson (1992a, p. 216) feels that

> "... *levels of achievement might be better defined and measured not in terms of the number of facts and procedures that a student can reproduce (i.e., test score as counts of correct items) but in terms of best estimates of his or her levels of understanding of key concepts and principles underlying a learning area.*"

Webb and Romberg (1992, p. 38) state that assessment should be a

> "... *process of understanding the meaning students give to mathematics - its concepts, its procedures, and the ways problems are solved, their reasonings used, the means of communication, as well as how one comes to appreciate the mathematical enterprise.*"

According to Shepard (1992, pp. 41 & 42), assessment should

> "... *collect these types of evidence:*
>
> a) ***Coherence of knowledge.*** *Beginners' knowledge is spotty and superficial, but as learning progresses, understanding becomes integrated and structured. Thus assessment should tap the connectedness of concepts and the student's ability to access "interrelated chunks."*
>
> b) ***Principled problem solving.*** *Advanced learners ignore the surface features of a task and recognise underlying principles and patterns needed to solve the problem.*
>
> c) ***Knowledge use.*** *Complete understanding also means knowing the conditions that mediate the use of knowledge.*
>
> d) ***Automatized skills.*** *Basic component skills must be automatized so as to be integrated into total performance. (This is the only indicator that resembles today's by-rote measures of skills.)*
>
> e) ***Metacognitive or self-regulatory skills.*** *Assessment should determine whether students are able to monitor their own understanding, use strategies to make questions comprehensible, evaluate the relevance of accessible knowledge, and verify their own solutions.*"

Many authors include in their definitions of good assessment a list of types of tasks or methods that are appropriate to good assessments. Sound assessment requires multiple sources and types of data (Ferrara & McTighe, 1992; Haney & Madaus, 1992; Lambdin, 1993; Ministry of Education, 1994) such as student products (Ferrara & McTighe, 1992; Haney & Madaus, 1992), performances (Ferrara & McTighe, 1992; Haney & Madaus, 1992), class participation (Ferrara & McTighe, 1992), homework (Ferrara & McTighe, 1992), conferences (Ferrara & McTighe, 1992), and teacher judgments (Haney & Madaus, 1992). Information can be gathered through biographical data, interviews, peer evaluations, self-assessments, reference checks, grades, and expert judgment (Haney & Madaus, 1992; Ministry of Education, 1994).

Haney and Madaus (1992) recommend that the effects of any new forms of assessment on education should be monitored and that educators should be wary of reliance on any single technology of assessment. They also recommend that testing programmes faithfully reflect the objectives sought by the educators, and that assessors select different kinds and mixes of assessments as appropriate for different purposes.

The use of context in assessment is arguably of importance in making a good assessment. Galbraith (1993) notes a strong move to ensure mathematics is taught in context, but continued use of generalized and abstracted test questions. Galbraith states that

> *"support for ... national, or international testing rests upon conventional assumptions that some generalizable mathematical truths usefully will emerge and can be validly measured and compared across contexts. ... Constructivists would regard the process as invalid since mathematics in any place is viewed as contextually constructed."*
>
> (Galbraith, 1993, p. 77).

De Lange (1992) writes of the importance of context which: allows a natural and motivating access to mathematics; gives a firm hold for learning the formal operations, procedures, notations and rules; allows applicability to the real world; and allows the exercising of specific abilities in applied situations. De Lange examines good and bad

examples based on the reality of the context, showing that some problems are just dressed in context and are not real in terms of what might be encountered outside school. De Lange points out that reality has its pitfalls such as students encountering material that might be upsetting for them (such as traffic accident statistics) or overtly political (such as health funding). Clarke, Forbes, and Blithe (1993) demonstrated that particular contexts are differentially effective with and attractive to males and females.

Griffin and Griffin (1996), using conventional tests and performance assessments, studied the effects of situated cognition and cognitive style on students' learning. The definition of situated cognition, sometimes called situated learning, is that learning takes place within the context in which it is constructed. In their study, Griffin and Griffin found that the conventional-instruction group performed better than the situated-cognition group on all tests. These tests included performance assessment and written assessments both immediately after instruction and five months later. This finding contradicts the idea that situated cognition would be better for the students and produce more well-developed and useable knowledge than conventional instruction. However, a number of other factors may have influenced this result. Firstly, students in the conventional-instruction group were able to work in pairs unlike the situated-cognition group. This may have outweighed the effect of context. Secondly, the time spent on instruction was very short. Perhaps taking students out of their comfort zone, and asking them work in a different way from what they were used to, may have adversely affected the situated-cognition group.

Despite the possible drawbacks of context, teachers are encouraged to teach mathematics in context.

> *"... the curriculum statement stresses the need for mathematics to be taught and learned within the context of problems which are meaningful to students and which lead to understanding of the way mathematics is applied in the world beyond school."*
>
> (Ministry of Education, 1992, p. 5).

2.7 The trend towards the use of Performance Assessment tasks

Performance tasks are not new tasks in the teaching of mathematics. Physical activities such as geometric constructions or measuring have long been part of mathematics, with the use of manipulatives such as geoboards and base-ten blocks a more recent addition to mathematics classrooms (Hambleton & Murphy, 1992; Kulm, 1994). If these types of performances are part of learning activities then it seems fair to conclude they should be part of mathematics assessment also. Hambleton and Murphy (1992) state that the motivation of advocates of performance assessments is to bring testing more in line with classroom instruction. This idea is what Rowntree (1987, p. 162) calls educational relevance – "*does a particular assessment method seem to 'go with' the content and style of the teaching and learning experienced by our students?*"

In order to assess those skills not easily assessed with pencil-and-paper tests, teachers, educators, psychometricians, and education systems are now turning more towards performance assessment tasks (Baker, Linn, & Herman, 1996; Brown et al, 1996; Griffin and Nix, 1991; Gronlund, 1993; Hambleton & Murphy, 1992; Kulm, 1994; Ministry of Education, 1992; Wang & Lane, 1996). As an example, Rowntree (1987, p. 139) states that "*... one cannot expect to learn the truth about how a student would ... tie his own shoelaces, by asking him to write an essay or answer some multiple-choice questions about how he would do it.*"

Performance assessments are promoted as able to positively affect teaching and learning (Baggett & Ehrenfeucht, 1996; Kulm, 1994; Ministry of Education, 1992). In England and Wales, where performance assessment tasks have been used in the national testing programmes, it has been found that there has been an improvement in teaching and learning and a broadened range of skills have been appraised (Gipps, 1994; Gipps, Brown, McCallum & McAlister, 1995; Nuttal, 1992).

> *"... teachers have redirected the focus of their teaching and this has been reflected in improved national assessment results in the 'basic skills'. Greater care in planning, close observation of children and a more detailed understanding of individual progress impacting on teaching were reported ... teachers have moved*

away from intuitive approaches to assessment towards more systematic, evidence-based techniques."

(Gipps, Brown, McCallum & McAlister, 1995, p. 187).

Proponents of performance assessment assert that it can change curriculum, teacher and student behaviour, as well as the communities attitude toward schools if effectively implemented (Clarke, 1992; Khattri & Sweet, 1996). It has been noted that there is equal or greater student satisfaction with performance assessments over multiple-choice tests (Elliot and Fuchs, 1997; Nuttal, 1992). Students find performance assessments less threatening than traditional examinations (Dochy & Moerkerke, 1997; Nuttal, 1992). Performance assessments can also enhance student motivation (The National Council of Teachers of Mathematics, 1991).

Performance assessment tasks can allow us to examine not only the product, but also the process used by the student to arrive at the answer (Bateson et al 1991; Gronlund, 1993).

"The use of problem-solving activities reveals the mathematics which our pupils choose to use, rather than the mathematics they can demonstrate on request. This distinction is a crucial one, and represents a new recognition of the extent to which our past assessment strategies have mislead us by focusing on explicitly cued facts and procedures. It is in problem-solving activities that the mathematics classroom most closely approximates the real world, and in which we are most likely to see the effective outcomes of our instruction."

(Clarke, 1992, p. 164).

Performance assessment tasks are seen as giving the opportunity to assess higher-order thinking skills (Gipps, 1994; Stiggins, 1991), that is, the ability to solve problems, to reason, to think critically, to interpret and refine ideas, as well as to apply them creatively (Bateson et al, 1991; Doig & Cheeseman, 1997). Thus there is more complete information about students' misconceptions or errors available and this in turn allows teachers to make decisions about instruction (The National Council of Teachers of Mathematics, 1991). Students can also learn that mathematics is a process that enables

people to solve problems, not just a "... *bunch of rules to memorize and follow*" (The National Council of Teachers of Mathematics, 1991, p. 13).

Performance assessment tasks are perceived as fairer assessments (Dochy & Moerkerke, 1997). A particular advantage of performance assessment tasks over pencil-and-paper tasks is that they can be useful for students with poor reading skills (Griffin & Nix, 1991; Gronlund, 1993). The TIMSS study encouraged students to ask for help with reading whenever they needed it, as this was not intended to be a test of their reading skills. There is also an opportunity in performance assessments for students who are excellent mathematicians but do their thinking slowly, to display a range of their abilities not just speed and accuracy (The National Council of Teachers of Mathematics, 1991).

> "*Practical tests offer the advantages of the provision of short-term learning goals, enhanced motivation, immediate and unambiguous feedback, and a high degree of assessment validity, since the skills are assessed in practice in the manner in which they were learned and as they will be applied.*"
>
> (Clarke, 1992, p. 160).

Linn and Burton (1994, p. 5) sum up the appeal of performance assessments: *"...they have appeal as assessments that better reflect good instructional activities, are often thought to be more engaging for students, and are better reflections of the criterion performances that are of importance outside the classroom (i.e., they are said to be more authentic)."*

2.8 Changes needed

There are differing opinions in the literature as to whether reform should be driven by assessment issues or instructional issues. Newmann (1992) asserts that restructuring is needed to make schools more appropriate and motivating for students – not structuring from the system down to the students but from the educational outcomes for the students up. The system should be restructured to include outcomes based around authentic achievements.

> *"... the idea of authentic achievement requires students to engage in disciplined inquiry to produce knowledge that has value in their lives beyond simply proving their competence in school. Mastery of this sort is unlikely to be demonstrated in familiar testing and grading exercises. Instead, such mastery is more often expressed in the completion of long-term projects that result in the creation of discourse, things, and performances of interest to students, their peers, and the public at large."*
>
> (Newmann, 1992, p. 139).

Newmann argues that teachers should structure students' work to: stimulate collaboration, give increased access to knowledge beyond the teacher enhanced by greater use of technology, give increased flexibility in planning and execution of work, allow for long-term projects, and be assessed by examining production of discourse, things and performances.

Restructuring the system so that the central focus is teaching children to think means that we need to think about how we can assess their thought processes (Ferrara & McTighe, 1992). The CAM project (Santos, Drisoll, & Briars, 1993) was established to answer the questions: What do students know? How can we best assess what they know? How can assessment improve instruction? Tasks developed for the project by teachers required students to explain their thinking, provide the most appropriate representation, and justify their responses. These tasks challenged both teachers and students. "*Many students have been in classroom settings and homes that do not ask*

them to think, analyze, argue, plan, and revise when trying to solve a problem." (Santos et al, 1993, p. 222) Teachers also are in "*transition from a belief system that has at its centre the goal that students remember what is told or shown toward one where what is valued is critical thinking and the construction of new knowledge by students*" (Santos et al, 1993, p. 222). Lambdin (1993) indicates that teachers cannot be blamed for the past focus. In the main, teachers have concentrated on discrete lower-level mathematical skills, demonstrated by the teacher and remembered by the student, in response to society's demands.

Classrooms, curricula and assessments should be restructured with the view that the teacher is in the class to find out how the children think and not to tell the children how the teacher thinks (Chambers, 1993; Nagasaki & Becker, 1993). Assessment should be ongoing during instruction. Instructional activities should be modified for those assessed as having great difficulties and augmented for those assessed as having great ease doing the activities. In the Japanese classrooms studied by Nagasaki and Becker (1993), teachers had the motto: "*We should learn from our students*". Much assessment is through observation of the strategies employed by students. Mathematics education in Japan is of great interest to mathematics educators because of the Japanese students' high ranking in international studies of mathematical performance.

However, Dochy and Moerkerke (1997) question whether we should aim at assessment-driven instruction or at instruction-driven assessment in order to have a system of instruction that maximizes an individual's potential for success. They feel that the use of appropriate assessments of higher-order skills will lead to these skills being a focus of teaching. They also feel that assessment can be reformed by using items that directly relate to instruction. Thus assessment and instruction can work together to reform each other and improve education.

New Zealand educationalists (for example, Ministry of Education, 1992) would agree with the outcomes of the reforms indicated here, but would also agree that many teachers, students, parents, and members of the wider community are still in a state of transition into a belief system that values critical thinking. Indeed there are still

members of the wider community who see the solution to students' lack of abilities as a return to teaching the *basics* (Lambdin, 1993).

2.9 Challenges faced in changing assessment

As mentioned previously, it is not solely a change in assessment instruments that is required, but also a change in attitudes and belief systems amongst both the school and the wider community (Goldin, 1992; Lambdin, 1993; de Lange, 1992; Santos et al, 1993). Stephens (1992, p. vi) summarises the challenges we face from the perspective of the problems with the current use of assessment:

> *"... a conception of mathematical knowledge as heirarchical and discrete; theories of learning which support this conception; methods of assessment and uses of statistics which are consistent with this conception; assessment practices which focus exclusively on the testing of skills and facts; the social purposes for which assessment information has been used; the disjunction between the intended curriculum in mathematics and what is assessed; the failure to use assessment to inform instruction; and the separation of doing and learning mathematics from its assessment."*

One of the characteristics of a good assessment is its ability to capture information about student thought processes. This has given prominence to the use of open-ended questions. However, Lamon and Lesh (1992) caution that giving the student the opportunity to detail a solution process is not a sufficient condition for creating a good assessment item. Baker (1993) and Wolf (1994) also note that the use of open-ended or authentic tasks rather than multiple-choice questions is not sufficient to ensure the assessment of higher-order thinking skills. Open-ended tasks may not "*elicit the kinds of cognitive performances that are expected or possible for a task* (Magone, Cai, Silver, & Wang, 1994, p. 318). Furthermore, Wang and Lane (1996, p. 176) feel that

> *"... evidence is needed to ensure reliable and valid assessments of students' high-level thinking skills. In particular, evidence is needed to ensure that inferences*

made from performance assessments are equally valid for different subgroups in the population."

Open-ended questions often lead to unexpected responses and examiners must be prepared to determine whether the unexpected responses are due to the question or the student. Examiners must also be aware that in analysing student communications there are often a number of different interpretations possible (de Lange, 1992; Peressini & Bassett, 1996).

Another pitfall with tasks that require students to communicate their thought processes or understanding of a mathematical concept is that examiners must be careful not to assume lack of understanding just because students have failed to communicate that understanding (Peressini & Bassett, 1996).

Tradition also impedes the implementation of alternative assessments as *society* still wants marks and grades to provide "*numeric evidence of student progress*" (Lambdin, 1993). Attitudes towards the results of assessments needs to change with the assessments (de Lange, 1992).

As with any reform, there is the possibility that, unintentionally, the goal of improving teaching and learning may not be realised. For example, problem solving may be taught to students by a set of rules or consigned to a Friday afternoon.

> *"**No matter how sophisticated the mathematical problems we may pose** (so that they seem to **require** higher-level thinking, strategic problem solving, and/or conceptual understanding for their solution), **it is conceivable to devise — and to teach — practical, rule-based noninsightful procedures for solving them.** Under these conditions successful performance does not reflect understanding. ... In schools, such misuses of assessment are most likely to occur when educators do not themselves understand what is actually being assessed, or why. These are major dangers in developing a new assessment framework, and they must be carefully avoided."*
>
> (Goldin, 1992, p. 67, emphasis in original).

Gipps et al. (1995) found that the national introduction of performance assessments in England and Wales had some negative aspects. Not only did the teachers find the process stressful and time-consuming, but some teachers reported that when they were challenged to improve instruction, they could not find the resources, particularly funding, to implement the ideas they had. Criticisms of performance assessment tasks frequently focussed on the expense and costs in terms of time (Harnisch, 1994; Nuttal, 1992). Many of the research subjects of Brown (1992, p. 55) in the United States said that they would "*invest in more innovative programs ... if inexpensive tests could be developed for them*". Wolf (1994) mentions the logistical difficulties that have been faced with IEA performance assessments including large amounts of physical resources required per student and difficulties with standardizing and keeping equipment operational. In response to the criticisms of the costs or expense of performance assessments, Monk (1996) cautions that in order to correctly analyse the costs a combination of cost and benefit analysis must be undertaken for the tasks under study and the possible alternatives. "*Costs are measures of what must be foregone to realize some benefit. For that reason alone, costs cannot be divorced from benefits. Expenditures, in contrast, are measures of resource flows regardless of their consequences*" (p. 120).

Brown (1992) found that most people the researchers talked to, including education administrators, principals, and teachers, believed that thinking, creativity, and problem-solving activities cannot be tested at all or cannot be tested objectively. There is also a belief that performance assessment tasks make examinations less rigorous and that there is the opportunity for input from parents, peers and other knowledgeable individuals in projects and portfolios (Nuttal, 1992). Observations and interviews are difficult to pursue in a normal classroom setting and the data is difficult to record (Lambdin, 1993).

Santos, Drisoll and Briars (1993) say that students who have previously enjoyed mathematics lessons because they relied less strongly on language than other lessons now struggle to accept the change that has been imposed by the change in teaching and assessment techniques. It has also been found that students do not easily accept any

content or teaching methods that are different from their normal methods (Cooney, Badger & Wilson, 1993).

However, despite the potential problems, Lambdin (1993) suggests we can have optimism for some change for a number of reasons. Firstly, public outcry over students' apparent lack of necessary abilities is driving the groundswell for change. Secondly, technology may help overcome the problems of observing and recording data. Thirdly, the impetus and energy for change comes both from those closely involved in the teaching programmes and those influencing policy.

2.10 Technical issues arising with performance assessments

The technical issues that need to be addressed in test construction include reliability, validity and generalizability. The advantage of standardized multiple-choice testing is that quality is controlled by addressing issues of "*test construction, test piloting, reliability and validity, and reporting formats*" (Adams & Wilson, 1996, p. 41). However, it is clear that concerns with efficiency, comparability, economy, and expediency have predominated when considering what form tasks should take, rather than valuing tasks because of their close alignment to the criteria of interest.

Standardized multiple-choice tests are perceived to have greater reliability: "*That is, they are composed of standardized tasks that: (a) are the same for all students, (b) can be scored using objective criteria, and (c) are congruent with existing psychometric models*" (Adams & Wilson, 1996, p. 45). However, traditional measures of reliability have been calculated assuming tasks are homogenous and that performance variations among tasks with different skill requirements are measurement errors (Wiley & Haertel, 1996). Students who perform one performance assessment task well may not perform another task equally well (Baxter, Shavelson, Herman, Brown, & Valadez, 1993).

> *"... formula-driven statistical measures of test reliability (for example, the Kuder-Richardson 20 reliability index) can be misleading in the sense that a mathematics test deemed to have high content validity by experienced teachers and a high reliability may not generate measures of mathematical understanding which are*

validated when students' levels of understanding are investigated by other means (like, for example, Newman interviews)."

(Clements & Ellerton, 1996, p. 152).

Lesh, Lamon, Behr, and Lester (1992, p. 392) contrast the reliability and response interpretation of traditional standardized tests with performance assessments. A score on a standardized test (or item) is reliable if:

"(i) students would get the same score if they did the item again, and (ii) experts consistently give the same score. The interpretation of a response only depends on task variables and on whether the students got the correct answer."

In contrast, for a performance assessment task, they state that:

"Because the tasks (and tests) that are emphasized are those that contribute to learning as well as to assessment, students who do them repeatedly would be expected to improve. Also, because complex performances are involved, it is expected that a given expert might assign different quality ratings to a given performance, e.g., depending on the purpose of the evaluation, or depending on the weights that are assigned to various attributes or sub-components of performance."

As Knight (1997) argues, it is not true that if you once understood something then you will always be able to understand it. He also argues that it is not true that you will always be able to remember something if you once understood it. The complexity of the relationship between memory and understanding calls into question the idea that a task is reliable if students get the same score when they repeat the task. Thus, traditional measures of reliability are usually inappropriate for performance assessment tasks.

Teacher developed tasks are perceived to have greater instructional validity; "*they are closer to the actual format and content of instruction, are based on the accumulated experience of teachers concerning their own students and allow maximum adaptation to local conditions*" (Adams & Wilson, 1996, p. 45).

Content of tasks should be consistent with the best current understanding of the subject matter and should involve performances that are an important part of the curriculum (Linn & Baker, 1996). In order to have evidential validity, performance assessments need to be clearly linked to the curriculum. It is clear that multiple-choice tasks lack this validity under new societal demands for complex thinking (Khattri & Sweet, 1996; Wiley & Haertel, 1996).

Tasks must be analysed in terms of their structure to identify the task demands and the possible performances, along with how these performances may be recorded and appropriately scored (Linn et al, 1991; Wiley & Haertel, 1996). The definition of a task should include a task specification, including physical environment, timing, equipment and information to be made available, and any communications directed to the person performing the task, as well as the task goal or performance goal of that task (Wiley & Haertel, 1996).

It is also important that the consequential basis of validity be considered when comparing forms of assessment if performance-based assessments are to realise the potential that the major proponents in the movement hope for. That is, assessments should be examined to find the consequences of their use for teaching and learning (Dochy & Moerkerke, 1997; Linn et al, 1991). One consequence of the reporting of the TIMSS study in the United States is the suggestion by Feld and Shepherd (1998) that parents buy a particular piece of computer software containing skills tests which parents might use to identify problem areas or giftedness in their child. Galbraith (1993) reports criticism of IEA studies, saying that as a consequence of being involved in the studies, some countries have lost the insightful aspects of their curriculum in the search for mastery of the routine formal mathematics, thus degrading the mathematical performance in these countries.

Construction of tasks so that they are fair to the students being assessed is also an important consideration. Tasks should be appropriate and fair to minority groups and students of different abilities in, for example, reading (Dochy & Moerkerke, 1997; Linn & Baker, 1996; Ministry of Education, 1994).

There has also been much discussion of the idea of generalizability. While a single performance task is not generalizable, multiple examples of student work on multiple performance tasks may be the answer to generalizability (Stephens & Sullivan, 1997). Large numbers of performance assessment tasks can be costly. However, Parkes (1996) noted that as the number of examinees increased, less tasks were needed for assurance of reliability and generalizability.

A technical issue outside the psychometric concerns is the problem of people interpreting *scores* generated from performance assessment under the old idea of learning of knowledge as discrete packets (Khattri & Sweet, 1996). Reporting of results from performance assessments needs to be done in such a way as to ensure useful interpretation to students as well as other interested parties.

Masters and Doig (1992) suggest a way of analysing the results of assessment using three maps:

(1) The curriculum map comprising the analysis of the results of a number of students on questions of different types. It allows the analysis of the relative difficulty of the questions and hence the order information can be *taught* to students.

(2) The individual map where a particular student's responses are mapped onto the curriculum map with correct responses on the left hand side and incorrect responses on the right hand side. This allows teachers to determine where particular problems lie, or where the student lies on the continuum.

(3) The response map displays the most common answers the students give to a particular item in a test.

These maps will be most useful in informing teachers rather than students or the general public.

2.11 Performance in mathematics - assessment and the educationally disadvantaged.

In 1993 the Ministry of Education published its vision for the future entitled *Education for the 21st Century.* One of its aims was "*Equality of educational opportunity for all to reach their potential and take their full place in society*" (p. 34). The desirable outcome of this aim was "*An education system in which no group experiences unfair outcomes in terms of participation or success*" (p. 34). They stated that "*What is taught, how it is taught, how achievement is assessed, the ways in which teachers interact with children, and the ways in which classes and schools are organised all have an effect on how well particular groups of students achieve*" (p. 34).

The Ministry of Education (1994) asserts that traditional tests are not always suitable for assessing Maori students. They provide examples of alternative assessment strategies which include performance tasks. However, Clark, Forbes, and Blithe (1994, p. 177) note that, "*A limited amount of research has been done comparing assessment methods to determine those which may best suit women, Maori or ethnic minorities.*" Also Baker (1993, p. 16) warns that:

> "*Early findings suggest that the performance of educationally disadvantaged groups does not miraculously improve because the form of assessment changes. In fact, they may be even further disadvantaged, at least until the education delivery system catches up to the new challenges and goals set by the assessments.*"

The educationally disadvantaged groups commonly discussed include: females, those of minority ethnic groups, and those who are economically disadvantaged.

Much research has been concerned with differences between males and females in relation to their success in mathematics. Early studies, that indicated boys were better at mathematics than girls, prompted many to find out why this was so. However, the literature includes several warnings on accepting conclusions about gender differences. Grossman and Grossman (1993) caution that many *gender* findings do not control for other influencing factors such as race and socio-economic status, or try to generalize results from a racially or socio-economically biased sample to the whole of the

population. Also Fennema (1983) found that many studies, that reported large differences, did not control for opportunity to learn maths. When this factor was introduced, so that both girls and boys had had the same amount of time spent learning maths, then few significant differences were found.

Recent studies indicate that girls often do at least as well as boys on mathematics tests if not better (Crooks & Flockton, 1996; Flockton & Crooks, 1998; Hanna, 1994). Young Loveridge (1994, cited in Clark, 1996) found little difference between girls and boys classified as high-achievers; but for those classified as low-achievers boys were much more likely to overcome early difficulties than girls. However, at the senior level in New Zealand secondary schools, small differences in achievement still exist in favour of the boys (Baker, 1994; Forbes, 1996). While there is only a small achievement difference, there is a much higher difference in the retention of girls in mathematics courses (Baker, 1994). Clark (1996) reports that at primary school small differences in achievement are noted with girls doing better in computational areas and boys in word problems.

Fennema (1983) has investigated spatial visualisation skills as an important factor in success or failure in maths, and has become less convinced that there is a direct causal link. Tartre (1990) found that spatial skill level may identify some of the girls who have difficulty in mathematics but is less likely to identify the boys who have difficulty.

Affective factors such as beliefs, attitudes, and confidence have been used to explain differences in performance between boys and girls (Fennema, 1983). Leder, Forgasz and Solar (1996, p. 958) stated in a review of research that "... *males hold more functional beliefs about themselves as learners of mathematics than do females. ... External influences can differentially influence students' beliefs: for example, parents, the peer group, socialisation patterns, and the media.*" Females report more anxiety and less confidence toward maths than do males (Fennema, 1983). "*Compared to males, females are less likely to attribute mathematical success to ability and failure to lack of effort, and are more likely to attribute failure to lack of ability*" (Leder, Forgasz & Solar, 1996, p. 958).

Changes in teaching practices suggested by motivation research (Stipek et al., 1998) have been found to positively influence students' beliefs about their own ability, and their attitudes towards the mathematics they studied, as well as their orientation to learning.

Fennema (1983) points out that how useful students see maths and whether they see it as a male domain or not, would influence students' willingness to study maths – the most influential variable for success in maths. Leder, Forgasz and Solar (1996, p. 958) found that while "*Mathematics continues to be viewed as a male domain, more so by males than by females. ... Perceptions of the usefulness of mathematics may not be as strongly implicated in gender-differentiated mathematics learning outcomes as was earlier believed.*" In fact, Boaler (1997) found that the girls in her study related their anxiety to their experience of school mathematics rather than to the nature of mathematics as a subject. "*They want to be able to understand mathematics and they will not accept a system which merely encourages rote learning of symbols and equations that mean little or nothing to them*" (p. 303).

Grossman and Grossman (1993) reported that females prefer cooperative learning environments whereas males respond better to competitive situations. Males prefer learning environments that involve working independently, actively manipulating materials and using numbers, logic and computers.

Boaler (1997) points out that much of the research lays the blame upon the girls for their problems in and with mathematics rather than examining the reasons for their actions or potential problems with mathematical epistemology, pedagogy and practice.

> *"A realisation has emerged, ... that the tendencies of girls to avoid or underachieve at mathematics were not at all 'maladaptive' ... nor were they indicative of some internal deficiencies of their own. Rather, the girls' responses were due to their rejection of a mathematics that made little sense to them, was often taught badly and that seemed to be largely irrelevant."*
>
> (p. 286).

Magone's study (1996) examined gender differences in mathematical proficiency for students who were taught mathematics with an emphasis on: problem solving, reasoning conceptual understanding, and communicating mathematically. The performance assessment instrument consisted of thirty-five extended constructed-response tasks. Results indicated that males and females seemed to have somewhat differing response styles to the assessment; females attempted more tasks, and displayed their work and justifications more completely.

Different test-taking strategies have been suggested as possibly causing differences in performances of boys and girls. In a review of research, Meyer (1992) notes that increasing the amount of time given to complete test items did not differentially improve the performance of any of the groups studied. However, test anxiety or context could interact with time pressure to result in differences for males and females. Also a greater tendency by females to change their answers could result in a female disadvantage on timed tests. No gender differences were found with different item arrangements. It appeared that the use of multiple-choice format could result in a male advantage, because of the greater tendency of males to guess, independent of ability.

Blithe, Clark, and Forbes (1993, p. 21) found that "*there are indications that females perform better in mathematics when they are internally assessed, rather than when they sit traditional single three-hour end-of-year examinations.*" They also found that the context and order of questions was important to males and females when they chose questions during examinations. Girls tended to prefer statistics rather than mathematics, and chose questions whose context was people-oriented rather than materials-oriented. There was also a tendency for test takers, both male and female, to do the earlier questions in the paper rather than the later ones, even if they had an opportunity to do only a small proportion of them.

Wang and Lane (1996) note that differences in male and female student performances also need to be examined with respect to their thinking and reasoning. Data could be examined to determine whether "*differences between male and female students are*

related to their solution strategies, representations, mathematical explanations, mathematical errors, or all of these" (p. 197).

Taking all the research on gender effects into account we might expect a smaller difference between the achievement of males and females on the performance assessment tasks than on the multiple-choice tasks. While current research suggests that the difference favours the males on multiple-choice tasks, on the whole, we might also expect few significant differences between the males and the females.

As previously mentioned, traditional tests have not been suitable for assessing Maori students in the past. The National Education Monitoring Project found that Maori students generally scored lower than non-Maori students in a range of assessment tasks, including interviews and performance tasks (Crooks & Flockton, 1996; Flockton & Crooks, 1998).

While performance assessments are lauded as being more appropriate than standardized assessments for children from minority groups, Baker and O'Neil's (1996) experiences showed that members of minority groups in the US were suspicious of the change to these type of assessments. In particular, their fears included:

a) the belief that the ground was shifting just as minority results were catching up to the majority;
b) the use of the label alternative implying a lowering of standards; and
c) the need for greater language proficiency in order to explain or record performances.

The authors also point out that making tasks contextually appropriate requires that there be an element of choice for minority students.

Linn et al (1991, p. 17) also agree that we should not assume that a change in test format from fixed-response to performance-based will "*obviate concerns about biases against racial/ethnic minorities or that such a shift would necessarily lead to equality of performance.*" They feel that differences in familiarity, exposure, and motivation cause gaps in performance among groups and so instructional strategy and resource allocation must change with the change in assessments. Language intensive tasks may

disadvantage students whose primary language is not that of the test (Elliot & Fuchs, 1997).

Zevenbergen (1997) notes that poverty is the greatest indicator of failure in mathematics. Crooks and Flockton (1996; Flockton & Crooks, 1998) found that when they compared the performance of students in schools with a low decile rating, that is those in a low socio-economic area, these students, in general faired worse than those in medium and high socio-economic areas. Students from poor backgrounds come into school with less knowledge and fewer skills than their more affluent peers. According to Zevenbergen, teachers in schools in poor areas identify lack of resources as compounding the problem of teaching these disadvantaged students. Because of the lack of resources, there is often an emphasis on rote learning in classrooms in poor areas rather than learning in context. Thus, poor students start behind the others and have difficulty catching up.

While it is hoped that a change in assessment practice will aid educationally disadvantaged students, it still remains for research to show that it makes a difference.

2.12 Summary

In theory, performance assessments, in their many guises, are the ideal assessments. They are well aligned with good instructional practice, and thus should positively affect the teaching and learning process for all students. However, the problems with the practicalities of performance assessments mitigate against their wholesale introduction into classroom, national, and international assessment practices. Despite the difficulties with performance assessments, it is the fervent hope that the desire to improve teaching and ultimately the learning of all students will continue to drive the process of reform.

3. ASSESSMENT AND THE TIMSS STUDY

3.1. What is TIMSS?

The Third International Mathematics and Science Study (TIMSS) is an international comparative study designed to provide information about students' knowledge and abilities in mathematics and science and about the cultural environments, teaching practices, curriculum goals, and institutional arrangements that are associated with student achievement. New Zealand is participating in TIMSS because international studies in education are an important part of our efforts to improve the teaching and learning processes. Such studies provide an opportunity to examine a wide variety of teaching practices, curriculum goals and structures, and institutional arrangements, and provide researchers and educators with alternatives and new ideas in these facets of education. Through TIMSS it is hoped that all the countries involved will gain more explicit understanding of effective teaching and learning practices.

The information collection activities that form TIMSS are wide ranging and examine the many facets of mathematics and science education. Broadly speaking the activities fall into three levels, national, school, and student based activities. At the national level, the Ministry of Education have collected and collated information on the structure of, and student participation in the education system, the position of mathematics and science in the education system, and the intended mathematics and science curricula. At the school level, teachers have provided information on the way mathematics and science curricula are implemented in their school, and their opinions about and attitudes toward these subjects. At the student level, students have been assessed on their knowledge and skills in mathematics and science using multiple-choice questions, free-response questions, and performance tasks. They have also responded to a background questionnaire which provides information on their opinions and attitudes to mathematics and science.

The first phase of testing for TIMSS took place in 1994. Details of the sampling, tasks, materials, and some reliability information are given in a number of published and

unpublished documents (for example, Garden, 1997) and are presented here in summary form. This current study focuses on the performance assessment component of TIMSS but also uses information gathered in the main study, hence summary details of the main study are also presented.

3.2. Sampling

In October 1994, 421 New Zealand schools, 873 teachers, and approximately 12300 students participated in the Third International Mathematics and Science Study. A sub-sample of the students took part in the performance assessment component of TIMSS. The performance assessment sub-sample of students is the sample group analysed in this thesis.

The TIMSS study was designed to look at three different groups of students, nine-year-olds, thirteen-year-olds and final-year secondary students along with their teachers and schools. The population definitions used in the study called for selection of students in adjacent class levels at the nine-year-old and thirteen-year-old level. Specifically, population one students were *all students enrolled in the two adjacent grades that contained the largest proportion of students in the age nine cohort (at the time of testing*); standard two and standard three were the two class levels selected in New Zealand. Similarly, population two students were *all students enrolled in the two adjacent grades that contained the largest proportion of students in the age thirteen cohort (at the time of testing*); form two and form three were the two class levels selected. Population three students were *all students enrolled in their final year of schooling;* form 6 and 7 students were selected in New Zealand. The overall sample of 421 schools were randomly selected from a stratified sampling frame with probability proportional to the size of the population cohort within that school.

The performance assessment component of TIMSS required a sub-sample of the selected nine-year-old students and thirteen-year-old students to attempt some hands-on mathematics and science tasks. For the performance assessment, this sub-sample comprised students from 100 schools, 50 primary schools and 50 secondary schools.

The students selected were only from the upper class level in each population, that is standard three students from population one and form three students from population two. Population three students did not take part in the performance assessment.

Performance assessment tasks were administered to nine students at a time. One or two clusters of nine students were selected from each class list, depending on the class size. As a result of the cluster selection process, 1449 students were selected — 621 students from standard three and 828 students from form three. Replacement students were available at all schools for any absentees so that these numbers also represented the achieved sample. Further details on the sampling process may be found in the Technical Report of the TIMSS Study in New Zealand (May & Udy, in press).

According to Garden (1997), the achieved performance assessment sample did not differ significantly in mean mathematics score from the other students who took part in TIMSS. Thus it can be assumed that the performance assessment sub-sample is representative of the whole of the TIMSS sample.

3.3. Administration

The performance assessment was conducted in each school by means of a *circus* administration system in which the tasks are set up at nine individual stations arranged around the room in a circuit. Each student visited three stations during the administration. At each station, students attempted either a combination of two short (15 minute) routine tasks, one science and one mathematics task, or one long (30 minute) investigation. Thus the total testing time for the performance assessment was 90 minutes.

The 12 tasks administered were presented at nine different stations according to the design in Table 3.1. Each station required 30 minutes to complete. During the administration, the stations were designated by signs: Station A, Station B, etc.

Table 3.1 Assignment of Tasks to Stations

Station	Task Name(s)	Content
A (2 tasks)	Pulse Dice	Science – Routine Task Maths – Routine Task
B (2 tasks)	Magnets Calculators	Science – Routine Task Maths – Routine Task
C	Shadows	Science and Maths – Investigation
D (2 tasks)	Batteries Folding and Cutting	Science – Routine Task Maths – Routine Task
E	Rubber Band	Science – Investigation
F	Packaging	Maths – Investigation
G	Solutions (Pop 2) Containers (Pop 1)	Science – Investigation
H	Around the Bend	Maths – Investigation
I	Plasticine	Science and Maths – Investigation

As each student visited only three of the nine stations, a scheme was devised for assigning the students to stations. In order to be able to investigate inter-task relationships, it was necessary to assign students to different combinations of stations. Using eighteen different combinations of stations, that is two rotation plans of nine combinations of stations each, each of the stations was paired with each other station at least once.

During the administration, each student was provided with a routing card containing the student's name and identification number along with the stations to go to and the order to go to the stations. At each station was a booklet, or booklets and some equipment. The booklet contained both instructions and spaces in which to write answers. Students were provided with labels to attach to their booklets which identified the student by

number, but not name. The administrator checked that the student went to the correct stations, labelled each booklet and attempted each task.

The tasks were administered by teachers who had been selected from a neighbouring school and were unfamiliar to the students. In order to standardize task administration, the teachers were trained in task administration and were provided with a manual of instructions and a sample introductory script. Students were told that the teachers were able to help them with the reading of the task instructions but that the teachers could not explain how they should work out the problems. Teachers were asked not to give additional help except in clearly specified circumstances, in which case the provision of the additional help was to be recorded on the student's response sheet. Task equipment was provided by the national centre to the teachers who took the equipment to each school.

Prior to training, the teachers had very little or no knowledge of the study. They responded to an advertisement published in the New Zealand Education Gazette, and were selected based on their geographical location and teaching experience. At the training meeting teachers had the opportunity to become familiar with the procedures and the tasks. They were also able to check the equipment supplied to ensure it was complete and in good working order prior to their first administration.

As part of the larger TIMSS study, all students completed a questionnaire giving information about themselves, their homes, activities, attitudes and beliefs with a primary focus on mathematics and science. The questionnaire was presented in a booklet and administered to the students by their class teacher. The class teacher was instructed to help the students complete the questionnaire as fully as possible.

All the TIMSS students also completed a written test containing multiple-choice and free-response mathematics and science questions. Because of the large number of items needed for curriculum coverage, and the requirement to keep testing time relatively short, eight booklets of written test items were assembled, but only one of these booklets was administered to each student on the basis of a rotation scheme similar to

that used for the performance assessment. The written test was administered to students by the class teacher usually with a rest break part way through the administration. The teacher attended a training session in administration procedures at which they received an administration manual containing a script to be read during the test and questionnaire administration. The tests were conducted under normal test conditions of silence with the test administrator allowed to answer questions on the instructions to be followed but not on how to solve or answer any of the test questions.

Further details on the administration of the written tests and questionnaires may be found in the reports on New Zealand performance in TIMSS (Garden, 1996; Garden, 1997).

3.4. Tasks and Materials

Twelve performance assessment tasks were selected after a pilot study conducted earlier in the year. Five of these tasks were primarily mathematics tasks, two were combined mathematics and science tasks, and the remaining five tasks were science tasks. Two of the mathematics tasks, named dice and packaging, are examined in this thesis.

All of the mathematics tasks were administered to both populations one and two, although in some cases the directions and response sheets for the tasks differed between the two populations, with population two students being asked additional or more difficult questions than population one students. Twenty-one of the individual items (sub-questions of the tasks) were identical for populations one and two. Table 3.2 compares the population one and two tasks for similarity in task content.

Table 3.2 Comparison of population one and two tasks for identical items.

Task name	Number of items identical for populations 1 and 2	Number of population 1 only items	Number of population 2 only items
Dice	6	0	0
Calculator	5	0	2
Folding and Cutting	3	0	1
Around the Bend	0	4	6
Packaging	3	0	0
Shadows	0	7	6
Plasticine	4	4	4

The performance assessment tasks were designed to focus on the types of performances not covered in the written tests such as using equipment and performing investigations or experiments. The performances expected are summarised in Table 3.3. Expectations that relate only to the questions for the form three pupils are indicated by the label "(pop. 2 only)".

Table 3.3 Summary of Expected Performances for the Mathematics Performance Assessment Tasks

Task Name	Expected Performances
Dice	Performing routine[1]* and complex[2] procedures; conjecturing; using equipment; describing and discussing.
Calculator	Using equipment; performing routine and complex procedures; predicting; developing and describing strategy; describing and discussing; recalling mathematical objects and properties (pop. 2 only); recognizing mathematical equivalents (pop. 2 only).
Folding and Cutting	Problem solving; performing complex procedures; using equipment; predicting (pop. 2 only).
Around the Bend	Problem solving; performing routine and complex procedures; using equipment; recognising equivalents; conjecturing (pop. 2 only); generalizing (pop. 2 only).
Packaging	Problem solving; performing routine and complex procedures; using equipment.
Shadows	Performing routine and complex procedures; problem solving; conjecturing; generalizing; using equipment; describing and discussing; recalling mathematical objects and properties.
Plasticine	Performing routine procedures; describing and discussing; problem solving; developing and discussing strategy; using equipment.

[1] Routine procedure is further detailed in the framework as: counting and routine computations; graphing; transforming one mathematical object into another by some formal process; measuring.

[2] Complex procedure is detailed as: estimating to arrive at an approximate answer to a question; collecting, organizing, displaying, or otherwise using quantitative data; comparing and contrasting two mathematical objects, quantities, representations, etc.; classifying objects or working with the properties underlying a classification system.

Each task was presented in a separate booklet. The booklet contained a list of equipment, a task summary, instructions to follow, and space in which to record answers and results. Students were able to ask the teacher for help with reading. However, it was assumed that most or all students could read with some proficiency because administrators could only be available to one student at a time to provide help with reading.

3.5. Coding

Coding rubrics for the student results were developed internationally and representatives from each country were trained in the use of the rubrics. The scheme used two digits to code each response. The first digit of the code indicated the degree of correctness. Where the answer was correct, the second digit indicated the method or approach which was used to arrive at the answer, and where the answer was incorrect, it indicated the misconception or error that was made. The maximum correct value ranged from 1 to 3 depending on the complexity of the item. If the response was incorrect the first digit was a 7 and if there was no response the first digit was a 9. These values of 7 and 9 were considered to be equivalent to a score of zero. The second digit could take a value from 0 to 9. The descriptive terms associated with the second value were developed from sample student responses in the pilot study and reflected the different methods students had used. Each method used by at least five percent of the students was assigned a value from 0 to 6. The value of 9 was used for any other response that did not fit into the descriptive terms listed above it. A special code of 76 was assigned to responses that repeated information given in the question or introductory material without answering the question. Table 3.4 shows a typical code allocation for a question with maximum correctness score 2.

Table 3.4 Typical code allocation for a question with maximum correctness score 2

Code	Response
20	correct response, answer category or method number 1
21	correct response, answer category or method number 2
22	correct response, answer category or method number 3
29	correct response, some other method used
10	partially correct response, answer category or method number 1
11	partially correct response, answer category or method number 2
12	partially correct response, answer category or method number 3
19	some other partially correct response
70	incorrect response, common misconception or error number 1
71	incorrect response, common misconception or error number 2
76	incorrect response, information in stem repeated
79	incorrect response, some other error made
90	crossed out or erased, illegible, or impossible to interpret
99	blank

3.6. Validity and Reliability

In April 1994 twenty-two tasks were trialed for potential inclusion in the performance assessment. At this stage, the sample for the main TIMSS study had been chosen so schools not included in the sample were selected as trial schools for the performance assessment. All procedures were trialed including coding of student answers. From the information on how the administration proceeded in each country during the pilot, and recommendations on modification and deletion of tasks, twelve tasks were selected and testing procedures modified.

Reliability tests on the coding of student answers have already been conducted at the international centre. As reported by Garden (1997) no inter-rater reliability scores could be obtained from New Zealand data because double coding procedures were not followed in New Zealand. While there were two people coding responses, they never worked on the same task so that each entire task was coded by one person. New Zealand was the first country to test students and code their responses and as such standard procedures were still being developed as the coding proceeded. New Zealand students' responses along with the coding of these responses were used to train coders in other countries in the use of the coding manual and no inconsistencies or problem codes were reported back to the New Zealand coders from this training.

4. METHOD

This thesis is a secondary analysis of some of the New Zealand data collected in the TIMSS study. Thus this methodology section does not contain information explaining and justifying the primary data collection techniques, although some of these have been explained in the previous chapter. My most recent role in the TIMSS study was to coordinate the performance assessment component in New Zealand, and my main focus in this analysis is examining the student performance on the performance assessment tasks.

4.1. Task Selection

A large amount of data was collected by TIMSS, but only a selection of tasks and background variables were examined in this thesis. Two performance assessment tasks, *Dice* and *Packaging,* are examined in detail, and were selected to represent the two types of performance tasks used in TIMSS. Dice was a routine task and was paired with the Pulse task at station A. It was expected that students would spend about 15 minutes doing the dice task. The packaging task was an investigation task at station F and was expected to take the students about 30 minutes to complete. These were the only two tasks which had all questions identical (see Table 3.3) for both standard three and form three students.

As part of the analysis, the performance assessment tasks were compared and contrasted with the multiple-choice and free-response tasks used in TIMSS. Multiple-choice and free-response tasks, that covered the same mathematics content area as the two performance assessment tasks, were selected and reviewed. The selection was based on the content description provided with each task by the International Coding Committee. Some of the multiple-choice questions are not yet available for publication and where these presented information that was different from other multiple-choice tasks they have been reported in such a way as to give the general idea without reporting specifics.

4.2. Data Analysis

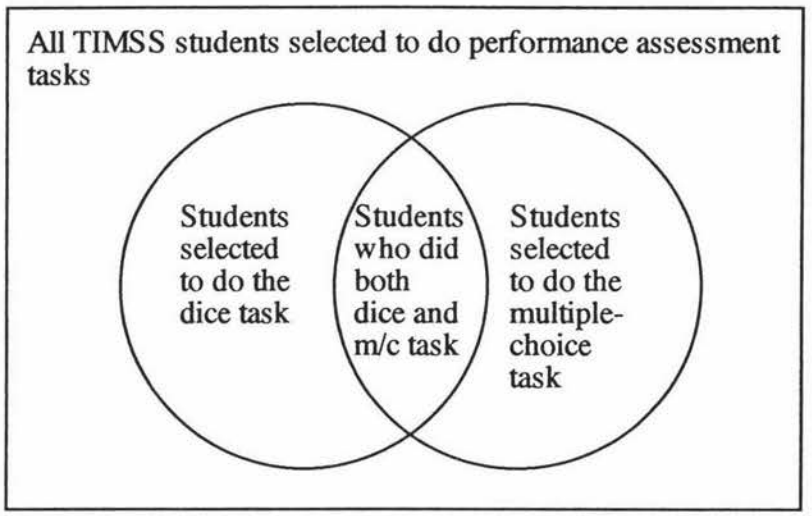
There is much international debate on the use of the reliability and validity procedures developed for multiple-choice tests on other types of tests (see Section 2.10). However, it is agreed that one important type of task validity to examine is evidential validity; that is whether the tasks used match curriculum expectations. Because of the international nature of the TIMSS study, the tasks were selected to try and match expectations in all the countries involved. In considering evidential validity, countries were asked during the pilot study to rate each task as to whether it had: instructional familiarity, assessment familiarity, whether it was taught, and the difficulty of the task for the students. However, the question of evidential validity was re-examined in this thesis by matching each question to the New Zealand Mathematics Curriculum, looking at both the intentions and the wording of the task.

It seems clear from discussions in the literature (see Section 2.10) that as well as evidential validity, the consequential validity of tasks needs to be addressed. What are the consequences of the use of performance assessment tasks for students and teachers in New Zealand? If the advantages and importance of performance assessments can be proven and communicated to test makers and consumers, then the literature suggests that the consequences for students and teachers will be positive. Examining specific consequences of the use of performance assessment tasks in the TIMSS study was outside the scope of this thesis, but speculation on the possible consequences was indulged in during the discussion of the results.

Basic information for the performance assessment tasks have already been published (Caygill, 1995; Garden 1997), as have the grouped and some individual results for multiple-choice and free-response questions (Garden 1996). As such, basic frequencies were only included in this thesis where they were of interest in comparisons of results. Note that in some cases the proportions of students reported in this thesis differ slightly from those reported by Garden because of different numbers of students selected for analysis.

The results for the performance assessment, multiple-choice, and free-response tasks, or parts of tasks, that have corresponding content were compared to see if there was any difference in performance. In general, the results for all the students selected to do the task under consideration were examined. However, because rotated booklets were used in the administration of both the performance assessment and the written tasks, there is only a partial overlap between the group of students that did any two tasks.

Figure 4.1 Representation of the partial overlap between performance assessment tasks and the associated multiple-choice tasks.



This overlap group was often a very small number of students, so in order to have enough students in each subgroup for the results to be meaningful, the larger group was used for each task. However, the overlap group was also examined to see if the same trend represented by the whole group was also in the overlap group. Where the trend differed these results are presented.

In considering how to determine significance in the differences in performance between task types or subgroups, a number of statistical tests were considered. Firstly, the way the data was scored needed to be considered. (See Table 3.4 for a description and sample of the codes used in scoring.) The scores exist as both nominal (two-digit code) and ordinal (first digit of the two-digit code) data. That is, the two-digit code should be considered nominal as a code of 22 is not higher than a code of 21. However, a score of

2, as in a code of 20, is considered higher than a score of 1, as in a code of 10 and so the first digit is ordinal. Note that a code starting with 7 or 9 is interpreted as having a score of zero. In general the ordinal scores were used in comparisons and the nominal scores were described and used where they added further information. It could be argued that the first digit of the code is an interval measurement since there is a zero point, and possible measurements are one unit apart, with some meaning to the distance between the measurements. Any interval measurement is also ordinal and so we may use the word *ordinal* to describe interval data but not vice-versa.

The comparisons between the multiple-choice and performance assessment tasks were intended to examine firstly whether the tasks were comparable in difficulty and content, and secondly, if they were similar, whether it was the same sub-group of students who could successfully complete each of the tasks. An example of an expected table if there was no difference in performance is given in Table 4.1.

Table 4.1 Example of expected table where there is no difference between the scores on two tasks.

	scores 0 in PA task	scores 1 in PA task	scores 2 in PA task
scores 0 in MC task	most of this row	some	very few
scores 1 in MC task	very few of this row	some	most of this row

The comparisons between sub-groups of students were intended to examine firstly whether these sub-groups performed similarly, that is, did the same proportion of students in each group successfully complete the task, and secondly, if there was a difference, how large was it. An example of an expected table with no difference is given in Table 4.2.

Table 4.2 An example of the expected table if there was no difference between groups.

	score 0	score 1	score 2
female	proportion a	proportion b	proportion c
male	same as or similar to a	same as or similar to b	same as or similar to c

This data is ordinal in the conservative view and it has been suggested (Eley, 1998, personal communication; Stewart, 1998, personal communication) that some non-parametric test should be used, preferably one that doesn't make too many assumptions about the data. Two possibilities are the Spearman or the Mann-Whitney tests. Both tests require replacing scores with ranks (Conover, 1971; Daniel, 1990). However, there are too few ranks for each task, at most three, and as a consequence there are too many observations with the same rank, the significance of the data is lost, and the tests are inconclusive.

The test used on a table like Table 4.1, comparing multiple-choice and performance assessment tasks, is the chi-square test for independence (Conover, 1971). This test only assumes that the sample is a random sample and that each observation may be classified into exactly one cell in the contingency table. The hypothesis for this test can be stated as *the rows and columns of the contingency table represent two independent classification schemes.* If this hypothesis is rejected, then the same students successfully completed each question.

To test the assumption that all the proportions in any column of a table like Table 4.2 were similar to each other, the Chi-Square Test for Differences in Probabilities, $r \times c$ (Conover, 1971) was used. This test assumes that the sample is random, the sub-groups are independent of each other, and that each observation may be categorised into exactly one cell of the contingency table. This test hypothesises that the probabilities for sub-group members being classified in a particular column are equal to each other for all columns.

This test requires data to be arranged in an r × c contingency table in which the subgroups are in the rows and the scores are in columns as in Table 4.2. If we denote the values in each cell (i,j) as O_{ij}, the row totals as n_i, the column totals as C_i then the test statistic T is given by

$$T = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \quad where \ E_{ij} = \frac{n_i C_i}{N}$$

The chi-square test is not very sensitive when expected values differ markedly for only a few cells in a table. For example, look at Table 4.3.

Table 4.3 Observed and expected proportions of girls and boys achieving each score on a hypothetical task

	girls		boys		
score	obs.	expt.	obs.	expt.	row total
0	28	34	40	34	68
1	36	33	30	33	66
2	36	33	30	33	66
column total	100	100	100	100	200

The chi-test would tell us that there is no statistical difference between the girls and the boys, that is, that there is not much difference between the expected and the observed values for the table. However, by observation we can see that a far higher proportion of boys than girls had a score of zero for this task. If we collapse Table 4.3 so that it contains only two rows, those with a score of zero and those with a score greater than zero, we then find that this is a statistically significant result at the ten percent level.

Table 4.4 Observed and expected proportions of girls and boys achieving a score of zero or more than zero on a hypothetical task

	girls		boys		
score	obs.	expt.	obs.	expt.	row total
0	28	34	40	34	68
1 or 2	72	66	60	66	132
column total	100	100	100	100	200

We can also see, in Table 4.3, that a difference of six percent between the observed proportions for girls and boys with a score of one, leads to a three percent difference between the observed and expected values. Since it is a difference between the observed and expected values that is examined in the chi-square statistic, this has the effect of making the difference between subgroups seem smaller. Daniel (1990) notes when discussing statistical tests that "*sufficiently large samples will reveal any difference, no matter how small, but it may be that only a fairly large difference is of any practical value. Likewise, small samples may fail to detect population differences (or relationships) that are of practical significance*" (p. 12). Thus although some differences were not found as statistically significant by the chi-square test, if the difference was greater than ten percent they were still reported in the results.

In order to compare the results on the performance assessment tasks of the standard three and form three students, the proportion of students with each total score was presented in graphical form. This total score was a simple total of points gained during the entire task.

4.3. Selection of background variables

The criticisms of standardized or traditional written tests as biased against some kinds of students (Griffin & Nix, 1991; Hambleton & Murphy, 1992; Haney & Madaus, 1992; Ministry of Education, 1992), suggested the selection of the following background variables to examine in relation to the student performances: gender, ethnicity, use of English at home, and items found in their home.

Students were asked in the background questionnaire which ethnic group they identified with from a list of twelve, and were allowed to select more than one group. The groups listed were: New Zealand Maori, New Zealand Pakeha/European, Other European, Samoan, Cook Island Maori, Tongan, Niuean, Other Pacific Island, Chinese, Indian, Other Asian, Other. Because the sizes of minority groups in the sample were small, the decision was made to look at a majority ethnic group versus a minority ethnic group. The *majority* ethnic group is defined as those who identified themselves as "*New Zealand Pakeha/European*" or "*Other European*" or both, and selected no other ethnic group. The *minority* group included all students who identified with any of the non-European ethnic groups even if they also identified themselves as "*New Zealand Pakeha/European*" or "*Other European*". It was not possible to examine Maori students as a separate group because the rotation of booklets caused this group to have less than ten students in it for some multiple-choice questions.

In the background questionnaire students were asked how often they spoke English at home. This question was designed to elicit whether English was the main language they used or not. Students were grouped into two groups based on their answer to this question. One group contained all those that responded "*always or almost always*". The other group contained all those who responded "*sometimes*" or "*never*". Those who responded "*never*" were grouped with those who responded "*sometimes*" because they were small in number and both lots of students can be said to be disadvantaged by a test in English. These groups were labelled as *the students from English-speaking homes* for those who responded "*always or almost always*" and *the students from non-English-speaking homes* for the rest of the students.

Students in the TIMSS study were asked to indicate, from a list of 16 items, which items could be found in their home. The items on the list were: calculator, computer, study desk/table, dictionary, television, musical instruments, CD player, video camera, video games, microwave, clothes dryer, dishwasher, two bathrooms, second car/motorbike, encyclopaedias, mobile or cellular phone. This question was designed to provide a surrogate measure of socio-economic status. Socio-economic status can impact on education because students from low socio-economic families are unlikely to

be surrounded by a range of out-of-school learning experiences which require materials or equipment to be purchased. Schools in areas where the majority of students come from low socio-economic families may also have limited extra funds for materials and equipment.

Initially, the students were placed in five subgroups based on the number of items they had identified as present in their home from the list. Subsequent to this subgrouping, the first three of these five subgroups were regrouped together into one group because of the small number of students in the first two groups. Thus at the conclusion of this process, the first group, the low socio-economic group, were those students that identified zero to ten of the listed items as being present in their homes. The second group, the middle socio-economic group, were those that identified eleven to thirteen of the listed items as being present in their homes, and the third group, the high socio-economic group, were those that identified fourteen to sixteen of the listed items as being in their homes.

In addition to the background questions, two attitudinal questions related to the value of mathematics were examined in relation to performance. They were:

(a) I think it is important to do well in mathematics at school

and

(b) Mathematics is important to everyone's life.

Students were able to *strongly agree, agree, disagree,* or *strongly disagree* with these statements. However, only a small percentage (three percent and five percent respectively) of students disagreed, that is they stated they disagreed or strongly disagreed, with each of these statements. Since this group of students who disagreed with each statement might be thought to do less well than the students who agreed, the data was examined to see if this was so.

4.4. *Limitations of this analysis.*

There are a number of limitations in the design of this study. The nature of secondary analysis is that the data has already been collected and methods of analysis have to be formulated around the available data. The two performance assessment tasks examined in this thesis did not have many multiple-choice tasks and even fewer free-response tasks associated with them. The method of rotating booklets amongst students resulted in there being only a small number of students selected to do both the performance assessment task and the multiple-choice task examined in many cases. Another difficulty with secondary analysis is that questions can not be reformulated to take account of gaps discovered during preliminary analysis.

There were also a number of limitations in the data analysis. In some instances, information supplied by students or scores created by markers were aggregated. There is always a danger that in aggregating information into single factors or scores, some important information might be masked or lost. No attempt was made in this thesis to do a multi-variate analysis, finding factors and stating how much each factor contributed to variance in results. The aim in examining different subgroups of students was not to find out whether a factor such as ethnicity contributed to differences in results, but rather whether differences were smaller with performance assessment tasks than with multiple-choice tasks and whether a greater range of skills could be displayed by the subgroups during the performance assessment tasks than with the multiple-choice tasks.

It was difficult to find a statistical technique suitable for determining the statistical significance of differences between subgroups and differences between questions. The chi-square test was the most appropriate technique found, although it was limited by its lack of sensitivity.

4.5. Summary

In summary, the analysis considered data from two performance assessment tasks and their associated multiple-choice tasks in relation to the curriculum, each other, and background variables using proportions and the chi-square test. The background variables were gender, ethnicity, use of English at home, socio-economic status, and the personal value of mathematics.

5 DICE

5.1 The dice[1] task

Your task:

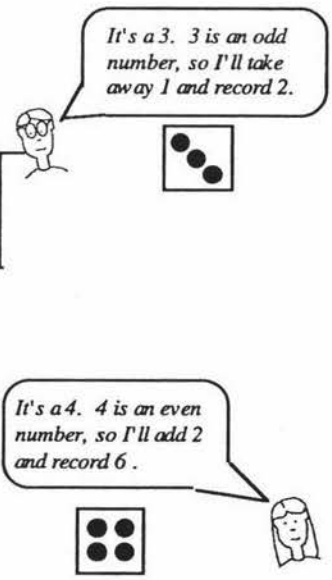
Find out what happens when we use a rule to change the numbers that turn up when a dice is thrown.

The rule for changing the numbers is:

- If an ODD number turns up, take away 1 and record the result.
- If an EVEN number turns up, add 2 and record the result.

1. In the table below, two examples have already been recorded for you. Use the rule to find out what the other changed numbers will be. Complete the table.

Number on dice	Changed numbers
1 	
2 	
3 	2
4 	6
5 	
6 	



page 2

2. What do you notice about the numbers you recorded?

[1] Note that the word dice is used in this document as well as in the test booklet to refer to the single object rather than the correct but possibly uncommon term die.

3. Throw the dice 30 times. Each time you throw the dice change the numbers that turn up using the rule. Each time record the number on the dice and the changed numbers. Write the numbers in the tables below.

Number on dice	Changed number

Number on dice	Changed number

page 3

4. Look again at the table you filled in for question number 3. How many times did you record each of the following numbers in the "Changed Number" column?

Changed Number	Number of Times Recorded
0	
1	
2	
3	
4	
5	
6	
7	
8	

5a. What changed number did you record most?

5b. Why did it happen this way?

The excluded space on the presented task material is where there was white space for recording student answers. Page one was the cover sheet for the task which includes task identification information and space for student identification information.

Both standard three and form three students attempted this task. They were supplied with a dice, a shaker, and material for rolling the dice on to reduce the noise for other students. The list of equipment was given at the beginning of the task.

According to the TIMSS Performance Assessment Coding Committee (1994), this task was intended to measure the student's

- *number sense*
- *ability to*
  - *use an arbitrary rule or algorithm*
  - *recognize patterns*
  - *explain the patterns in the light of the algorithm used*
  - *use mathematical concepts such as probability to explain patterns.*

5.2 The multiple-choice and short-answer tasks associated with the dice task.

There were eight tasks in the multiple-choice section of the TIMSS study, and one short-answer question, that corresponded to the content in the dice task. Two of these tasks, H8 and G1 have not been released at the time of writing and so cannot be detailed here. The multiple-choice questions H8, J5, L13[2], and the short-answer question U4 required students to examine patterns in a number of different ways, find the next number in the sequence, describe the rule, or compare patterns. Questions K4, G1, and L10[3] required students to take information from tables. Questions N18 and M3[4] were about probability.

Question H8, an unreleased task, required standard three students to complete a number sequence given the first five numbers.

---

[2] This task was labelled L4 in the standard three booklet and L13 in the form three booklet. For simplicity it will be labelled L13 throughout this document.

[3] This task was labelled F5 in the standard three booklet and L10 in the form three booklet. For simplicity it will be labelled L10 throughout this document.

[4] This task was labelled L2 in the standard three booklet and M3 in the form three booklet. For simplicity it will be labelled M3 throughout this document.

J5. What do you have to do to each number in Column A to get the number next to it in Column B?

Column A	Column B
10	2
15	3
25	5
50	10

A. Add 8 to the number in Column A.

B. Subtract 8 from the number in Column A.

C. Multiply the number in Column A by 5.

D. Divide the number in Column A by 5.

L13. These shapes are arranged in a pattern.

○△○○△△○○○△△△

Which set of shapes is arranged in the same pattern?

A. ★□★□★★□□★★□□

B. □★□□★□□□★□□□□

C. ★□★★□□★★★□□□

D. □□★★□★□□★★□★

U4. These numbers are part of a pattern.

50 , 46 , 42 , 38 , 34 , ...

What do you have to do to get the next number?

Answer: _____

K4. Kylie and Ben are playing a game. The object of the game is to get the highest total of points. This chart shows how many points they each scored.

Score Card

Player	Kylie	Ben
Round 1	125	100
Round 2	125	125
Round 3	150	100
Round 4	50	150

Who won, and by how many points?

A. Ben won by 25 points.
B. Ben won by 100 points.
C. Kylie won by 25 points.
D. Kylie won by 175 points.

Question G1, an unreleased task, required form three students to read a table of frequency data and summarise a selection of data, greater than a particular data point.

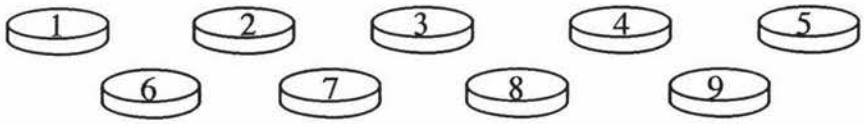
L10. This chart shows temperature readings made at different times on four days.

TEMPERATURES					
	6 a.m.	9 a.m.	Midday	3 p.m.	8 p.m.
Monday	15	17	20	21	19
Tuesday	15	15	15	10	9
Wednesday	8	10	14	13	15
Thursday	8	11	14	17	20

When was the highest temperature recorded?

A. Midday on Monday
B. 3 p.m. on Monday
C. Midday on Tuesday
D. 3 p.m. on Wednesday

N18. The nine chips shown are placed in a jar and mixed.



Melanie draws one chip from the jar. What is the probability that Melanie draws a chip with an even number?

A. $\frac{1}{9}$

B. $\frac{2}{9}$

C. $\frac{4}{9}$

D. $\frac{1}{2}$

M3. There is only one red marble in each of these bags.

10 marbles 100 marbles 1000 marbles

Without looking in the bags, you are to pick a marble out of one of the bags. Which bag would give you the greatest chance of picking the red marble?

A. The bag with 10 marbles

B. The bag with 100 marbles

C. The bag with 1000 marbles

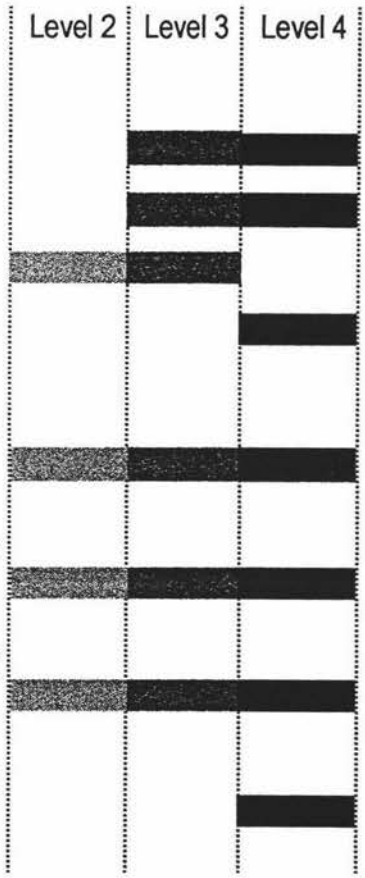
D. All bags would give the same chance.

5.3 Face validity - how do these tasks compare with curriculum expectations?

In examining the curriculum expectations in New Zealand, I have focussed on the expectations from levels 2 to 4 from the Mathematics in the New Zealand Curriculum document (Ministry of Education, 1992). Level 2 is the level in which most standard

three or year five children are expected to be working, with some moving in to level 3. Level 4 is the level in which most form three or year nine children are expected to be working. The processes in these levels relevant to the dice task are developing logic and reasoning and communicating mathematical ideas and the relevant objectives are summarised in Table 5.1.

Table 5.1 Achievement objectives from the Mathematical Processes strand associated with the dice task[5]



The achievement objectives from the other five strands in the curriculum document, which are relevant to the dice task, are summarised in Table 5.2.

[5] The shading in this table represents that used in the Mathematics in the New Zealand Curriculum Document

Table 5.2 Achievement objectives from the content strands associated with the dice task

	Level 2	Level 3	Level 4
Number			
• mentally perform calculations involving addition and subtraction	√		
Algebra			
• continue a sequential pattern and describe a rule for this	√		
Statistics			
• collect and display category data and whole number data in pictograms, tally charts, and barcharts, as appropriate	√		
• talk about the features of their own data displays	√		
• use their own language to talk about the distinctive features, such as outliers and clusters, in their own and others' data displays		√	
• use a systematic approach to count a set of possible outcomes		√	
• report the results of mathematical explorations concisely and coherently		√	
• predict the likelihood of outcomes on the basis of a set of observations		√	
• report the distinctive features (outliers, clusters, and shape of data distribution) of data displays			√

So how then do the dice task expectations match the curriculum expectations? Most standard three and form three children should be able to apply the algorithm from the perspective of addition and subtraction. Exploring even and odd numbers is given as a suggested learning experience at level 2 in the number strand so they would not be expected to have problems with this aspect of the algorithm. They would also be expected to be able to give a description of the pattern resulting from applying the algorithm, given that they have correctly achieved it in the first question, and follow the next set of instructions to throw the dice thirty times, recording the results of this

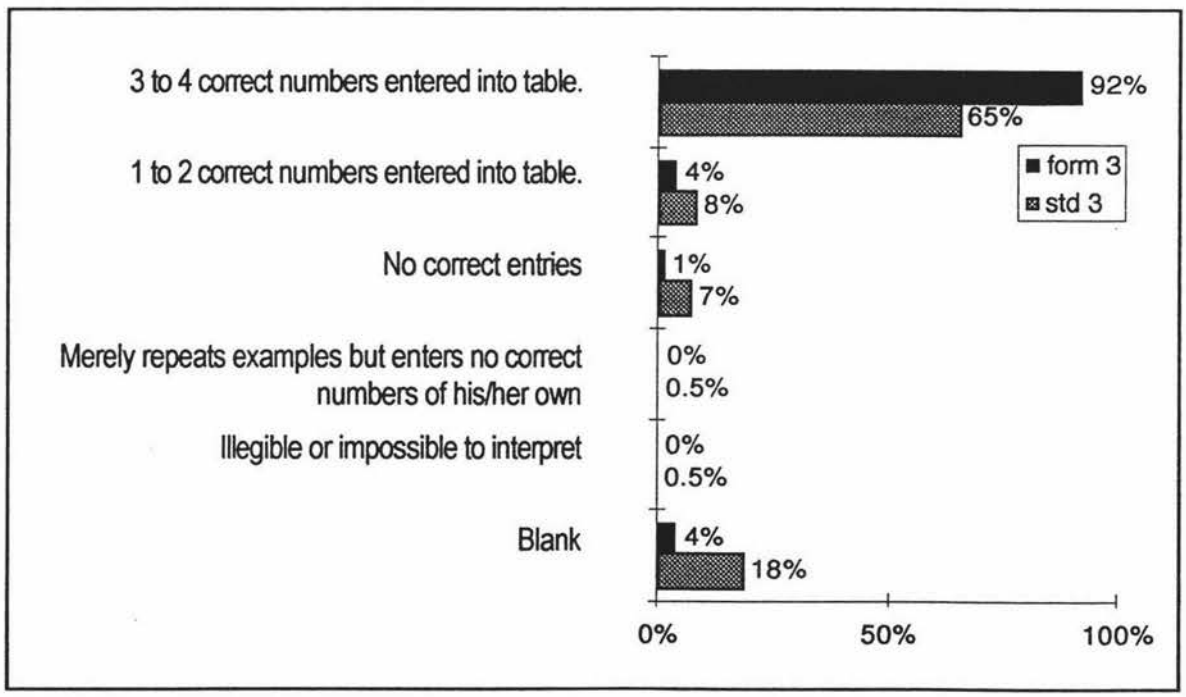
activity. Most of the children should be able to collect and display this data in the tally chart, but it may well be outside the expected ability of many of the standard three children to report the mode and analyse why this occurred.

For the multiple-choice questions, all the questions based on patterns, H8, J5, and L13 would be expected to be within the capabilities of most of the students who attempted them. Some of the standard three students might be expected to have some difficulties with question K4 on tables, because of the number of ideas that need to be integrated to obtain the result, but most of them should be able to select the correct answer for L10. The question involving probability, M3, might be outside the experience of some standard three students but most of the form three students would be expected to be able to select the correct option. Similarly, G1, L10 and N18, would be expected to be correctly answered by most form three students.

5.4 Results of the dice task.

Question one asked students to complete the table showing how the algorithm would affect each of the numbers on the dice. Two numbers were already entered in the table, so for full marks, three or four numbers needed to be correctly entered in the table. Half marks were awarded when only one or two numbers were correctly entered in the table.

Figure 5.1 Proportions of student responses to question one of the dice task.

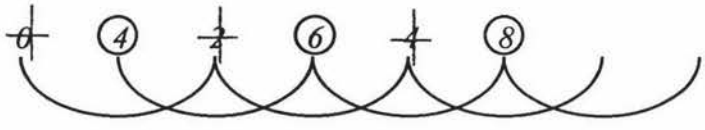


The errors students made in applying the algorithm included subtracting one instead of adding two or vice versa, that is, a mistake in choosing the odd or even part of the algorithm. In other cases the mistake could not be explained by a mistake in choosing odd or even. For example, one student had written in the table next to the dice number six a changed number of two, another student had added one to an even number instead of adding two. It was also observed during administration that some students would throw the dice until they obtained the number they wanted and then apply the rule to complete this table. This added activity was brought to the attention of one administrator by a student, who when asking for help, complained that they were having trouble throwing a six on the dice.

It is interesting to note that while eighteen percent of standard three students and four percent of form three students did not write anything in the table, over half of these standard three students (9 % over all), and all but one of these form three students, were able to correctly enter some or all numbers in question three, which relied on the results of question one. Thus if we add these students to those who correctly entered three or four numbers in the table, it would be fair to say that seventy-four percent of standard three students and ninety-five percent of form three students demonstrated that they were able to use the algorithm to change the numbers on the dice.

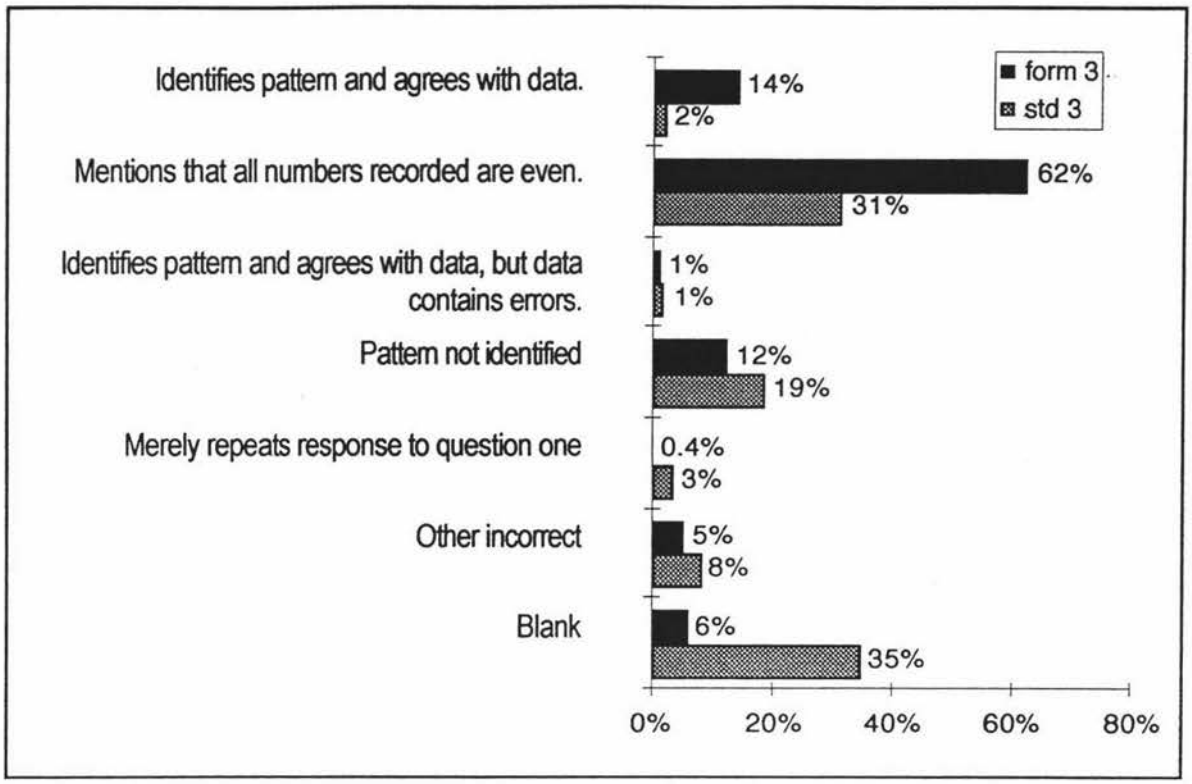
Question two asked students to describe the resulting pattern in question one. For the answer to be correct, a prose description consistent with the data in question one was required, thus allowing a number of possible responses. Some examples of patterns given, consistent with question one are:

- *every second number is the next in a sequence of even no's.*



- *The numbers are all even and the gaps between them go +4,-2,+4,-2*
- *they are all even numbers below 10 and 4 is doubled up*

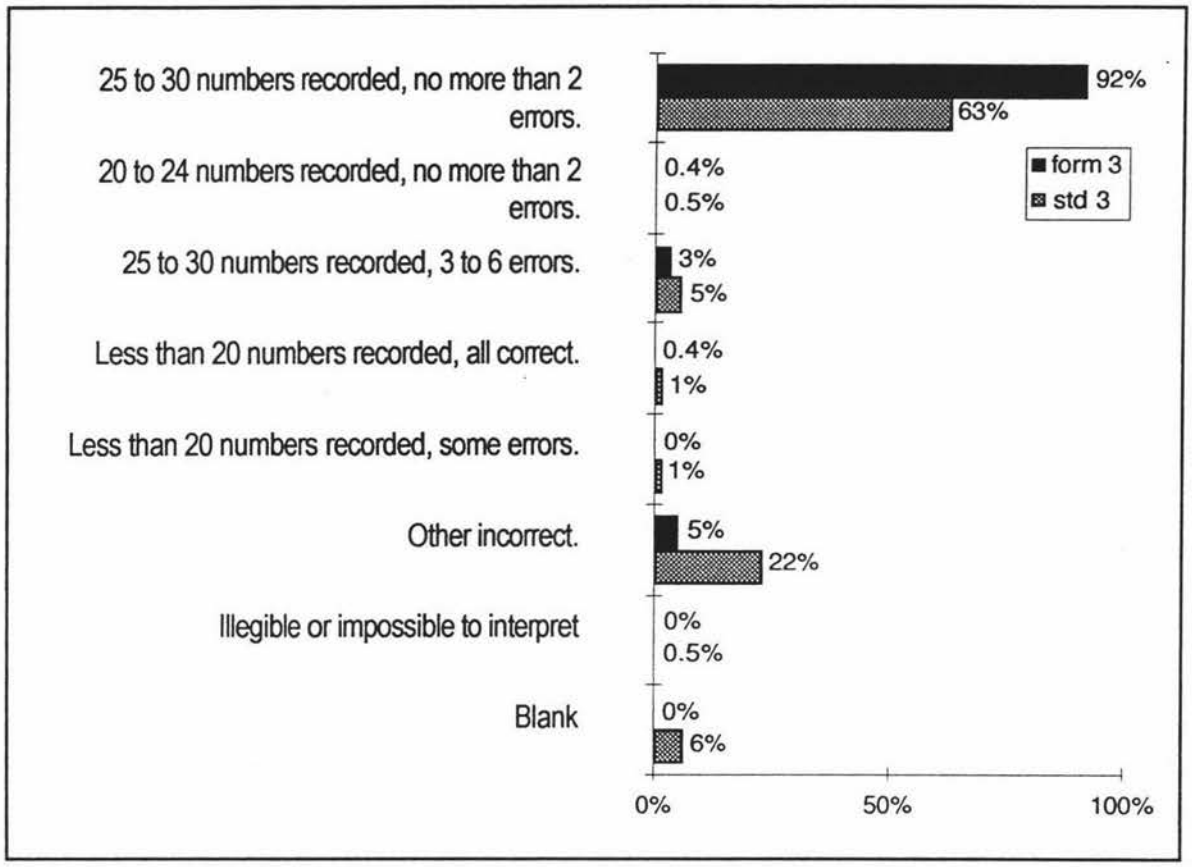
Figure 5.2 Proportions of student responses to question two of the dice task.



The low numbers of standard three students attempting this question raises the query of whether students were unable to do this question, possibly because of the need to provide a written explanation, or whether it seemed more interesting to get on with the next part of the task where they actually got to do something physical. This is especially pertinent when we note that the next part of the task, question three, was left unattempted by only six percent of the standard three students.

Question three asked them to throw the dice thirty times, each time recording the number on the dice and the changed number resulting from the application of the algorithm. For this question to be correctly completed, students had to complete at least twenty-five throws of the dice and correctly apply the algorithm.

Figure 5.3 Proportions of student responses to question three of the dice task.

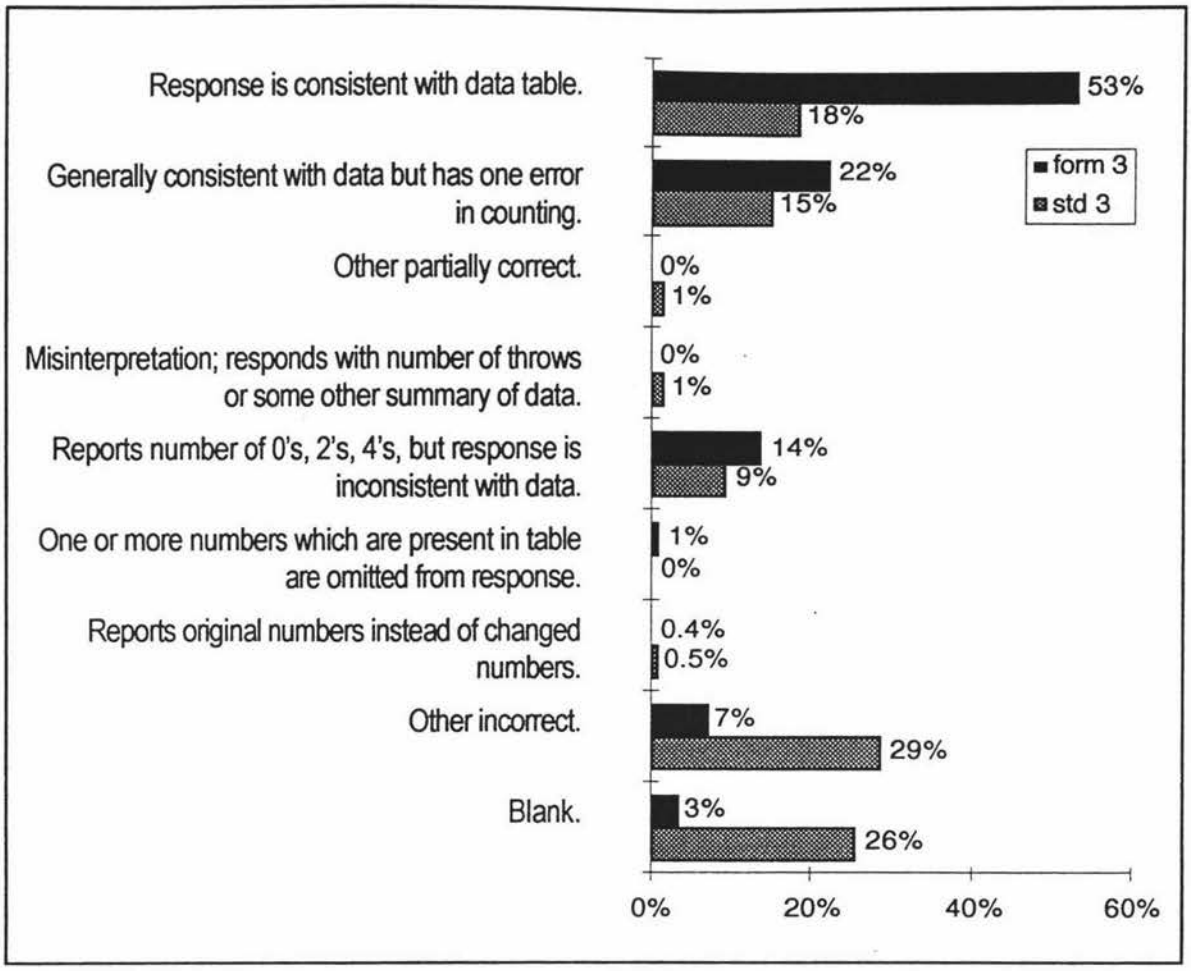


Note that two percent of standard three students and one percent of form three students correctly filled in part of the table but ran out of time to complete the task, that is did not complete the rest of the booklet. One percent of the standard three students filled in part of the table, making at least one error, but ran out to time to complete the task.

The possible causes of the errors in question three included having incorrectly applied the rule in question one and carrying this error through, or having introduced a new error in applying the rule at this stage. In a small number of cases all the numbers entered in the table appear to be results of dice throws with no application of the rule, with some of the results of dice throws entered in the *changed number* column.

Question four asked students to count how many times each of the possible changed numbers occurred in their experiment and record it in another table. For a complete response to question four, the entries in the table had to be consistent with the data in the previous question.

Figure 5.4 Proportions of student responses to question four of the dice task.



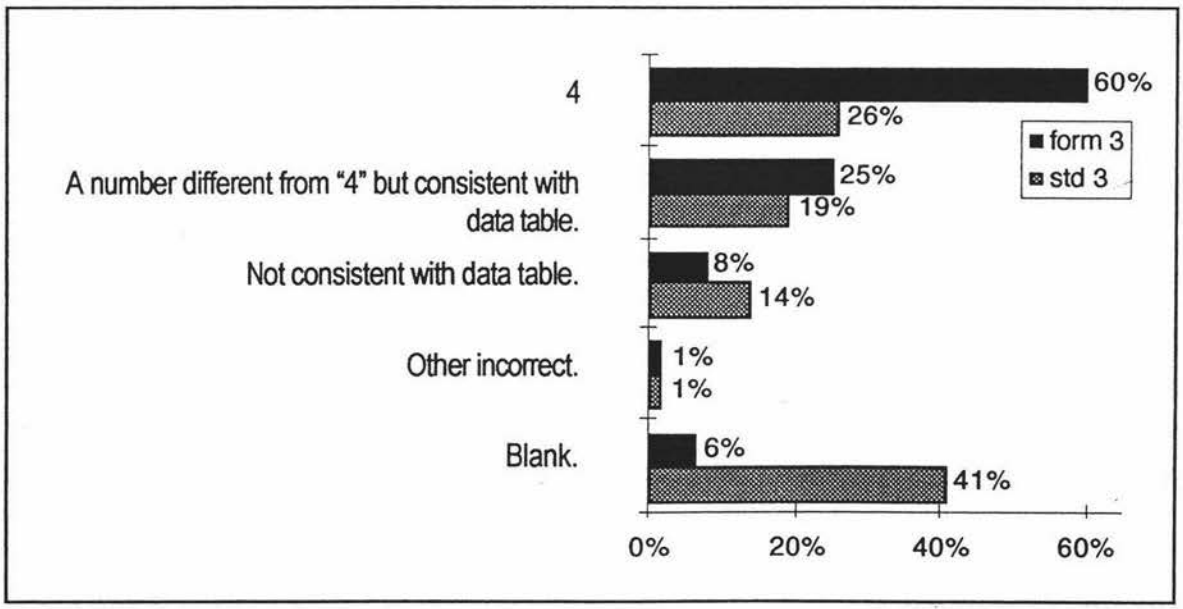
The types of errors encountered in this task included making more than one error counting the changed numbers, applying the algorithm to the table entries in this question, counting both the changed numbers and number on the dice or some combination of both. For some students, their table contained entries that did not appear to relate to question three, particularly the entries adjacent to seven or eight or both. It was not clear how students had arrived at these table entries. One standard three student merely ticked the numbers which were found in the table in question three, two students reported original dice throw numbers (one standard three student and one form three student), and two form three students omitted reporting a total count for one or more of the numbers in the table in question three.

Of those that attempted this question, forty-seven percent of standard three students and seventy-nine percent of form three students demonstrated that they knew how to take the data from the previous question and form the summary table required. It could

be argued that those students who took numbers from both columns and summarised all sixty numbers, were able to summarise the data also, but these students have been excluded from this figure because they did not discriminate between the two different sets of numbers. An aspect that may have caused trouble for some students was the fact that the table in question three was split into two sections. It would be interesting to see if having a single long table in question three would have facilitated answering question four. It would also be interesting to see if students would have been better able to complete the table in question four if the two tables were on the same page.

Question five asked students to state which number occurred most often, and propose a reason for this occurrence. For a complete response, the most often occurring number had to be consistent with the data presented in the previous questions, and the reason had to account for the predominance of one number.

Figure 5.5 Proportions of student responses to question 5 part (a) of the dice task.

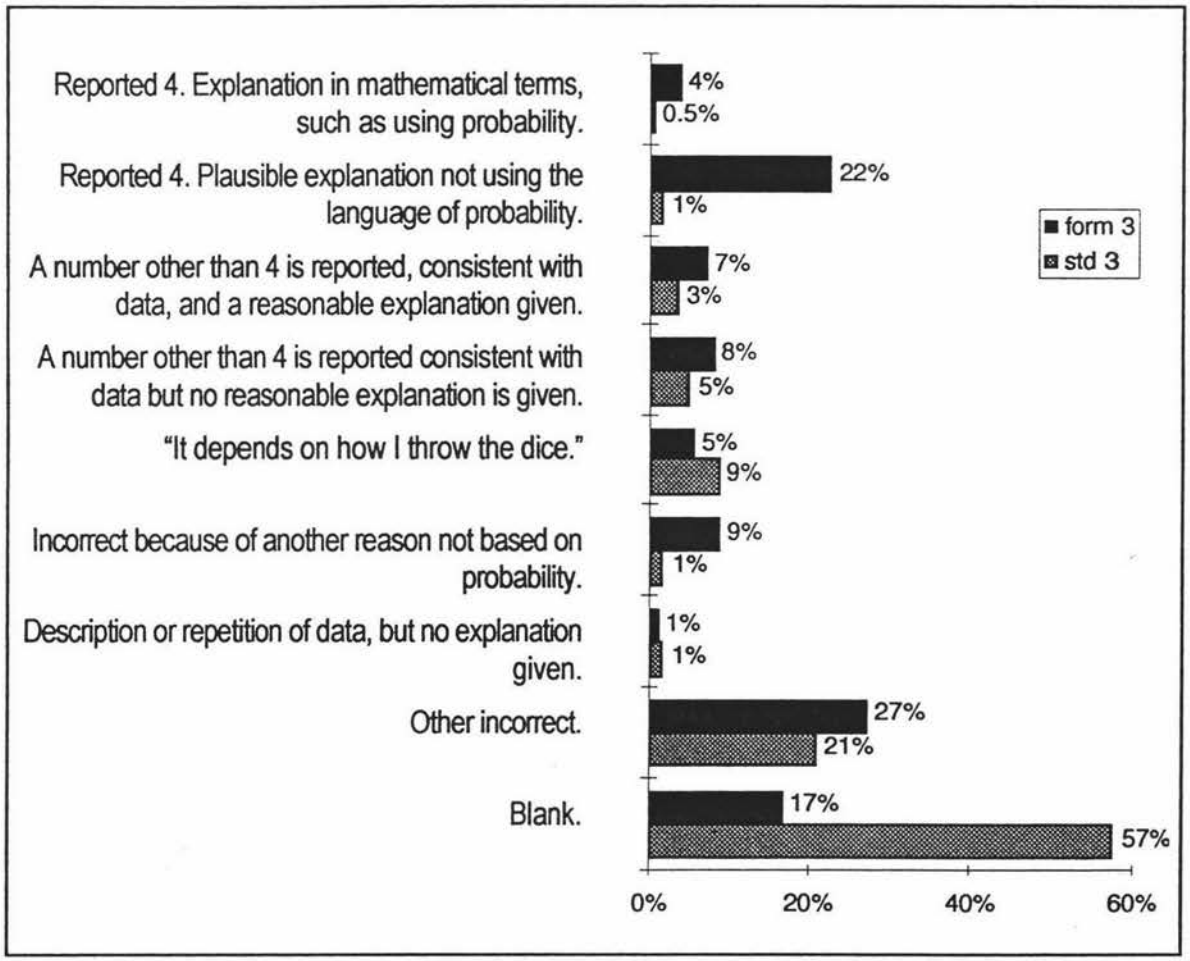


The inconsistencies included slip-ups in reading the table such as choosing the second most common number, or both the second most common as well as the most common number. A few students reported the frequency with which they recorded the most common changed number rather than identifying the most common changed number. Some students selected the most common number in the "*number of times recorded*" column of the table in question four. A couple of students reported that all the changed

numbers are even and some who had done question four incorrectly went back to question three to find the most common number thrown on the dice. For a few students it was not clear what they had done in this question..

Excluding these students who did not do the question from the calculations, seventy-five percent of standard three students and ninety percent of form three students reported the most common changed number consistent with the data in question four.

Figure 5.6 Proportions of student responses to question 5 part (b) of the dice task.



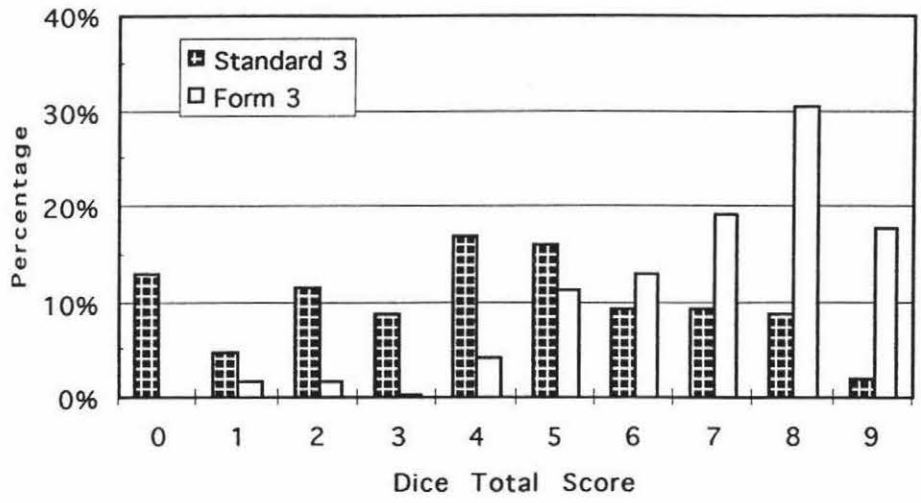
The types of errors made by students at both levels included saying that the reason for this result lay in the way they threw the dice, without mentioning specifics, or just luck without saying what they meant by that. Some students stated the fact that it was an even number, or that you could only get even numbers because of the rule, as being the reason for four being the most common changed number and some thought four was the most common changed number because it was the middle number between zero and

eight. As with part (a) of this question, some students gave reasons that were difficult to interpret.

Many students attributed the most common changed number to throwing a specific number on the dice most often. In some cases, it was the number four they said they had thrown lots of times, clearly confusing the changed number and the original dice throw. In other cases, they either said it was because they had thrown the number two lots of times or they said they had thrown the number five lots of times. Reviewing their data it was often found that this was not the case; these students had not recognised that both two and five contributed to a changed number of four. Other students, stated that they had thrown lots of twos and fives. These students appeared not to recognise that because both dice throws of two and five contributed towards a changed number of four there was as greater chance of getting a changed number of four.

While this task should have been able to be completed by most form three students and completed to question five part (a) by most standard three students it is clear that it was harder for the standard three students to complete, and achieve well in this task. Figure 5.7 presents the total scores for the dice task showing most form three students were able to do this task whereas many standard three students had difficulty completing it. The combination of time pressure, the unusual testing situation, and both being out of their normal classroom and with a teacher unknown to them, would have contributed to the difficulty in completing the task.

Figure 5.7 Total scores for the dice task



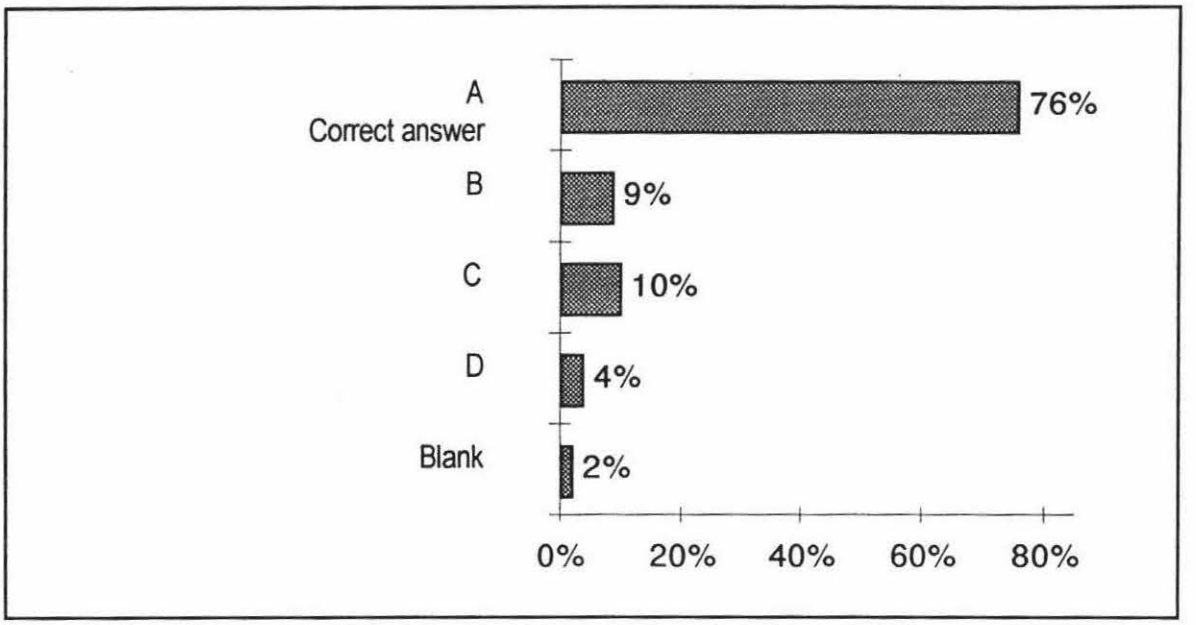
5.5 Results of the multiple-choice tasks

For the analyses of the multiple-choice questions (see section 5.2), the information presented in this chapter draws on the performances of all students in the performance assessment component of the study who had the opportunity to do each question. For comparison purposes with the dice task, we might prefer to look at only those students who did the multiple-choice task and the dice task, but because this was such a small group, any differences are exaggerated when the numbers are transformed into percentages. Instead, the results presented here are the values from the larger group of all respondents, as these have a similar distribution and are thus representative of the smaller group, all respondents who did both the dice and the multiple-choice question.

Question H8, an unreleased task[6], required standard three students to complete a number sequence given the first five numbers. This task required students to find the relationship and then apply it. The relationship was easier to calculate than that in question one of dice, requiring no knowledge of odd or even.

[6] As mentioned in Section 4.1, some TIMSS multiple-choice tasks are not yet available for publication.

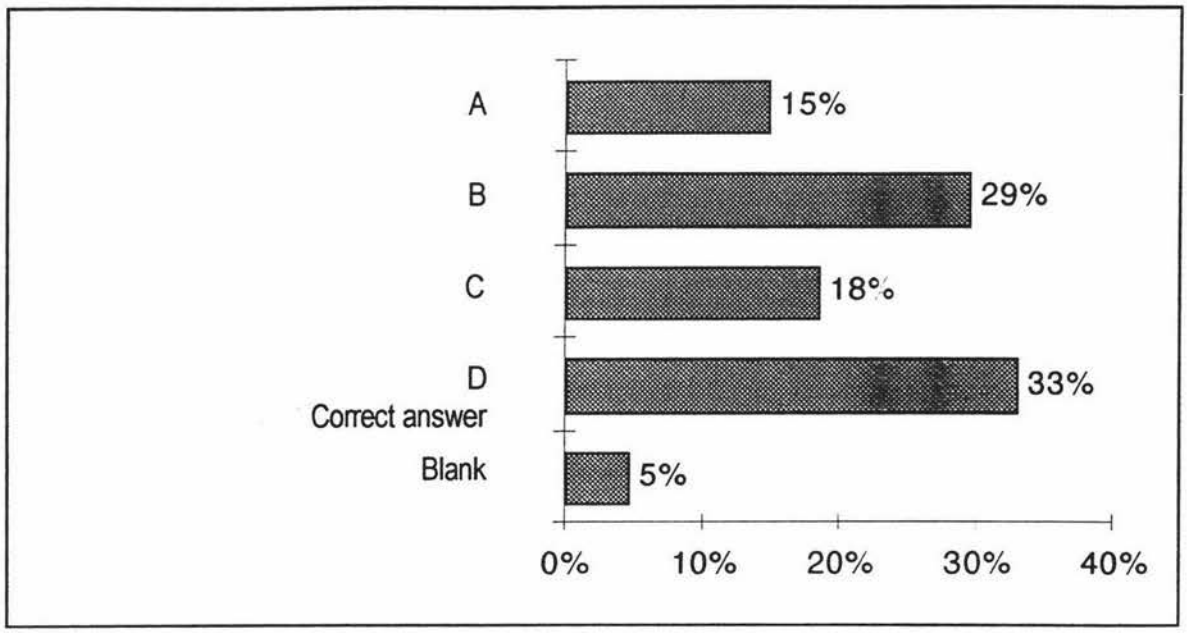
Figure 5.8 Proportions of student responses to question H8.



Fewer students correctly answered question one of dice, than of H8, although if we take into account the results of question three, about the same proportion of students could correctly apply the algorithm in the dice task. Question one of dice gives us slightly more information about the error pattern, but further questioning is needed to elicit the reasons why students have made these errors for both questions.

Question J5 required standard three students to find the rule given number pairs, and the correct option was D. We would expect answer B, the most commonly chosen of the incorrect options, if students only identified the relationship between the first pair of numbers.

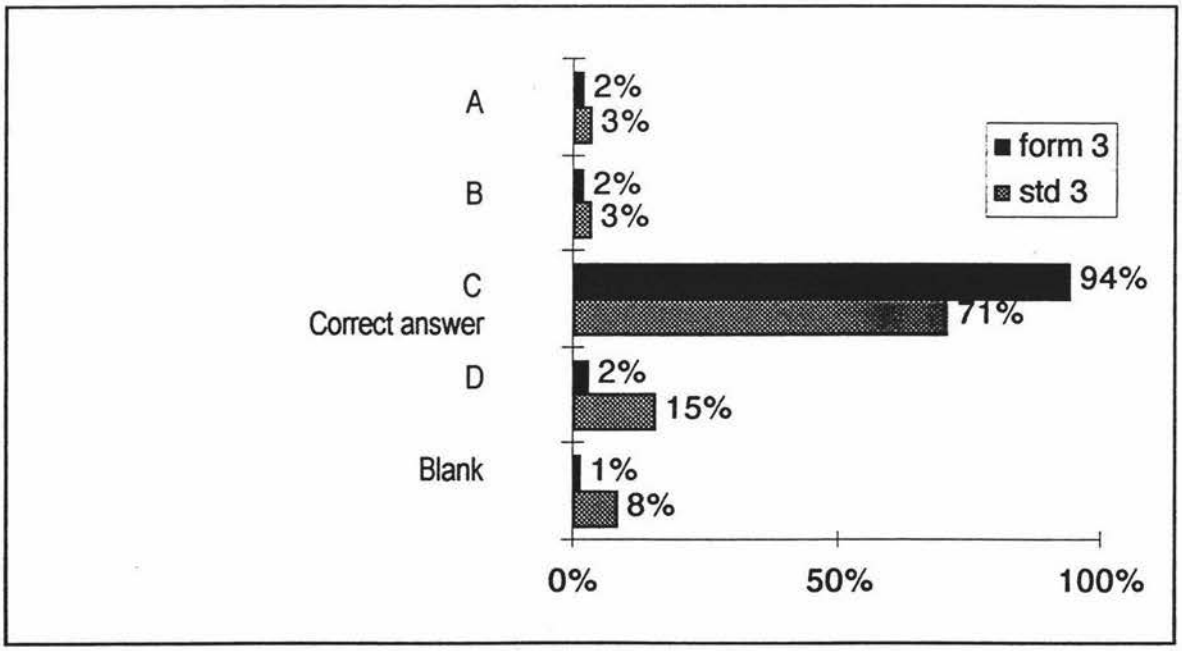
Figure 5.9 Proportions of student responses to question J5.



This task relates to question two in dice, because for both questions students have to examine the numbers and describe the relationship between them. In comparison with J5, more of the students correctly identified the numbers as all even in dice question two, although none gave a description of the pattern. However, the wording of question two was more general in nature than that of J5.

Question L13 required standard three and form three students to recognise similarity in two patterns and had a correct option of C.

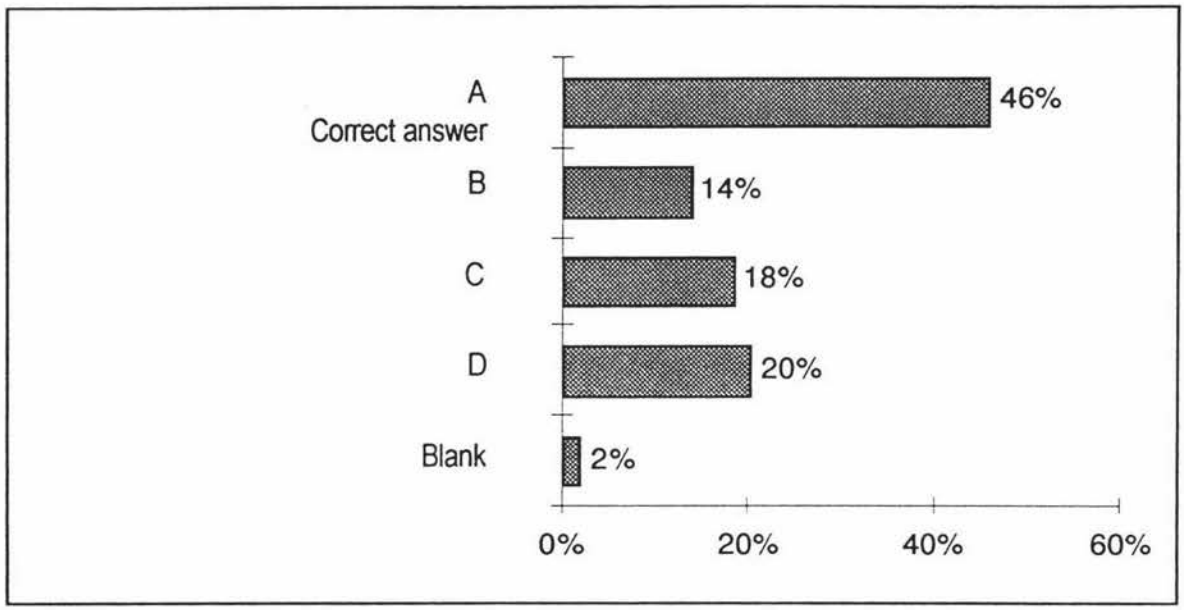
Figure 5.10 Proportions of student responses to question L13.



It is clear that the students found it easier to recognise that two patterns were the same than to describe a pattern in J5 or question two of dice.

Question K4 required standard three students to take information from a table and draw conclusions from it. The correct option for this question was option A.

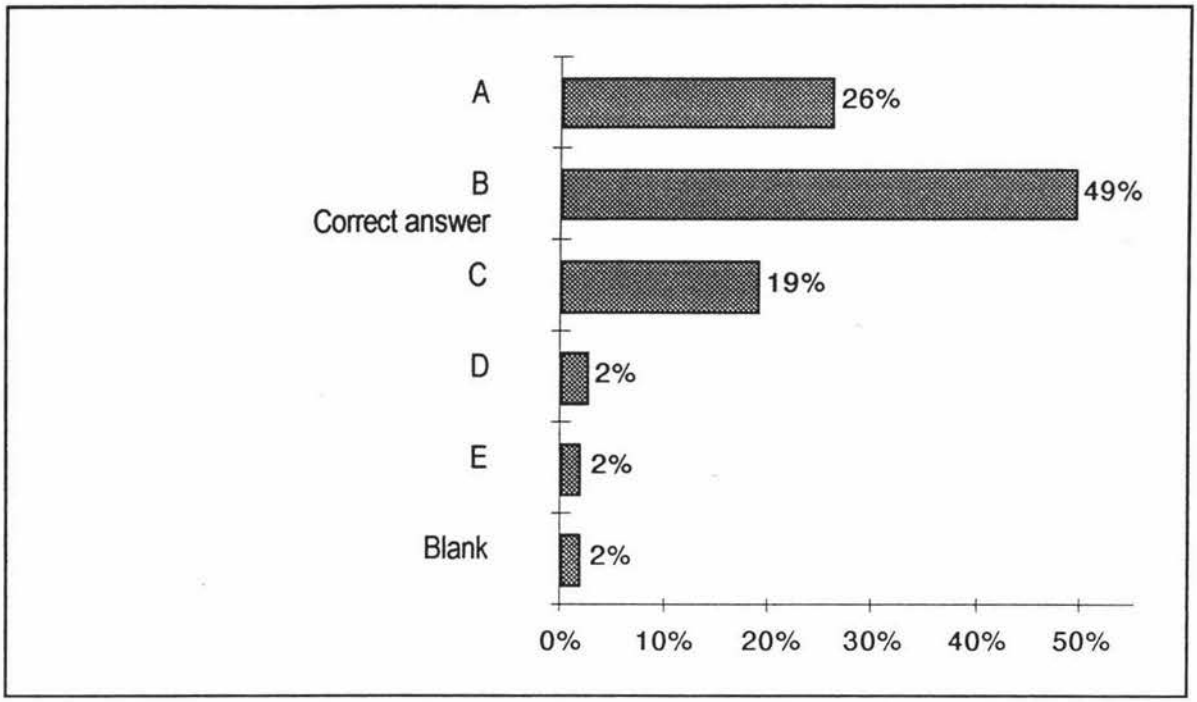
Figure 5.11 Proportions of student responses to question K4.



This task requires some skills that were also required for question four and question five (a) of dice. Question four required students to take information from one table and summarise it in another table and question five (a) required students to find the most common number in their table. Fewer students correctly summarised the information or made one counting error in question four. About the same proportion of the students were able to correctly identify the number with the highest count in question five (a). These questions all ask for different levels of table reading and interpretation and each one gives us different information about the students' abilities. The dice task however would give us more information about where problems or difficulties lie for students.

Question G1, an unreleased task, required form three students to read a table of frequency data and summarise a selection of data, greater than a particular data point. The correct answer for this question was option B.

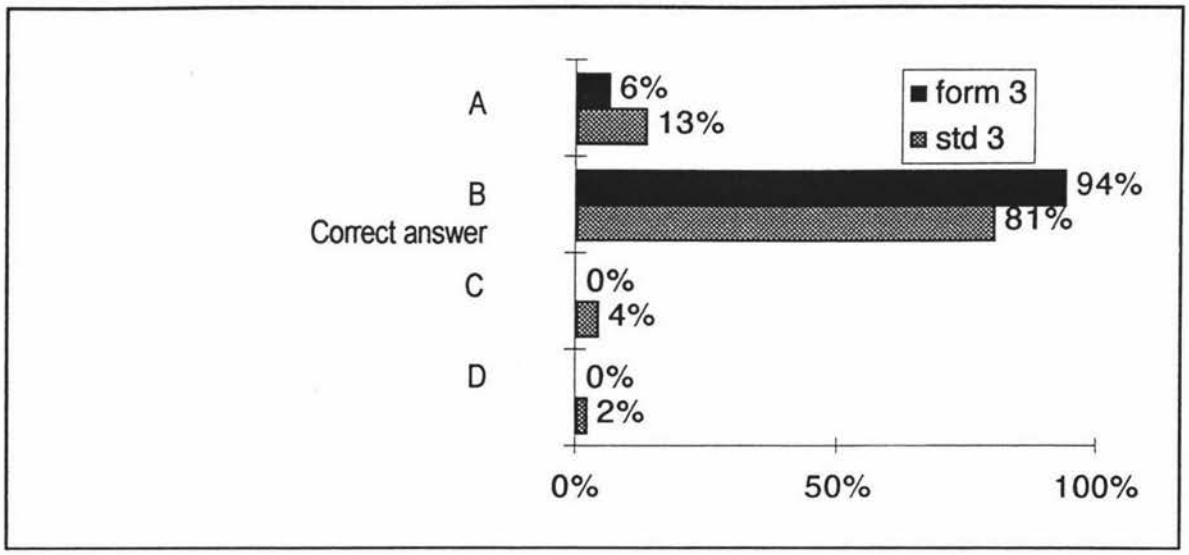
Figure 5.12 Proportions of student responses to question G1.



As for K4, this task requires some skills that were also required for question four and question five (a) of dice. A higher proportion of students correctly summarised the information or made one counting error in question four. Similarly, a higher proportion of students were able to correctly identify the number with the highest count in question five (a).

Question L10 required standard three and form three students to read a table of temperatures and choose the option that gave the time and day of the highest temperature recorded in the table. The correct option for this question was B.

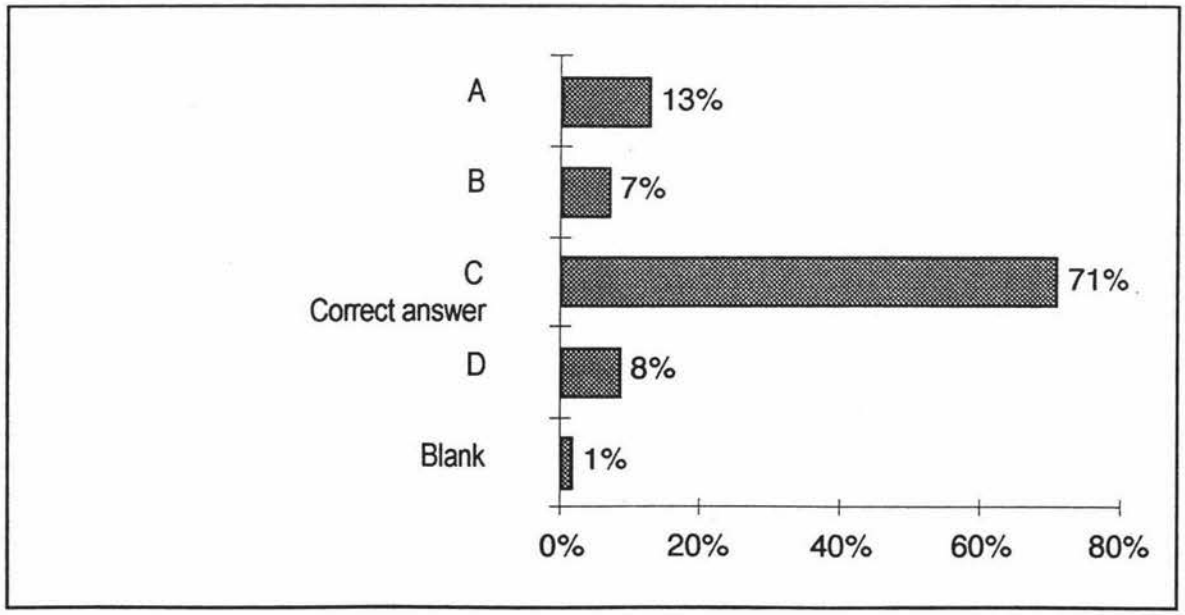
Figure 5.13 Proportions of student responses to question L10.



Question five (a) of dice required students to find the changed number that occurred the most often from the table they had made. Fewer students correctly gave the most common changed number, although a far higher proportion of students did not attempt question five (a).

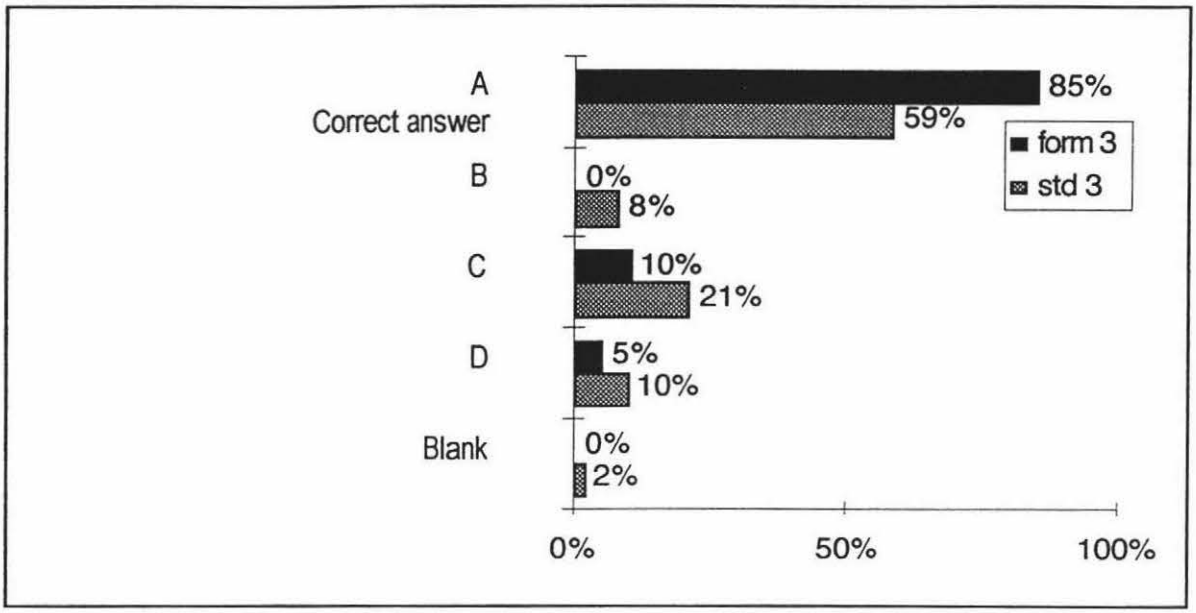
Question N18 required form three students to give the probability of drawing an even number from a jar containing nine chips labelled one to nine.

Figure 5.14 Proportions of student responses to question N18.



Question M3 required standard three and form three students to indicate which of three bags gave the greatest chance of drawing a red marble. The correct option for this question was option A.

Figure 5.15 Proportions of student responses to question M3.



In comparison with questions N18 and M3, the only part in the dice task which required probability information, question five (b), was made more difficult by having the contributing factor of the dice throws. Therefore students appeared to have difficulty recognising that the changed number four had a higher probability because they wrongly (in most cases) attributed the higher number of fours to throwing a larger number of twos, fives, or both twos and fives on the dice. Fewer students answered question five (b) in a way appropriate to their data using the ideas of probability.

We could say that these multiple-choice questions give greater opportunity for students to show their understanding of probability than does the dice task. However, in the 'real world' people often believe other factors to be more important than probability and the dice task provides an opportunity to expose and explore these beliefs.

Overall, this information appears to show that students have more success on the multiple-choice questions. Only question G1, when compared with the dice task, had a lower proportion of students correctly answering the question. There was also a higher proportion of students who did not attempt a number of the dice questions in

comparison to the multiple-choice questions. It is important to note though when judging the dice task, that question two of dice had more students scoring a mark than J5, and the opportunity for half marks in other questions resulted in students getting some credit for their work even if they ran out of time to finish.

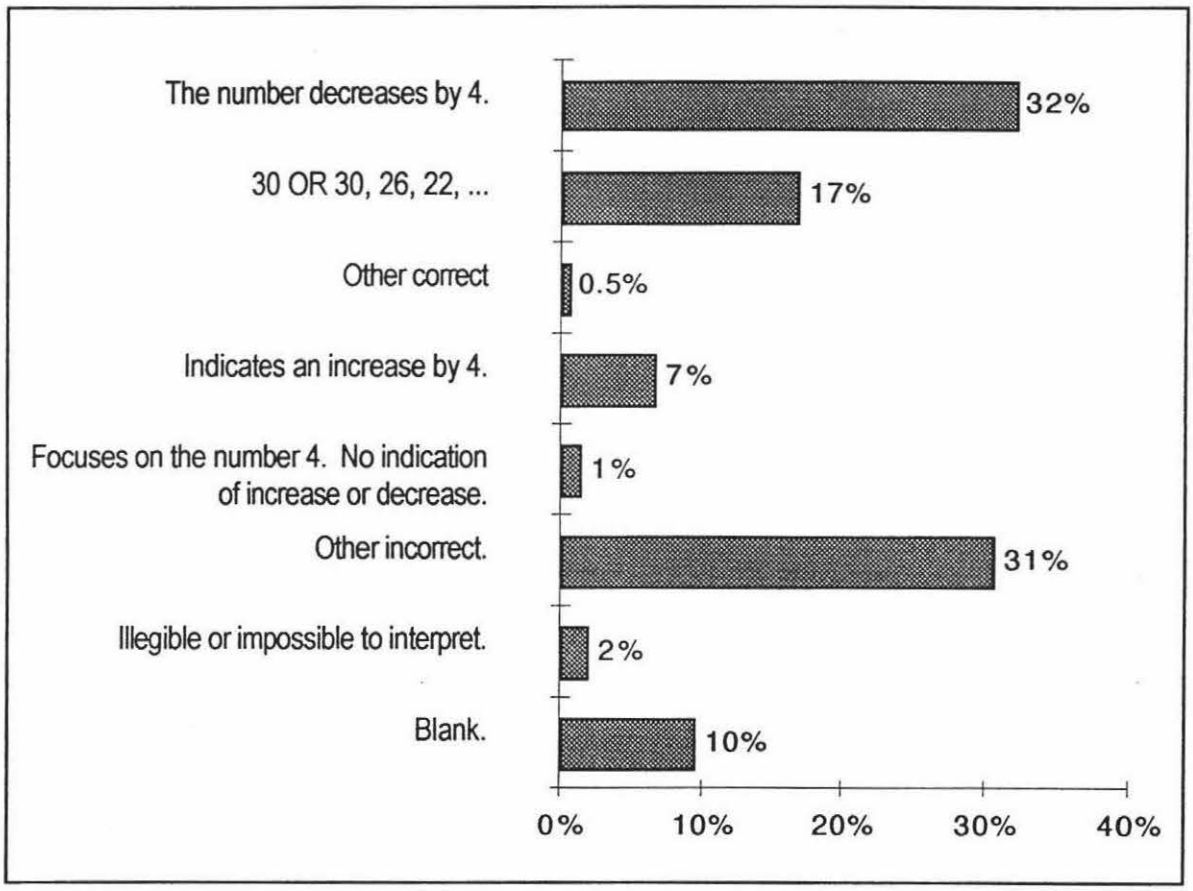
By examining the results of the dice task, it was, however, easier to determine when students had made a computational error rather than misunderstanding the relevant mathematical concepts. In this respect the dice task was more useful to people involved in seeing the results of the task, and would be more useful in a classroom also.

A positive outcome of the use of multiple-choice questions might be that if students get a result from their calculations that is not listed in the multiple-choice options, they may be prompted to reattempt the question and so discover their mistake.

5.6 Short-answer questions that cover similar mathematical content and their results

There was only one short-answer question, U4, that had similar content to the dice task. Standard three students were required to describe the relationship between the numbers in a sequence.

Figure 5.16 Proportions of student responses to U4



Half of the students gave an acceptable answer to this question, a greater proportion than the proportion of students who gave an acceptable answer for question two of dice. Question U4 is similar to dice question two but is much more specific. Thus the students are asked to look specifically at the relationship between the numbers in the series and are more likely to describe this relationship rather than giving an overall impression of the numbers, by saying that they were all even.

5.7 Comparison of the results of the dice task and the multiple-choice and short-answer tasks.

The dice task gave us more information than the multiple-choice and short-answer tasks about the abilities of students to perform a systematic experiment, record their results, and summarise their results in the table provided. It also gave us more information about how students can integrate their knowledge from a number of different aspects of the mathematics curriculum, particularly when explaining the reason for the most commonly occurring number.

Many of the multiple-choice questions identified in section 5.2 as similar to the content in the dice task questions were not independent of the associated dice questions by chi-square analysis. This result also held for U4, the short-answer question. This means that most of the students who correctly answered the dice question also correctly answered the multiple-choice question, and similarly, most of the students who gave an incorrect answer to one also gave an incorrect answer to the other. Rather than say these tasks were not independent by chi-square analysis, for the rest of this section I will use the phrase *associated by chi-square analysis*.

For questions one and two of dice, the associated questions by content analysis on application of algorithms and pattern recognition were H8, J5, L13, and U4. Of these questions[7], J5, U4, and L13, for the standard three students, were found to be associated to question two, by chi-square analysis also. Question U4 was found to be associated to question one as well by chi-square analysis. No questions were associated with question three of dice by content analysis, and only U4 was found to be associated by chi-square analysis. Since question three of dice was related to question one then the relationship of U4 with both of them is understandable. At the form three level, no multiple-choice questions were associated with the dice questions one, two or three by chi-square analysis despite associations by content analysis.

---

[7] The number of standard three students included in the analysis of each multiple-choice question and the dice task was 42 for H8, 42 for J5, 22 for L13, and 80 for U4. 35 form three students were included in the analysis of L13 with the dice task.

For questions four and five part a of dice, the associated questions by content analysis were K4, G1, and L10. Of these questions[8], K4 and G1 were found to be associated to both questions four and five part a, by chi-square analysis also. L10 was found to be independent of the dice task for all questions at both levels.

For question five (b) of dice, the associated questions[9] by content analysis on probability were M3, and N18. Since few students appeared to recognise the need to think of ideas of probability for question five (b) it is of little surprise that this question is independent of the results of M3 and N18 for both standard three and form three students.

The associations in this section may lead us to believe on casual observation that the multiple-choice questions tell us the same thing about the students as the dice task, but several things give us pause for thought. Firstly, there were no associated tasks with question four on content, and although M3 and N18 appeared to have the same content as question five (b), they actually exposed different aspects of student abilities and beliefs. Secondly, although the chi-square test revealed some associations between the results of the dice task and the multiple-choice tasks associated by content, there were still students who performed differently on the two associated tasks. For example, although all but one of the students who got G1 correct also got question five (a) correct, many of the students who got G1 wrong correctly answered question five (a). Similar sorts of patterns held for the other questions with a statistical association of results although in some cases it was many of the students who correctly answered the multiple-choice question, that incorrectly answered the dice question examined.

---

[8] At the standard three level, 58 students were included in the analysis of K4 and 81 students were included in the analysis of L10 with the dice task. At the form three level, 79 students were included in the analysis of G1 and 35 students were included in the analysis of L10 with the dice task.

[9] At the standard three level, 22 students were included in the analysis of M3 with the dice task. At the form three level, 22 students were included in the analysis of M3 and 9 students were included in the analysis of N18 with the dice task.

5.8 How do the sub-groups of students fare in the dice task and its associated tasks.

5.8.1 Gender differences.

All gender differences in the maximum score achieved for the multiple-choice, short-answer and dice questions, where the difference is greater than six percent, favoured the female students (see Figure 5.17 and Figure 5.18). This result is in contrast to much of the literature which has gender differences largely favouring the boys in nearly all aspects of mathematics. The maximum difference favouring the boys in the dice task was two percent (question three, standard three) and for the multiple-choice tasks was just under six percent (L10 standard three). Presented below are the graphs (Figure 5.17 and Figure 5.18) showing the percentages[10] of each group, girls and boys who were given the question to do, who scored the maximum score.

In some cases the difference in achievement of maximum score in the dice task was because the boys got only half marks. For example, for question one, thirteen percent of standard three boys compared with four percent of girls achieve half marks. However, in other cases, particularly question four at both the standard three and form three level, a higher proportion of the girls scored the maximum score and a higher proportion of them scored half marks. Only one mark was given for the multiple-choice and the short-answer question so half marks did not apply.

[10] The percentage is calculated by including the students who didn't attempt the question and the percentages for the multiple-choice tasks includes all the students who were involved in the performance assessment not just those who did the dice task as mentioned in chapter four.

Figure 5.17 Proportions of standard three boys and girls who scored the maximum score.

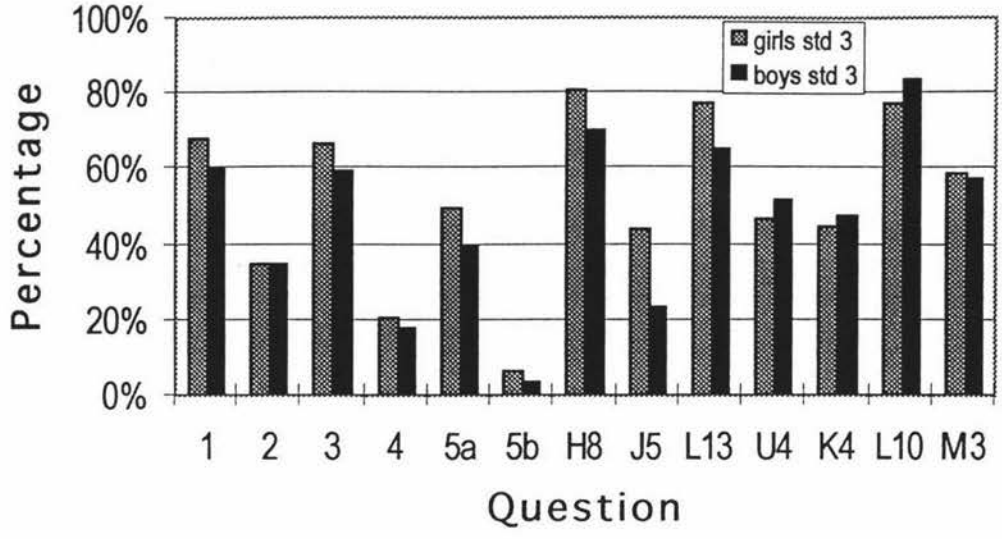
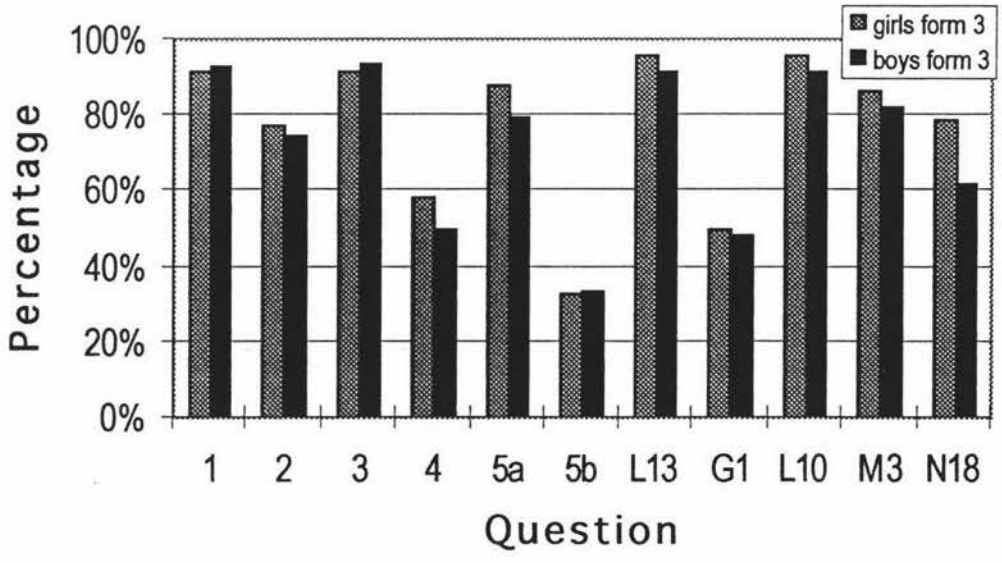


Figure 5.18 Proportions of form three boys and girls who scored the maximum score.



It is interesting to note the higher proportion of non-response from the boys than the girls for the majority of questions, both for the dice task and the multiple-choice tasks (see Figure 5.19 and Figure 5.20). The differences were particularly high, and statistically significant by chi-square analysis, for question two at the standard three level, question five (a) at the form three level, and question five (b) at both levels. As

the percentages of students who missed doing the question were much higher for the dice task this higher non-response for the boys is particularly interesting. This contradicts other research findings that state that boys are more likely to take a guess at the answer than the girls, particularly in multiple-choice questions. Possible reasons for the higher non-response from the boys could be that they spent longer on other questions or the other task paired with the dice task or that they gave up more easily.

Figure 5.19 Proportions of standard three girls and boys who did not respond to each question.

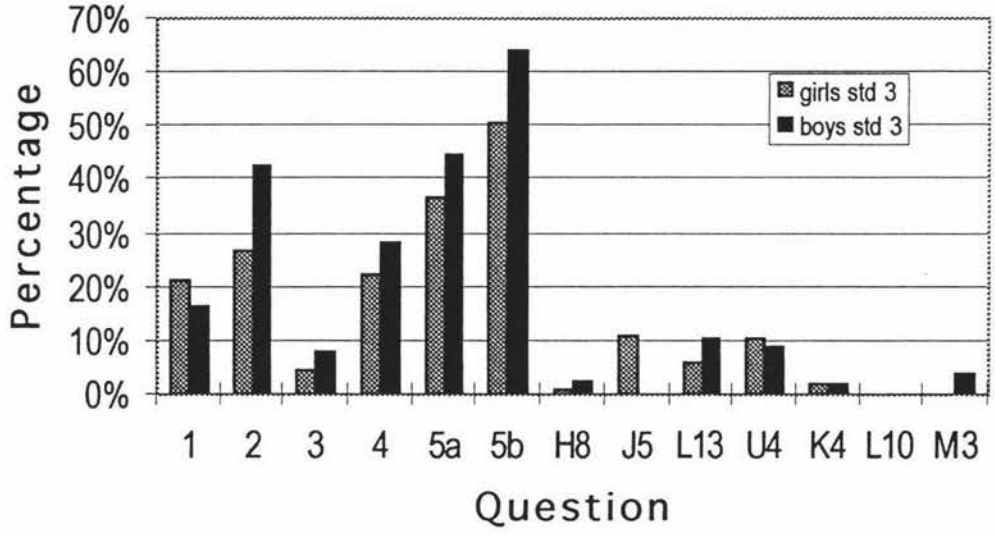
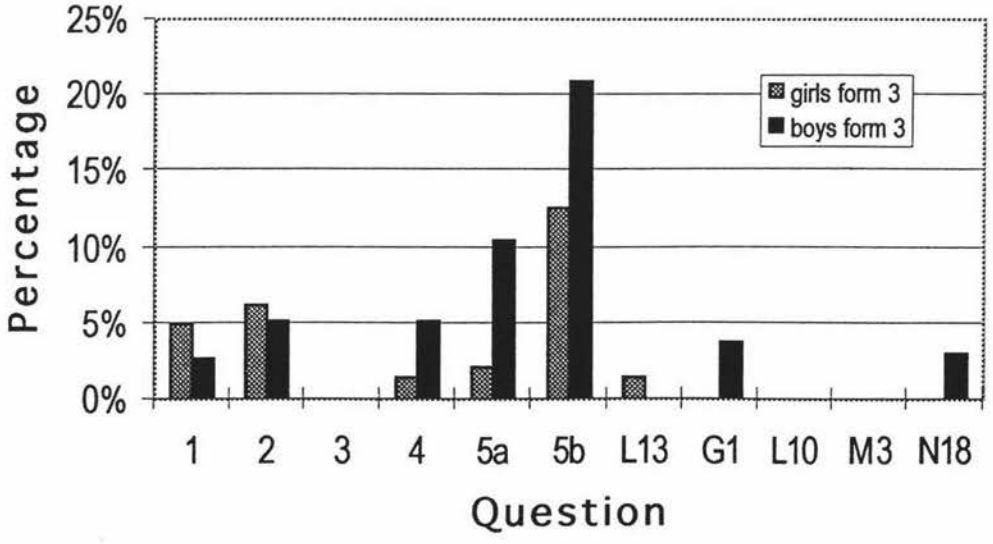


Figure 5.20 Proportions of form three girls and boys who did not respond to each question.



Contrary to studies found in the literature the majority of questions examined here were correctly answered by a higher proportion of female than male students, and were also more likely to be answered by the females. This pattern did not appear to differ depending on type of question.

5.8.2 Ethnic differences.

For this analysis students were categorised into two groups: the *majority* ethnic group and the *minority* ethnic group. The majority ethnic group comprised those students who identified themselves as only being "*New Zealand Pakeha/European*" or "*Other European*". The minority group included all other students including those who partially identified themselves as "*New Zealand Pakeha/European*" or "*Other European*". A comparison of these two groups is appropriate, because those students who identified themselves as Pakeha or of European extraction, are the majority of the population of students as well as the majority of teachers and examination writers. Thus the education system might be said to biased towards the majority students.

In general, a higher proportion of the students in the majority group achieved a maximum score than the minority group (see Figure 5.21 and Figure 5.22). The differences favouring the majority group were statistically significant by chi-square at the standard three level for question one in dice, H8, U4, and L10, and at the form three for question N18. The maximum difference favouring the minority group in both the dice task and the multiple-choice tasks was eight percent (question two, form three and L13, form three).

Figure 5.21 Proportions of standard three students in the minority and majority ethnic groups who scored the maximum score.

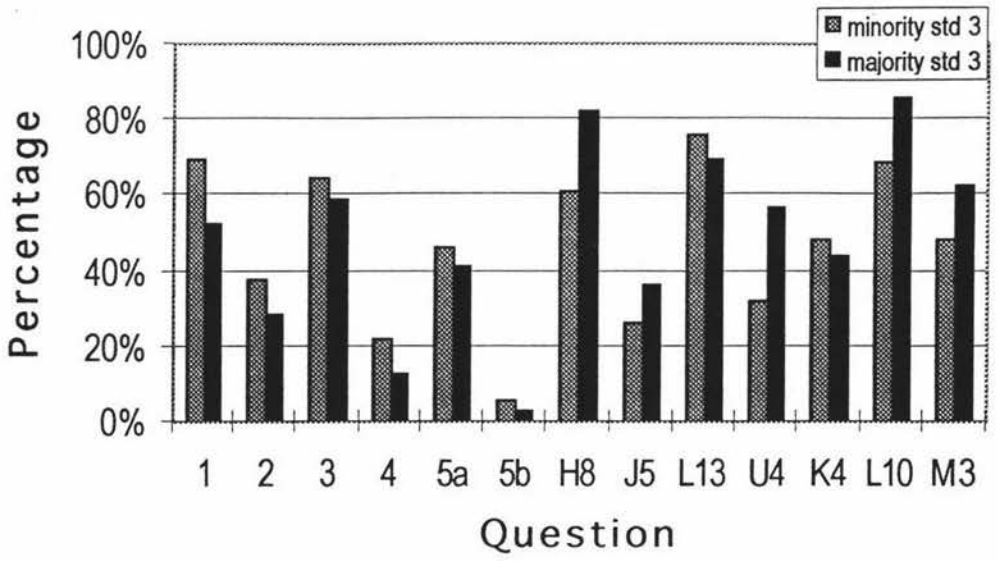
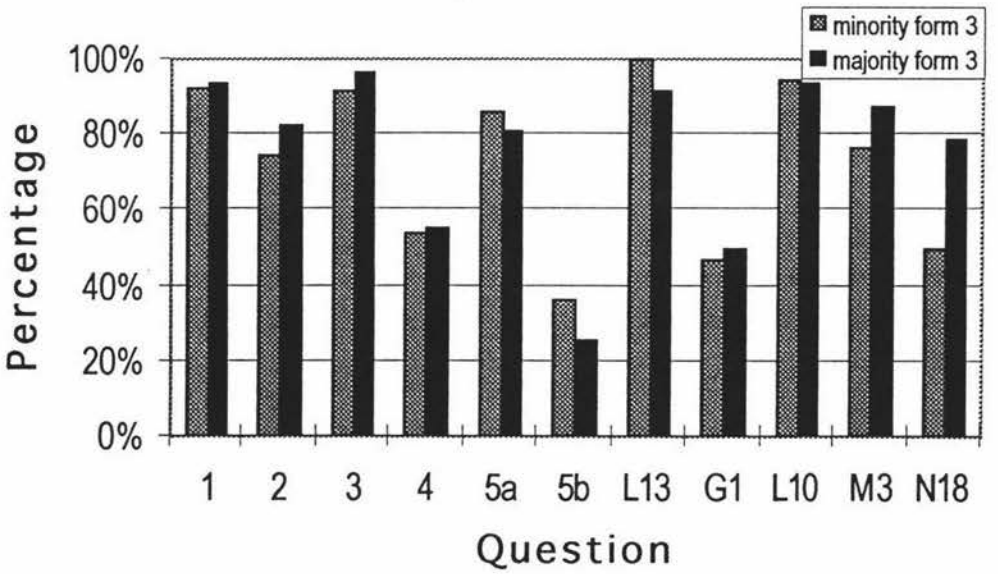


Figure 5.22 Proportions of form three students in the minority and majority groups who scored the maximum score.



For each of the dice questions, a higher proportion of the minority standard three students gave no response, whereas at the form three level a higher proportion of majority students gave no response to the question (see Figure 5.23 and Figure 5.24). Only the difference between the proportion of non-response of the minority and majority students for question five (b) was statistically significant by chi-square. For

the short-answer question U4, a statistically significant higher proportion of the minority students at the standard three level did not respond to the question.

For most of the multiple-choice questions, with the exception L13, the proportion of students who did not respond to the question was less than four percent. While this difference was low, it was generally a greater proportion of the majority students who did not respond to the question.

Figure 5.23 Proportions of standard three students in the minority and majority ethnic groups who did not respond to each question.

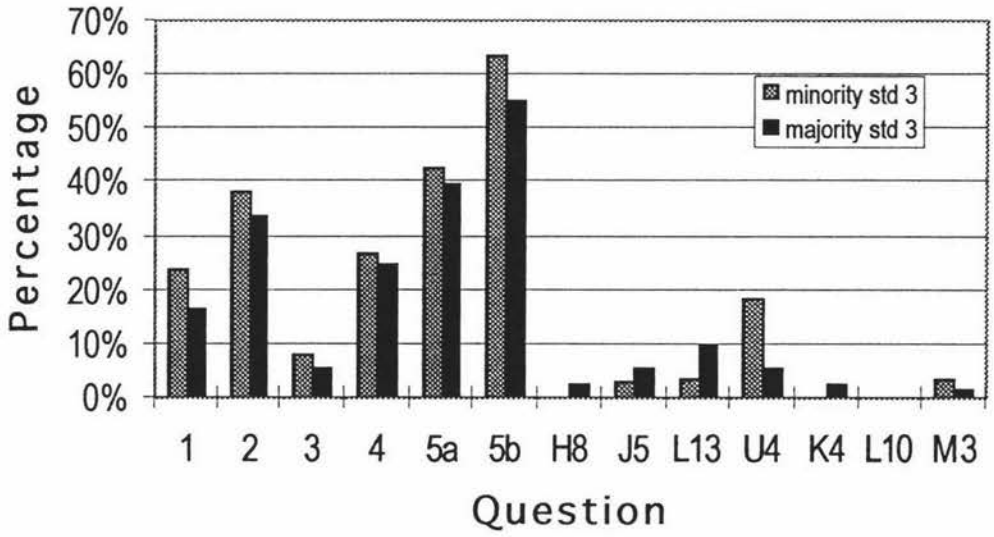
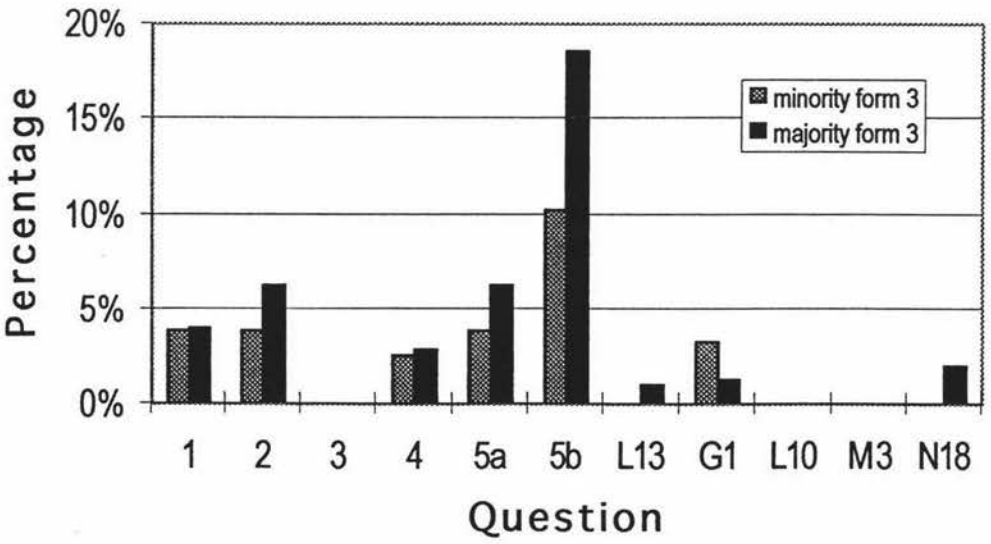


Figure 5.24 Proportions of form three students in the minority and majority ethnic groups who did not respond to each question.



It is interesting that for the standard three students, the questions that asked students to supply their own answer had a higher proportion of minority students not responding to the question; but for the multiple-choice questions a higher proportion of the majority students did not respond to the question. At the form three level, however, a higher proportion of the majority group did not respond to questions requiring a student-generated response. For the multiple-choice questions, the differences were all small, as was the overall non-response.

For standard three students a greater proportion of the majority students correctly answered the question for most of the multiple-choice questions and all the questions requiring the students to make a response. At the form three level this pattern had changed so that around half of the questions had a difference favouring the majority students, and the other half favoured the minority students. Four of the six questions with a difference favouring the form three minority students, that is where a higher proportion of the minority students correctly answered the question, were questions requiring students to generate their own response.

It would appear from these results that the minority students are disadvantaged at the standard three level when they have to generate their own response to the questions, but by the form three level this apparent disadvantage disappears.

5.8.3 Comparisons between students from English-speaking homes and those from non-English-speaking homes.

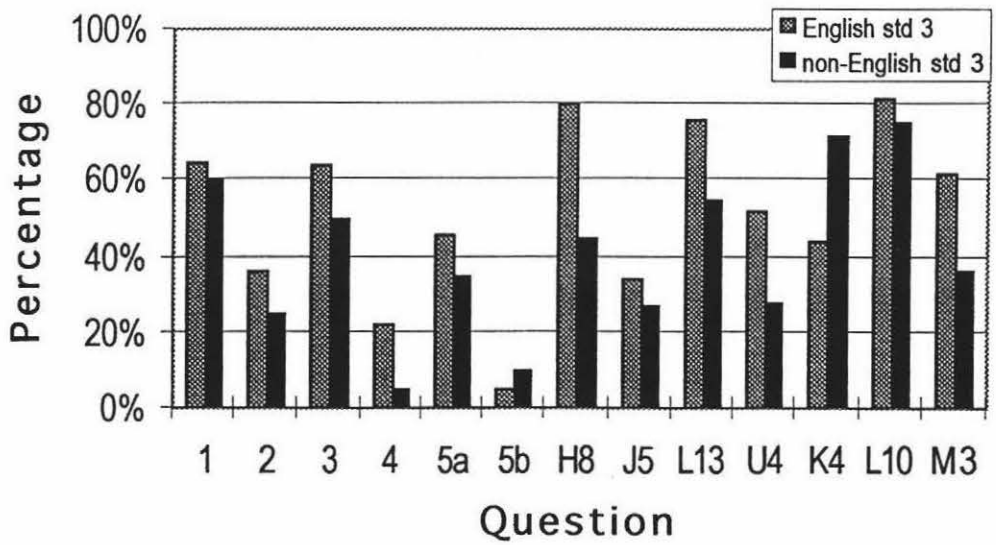
Students for whom English was not their first language may be disadvantaged by a test written in English, requiring answers to be written in English also. Students in two groups, "*the students from English-speaking homes*" and "*the students from non-English-speaking homes*" were compared (for further information on this grouping see section 4.3).

The standard three students from English-speaking homes generally had a higher proportion of students who achieved a maximum score, in both the dice task and the multiple-choice questions, than those students from non-English-speaking homes (see Figure 5.25 and Figure 5.26). The only exceptions with a higher proportion of students

from non-English speaking homes achieving a maximum score were K4 and question five (b). The difference in performance between the two groups for questions four, H8, and U4, favouring the students from English-speaking homes, was statistically significant by chi-square.

However, for the dice task at the form three level, a higher proportion of the students from non-English-speaking homes achieved the maximum score for all questions except question five (a). None of the differences in proportions of form three students achieving maximum score on the dice task were statistically significant, even though the difference favouring the non-English-speaking group for question four was thirteen percent. The multiple-choice questions at the form three level showed no consistent pattern, although questions G1, M3, and N18 all have statistically significant differences favouring the students from English-speaking homes. The maximum difference favouring the students from non-English-speaking homes, in proportions gaining the maximum score, for the dice task was thirteen percent (question four, form three) and for the multiple-choice tasks was twenty-eight percent (K4, standard three).

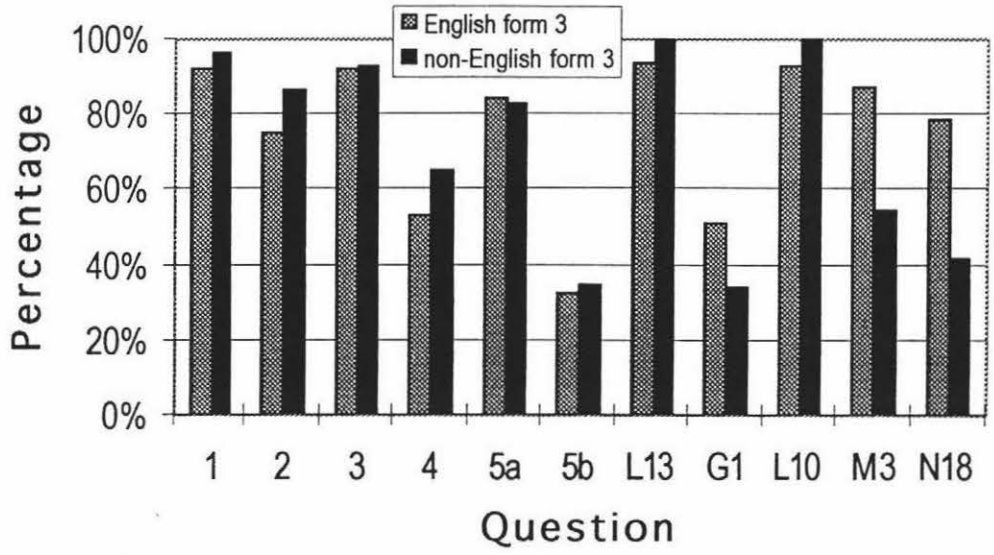
Figure 5.25 Proportions of standard three students from English-speaking and non-English-speaking homes who scored the maximum score.



For question two of dice, the most commonly given correct answer was to state that the changed numbers were all even. For this question, a higher proportion of the students

from non-English-speaking homes (twenty-eight percent versus thirteen percent of students from English-speaking homes), gave a correct answer to question two in dice that did not concentrate on the evenness of the question, but rather detailed some other pattern in the changed numbers. Since the most common answer was to recognise the fact that all the changed numbers were even, it could be said that a higher proportion of the students from non-English-speaking homes answered with an uncommon or untypical answer.

Figure 5.26 Proportions of form three students from English-speaking and non-English-speaking homes who scored the maximum score.



For all the dice questions, a higher proportion of the standard three students from non-English-speaking homes gave no response to the question. At the form three level however, a higher proportion of the students from English-speaking homes gave no response to the question. For the multiple-choice and short-answer questions it was usually a greater proportion of the students from non-English-speaking homes who missed doing the question; and where this pattern differed, the proportions for each group were similar. Only the differences for questions two, at the standard three level, and G1, at the form three level, were statistically significant by chi-square.

Figure 5.27 Proportions of standard three students from English-speaking and non-English-speaking homes who did not respond to each question.

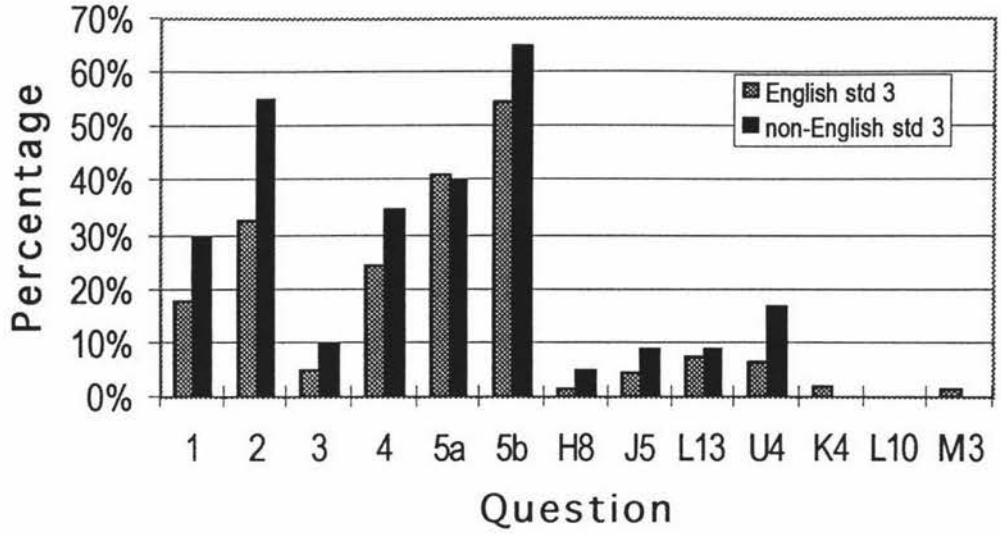
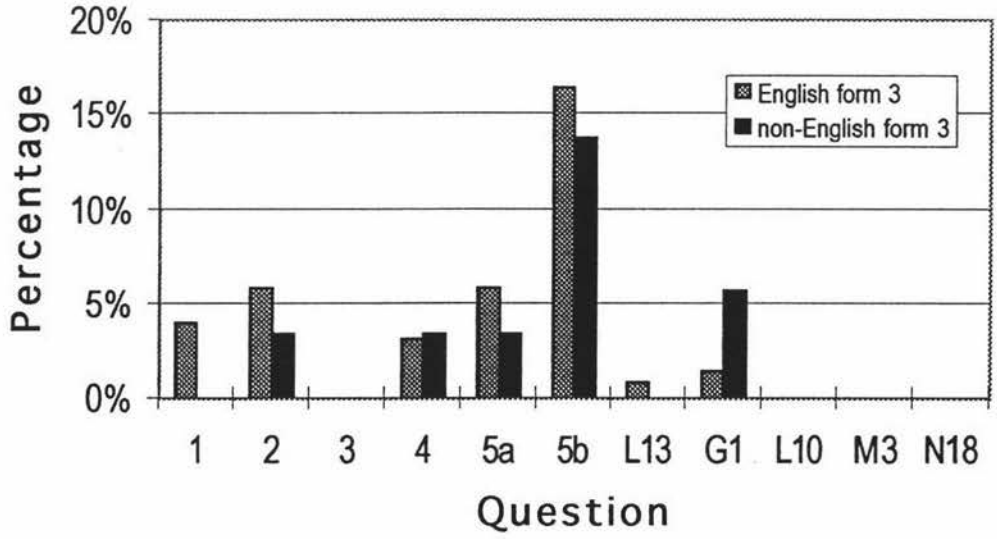


Figure 5.28 Proportions of form three students from English-speaking and non-English-speaking homes who did not respond to each question.



It would appear from these results that the students from non-English-speaking homes are disadvantaged at the standard three level but by the form three level this apparent disadvantage disappears when they have to generate their own response to the questions. However the form three students from non-English-speaking homes appear to be disadvantaged by multiple-choice questions. This may be because the multiple-

choice questions require greater decoding in order to choose the correct answer and have a greater potential to seduce students into selecting an incorrect answer.

5.8.4 Differences between students from differing socio-economic status groups.

Socio-economic status can impact on education because students from families with a low socio-economic status are unlikely to have available a range of materials or equipment to enhance out-of-school learning experiences. Schools in areas where the majority of students come from families with a low socio-economic status may also have limited extra funds for materials and equipment. Students were divided into three groups for this analysis, those of low, those of middle, and those of high socio-economic status.

The group of students from high socio-economic homes generally had a higher proportion of students who achieved a maximum score than those students from the middle or the low socio-economic groups (see Figure 5.29 and Figure 5.30). A number of the differences in performance that favoured the high socio-economic group were statistically significant, those for questions two and L10 at the standard three level, and those for questions three, five part a, and G1 at the form three level. Question H8 at the standard three level had a statistically significant higher proportion of students from the middle socio-economic group gaining the maximum score than those students from the other two groups. There were small differences, less than two percent, favouring the low socio-economic group for questions K4 at the standard three level, and M3 and dice question one at the form three level.

In the cases in the dice task where there were two marks, the socio-economic group with the lowest proportion attaining two marks had the highest proportion gaining one mark. Thus much of the differences presented were in whether the students attained one or two marks for their answer.

Figure 5.29 Proportions of standard three students from each socio-economic group who scored the maximum score.

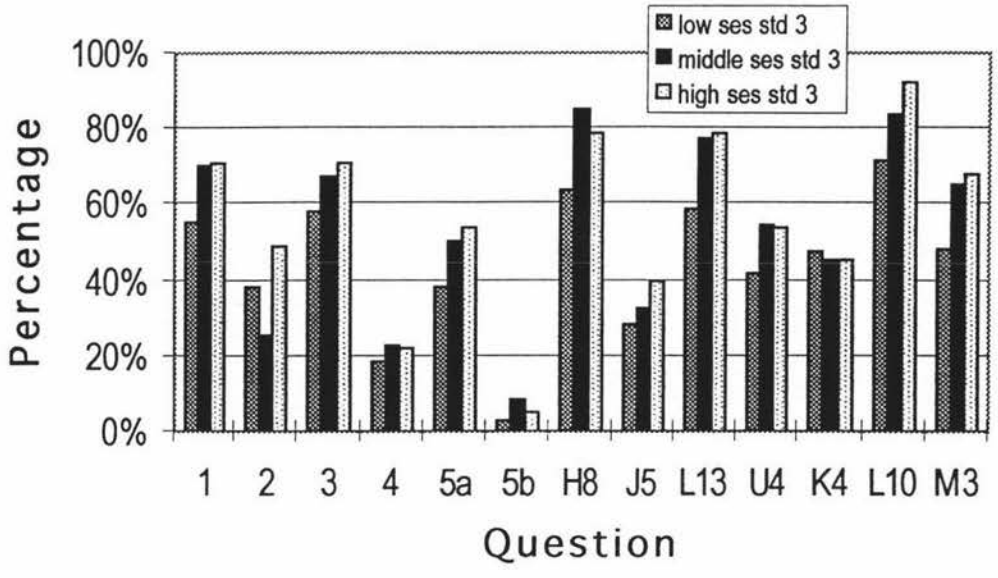
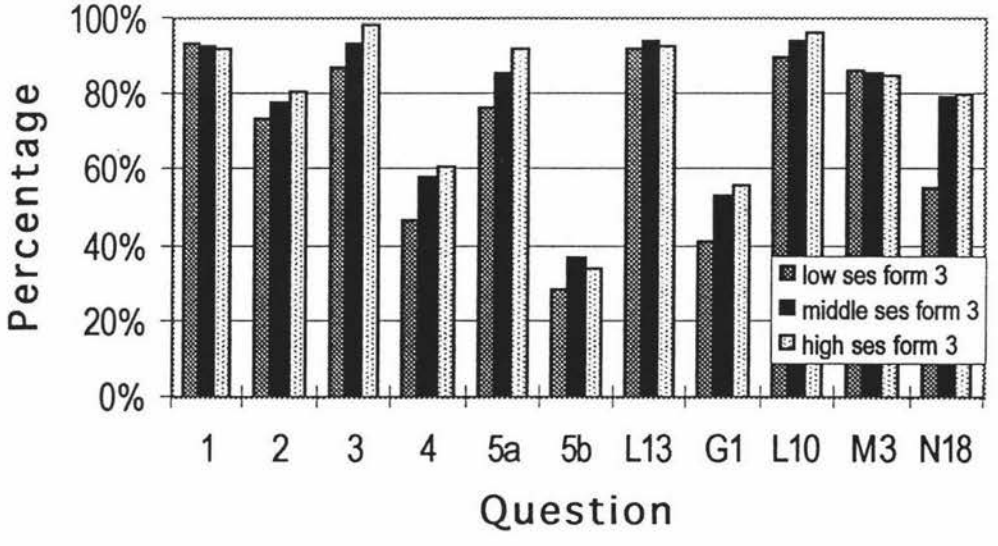


Figure 5.30 Proportions of form three students from each socio-economic group who scored the maximum score.



For the multiple-choice and short-answer questions it was usually a greater proportion of the students from low socio-economic homes who did not respond to the question (see Figure 5.31 and Figure 5.32). For most of the multiple-choice questions with a higher proportion of middle or high socio-economic students who did not respond to the question, the difference was less than three percent. For the dice task, there was no

consistent pattern in the proportions of each socio-economic group who missed doing the question at the form three level, but for most questions at the standard three level it was generally the low socio-economic group who had the highest proportion of students who missed doing the question.

Figure 5.31 Proportions of standard three students from each socio-economic group who did not respond to each question.

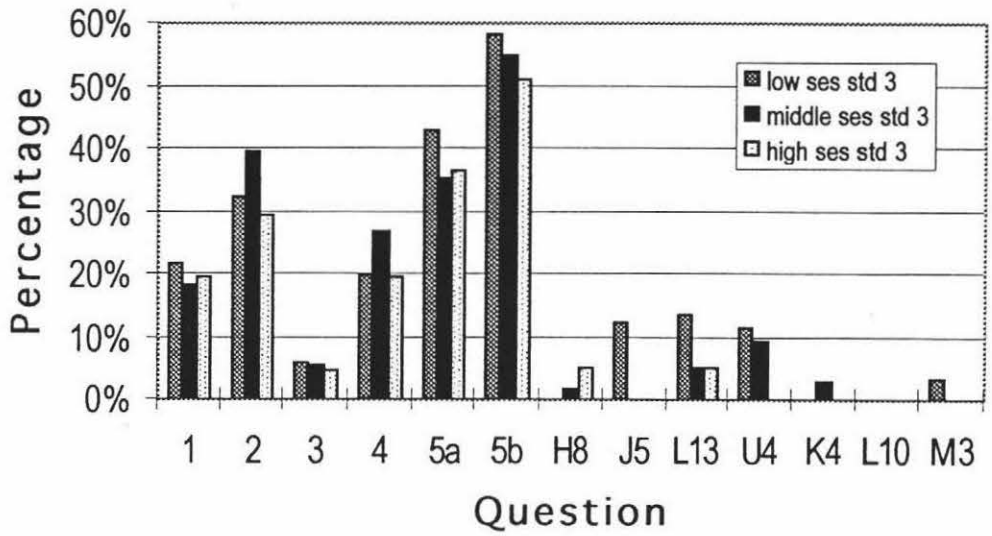
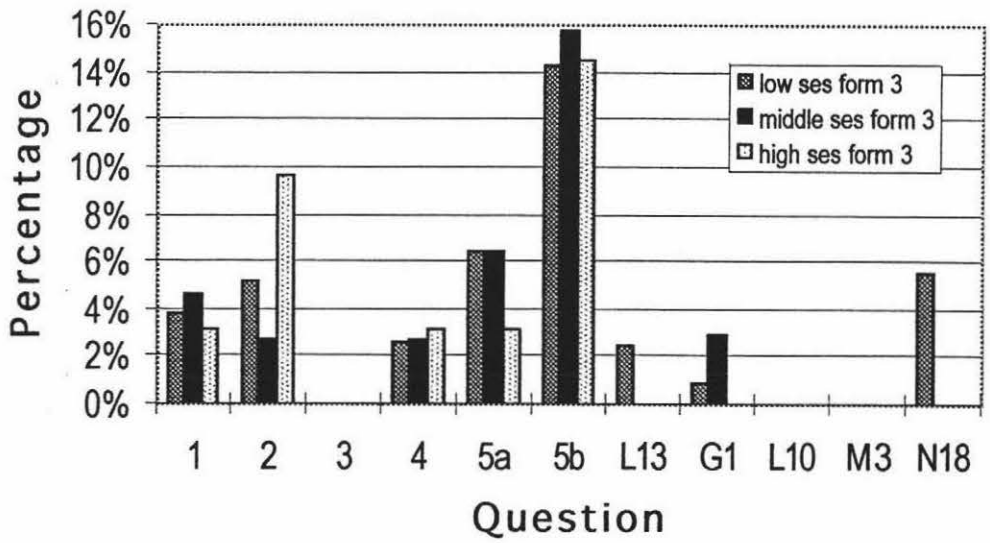


Figure 5.32 Proportions of form three students from each socio-economic group who did not respond to each question.



In summary, a lower proportion of the low socio-economic students, than of the higher groups scored maximum marks in both the multiple-choice and the dice task. However, it is notable that by the form three level there were no great differences in the proportions of students in each of these groups who did not respond to the questions. Where there were differences greater than five percent at the standard three level for the dice task, half of these differences were where the middle group had a higher proportion who did not respond to the question. It appears that the students in the low socio-economic group were prepared to attempt each question which is positive for a test which was low stakes for both participants and their teachers.

5.8.5 Personal value of mathematics

The performances of groups of form three students were examined, where the groupings were based on their responses to two statements indicating how much students valued mathematics. The two statements were:

(a) I think it is important to do well in mathematics at school

and

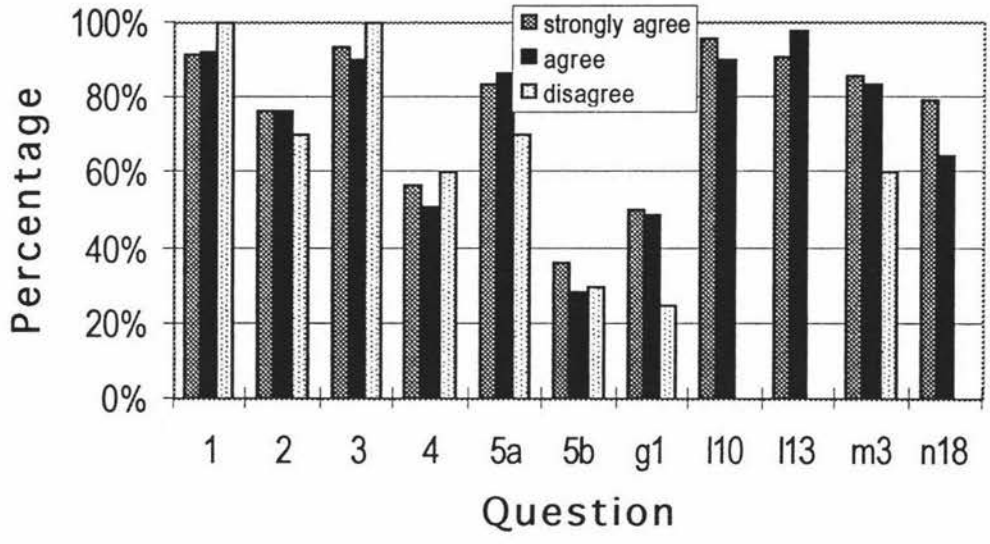
(b) Mathematics is important to everyone's life.

The three groups were those students who *strongly agreed, agreed,* and *disagreed* with these statements. Note that only a small percentage (three percent and five percent respectively) of students disagreed with each of these statements.

For the statement *I think it is important to do well in mathematics at school,* there is no consistent pattern when we examine which group had the highest proportion of students obtaining the maximum score in the dice task. For several of the multiple-choice questions, L10, L13, and N18, there were no students who disagreed with the statement. (Recall from chapter 4, that because of rotated booklets, the group of students examined for the dice task does not completely overlap with the multiple-choice tasks.) For all of the multiple-choice questions, except L13, a higher proportion of the students who strongly agreed with the statement achieved the maximum score for the task. This result tends to confirm the idea that those who think it is important to do well would therefore do well until we notice that for half the dice questions, a higher proportion of the

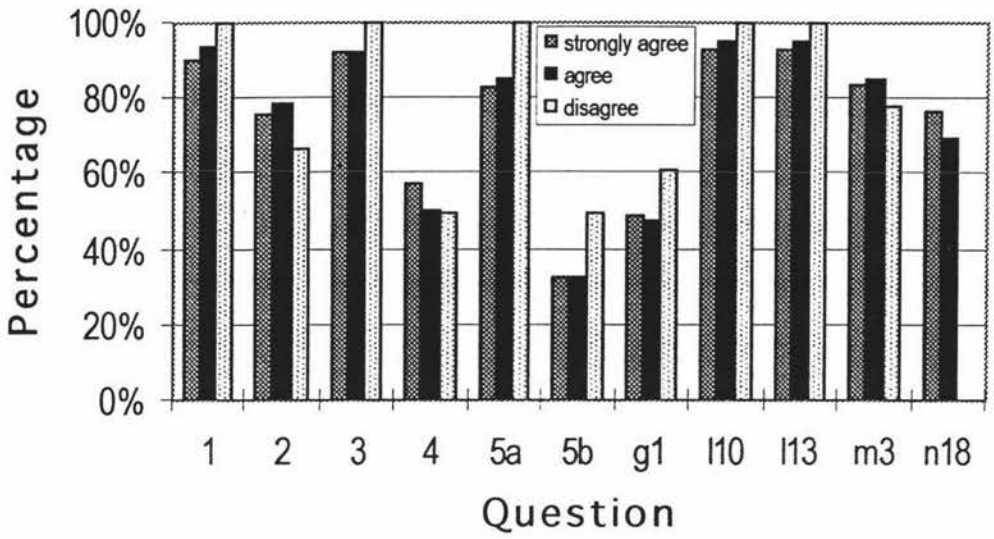
students who disagreed with the statement obtained the maximum score. While none of these three differences were statistically significant, the differences between those who agreed and those who disagreed were as large as ten percent for questions three and four. Maybe, as is suggested by some advocates of performance assessment, some of these students found the dice task more motivating than the multiple-choice tasks.

Figure 5.33 Proportions of students gaining the maximum score who agreed and disagreed with the statement *I think it is important to do well in mathematics at school*



For the statement *Mathematics is important to everyone's life,* it was generally a higher proportion of the students who disagreed with this statement who gained the maximum score for many of the multiple-choice and dice questions. The exceptions were questions two and M3 which favoured the students who agreed with the statement, and questions four and N18 which favoured the students who strongly agreed with the statement. None of the differences were found to be statistically significant by chi-square despite the difference between the disagreed group and the strongly agreed group being seventeen percent on question five (a).

Figure 5.34 Proportions of students gaining the maximum score who agreed and disagreed with the statement *Mathematics is important to everyone's life*



For the statement *I think it is important to do well in mathematics at school,* there was generally a higher proportion of the group of students who disagreed with the statement who did not respond to each of the dice questions. Despite the large differences between proportions of students in each group not responding, particularly for question five (b), none were statistically significant by chi-square, probably because of the small number of students who disagreed with the statement. Very few students did not respond to the multiple-choice questions.

In comparison to the earlier statement, a higher proportion of the students who strongly agreed that *Mathematics is important to everyone's life* did not respond to each of the dice questions. Two questions, five (a) and five (b), were found to have statistically significant results when comparing the proportions of students in each group who did not respond to the question.

Figure 5.35 Proportions of students who agreed and disagreed with the statement *I think it is important to do well in mathematics at school* that did not respond to the packaging and multiple-choice questions.

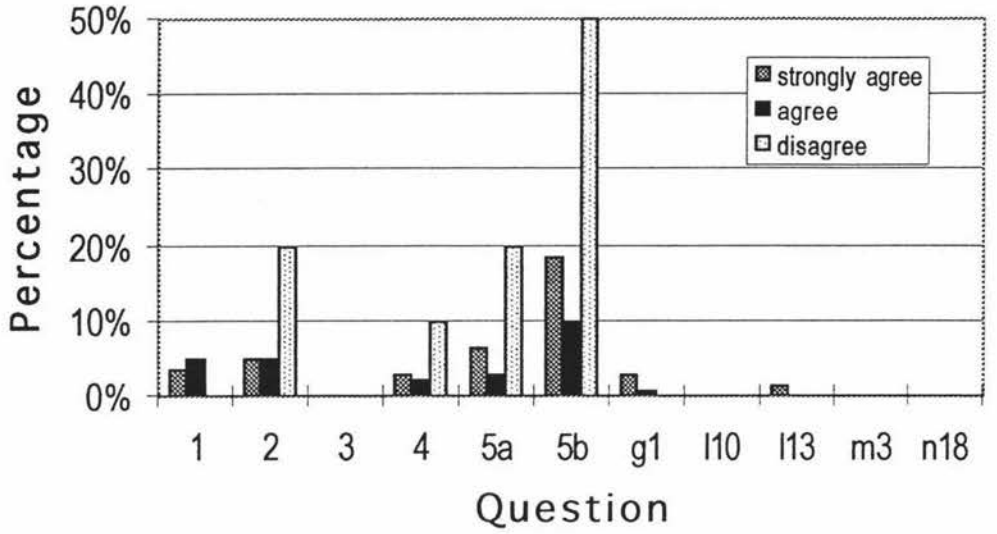
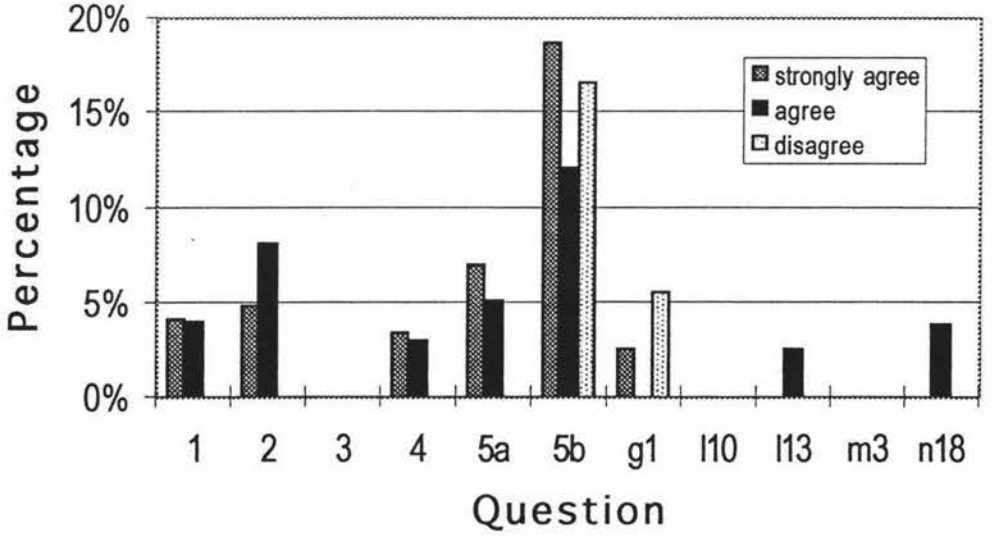


Figure 5.36 Proportions of students who agreed and disagreed with the statement *Mathematics is important to everyone's life* that did not respond to the packaging and multiple-choice questions.



The higher proportions of students, in the group who disagreed with each of the statements, who achieved the maximum score on many of the dice task questions leads us to speculate why this might be. Possibilities include the idea that the performance

assessment tasks could have a motivating effect on some students. We may also wonder at the reasons for students disagreeing with each of the statements, and what effect this has on their behaviour. However, it is possible that the small number of students in each group who disagreed with the statements caused the results to be exaggerated when they are transformed into proportions.

5.9 Summary

The purpose of the dice task was to measure students' number sense, ability to apply an algorithm, and ability to recognise and explain a pattern. It also revealed the students' abilities to carry out a systematic data gathering exercise, to record and to summarise their results. There were eight multiple-choice and one short-answer question found to be associated by content analysis with questions in the dice task. According to the mathematics curriculum (Ministry of Education, 1992) all of the questions were appropriate for the majority of form three students but questions five (both parts), K4, and M3 might cause difficulties for many standard three students. This is generally evidenced in the results, although questions four, J5 and G1 may have caused more difficulties for students than expected. Also standard three students did better than might have been expected on question M3.

The dice task gives us more information than the multiple-choice and short-answer tasks about the abilities of students to perform a systematic experiment, record their results, and summarise their results in the table provided. It also gives us more information about how students can integrate their knowledge from a number of different aspects of the mathematics curriculum, particularly when explaining the reason for the most commonly occurring number.

Many of the multiple-choice questions identified as similar in content to the dice task questions were also found to be similar by the chi-square analysis. That is, most of the students who correctly answered the dice question also correctly answered the multiple-choice question, and similarly, most of the students who gave an incorrect answer to one also gave an incorrect answer to the other. These similarities might lead us to believe that the multiple-choice questions tell us the same thing about the students as the dice

task, but although the chi-square test revealed some associations between the results of the dice task and the multiple-choice tasks associated by content, there were still students who performed differently on the two associated tasks. Also, there were no associated tasks with question four on content, and although M3 and N18 appeared to have the same content as question five (b), they actually exposed different aspects of student abilities and beliefs.

The girls in this study achieved better than the boys and were more likely to attempt the questions particularly for the dice task, but they also achieved better on the multiple-choice and the short-answer questions so this is not an artefact of the extended response questions.

Minority students, those from non-English-speaking homes, and low socio-economic status students did not fare as well as their peers from the majority group, English-speaking homes and middle or high socio-economic groups. This is particularly so for the multiple-choice questions and at the standard three level for the dice task. However the minority form three students and those form three students from non-English-speaking homes achieved just as well or better than those from the majority group and those from English-speaking homes on the dice task. This is not the case for the low socio-economic students who achieved just as poorly at this level as they do at the standard three level. However it is good to know that these low socio-economic status students are generally just as prepared to *have a go* at the question as their middle and high socio-economic status counterparts.

Surprisingly, students who disagreed with the two value statements *I think it is important to do well in mathematics at school* and *Mathematics is important to everyone's life* had a higher proportion of students who achieved the maximum score on many of the dice questions than the two groups who agreed with the statements. It is unclear whether this result is an artefact of the small number of students who disagreed with the statement or whether there is some other underlying reasons.

6 PACKAGING

6.1 The packaging task

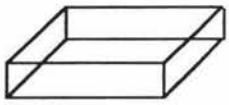
Your task:

Design different boxes which will just hold 4 plastic balls.

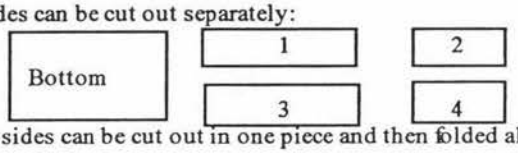
Read this before answering the questions:

The following shows what is meant by the net of a box.

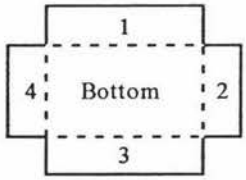
This box has a bottom and 4 sides.



The sides can be cut out separately:



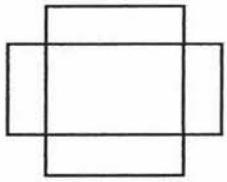
Or the sides can be cut out in one piece and then folded along the dotted lines like this:



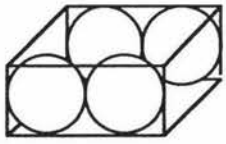
This is a net of a box.

page 2

This is the shape of a net of a box like the one which holds the 4 balls. It is not drawn to size but if it were, you could fold up the sides and make the box.



You have been given the box with the four balls just fitting in like this.



Other boxes with different shapes could be made so that the 4 balls would just fit in.

1. Use the balls to find 3 other boxes in which the 4 balls will just fit. Make a drawing of each box showing the 4 balls in it.

2. Now make a drawing of the net for each box.

3. Choose ONE of the box designs you have drawn. Take a piece of plain card. On this card draw the net of the design you have chosen. Draw it to the correct size so that if you made the box it would just hold 4 balls.

page 4

The excluded space on the presented task material is where there was white space for recording student answers. Page one was the cover sheet for the task which includes task identification information and space for student identification information.

Both standard three and form three students attempted this task. They were supplied with four plastic balls packed into a box; blu-tack to stop the balls from rolling around; some thin card to make a package for the balls; a compass; a 30 cm ruler; two pieces of thick card to help measure the balls; scissors; sellotape; and paper clips. The list of equipment was given at the beginning of the task.

According to the TIMSS Performance Assessment Coding Committee (1994), this task was intended to measure the student's sense of spatial relations. This was evidenced by the student's ability to:

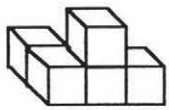
- *visualize or work out different arrangements of objects in packaging*
- *visualize the 2-dimensional net for a 3-dimensional object*
- *translate a 2 or 3-dimensional sketch into its corresponding 2-dimensional net*
- *translate the sketch and actual size of the objects (balls) into an actual size net.*

6.2 The multiple-choice tasks associated with the packaging task.

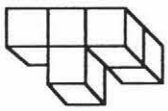
There are only three tasks in the multiple-choice section of the TIMSS study, and no short-answer questions, that correspond to the content in the packaging task. One of these tasks, B11 has not been released at the time of writing and so cannot be detailed

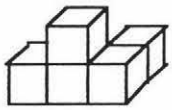
here. Task K3[1], used at both the standard three and form three level, and task L5, used at only the standard three level, required students to visualise a 2-dimensional picture as the 3-dimensional object they represent. For task K3, students needed to visualise how the shape would look after rotation. For task L5, students had to realise that the 2-dimensional representation of the 3-dimensional solid does not show the entire solid and that there are more edges in the solid than represented in the diagram.

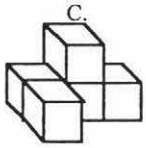
K3. This figure will be turned to a different position.

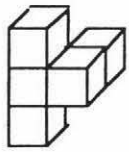


Which of these could be the figure after it is turned?

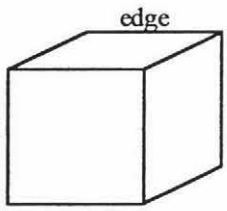
A. 

B. 

C. 

D. 

L5. This picture shows a cube with one edge marked. How many edges does the cube have altogether?



A. 6
B. 8
C. 12
D. 24

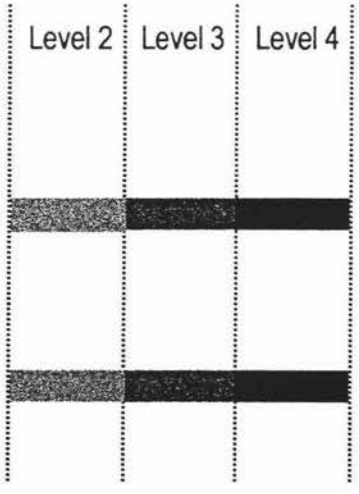
[1] H5 was the label used at the standard three level but for simplicity it will be labelled K3, its label at the form three level, throughout this document

Task B11, used at the form three level, showed a net with various potential faces identified uniquely. Students were asked to say which of four folded-up solids could be made by the net. To do this task, students needed to visualise where each face would be in relation to the other uniquely marked faces, when the net was folded into a solid.

6.3 Face validity - how do the packaging task and the associated multiple-choice tasks compare with curriculum expectations?

In examining the curriculum expectations in New Zealand, I have focussed on the expectations in levels 2 to 4 from the Mathematics in the New Zealand Curriculum document (Ministry of Education, 1992). Level 2 is the level in which most standard three or year five children are expected to be working, with some moving on to level 3. Level 4 is the level at which most form three or year nine children are expected to be working. The processes in these levels relevant to the packaging task are problem solving and communicating mathematical ideas and the relevant objectives are summarised in Table 6.1

Table 6.1 Achievement objectives from the Mathematical Processes strand associated with the packaging task[2]



Problem Solving

- use equipment appropriately when exploring mathematical ideas

Communicating Mathematical Ideas

- use their own language, and mathematical language and diagrams, to explain mathematical ideas

[2] The shading in this table represents that used in the Mathematics in New Zealand Curriculum document.

The achievement objectives from the other five strands in the curriculum document, which are relevant to the packaging task, are summarised in Table 6.2.

Table 6.2 Achievement objectives from the content strands associated with the packaging task

	Level 2	Level 3	Level 4
Measurement			
• carry out practical measuring tasks, using appropriate metric units for length, mass, and capacity	√		
Geometry			
• make, ... everyday shapes and objects	√		
• make clockwise and anticlockwise turns	√		
• design and make containers to specified requirements;		√	
• model and describe 3-dimensional objects illustrated by diagrams or pictures		√	
• draw pictures of simple 3-dimensional objects		√	
• design the net and make a simple polyhedron to specified dimensions			√

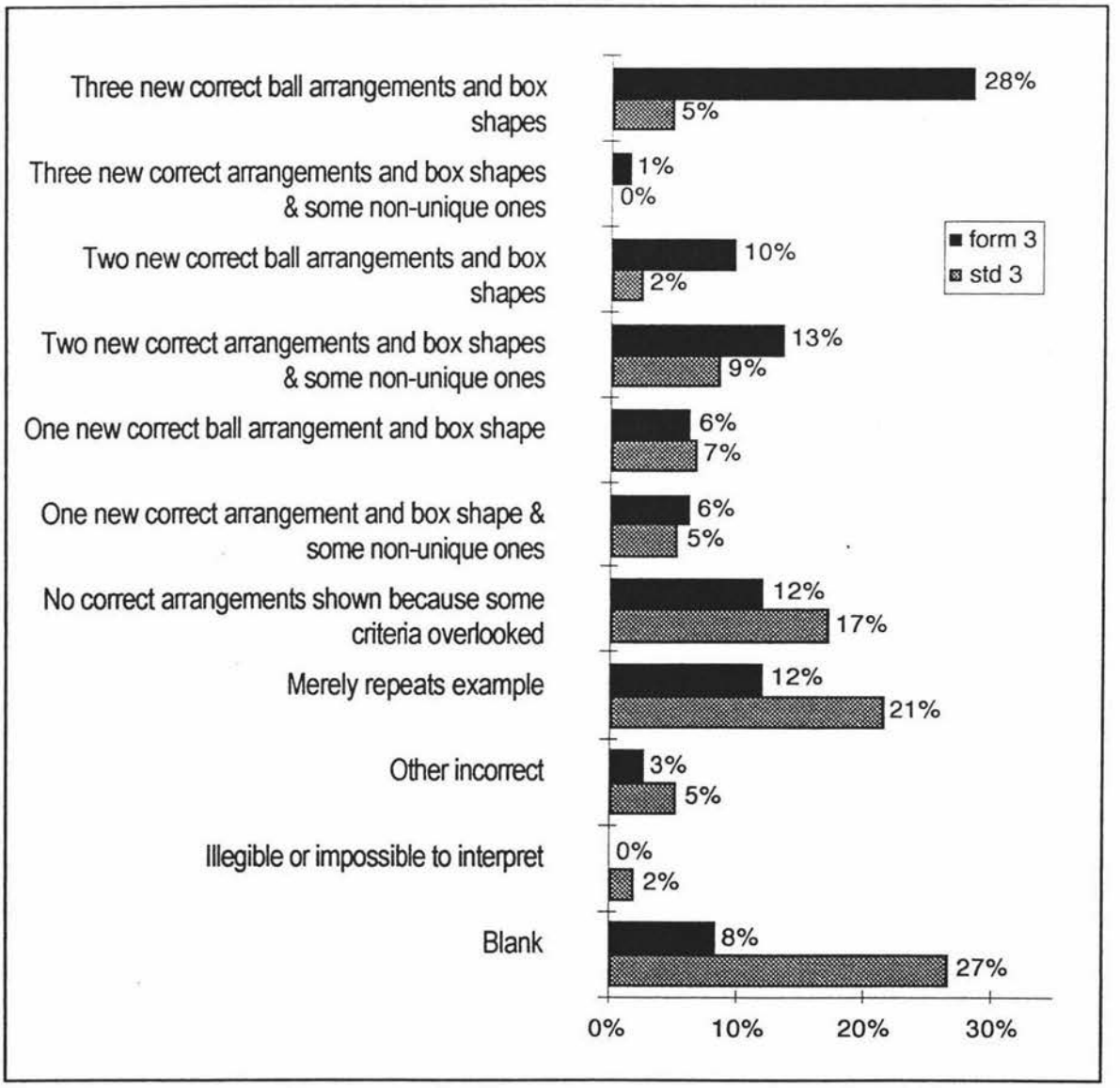
The packaging task expectations match the curriculum expectations well for most form three students but many standard three students may have had little exposure to the requirements of this task. One of the suggested learning experiences at level 2 is to explore common cardboard boxes and packages and the nets used to make them, so students at the standard three level may have experience with making the boxes up given a net. It is unclear from the objectives whether the standard three students would have the required content exposure to do the multiple-choice tasks. There seems to be a concentration on plane shapes in terms of rotation during level 2, although this can only be inferred from the suggested learning experiences. We could reasonably expect most

form three students to do the associated multiple-choice tasks assigned to them based on the objectives and suggested learning experiences in the curriculum.

6.4 Results of the packaging task

For **question one** to be completed correctly, students needed to draw three unique ball arrangements and box shapes. All four balls needed to be shown in each drawing, in a tightly packed formation.

Figure 6.1. Proportions of student responses to question one of the packaging task.



The different possibilities presented by the students included:

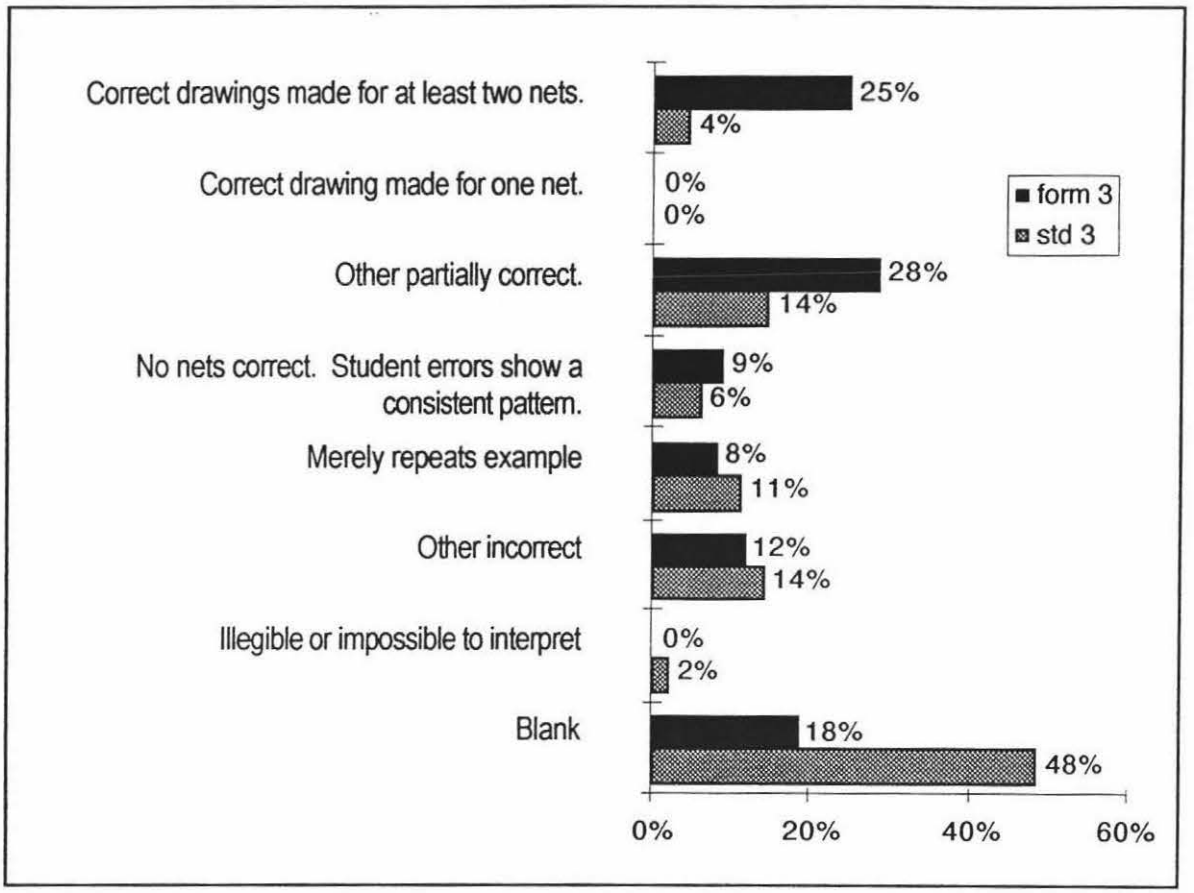
- two rows of two balls stacked vertically in a rectangular-based box;
- four balls in a line in a rectangular box;
- four balls in a column in a square-based box;
- four balls in a column in a cylindrical box;
- a single ball placed on top of a row of three balls in a box with L-shaped sides;
- four balls arranged in a box with an L-shaped base;
- four balls arranged in a T-shape in a triangular-based box;
- a single ball placed on top of the centre ball in a row of three balls in a box with T-shaped sides;
- four balls arranged as a tetrahedron in a pyramidal box;
- four balls arranged as a rhombus in a rhomboid box;
- four balls arranged in a circle with a circular-based box;
- four balls arranged in two rows of two balls with one row offset from the other so that the base is a squared Z;
- four balls arranged in a box with a U-shaped base, the U has sloping sides.

Some students gave more information than was asked of them by providing measurements on their sketches.

The most common errors for question one were drawing a diagram of the box and balls provided with the task rather than creating a different shaped box and ball arrangement; and not showing the balls in a tightly-packed formation in their diagrams.

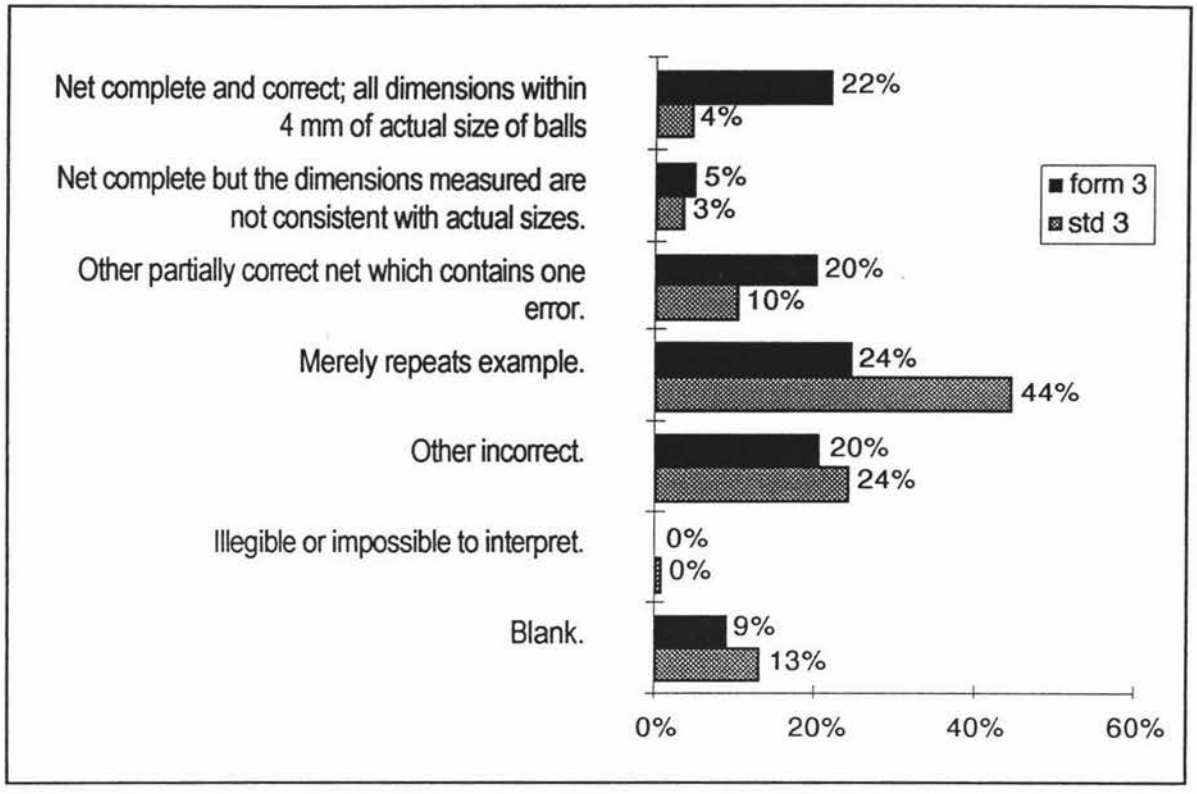
For **question two** to be completed correctly, students needed to draw at least two nets consistent with the ball arrangements shown in question one. The nets needed to show correct proportions and shapes required to appropriately constrain the balls in a tightly-packed formation.

Figure 6.2 Proportions of student responses to question two of the packaging task.



For **question three** to be completed correctly, students needed to construct a net from a single piece of card paper consistent with a net drawn in question two. The net needed to have correct dimensions, within 4 mm of actual size, and had to include a base with side flaps which would constrain the balls in a tightly-packed formation.

Figure 6.3 Proportions of student responses to question three of the packaging task.



Despite the range of answers given in question one, the range of different shapes successfully constructed in question three was much more limited. The most common correct response to this question, for both standard three and form three students, was a net for the four balls in a line with a rectangular-shaped base. Only three other shapes of boxes were selected and correctly constructed by form three pupils. They were:

- two rows of two balls stacked vertically in a rectangular-based box;
- four balls in a column in a square-based box;
- four balls arranged in a box with an L-shaped base.

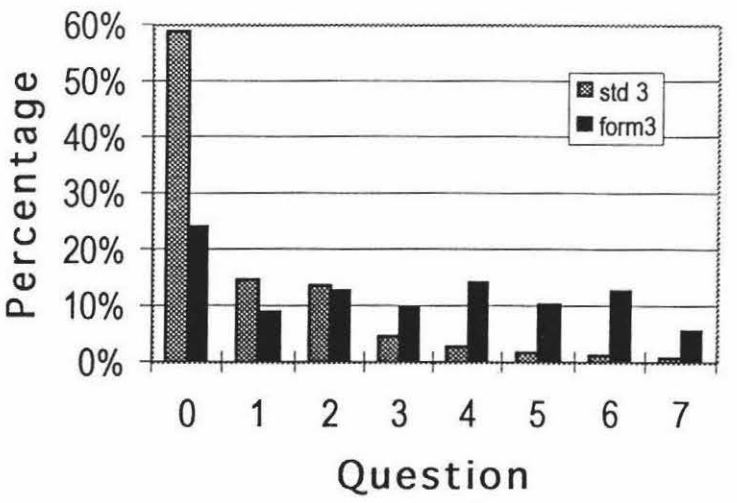
Only the first two shapes listed above were selected and correctly constructed by standard three students. One student at each level attempted to make a circular box from one piece of card.

The errors for question three were similar to those for question one in that students often constructed a replica of the box supplied with the task or did not construct an appropriately sized box to fit and constrain four balls in a tightly packed formation. It

was also fairly common for students to cut out each side of their proposed box separately rather than construct a net from one piece of cardboard. This method of box construction from five separate pieces of cardboard was suggested in the information section and although this method was not referred to again in the body of the task some students may have thought that this was the preferred method.

It is clear from Figure 6.4 that most standard three students were not able to demonstrate a good sense of spatial relations with this task. However, this may not necessarily mean that the students had poor spatial abilities. The marking criteria which required that sketches of box shapes show the balls in a tightly-packed formation, and the tendency for students to repeat the given example rather than find unique box shapes and ball arrangements, may have contributed to this poor demonstration of abilities.

Figure 6.4 Graph of score totals for the packaging task

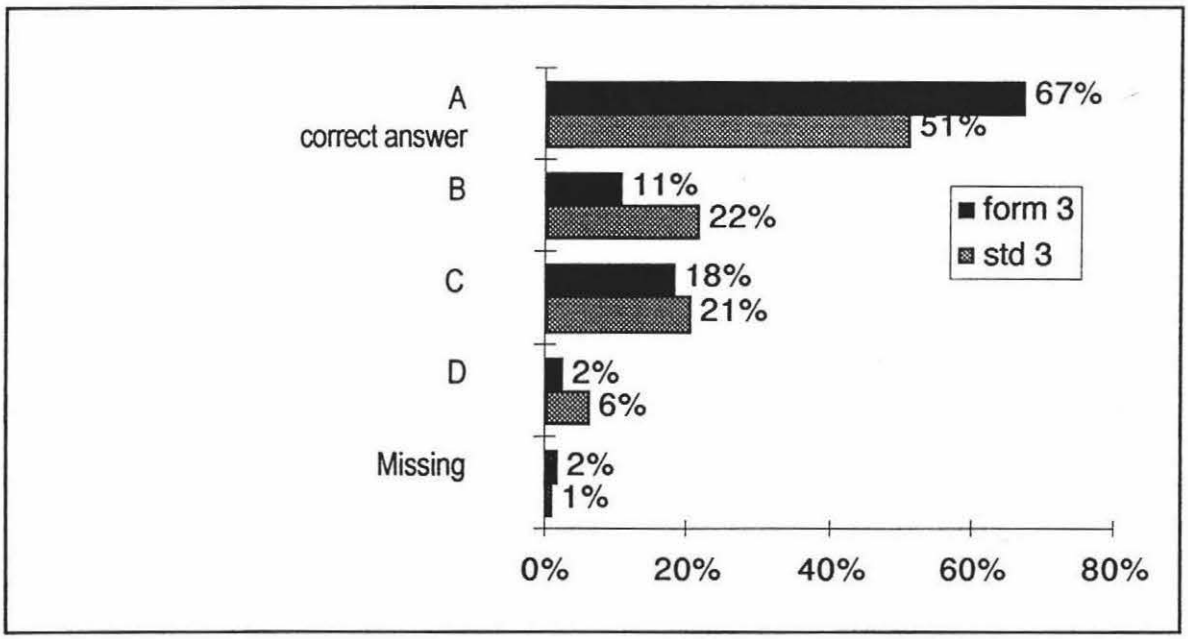


6.5 Results of the multiple-choice tasks

For the analyses of the multiple-choice questions, the information presented in this chapter draws on the performances of all students in the performance assessment component of the study who had the opportunity to do each question. For comparison purposes with the packaging task, we might prefer to look at only those students who did the multiple-choice task and the packaging task, but because this was such a small group, any differences are exaggerated when the numbers are transformed into percentages. Instead, the results presented here are the values from the larger group of all respondents, as these have a similar distribution and are thus representative of the smaller group, all respondents who did both the packaging and the multiple-choice question.

For **question K3** the correct answer is option A. Half of the standard three students and two-thirds of the form three students correctly identified the appropriate shape as the rotated solid.

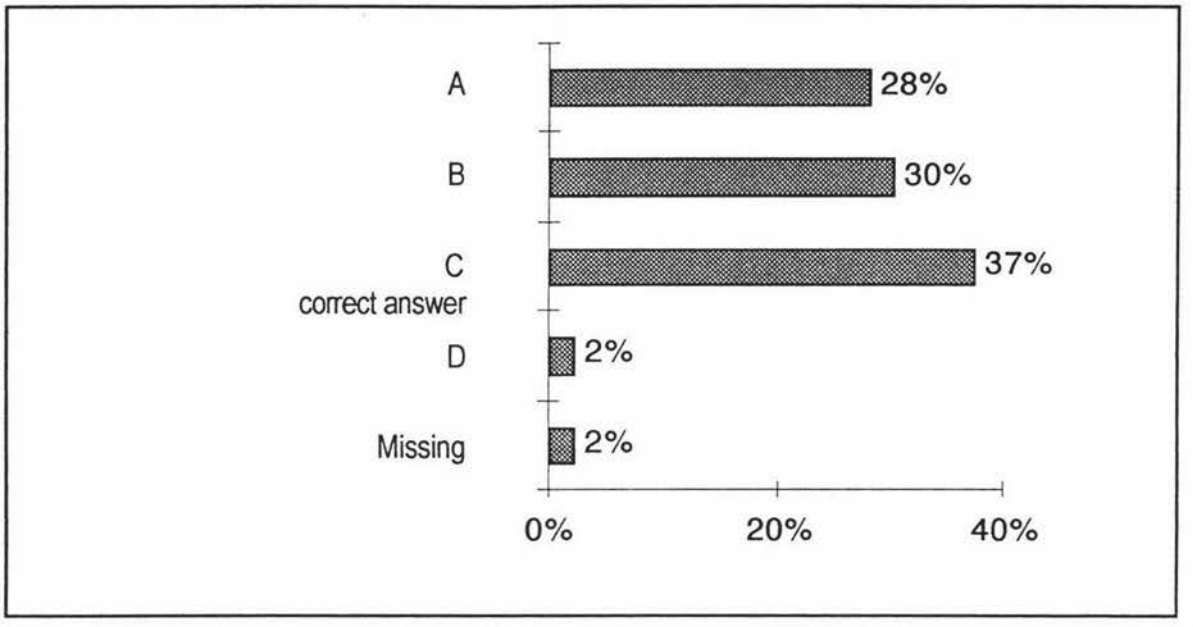
Figure 6.5 Proportions of student responses to K3.



For **question L5** the correct answer is option C. The distribution of the results over options A, B and C is very similar to the distribution expected by chance if option D

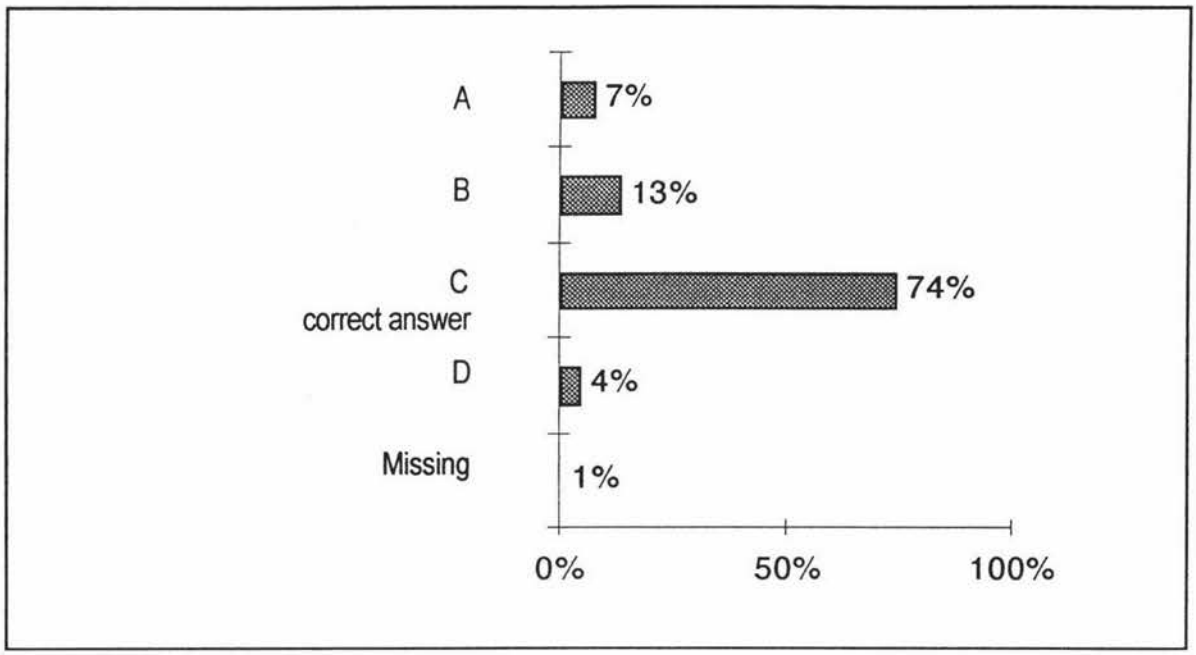
was excluded. Because the answer "9", which would be given if the student counted up the edges shown in the picture, is not presented as an option, students were forced to select some other answer. If we asked students the reason why the option they selected was correct, this would give us greater insight into their spatial abilities. It is not clear from the answers given whether students, even those who selected option C, recognised that the picture represented a 3-dimensional object with some edges not illustrated.

Figure 6.6 Proportions of responses to L5.



For **question B11** the correct answer is option C. Three quarters of form three students were able to correctly identify how the faces related to each other in the completed solid from the uniquely identified faces in the net.

Figure 6.7 Proportions of responses to B11.



6.6 Comparison of the results of the packaging task and the multiple-choice tasks.

The packaging task certainly gives us more information than is possible with any multiple-choice task, particularly about creativity of the students and their ability to present information and ideas pictorially. We can also find out how precise they are in their constructions, although it is not clear for those students who copied the example how precise they could be if they constructed their own net.

For both multiple-choice questions given to standard three students[3], K3 and L5, the results are independent from the results of the packaging task (by chi-square analysis). That is, those students who got the multiple-choice question correct, as often as not, did not get the packaging task question correct and vice versa. Thus the multiple-choice questions are not suitable indicators of standard three student ability with the performance task.

[3] 74 standard three students are included in the analysis of K3 with the Packaging task; and 24 standard three students in the L5 analysis.

For the form three students[4], question one of the packaging task is not independent of the multiple-choice tasks, K3 and B11, by chi-square analysis. For these multiple-choice tasks, most of the students who got it correct also got question one correct. Question two is not independent of B11 but is independent of K3, and question three is independent of B11 but not independent of K3. Those students who got B11 correct, also correctly answered question two of packaging but not necessarily question three.

B11 required students to visualise how a 2-dimensional net would look in 3-dimensions so this question is more related to questions one and two in the packaging task. Half of the students who got B11 wrong at least partially correctly answered question one of the packaging task. If only the multiple-choice question, B11, was given to these form three students, the half of the students who got the multiple-choice question wrong would have been deprived of the opportunity to show their abilities.

K3 required students to visualise how a 3-dimensional object would look after rotation. The ability to visualise 3-dimensional shapes is possibly a feature of question one in the packaging task, however because of the availability of the four balls it was not necessary for students to imagine shapes as they could construct them with the balls. Maybe question one can be seen as more closely related to K3 than the other questions in the packaging task but it is not clear why it is independent for standard three students and not for form three students. However, only a small number of form three students had the opportunity to do both K3 and the packaging task so maybe if the sample size was increased we would not see any relationship.

It is clear from this analysis that the multiple-choice tasks must be quite similar in content, not just in the same content area of the curriculum, to give a similar indication of student ability. Even then, one half of the students who got the multiple-choice question wrong were able to demonstrate some spatial visualisation skills with the performance assessment task.

---

[4] 106 form three students are included in the analysis of K3 with the Packaging task; and 19 form three students in the B11 analysis.

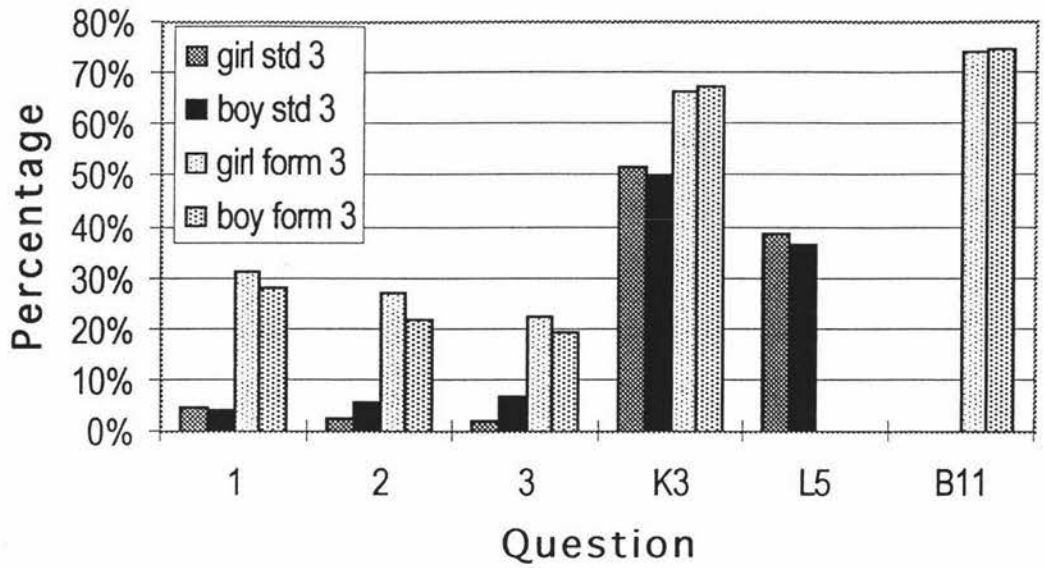
6.7 How do sub-groups of students fare in the packaging task and its associated tasks

6.7.1 Gender differences

There was no consistent overall pattern in the gender differences for these tasks. The differences between the proportions of girls and boys achieving the maximum score for each of the multiple-choice tasks was almost zero; no differences were greater than two percent, and so they were not of practical significance, and no differences were statistically significant. However, when the students who attempted the packaging task were examined, the percentage differences between the proportions of boys and girls gaining the maximum score for the packaging task are larger than those for the multiple-choice tasks despite the overall proportions being smaller (see Figure 6.8).

For the standard three students, the proportions of girls and the proportions of boys scoring the maximum score on the packaging task were less than ten percent, so the differences between these two groups were less than five percent. Standard three boys were more likely to gain the maximum score than for the standard three girls for questions two and three of the packaging task, and had similar results for question one. The difference for question one was less than one third of a percent. Only the difference of five percent between the standard three girls and boys on question three was statistically significant by chi-square (at the 10% level) for the packaging task. In contrast to the standard three students, a higher proportion of the form three girls gained the maximum score for each of the questions in the packaging task. However, the largest difference between the boys and the girls was just under five percent, and was not statistically significant.

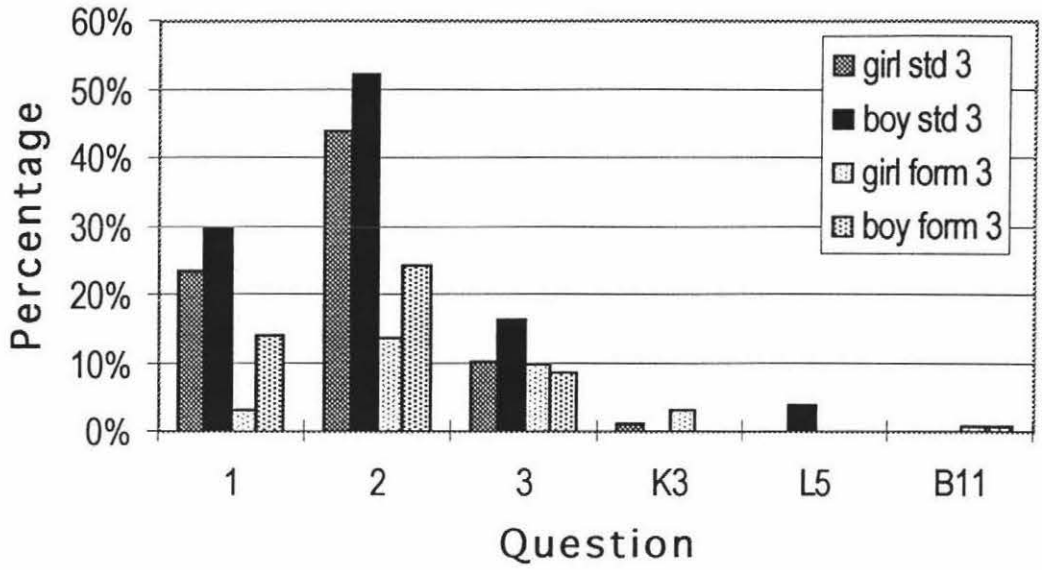
Figure 6.8 Proportions of girls and boys who scored the maximum score for each question



While the differences in maximum score were small, the differences between the proportions of girls and boys who gained partial marks were larger, as were the differences between those who did not give any response to the question. For all questions at the form three level, a higher proportion of girls gained partial marks, but at the standard three level, a higher proportion of boys gained partial marks for questions one and two. The differences in partial marks followed the same pattern as the differences in maximum score, and so a practically significant higher proportion of standard three boys gained at least partial marks on questions one and two of packaging. Similarly, a higher proportion of form three girls gained at least partial marks on all questions in packaging.

For most questions, a higher proportion of boys did not give any response to the question. The differences between the proportions of form three boys and girls who did not give any response are statistically significant by chi-square for questions one and two of packaging. No other difference represented in Figure 6.9 was statistically significant, although the difference between the standard three boys and girls on question two was eight percent.

Figure 6.9 Proportions of girls and boys who did not give any response to each question.



The only other difference between boys and girls worth noting was at the standard three level. For all the packaging questions, standard three girls were more likely than the boys to replicate the example. While the difference was only statistically significant for question one and two, all the actual differences were greater than ten percent.

Overall, gender differences in performance were very small for the multiple-choice questions, and never statistically significant. A slightly higher proportion of standard three boys than girls gained the maximum score for the packaging task, but this pattern was reversed at the form three level. The same pattern could be seen when partial marks were examined for the packaging task. For most questions, a higher proportion of boys did not give any response to the question, and a higher proportion of girls replicated the example. Both of these types of responses are incorrect, but are possibly different reactions to not knowing what to do. In order to determine reasons for their differential behaviours we would need to interview these students.

6.7.2 Ethnic differences

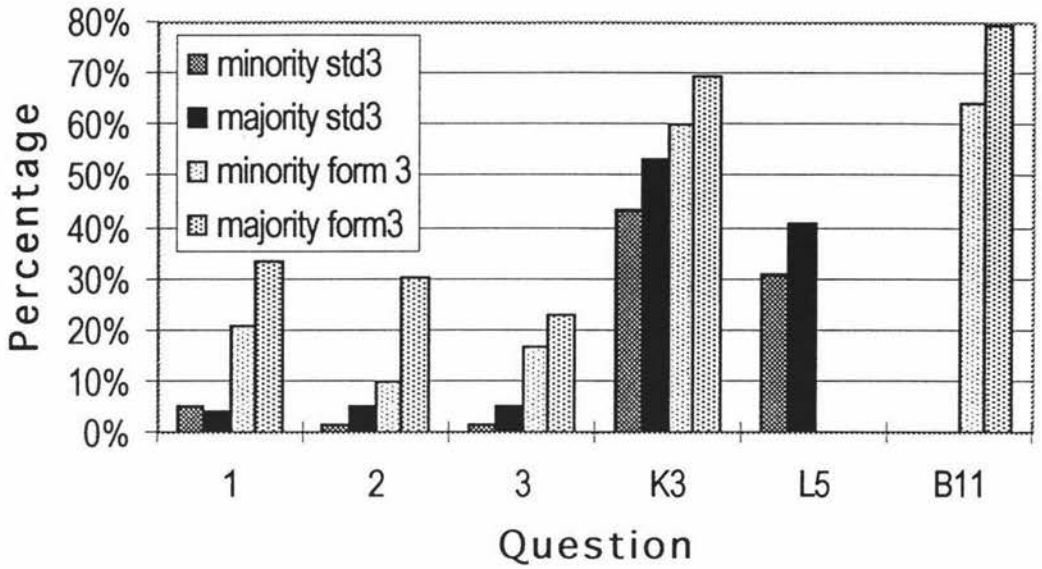
For this analysis students were categorised into two groups: the *majority* ethnic group and the *minority* ethnic group. The majority ethnic group comprised those students who identified themselves as only being "*New Zealand Pakeha/European*" or "*Other European*". The minority group included all other students including those who partially identified themselves as "*New Zealand Pakeha/European*" or "*Other European*". A comparison of these two groups is appropriate, because those students who identified themselves as Pakeha or of European extraction, are the majority of the population of students as well as the majority of teachers and examination writers. Thus the education system might be said to biased towards the majority students.

In general, a larger proportion of the students in the majority ethnic groups achieved the maximum score for each of the questions, with the only deviation being question one for the standard three students, and here the difference was less than one percent. As for the gender differences, the differences between the proportions of students in the minority and majority ethnic groups gaining the maximum score for the packaging task were larger than those for the multiple-choice tasks despite the overall proportions being smaller (see Figure 6.10).

For the standard three students, the proportions of students in the minority and majority ethnic groups scoring the maximum score on the packaging task were less than ten percent, so the differences, all of which were less than four percent, were not statistically significant. Despite the differences being almost ten percent between the two multiple-choice questions at the standard three level, these also were not statistically significant by chi-square, probably because the numbers of minority students were small. Most of the form three differences between the minority and majority ethnic students were statistically significant by chi-square (at the ten percent level). The exceptions, question three and K3, had differences of six percent and ten percent respectively. In general, where there was any practical significance between the proportions of students who scored part marks on the packaging task, it was in favour

of the majority group which also had a higher proportion of students who scored the maximum mark.

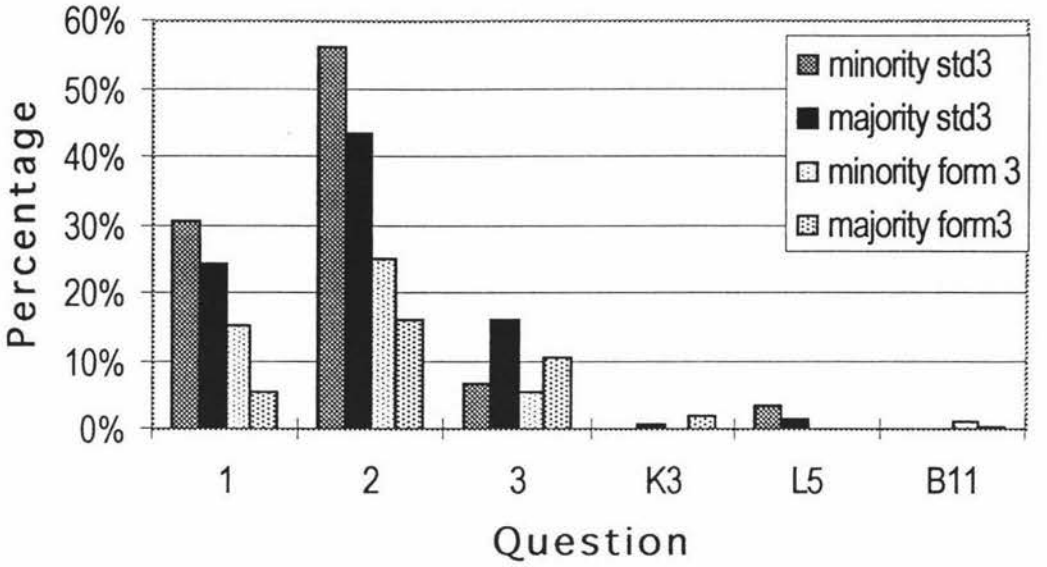
Figure 6.10 Proportions of students in the minority and majority ethnic groups who scored the maximum score for each question



In contrast to the differences in maximum score, the differences between the proportions of minority and majority ethnic students who did not give any response to each question were not consistently in the same direction (see Figure 6.11). The differences between the proportions of minority and majority ethnic students who did not give any response to the multiple-choice questions were very small (less than two percent), as were the overall proportions, and were not statistically significant. A higher proportion of the minority group gave no response to question one and two and this was statistically significant for the form three students, but not the standard three students. Even though not statistically significant, thirteen percent more of the minority standard three students did not respond to question two. For question three, a higher proportion of the majority students gave no response to the question, and this was statistically significant for the standard three students, but not for the form three students. Question three in the packaging task was often done by students who omitted to do questions one and two and who replicated the box given as an example. A statistically significant

higher proportion of the minority students, both form three and standard three, made a replica of the example box rather than a box of their own design.

Figure 6.11 Proportions of the minority and majority ethnic groups who did not give any response to each question.



Overall, a larger proportion of the students in the majority ethnic groups gained the maximum score for each of the questions. Most students responded to the multiple-choice questions. However, it was fairly common for students not to respond to questions one and two on the packaging task, and non-responses involved a larger proportion of the minority students than of the majority students. A higher proportion of the minority students replicated the example box, but a higher proportion of majority students did not attempt to create the net for question three.

6.7.3 Differences found between students from English-speaking homes and those from non-English speaking homes

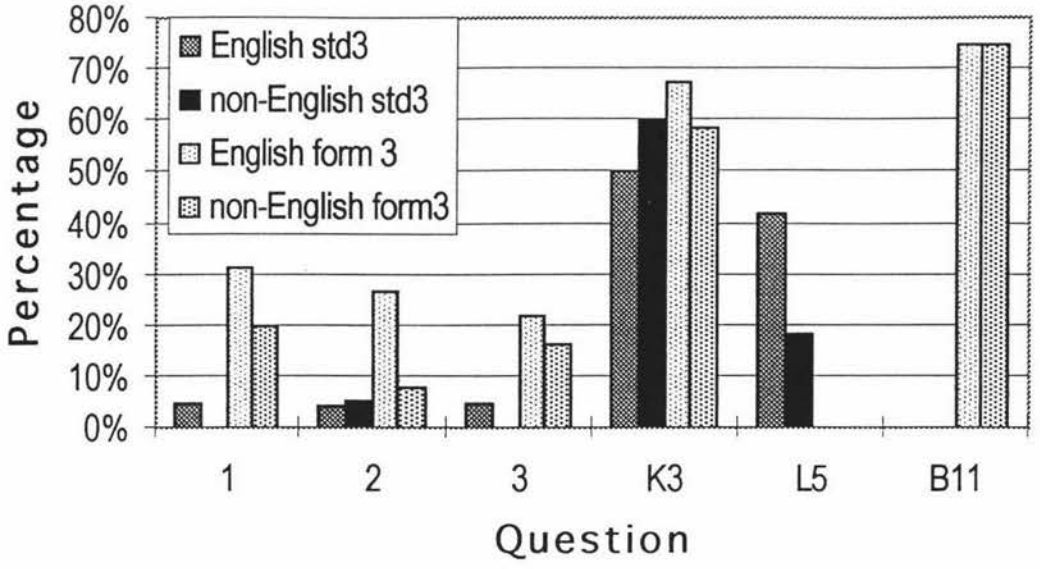
Students for whom English was not their first language may be disadvantaged by a test written in English, requiring answers to be written in English also. Students in two groups, *the students from English-speaking homes* and *the students from non-English-speaking homes* were compared (for further information on this grouping see section 4.3).

In general, a larger proportion of the students from English-speaking homes achieved the maximum score for each of the questions, with the only deviations being question two and K3, both for the standard three students, where the differences were less than one percent and ten percent respectively (see Figure 6.12).

For the standard three students, the proportions of students from both the English-speaking homes and the non-English-speaking homes scoring the maximum score on the packaging task were less than ten percent, so the differences of less than four percent were not statistically significant. Despite the large differences in performance between the two multiple-choice questions at the standard three level (ten percent and twenty-four percent), these were not considered statistically significant by chi-square, probably because the number of students from non-English-speaking homes was small for each question. Similarly, at the form three level, only question two of the packaging task had a statistically significant result favouring the students from English-speaking homes, despite all of the differences other than that of B11 being greater than six percent and also favouring the students from English-speaking homes.

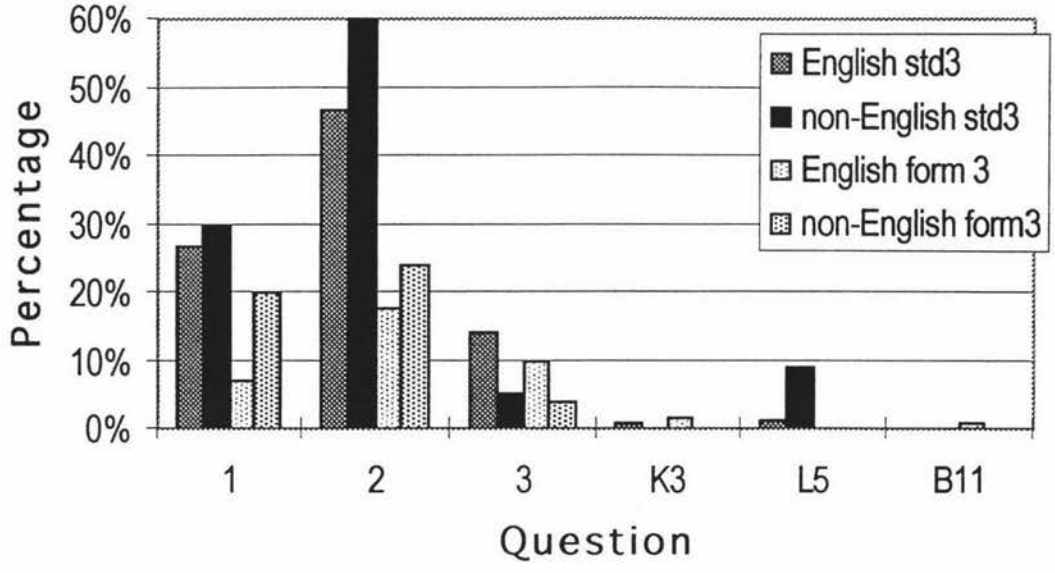
For the packaging task, question two at the standard three level had a far higher proportion of students from English-speaking homes who scored partial marks, as did question three at the form three level. Overall, a higher proportion of students from English-speaking homes scored at least partial marks for all packaging questions at both form three and standard three.

Figure 6.12 Proportions of students from English-speaking and non-English-speaking homes who scored the maximum score for each question



Like the differences between the minority and majority ethnic students, the differences between the students from English-speaking and non-English-speaking homes who did not give any response to each question were not consistently in the same direction (see Figure 6.13 and section 6.7.2). The differences between the proportions of students from English-speaking and non-English-speaking homes who did not give any response to the multiple-choice questions were very small (less than two percent or only one student per group) as were the overall proportions and were not statistically significant. For questions one and two, a higher proportion of the students from non-English-speaking homes gave no response to the question and this is statistically significant for the form three students, but not the standard three students. Even though it is not statistically significant, thirteen percent more of the standard three students from non-English-speaking homes did not respond to question two. For question three, a higher proportion of the students from English speaking homes gave no response to the question, and while the differences for standard three students and form three students were nine percent and six percent respectively, these were not statistically significant.

Figure 6.13 Proportions of students from English-speaking and non-English-speaking homes who did not give any response to each question.



As mentioned in section 6.7.2, question three in the packaging task was often done by students who omitted to do questions one and two and replicated the given box. A statistically significant higher proportion of the students from non-English-speaking homes, both form three and standard three, made a replica of the example box rather than a box of their own design. Also, at the form three level, there was a higher proportion of students from non-English-speaking homes who repeated the example information in questions one and two. This raises questions as to whether the students were unable to understand the instructions or were confused by the amount of information provided. We could only ascertain this by interviewing students or having them explain their thoughts as they follow through the task.

Overall, the results for the students from non-English-speaking homes were similar to those from the minority ethnic groups. This could be expected from the sample since a larger proportion of the students from non-English-speaking homes, although not all of them, also identified with the minority ethnic group. A larger proportion of the students from English-speaking homes gained at least partial marks for each of the questions. Most students responded to the multiple-choice questions. However, many students did not respond to questions one and two on the packaging task, and a larger

proportion of the students from non-English-speaking homes did not respond to these two questions. A higher proportion of the students from non-English-speaking homes replicated the example box or the information in the stem, but a higher proportion of students from English-speaking homes did not attempt to create the net for question three.

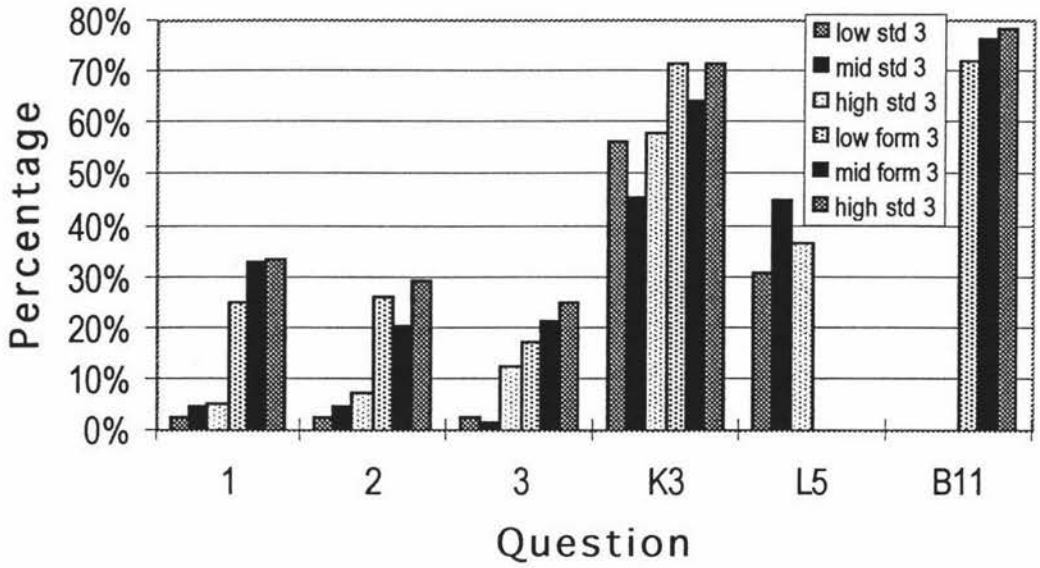
6.7.4 Differences between students from differing socio-economic status groups

Socio-economic status can impact on education because students from families with a low socio-economic status are unlikely to have available a range of materials or equipment to enhance out-of-school learning experiences. Schools in areas where the majority of students come from families with a low socio-economic status may also have limited extra funds for materials and equipment. Students were divided into three groups for this analysis, those of low, those of middle, and those of high socio-economic status.

There was no consistent difference in proportions of students from each socio-economic grouping who gained the maximum score. In general, a larger or equal proportion of the students from the high socio-economic group achieved the maximum score when compared with the low socio-economic group. However the middle socio-economic group varied in position relative to the other two groups, usually lower than or equal to the high socio-economic group, but sometimes lower than the low socio-economic group (see Figure 6.14). Only for question three of the packaging task were the differences between the three socio-economic groups statistically significant (at the five percent level) and this was only for the standard three students. The differences between the high and low socio-economic groups were less than eight percent, for all other questions.

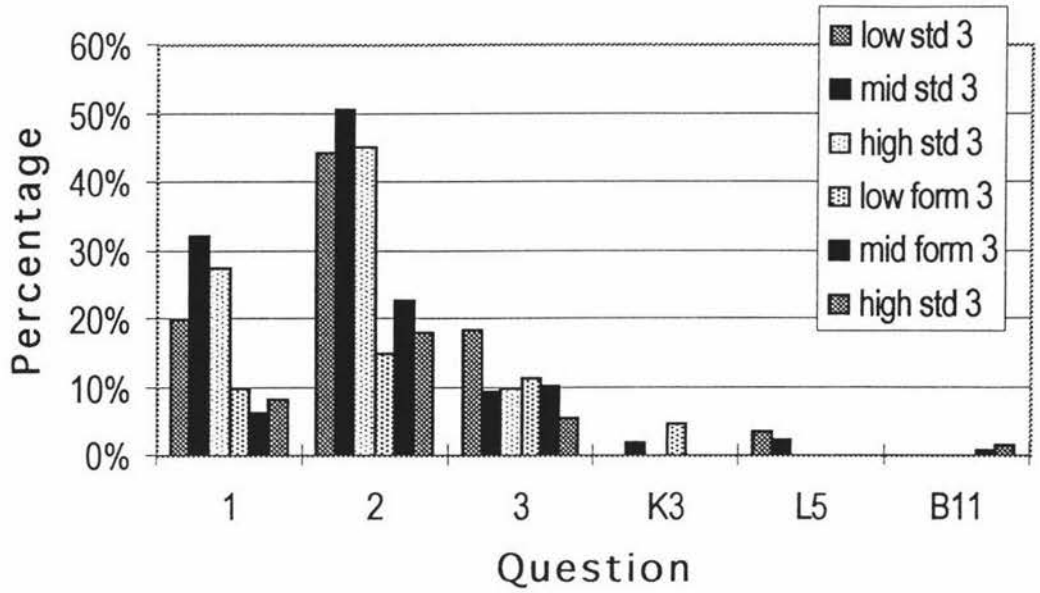
At the standard three level, a higher proportion of students from the high socio-economic group gained partial marks, whereas at the form three level, a higher proportion of students from the low socio-economic group gained partial marks.

Figure 6.14 Proportions of students from each socio-economic group who scored the maximum score for each question



Like the differences between the socio-economic groups for maximum score, there was no consistent pattern between these groups for non-response to the questions. (see Figure 6.15). However, unlike the differences for maximum score there was not even a consistent pattern between the proportions of students not responding from the low socio-economic group and the high socio-economic group. All differences were less than ten percent except for the difference between the low and middle groups at standard three for question one.

Figure 6.15 Proportions of students from each socio-economic group who did not give any response to each question.



As mentioned in section 6.7.2, question three in the packaging task was often done by students who omitted to do questions one and two and who replicated the given box. Although not statistically significant, a lower proportion of the students from the high socio-economic group, both form three and standard three, made a replica of the example box rather than a box of their own design. The percentage was lower for the high socio-economic group, when compared to the other two groups, by at least sixteen percent at standard three, and at least four percent at form three.

Overall, the results for socio-economic groupings, as they have been grouped for this thesis, for the packaging task and its associated multiple-choice tasks was inconclusive. We cannot say whether socio-economic status has any bearing on the results of these tasks, or whether the surrogate measure was inappropriate, or whether the subsequent grouping was inadequate.

6.7.5 *Personal Value of Mathematics*

The performances of three groups of form three students were examined, where the groupings were based on responses to two statements:

(a) I think it is important to do well in mathematics at school

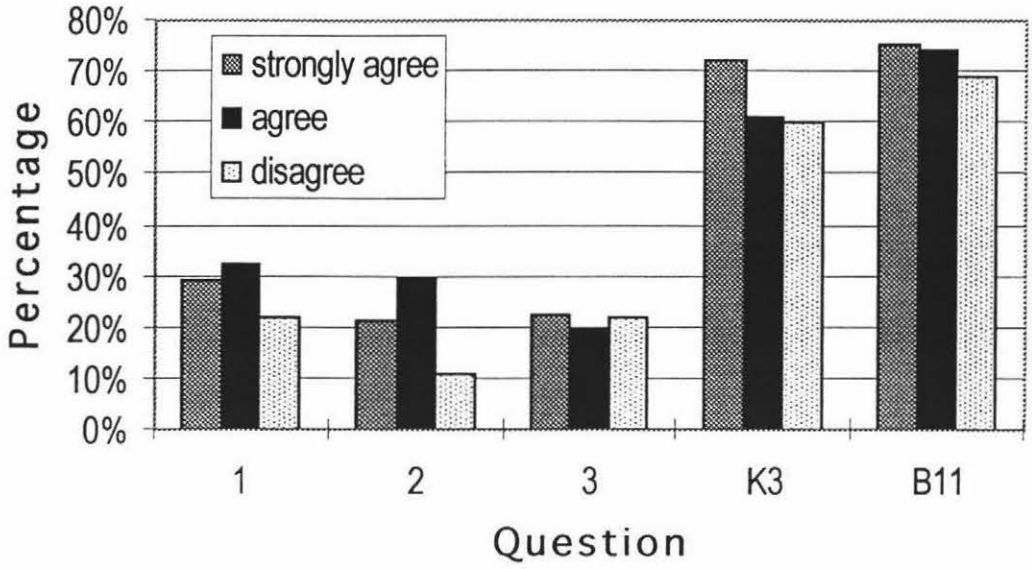
and

(b) Mathematics is important to everyone's life.

The three groups were those students who *strongly agreed, agreed,* and *disagreed* with these statements.

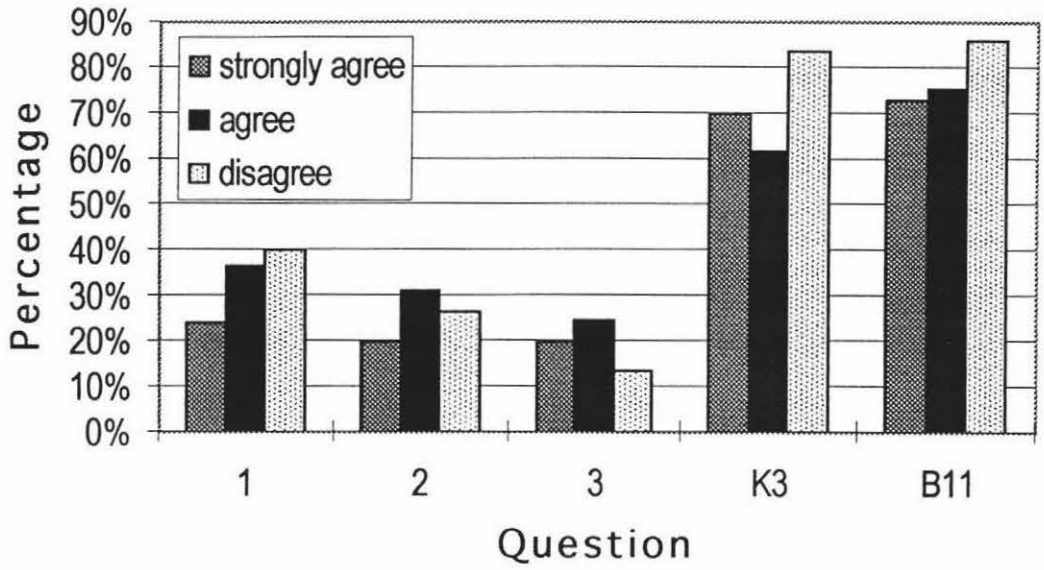
The overall pattern presented in Figure 6.16, of the proportions of students gaining the maximum score grouped by their agreement with the statement *I think it is important to do well in mathematics at school,* was not consistent across tasks. However, the group of students disagreeing with the statement was mostly the group with the smallest proportion gaining the maximum score for the task. Although the largest difference, the difference between the simply agreeing and disagreeing students for question two, was nineteen percent, none of these results was statistically significant. This was because of the small number of students (three percent) who disagreed with the statement.

Figure 6.16 Proportions of students gaining the maximum score who agreed or disagreed with the statement *I think it is important to do well in mathematics at school*



Interestingly, when we look at the other statement, *Mathematics is important to everyone's life,* the pattern is quite different (see Figure 6.17). Often it was a higher proportion of the students who disagreed who gained the maximum score compared to those who strongly agreed. However, no consistent pattern can be discerned between those who simply agreed and those who disagreed. Again because of the small number of students (five percent) in the disagreeing group, these differences were not statistically significant despite the size of the differences.

Figure 6.17 Proportions of students gaining the maximum score who agreed and disagreed with the statement *Mathematics is important to everyone's life*



When the proportions of students in each group who did not respond to each question are examined for both statements on the value of mathematics we find no consistent pattern. Only one difference in proportions was statistically significant, that for question two, where eleven percent more of the students who strongly agreed with the statement *Mathematics is important to everyone's life* did not respond to the question.

Figure 6.18 Proportions of students agreeing and disagreeing with the statement *I think it is important to do well in mathematics at school* who gave no response to each question.

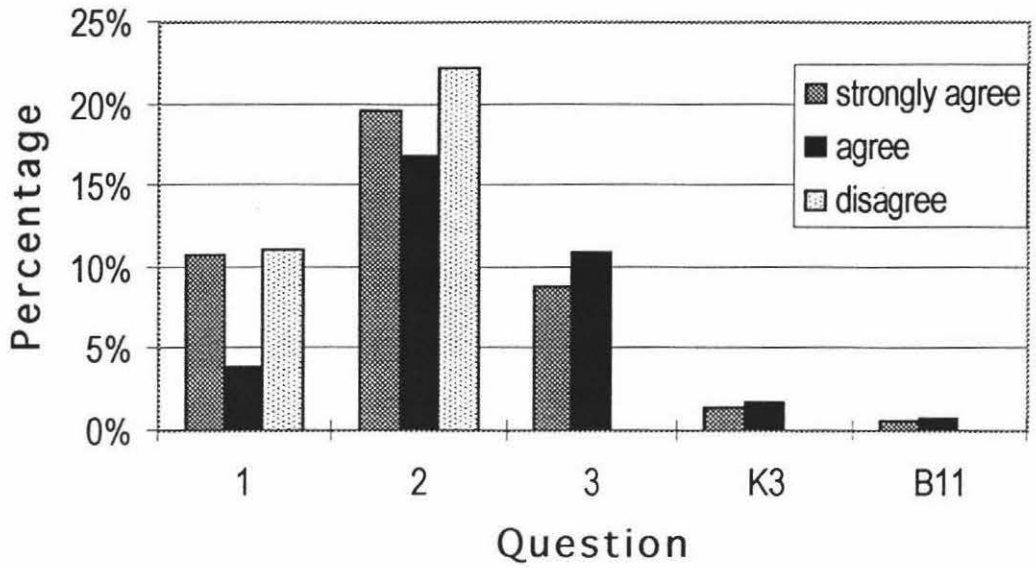
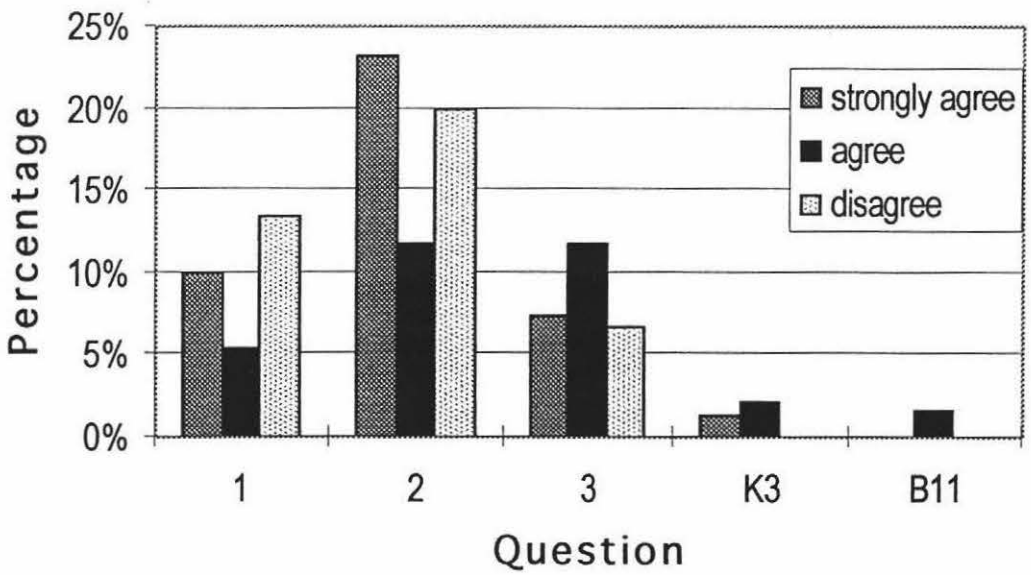


Figure 6.19 Proportions of students agreeing and disagreeing with the statement *Mathematics is important to everyone's life* who gave no response to each question.



It could be argued because of the lack of consistent pattern and statistical significance that we cannot say that the student values embodied in these statements have any effect

on student performance in either multiple-choice or performance assessment tasks. Possibly an increase in sample size might give us a chance to find differences that were statistically significant. However, since almost everyone thinks it is important to do well in mathematics and that mathematics is important to life, maybe it is the relativity of the importance of mathematics to other school subjects and life activities that would affect the outcomes.

6.8 Summary

The purpose of the packaging task was to measure the student's sense of spatial relations. Only three multiple-choice tasks were related to this task in some way. According to the mathematics curriculum (Ministry of Education, 1992), all these tasks would be appropriate to allow form three students to show their abilities but would possibly cause some difficulties for standard three students. This was evidenced in the results, although the marking criteria, the tendency of the students to replicate the examples, and the confusion caused for some children by reading the different methods of creating the box in the set-up information, may all have contributed to lowering the proportions of students who succeeded in the performance assessment task. The difficulties faced by students in interpreting the packaging task expose the difficulties of writing performance assessment tasks, particularly making them open-ended enough to allow students to reveal their ideas and abilities, while not leaving them too open to serious misinterpretations.

It is clear that despite the similar content on the two multiple-choice tasks at each level, the results of each multiple-choice task were quite different to each other and also quite different to those of the performance assessment task. The results also demonstrate that we could not get a full picture of students' knowledge and abilities using just the multiple-choice tasks on spatial visualisation. The packaging task certainly gave us more information about creativity of the students and their ability to present information and ideas pictorially. We could also find out how precise they could be in their constructions although it is not clear for those students who copied the example how precise they could be if they constructed their own net.

A much higher proportion of students gave no response to parts of the packaging task than to the multiple-choice tasks. In particularly, boys, those from minority ethnic groups, and those from non-English-speaking homes, were more likely to give no response to packaging task questions.

While there was practically no difference in achievement between the boys and the girls on the multiple-choice tasks, standard three boys did equally well or better than the girls, and form three girls did equally well or better than their male counterparts on the packaging task. The higher tendency on behalf of the standard three girls to replicate the example will have contributed to lowering their group achievement. The literature has shown a reduction of the gap between girls and boys mathematics achievement in the last ten years. As poor spatial visualisation abilities have been touted by some as a possible reason for the disparity between boys and girls, it is encouraging to see that with this group of students, particularly at the form three level, the girls are doing as well, if not better than, the boys.

It is of concern, however, that students from minority ethnic groups and those from non-English-speaking homes, in general, did not do as well as their majority and English-speaking counterparts. Of particular concern is that the differences were larger for the packaging task than the multiple-choice tasks, despite overall percentages for the maximum score being smaller. It could be argued that the opportunity to guess allows some students to get the multiple-choice questions correct where they are unable to answer the question, and thus could contribute to a smaller difference. It could also be argued that the high reading load, despite the possibility of help from the teacher, contributes to difficulties with the performance assessment tasks. The higher tendency of the minority ethnic groups and those from non-English-speaking homes to replicate the example, and thus show an incomplete understanding of the requirements of the task, will also contribute to difficulties demonstrating their abilities. These results are of particular concern because of the implications for these students of changes to assessment procedures to include more performance assessment type tasks. This tends to affirm the beliefs expressed by some writers that reforms may further disadvantage these groups.

The results of the analysis of high, low and middle socio-economic groupings was inconclusive. If the low socio-economic status of individuals affects the expression of their abilities, then we would expect that the low group would do worst overall, and the middle group would do better than the low group but less well than the high group. However, this was not consistently evidenced in the results of the multiple-choice or the packaging questions. Similarly the statements on the importance of mathematics were inconclusive, although in general, the students who disagreed with the statement *I think it is important to do well in mathematics at school* did less well than those who agreed with the statement.

7 DISCUSSION

7.1 Appropriateness of tasks examined

A criticism of large-scale international studies is that the questions used do not reflect the expectations of the New Zealand curriculum and the results do not accurately portray the abilities of our students. As mathematics is not a finite set of discrete skills and concepts, nor is mathematical knowledge absolute, any cross-cultural test must be a compromise between the nations involved.

All the tasks examined in this study, both multiple-choice and performance assessment tasks, were appropriate for form three students. However, a number of questions were considered to be outside the experiences of most standard three students. These questions were: all the packaging task, the two multiple-choice tasks associated with the packaging task, the final questions on the dice task, one multiple-choice task on table reading, and one multiple-choice task on probability. Despite the potential difficulties for standard three students, many of them enjoyed doing the tasks, and working with the equipment. It is useful to recall that "*Teachers know that students are capable of solving quite difficult problems when they are free to use concrete apparatus to help them think the problems through*" (Ministry of Education, 1992, p. 13), and many students performed at a higher level than expected.

7.2 Relationships between multiple-choice and performance questions.

According to Garden (1997), at the form three level, correlations between standardised scores on the written tests and performance assessments were moderate ($r=0.51$, $p<0.0001$) indicating that performance assessment is adding a substantial amount of information to that provided by the written test.

Only multiple-choice questions with similar content to the performance task questions were examined in this thesis. Two tasks were said to be similar in content if they could be characterised by the same achievement objectives from the mathematics curriculum. For example, one achievement objective from the algebra strand stated that "*students should be able to continue a sequential pattern and describe a rule for this*" (Ministry of

Education, 1992, p. 134). Several different questions could be characterised as examining whether students could meet this objective.

In this study, student performances on the multiple-choice questions that were found to have similar content to each of the performance assessment tasks, packaging and dice, were compared by the chi-square test of independence. Two tasks found to be independent, in general, did not elicit the same performance from each student. If we decided that we only wanted to use multiple-choice questions in a test, we could not use independent multiple-choice questions to substitute for the performance questions with similar content. More than half of the pairs of questions which were associated by content were found to be independent.

From the other point of view, only some of the multiple-choice questions at each level were found not to be independent of the performance assessment questions. For the pairs of questions found not to be independent, the student performances were similar on the two questions such that many of the students who correctly answered the multiple-choice question also correctly answered the performance assessment question. Similarly, many of the students who incorrectly answered the multiple-choice question also incorrectly answered the performance assessment question. However, examining the proportions revealed that, even for those questions found to be associated by both content and student performance, quite a few students incorrectly answered one of the paired questions and correctly answered the other. With this in mind, it is clear that the performance assessment questions give us a different appreciation of the abilities of the students when compared with the multiple-choice questions.

7.3 Non-response to questions

For both performance assessment tasks, students were more likely to give no response to the performance assessment tasks than the multiple-choice tasks particularly at the standard three level. While some of the questions were outside the expected experiences of the standard three students, this was the case for both the packaging task and the multiple-choice questions with similar content. Yet a much higher proportion of students did not respond to some part of the packaging task. It is possible that the time

constraints imposed on the performance tasks were the cause of the higher level of non-response, particularly for the dice task, and given more time, more of the students would have completed the task.

It could be argued that using performance assessment tasks gave us a better picture of student abilities than multiple-choice questions, because students were more likely to give no response to a question than make a random guess when they didn't know the answer. However, because it is easier to give no answer to a performance assessment task than it is to find an answer, students who could have made a correct or partly correct intelligent guess, or work it out, may have been more inclined to give no response.

7.4 Performances of disadvantaged sub-groups of students

When comparing the differences in proportions of the subgroups who achieved the maximum score, that is the subgroupings based on gender, ethnicity, language of home, socio-economic status, and attitude to the value of mathematics, on average, the differences were smaller for the dice task questions than the multiple-choice questions. However, for the packaging task and its associated multiple-choice questions there was no consistent pattern.

If only the dice task and its associated multiple-choice questions were examined in this thesis we might conclude from this result that it is preferable to use performance assessment tasks than multiple-choice tasks, as these appear to allow students a better opportunity to display their skills and abilities, and do not unfairly disadvantage girls, minority ethnic students, low socio-economic students – at least not as much as multiple-choice questions. While the results of the packaging task give us pause for thought, it is important to remember that the few multiple-choice questions associated by content were not always found to be associated by performance. Also, the multiple-choice questions presented students with alternatives rather than asked them to create their own ideas. The complex nature of the introduction to the packaging task, including methods which were marked as incorrect if duplicated by students, will have increased

the chance of students encountering difficulties, especially those disadvantaged by poor language abilities.

7.4.1 Boys and Girls

According to Harmon *et al.* (1997), the average percentage scores for the dice task and the packaging task were similar for boys and girls with no statistically significant differences. Despite the lack of statistical significance, it is interesting to note that with the exception of the standard three girls' performance on the packaging task, the average percentage score for the girls was higher than the boys for these two tasks.

Comparing the individual questions examined, with the exception of the performance of the standard three girls on the packaging task, in general, a higher proportion of girls obtained the maximum mark. Sometimes, a higher proportion of the boys gained half marks, where these were given. Where the proportion of students not responding to a question was of practical significance, a higher proportion of the boys at both levels did not respond to it. Standard three girls were more likely to replicate the example given in the packaging task.

There could be a number of reasons for these results favouring the girls. It was evident from the literature that the gap between the performances of boys and girls in mathematics has been narrowing in recent years. Actions to encourage and include girls in mathematics appear to have borne fruit in the form of improved performances, even to the extent of girls outstripping boys in performance. Another possible reason could be that the more integrated and purposeful tasks may motivate students and allow girls to demonstrate their skills and abilities to better effect. If this were so, we might expect to find that girls as a group performed better on the performance tasks and maybe on the more context based multiple-choice tasks. However, even for the more traditional forms of multiple-choice question, a higher proportion of girls obtained the maximum score.

Garden (1997) hypothesises that the reading and writing components of the tasks may have advantaged girls, given that girls generally achieve better than boys in the fields of reading and writing. In the Reading Literacy study (Wagemaker, 1993), the differences in performance between the standard three girls and boys, in favour of the girls, was

larger than that of the form three students. We would expect the reading requirements of the tasks to cause more problems for standard three boys. However, the packaging task had quite a long introduction that needed to be read and understood before students could embark on the appropriate actions. A higher proportion of boys, at the standard three level, correctly completed questions one and two of this task, which admittedly required no writing. As many of the multiple-choice questions had a reasonable reading load, some requiring decoding of distracters as well as the question, the reading requirements were not just specific to the performance tasks. Also, the students were able to obtain help with reading if required during the performance assessment administration, and so it is more likely that a lack of reading abilities would have been a disadvantage with the multiple-choice questions.

Examining the idea that the writing component of the tasks advantaged girls, we would then expect that questions involving an explanation, such as U4, and questions two and five of dice would be correctly answered by a higher proportion of girls than boys. We might also expect a larger difference in favour of the girls for these questions when compared with the other questions, but this was not so.

It appears from this analysis that the reading and writing components did not play a large factor in the relatively higher performance of the girls on many of the questions examined. However further research, including observations, interviews, or reading tests would clarify this supposition.

A higher proportion of boys gave no response to many of the questions examined, particularly for the performance assessment questions where there was a higher level of non-response than the multiple-choice questions. It is possible that the time limit imposed on the performance tasks caused this difference because the boys spent longer on individual questions, or in the case of the dice task, spent longer on the task paired with dice. It is also possible that boys were more likely, when they did not know how to proceed with a question, to give no response, whereas it appears that the girls were more likely to replicate the example given, particularly at the standard three level.

7.4.2 Ethnic groups and language differences

Students from minority ethnic groups and those from non-English-speaking homes, in general, did not do as well as their majority and English-speaking counterparts. The differences were larger for the packaging task than the associated multiple-choice tasks, despite overall percentages for the maximum score being smaller. However for the dice task and its associated multiple-choice questions, the differences were larger for the multiple-choice tasks. In contrast to this pattern, at the form three level, the minority students and those from non-English-speaking homes achieve just as well or better on the dice task than their counterparts.

Students from the minority group and those from non-English-speaking homes were more likely to replicate the example in the packaging task, and thus show an incomplete understanding of the requirements of the task. Except for where they replicated the example, students from non-English speaking homes were more likely to give no response to the performance assessment questions. When comparing the minority and majority students there was no consistent pattern of non-response.

For the students from non-English-speaking homes, difficulties with language comprehension must be the greatest barrier to successful completion of tasks. The reading component, despite the possibility of help from the teacher, contributes to difficulties with any tasks, either performance assessment or multiple-choice. It is clear that unless students can be tested in their most familiar language, any test with words is going to present difficulties.

It is of more concern that the minority ethnic students are not, in general, achieving as well as their counterparts in the majority ethnic groups, despite the results of the dice task apparently indicating that the minority students are catching up at the form three level. The packaging task did not have the same positive pattern for the minority students as the dice task. These results are of particular concern because of the implications for these students of changes to assessment procedures to include more performance assessment type tasks. The results here tend to affirm beliefs expressed by some writers that reforms will further disadvantage these students.

Although we know there is an overlap between the students from non-English-speaking home and those from minority ethnic groups, it is those minority students who are from English-speaking homes, approximately three quarters of the minority students, that are of particular concern. The more positive performance of the minority group on the dice task may lead us to believe that finding an appropriate context for these students will improve the educational outcome.

7.4.3 Socio-economic status

Over all the tasks examined, often the students from the low socio-economic group did not fare as well as their counterparts from the middle and high socio-economic groups on the dice task. However the results were not always consistent, with the middle socio-economic group sometimes having a smaller proportion scoring the maximum score when compared with the other two groups. There was no consistent pattern of non-response to the questions and students from the low socio-economic group were just as likely as the students from the other groups to give no response to each question.

The inconclusiveness of these results for socio-economic status gives rise to three conjectures as to the effect of socio-economic status on mathematics performance. Firstly, we could conjecture that socio-economic status has no effect on mathematics performance. Secondly, we could conclude that the surrogate measure of socio-economic status was inappropriate and we can conclude nothing about socio-economic status from these results. Thirdly, we could speculate that the grouping of results from the surrogate measure of socio-economic status was inappropriate and masked the effect of socio-economic status on the performance of students. Further research is needed to explore these issues.

7.4.4 Attitude to value of mathematics

The two value statements examined, *I think it is important to do well in mathematics at school* and *Mathematics is important to everyone's life* had few students disagreeing with them. Whether it was the small number of students disagreeing, or whether there was some other reason, the two performance tasks and their associated multiple-choice tasks did not consistently show one group with a higher proportion of correct answers than

the other. Neither was there any consistent pattern of non-response. Maybe some of the students who disagreed with the statements had considered the importance of mathematics to them and others around them and recognised that some people did not need to achieve well in school mathematics to achieve well in society. While they believed this, it did not stop them from being motivated to do well themselves in mathematics.

Another possible explanation for this result could be that students were motivated by questions which were different from their classroom activities – even if the only difference was that they were from an outside agency rather than their own teacher.

Because the TIMSS study also included science, the background questionnaire also included questions on the importance of science. This could have lead students to consider the relative importance of mathematics when compared with other school subjects. This in turn could lead these students to consider mathematics as less important to them than other subjects and hence they gave it a lower rating than they would for other subjects.

7.5 Positives and negatives of the different assessment activities - performance assessment and pencil-and-paper.

7.5.1 Test construction

The difficulties of question construction are evident for both types of assessment activities. It is often problematic to make a performance assessment question open-ended while still trying to ensure that the question assesses the knowledge and skills it is designed to assess. For example, in the packaging task, many students replicated some part of the introductory material or constructed a copy of the box given as an example, rather than finding a unique design of their own. Some students who found a unique ball arrangement, misinterpreted the information in the introductory section and incorrectly presented their design, for example, as a box made from four or more separate pieces. We cannot assume that an incorrect answer means that the student did not have the skills and knowledge to complete the question, and it is clear, when interpreting the

results, that the information and equipment provided to help students complete the task contributed to some of the incorrect answers.

When constructing both multiple-choice and performance assessment questions, alignment with the curriculum needs to be considered. TIMSS as a whole was designed to cover the curriculum with the performance assessment tasks covering the aspects not covered by the multiple-choice questions. It is clear that with performance assessment tasks, unlike multiple-choice tasks, specific performances, processes, and creativity can be tested. The dice task allows us to find out if students can carry out an experiment and record the data which cannot be done with the multiple-choice or the short-answer questions.

A further advantage of the use of performance assessment tasks is their alignment with the problem-solving approach encouraged in the mathematics curriculum. It is appropriate in a curriculum which emphasises a teaching approach based on problem-solving, and a desired outcome of a teaching programme embedded in this curriculum, to allow students to solve a problem in the way they see as most appropriate, that is, construct the response that they see as the best solution to the problem. However, it is important to consider the outcome of a question, when designing it. For example, for question two of dice, if we want to find out if students have the ability to see a complex pattern and describe it, we need to ask them a different question, and possibly more than one question, to allow them to display these skills.

7.5.2 Testing and Administration

There are a number of arguments that seem to be in favour of multiple-choice questions rather than performance assessment questions, when we consider the administration of tests. On initial glance, it would seem an advantage that multiple-choice and short-answer questions allow students to be more focussed on one particular problem, without the complicating factors of questions which build on results from previous questions and which require different knowledge and skills to complete. It appears that when a question has a narrow focus, it is easier for students to give the answer expected by the teacher, but when they solve an extended problem, more akin to a real life

problem, some students cannot display the same abilities. This was particularly noticeable for dice question five, where few students used the language of probability or probability concepts to explain the pattern in their results. Many of these same students could correctly answer the multiple-choice questions on probability. However, if the desired outcome of our education system is to produce students who can use the skills and knowledge they learn in the classroom when they encounter novel, extended, and integrated problems outside the classroom, then the multiple-choice questions give us no indication of this potential.

One of the arguments in favour of multiple-choice questions, is that they are quick to administer so we can ask more than one question to find out if students have a particular skill or knowledge. However, if a complex calculation is required to solve a problem, it will not be any quicker to complete if the question is written as multiple-choice rather than performance assessment. If, after completing the problem, the student finds they have an answer not given as an option, additional time will be needed for the student to decide which alternative to select, or to recalculate the answer.

The perception that the multiple-choice questions are quicker to administer than the performance assessment tasks may link with the criticism that the multiple-choice questions test only simple skills and knowledge whereas performance assessment tasks assess higher-order thinking skills. It will often take longer to test the higher-order skills than the recall of knowledge. Although the questions examined in this thesis have not been compared to a taxonomy of skills, and thus we cannot compare them on this basis, it is clear that for most students the multiple-choice questions required a higher level of response than simple recall, as did the performance assessment questions.

Both the multiple-choice questions and the performance assessment tasks were administered under time constraints. There was a much higher non-response to the performance assessment tasks, and it would seem that this is partly because of the time allowed. It would be interesting to give students as much time as they needed to complete each task and then compare the level of non-response. If under no time limit the level of non-response was still higher for the performance assessment tasks, it still

remains for us to consider whether this is a disadvantage of the performance assessment tasks as discussed in section 7.3.

A suggested advantage for students of performance assessment tasks is that the availability of concrete apparatus can help students to think through problems. However, we have no evidence, from this analysis of the packaging and the dice tasks, that the equipment helped with the thought processes. The use of concrete apparatus may also contribute to greater student motivation and satisfaction with the performance tasks. Many students informally commented to administrators that they enjoyed doing the performance tasks in the TIMSS study.

Although not evaluated in this thesis, the cost of supplying apparatus and test booklets for the performance assessment tasks was greater than the cost of printing the test booklets for the multiple-choice tasks in the TIMSS study. If the outcomes of the test process are not considered, multiple-choice questions seem to be easier to administer to students.

7.5.3 Evaluation

While the difficulties of test construction and administration often weigh heavily on the minds of administrators of assessments, the most important aspect for a student is the result of the test. Evaluating student responses is not without problems for both multiple-choice questions and performance assessment tasks. With performance assessment tasks, the use of open-ended questions allows students to construct their own response. This introduces problems of interpretation for the examiner, and careful consideration must be given to all responses supplied by students. For the packaging task, particularly for question one, the majority of students gave responses that did not meet the criteria for a fully complete and correct response. An examiner has to consider how these responses are to be handled, with particular consideration to how the results of the assessment are to be used.

For question one of the packaging task, one of the criteria required for a complete answer was: *the balls must be shown in a tightly packed arrangement.* Many students drew diagrams with only some or none of the balls touching and they were given no

credit for each of these incorrect diagrams. In order to fully describe each student's abilities, it may have been useful to include a specific category which took into account this particular deficiency in answers.

In comparison, the closed nature of multiple-choice questions also causes problems in interpreting whether students have the knowledge or skills the question is designed to elicit. We cannot know without interviewing the students whether their selection of the correct answer is based on knowledge, remembering the result from elsewhere, making a guess based on things they know, or making a correct guess by chance. Also, with multiple-choice questions, students are sometimes unable to choose the answer that they would give if they had the opportunity. For example, question L5 asks students how many edges there are on the cube. The picture shows only nine of the twelve edges, so there would be some students who would think that nine was the correct answer. However, nine is not one of the options given, so students are forced to select some other option, possibly selecting by chance the correct answer.

It is, perhaps, an argument in favour of multiple-choice questions, that if students get a result not listed in the multiple-choice options they may be prompted to attempt the question again and so discover their error. Unfortunately they have other options such as picking an answer that is close to their first answer or getting the right answer for the wrong reason. In any of these cases, an examiner has no way of finding out which of these methods were used by the student.

If an incorrect answer is presented to a multiple-choice question, there is no opportunity for the examiner to determine whether the error was arithmetical or procedural. In contrast, so long as the student has recorded their working or logic while completing a performance assessment question, an examiner has more understanding of the skills and knowledge students bring to a question. Part marks can be awarded to reflect the skills students show during the performance task.

For performance assessment questions requiring higher-order thinking, there is potential for responses that range in complexity. The use of ranges of descriptive terms or marks can be applied to the results of performance tasks to indicate this complexity. For

example, question five (b) of dice, has a number of descriptive terms. These descriptive terms allow us to determine which students use the language of probability, but also give credit to those students who do not.

To summarise this section, the advantages and disadvantages of each type of assessment activity are presented in Table 7.1.

Table 7.1 Advantages and disadvantages of the use of performance assessment and multiple-choice questions.

Performance Assessment		Multiple-choice	
Advantages	**Disadvantages**	**Advantages**	**Disadvantages**
Test Construction • can assess aspects of curriculum not easily assessed by multiple-choice • can assess performances, processes, and creativity • aligned with problem-solving approach in curriculum	• more opportunity for students to diverge from desired outcomes	• no chance for divergence from desired outcomes unless presented in options	• no opportunity to assess performances, processes, or creativity • less aligned with problem-solving approach of curriculum • more likely to confine method of solution
Testing and Administration • potential to test higher-order thinking • concrete apparatus may help students think through problem • may give greater satisfaction and motivation	• difficulties for students compounded by relations between questions • greater administration costs • may have higher non-response	• may be quicker to administer • cheaper to administer • lower non-response, chance to guess	• may test predominantly lower-order thinking
Evaluation • allows student to construct own response • allows flexible marking which can be used to reflect complexity of thinking and give credit for incomplete working • opportunity for process as well as product to be examined • more appropriate for educationally disadvantaged students	• difficulties interpreting responses and non-responses • requires flexible marking	• no difficulties interpreting responses • may cause students to reflect on thinking and produce appropriate response	• a correct answer does not indicate a correct and complete understanding • gives no indication of process used to determine answer • less appropriate for educationally disadvantaged students

7.6 Areas for future research.

There are a number of questions raised in this thesis which require further investigation. Because only one written question was associated with the performance tasks examined in this thesis, it was difficult to compare the efficacy of the performance assessment tasks with the written open-ended questions. Since multiple-choice questions are used less frequently by teachers within their classrooms, a comparison between written open-ended questions and performance assessment questions would be of more interest to them.

It could be argued that neither the dice nor the packaging task needed the performance aspect and that students would have performed similarly if the tasks had been written open-ended problems. For example, the dice task could have had the values of the dice throws already entered in question three, and the rest of the task could be administered unchanged. The packaging task could be administered without the example box, with a scale diagram of the balls used, and only a ruler required to complete the task. How much would these alterations to the task structure change the task and the results?

With any further investigations of performance assessment tasks, greater flexibility of time for completion of the tasks needs to be considered. Will this reduce the level of non-response to questions?

Questions used in any investigation that compares assessment activities should be evaluated in relation to a taxonomy of skills used. This would give a greater basis for comparing results. Do questions requiring higher-order thinking take longer to complete regardless of activity type? Do performance assessment tasks use more higher-order skills than multiple-choice questions?

Although it appeared that the reading and writing requirements of the performance assessment tasks did not negatively affect the results for most students, it would be useful to ascertain students' reading and writing skills prior to using performance assessment tasks. If the same tasks were used as in the TIMSS study, and the same assistance was given, would the poor readers find the performance assessment tasks

more difficult than the multiple-choice tasks? Would some questions be more difficult because of the amount of written explanations required?

It would also be worth interviewing students to discover their feelings after the assessment activities. Did they enjoy the performance tasks more than the multiple-choice or written tasks? Did having the concrete apparatus help them work through the problem? How did they work through each problem? What thought processes did they use?

Within these other areas of research, further research is needed on the use of performance assessment tasks with students who are seen as educationally disadvantaged. It is important to determine whether the use of performance tasks will further disadvantage these students or whether they will be better able to engage with these tasks. These investigations also need to compare performance tasks with written open-ended tasks.

7.7 Concluding thoughts.

While there are difficulties with the implementation of performance assessment tasks, there is evidence to suggest that the use of these activities will enhance teaching and learning and will add to the information supplied by written tests.

Many students were found to perform differently when their performances were compared in the multiple-choice and performance assessment questions that had similar content.

For some, but not all, of the performance questions there was a smaller difference between the educationally disadvantaged subgroups of students and their peers, when compared with the differences between them on the multiple-choice tasks. There is further research needed, and while the results on the packaging task were not as conclusive as the dice task, it appears that the reform in assessment activities is unlikely to increase the disadvantage to these students. In fact, performance assessment tasks may give greater freedom to express beliefs, knowledge, and skills, and may allow all students to receive credit for their ideas.

References

Adams, R. J., and Wilson, M., (1996). Evaluating progress with alternative assessments: a model for title 1. In Kane, M. B., and Mitchell, R., (eds), *Implementing Performance Assessment: promises, problems, and challenges, 39 - 59.* New Jersey: Lawrence Erlbaum Associates, Inc.

Aschbacher, P. R., (1991). Performance Assessment: state activity, interest, and concerns. *Applied Measurement in Education, 4(4), 275-288.*

Baggett, P., and Ehrenfeucht, A., (1996). *The role of calculators in developing children's problem solving skills.* Paper presented at American Educational Research Association Conference, New York.

Baker, E., (1989). Higher order assessment and indicators of learning. *CSE Technical Report 295.* Los Angeles: UCLA.

Baker, E. L., (1993). Questioning the technical quality of performance assessment. *The School Administrator, 50(11), 12 - 16.*

Baker, E., Linn, R., and Herman, J., (1996). CRESST: A continuing mission to improve educational assessment. *Evaluation Comment, Summer.* Los Angeles: UCLA.

Baker, E. L., and O'Neil, H. F. Jr., (1996). Performance assessment and equity. In Kane, M. B., and Mitchell, R., (eds), *Implementing Performance Assessment: promises, problems, and challenges, 183 - 198.* New Jersey: Lawrence Erlbaum Associates, Inc.

Baker, R., (1994). Including girls in mathematics and science programmes. *SAME papers: science and mathematics education papers, 40 - 55.*

Barnes, M., Clarke, D., and Stephens, M., (1995). Links between assessment and the teaching of mathematics in secondary schools: preliminary report. *Proceedings of the eighteenth annual conference of the Mathematics Education Research Group of Australasia (MERGA), 57 - 65.* Northern Territory University: Mathematics Education Research Group of Australasia

Bateson, D., Nicol, C., and Schroeder T., (1991). *Alternative assessment and tables of specifications for the Third International Mathematics and Science Study.* TIMSS Working Paper.

Baxter, G. P., Shavelson, R. J., Herman, S. J., Brown, K. A., and Valadez, J. R., (1993). Mathematics performance assessment: technical quality and diverse student impact. *Journal of Research in Mathematics education 24(3), 190-216.*

Begg, A., (1991). Assessment and constructivism. *The New Zealand Mathematics Magazine, 28(2), 14-19.*

Begg, A., (1996). Constructivism in the classroom. *The New Zealand Mathematics Magazine, 33(1), 3-17.*

Blithe, T., Clark, M., and Forbes, S., (1993). *The testing of girls in mathematics.*

Boaler, J., (1997). Reclaiming school mathematics: the girls fight back. *Gender and Education, 9(3), 285-305.*

Brown, R., (1992). Testing and thoughtfulness. In Burke, K., (ed), *Authentic assessment: a collection, 53 - 58.* Australia: Hawker Brownlow Education.

Brown, W., O'Gorman, K., and Du, Y., (1996). *The reliability and validity of mathematics performance assessment.* Paper presented at American Educational Research Association conference, New York.

Burton, L., (1992). Who assess whom and to what purposes? In Stephens, W. M., and Izard, J. F., (eds), *Reshaping assessment practices: assessment in mathematical sciences under challenge, 1 - 18.* Australia: Australian Council for Educational Research Ltd.

Carr, K., and Ritchie, G., (1991). Evaluating learning in mathematics. *SET: research information for teachers, 1, 15.*

Caygill, R. V., (1995). *Third International Mathematics and Science Study (TIMSS): Performance Assessment component - preliminary analysis.* Wellington: Research Section, Ministry of Education.

Chambers, D. L., (1993). Integrating assessment and instruction. In Webb, N. L., and Coxford, A. F., (eds), *Assessment in the mathematics classroom, 3 - 25.* Reston, Va: National Council of Teachers of Mathematics, Inc.

Clark, M., (1996). Mathematics, women, and education in New Zealand In Hanna, G., (ed), *Towards gender equity in mathematics education, 257 - 270.* Dordrecht: Kluwer Academic Publishers.

Clark, M., Forbes, S., and Blithe, T., (1994). The assessment of females in mathematics. *New Zealand Annual Review of Education, 4, 175 - 190.*

Clarke, D., (1992). The role of assessment in determining mathematics performance. In Leder, G. C., (ed), *Assessment and learning of mathematics, 145 - 168.* Victoria: The Australian Council for Educational Research Ltd.

Clements, M. A., and Ellerton, N. F., (1996). *Mathematics education research: past, present and future.* Bangkok: UNESCO Principal Regional Office for Asia and the Pacific.

Clements, M. A., and Ellerton, N. F., (1995). Assessing the effectiveness of pencil-and-paper tests for school mathematics. In *Proceedings of the eighteenth annual conference of the Mathematics Education Research Group of Australasia (MERGA), 184 - 188.* Northern Territory University: Mathematics Education Research Group of Australasia.

Cole, N. S., (1990). Conceptions of educational achievement. *Educational Researcher, 19(3), 2-7.*

Conover, W. J., (1971). *Practical nonparametric statistics.* New York: John Wiley & Sons, Inc.

Cooney, T. J., Badger, E., and Wilson, M. R., (1993). Assessment, understanding mathematics, and distinguishing visions from mirages. In Webb, N. L., and Coxford, A. F., (eds), *Assessment in the mathematics classroom, 237 - 247.* Reston, Va: National Council of Teachers of Mathematics, Inc.

Crooks, T., and Flockton, L., (1996). *Graphs, tables and maps: assessment results 1995.* Dunedin: Educational Assessment Research Unit.

Curran, J., (1994). *Gender difference in intermediate level mathematical problem solving*. Paper presented at the New Zealand Association for Research in Education conference.

Dager Wilson, L., and Chavarria S., (1993). Super-item tests as a classroom assessment tool. In Webb, N. L., and Coxford, A. F., (eds), *Assessment in the mathematics classroom, 133 - 142*. Reston, Va: National Council of Teachers of Mathematics, Inc.

Daniel, W. W., (1990). *Applied nonparametric statistics*. Boston: PWS-Kent.

de Lange, J., (1992). Assessment: no change without problems. In Stephens, W. M., and Izard, J. F., (eds), *Reshaping assessment practices: assessment in mathematical sciences under challenge, 46 - 76*. Australia: Australian Council for Educational Research Ltd.

Dochy, F. J. R. C., and Moerkerke, G., (1997). Assessment as a major influence on learning and instruction. *International Journal of Education Research 27(5), 415-432*. Great Britain: Elsevier Science Ltd.

Doig, B., and Cheeseman, J., (1998). 'The time has come,' the Walrus said, 'to talk of many things: of ... performance tasks'. *Australian Primary Mathematics Classroom, 3(1), 27-31*.

Doig, B. A., and Masters, G. N., (1992). Through children's eyes: a constructivist approach to assessing mathematics learning. In Leder, G. C., (ed), *Assessment and learning of mathematics, 269 - 289*. Victoria: The Australian Council for Educational Research Ltd.

Educational Testing Service, (1995). Capturing the Power of Classroom Assessment, *Focus 28*. Princeton: Educational Testing Service.

Ellerton, N. F., and Clements, M. A., (1997). Pencil-and-paper mathematics tests under the microscope. In Biddulph, F., and Carr, K., (eds), *People in mathematics education, 155 - 162*. Waikato: The Mathematics Education Research Group of Australasia Inc.

Elliot, S. N., and Fuchs, L. S., (1997). The utility of curriculum-based measurement and performance assessment as alternatives to traditional intelligence and achievement tests. *School Psychology Review, 26(2), 224-233*.

Feld, A. M., and Shepherd, C., (1998). Top achievement. *Machome, 6(7), 54*.

Fennema, E., (1983). Success in Mathematics. In Marland M., (ed), *Sex Differentiation and Schooling, 163 - 180*. London: Heinemann Educational Books.

Ferrara, S., and McTighe, J., (1992). Assessment: A thoughtful process. In Burke, K., (ed), *Authentic assessment: a collection, 157 - 177*. Australia: Hawker Brownlow Education.

Flockton, L., and Crooks, T., (1998). *Mathematics: assessment results 1997*. Dunedin: Educational Assessment Research Unit.

Forbes, S. D., (1996). Curriculum and assessment: hitting girls twice? In Hanna, G., (ed), *Towards gender equity in mathematics education, 71 - 92*. Dordrecht: Kluwer Academic Publishers.

Frary, R. B., (1985). Multiple-choice versus free-response: a simulation study. *Journal of Educational Measurement, 22(1), 21-31*.

Galbraith, P., (1993). Paradigms, problems and assessment: some ideological implications. In Niss, M., (ed), *Investigations into assessment in mathematics education, 73 - 86.* The Netherlands: Kluwer Academic Publishers.

Garden, R. A. (ed), (1996). *Mathematics performance of New Zealand form 2 and form 3 students: national results from New Zealand's participation in the third international mathematics and science study.* Wellington: Research and International Section, Ministry of Education.

Garden, R. A. (ed), (1997). *Performance assessment in the third international mathematics and science study: New Zealand results.* Wellington: Research and International Section, Ministry of Education.

Gay, S., and Thomas, M., (1993). Just because they got it right, does it mean they know it? In Webb, N. L., and Coxford, A. F., (eds), *Assessment in the mathematics classroom, 130 - 134.* Reston, Va: National Council of Teachers of Mathematics, Inc.

Gipps, C., (1994). *Beyond testing: towards a theory of educational assessment.* London: Falmer Press.

Gipps, C., Brown, M., McCallum, B., and McAlister, S., (1995). *Intuition or Evidence? : teachers and national assessment of seven year olds.* Buckingham: Open University Press.

Goldin, G. A., (1992). Toward an assessment framework for school mathematics. In Lesh, R., and Lamon, S. J., (eds), *Assessment of authentic performance in school mathematics, 63 - 88.* Washington DC.: American Association for the Advancement of Science.

Gordon, E. W., and Bonilla-Bowman, C., (1996). Can performance-based assessments contribute to the achievement of educational equity? In Boykoff Baron J., and Palmer Wolf D., (eds), *Performance-based student assessment: challenges and possibilities, 32 - 51.* Chicago: The National Society for the Study of Education.

Griffin, M. M., and Griffin, B. W., (1996). Situated cognition and cognitive style: effects on students' learning as measured by conventional tests and performance assessments. *Journal of Experimental Education, 64(4), 293-308.*

Griffin, P., and Nix, P., (1991). *Educational assessment and reporting: a new approach.* Sydney: Harcourt Brace Javanovich.

Gronlund, N. E., (1993). *How to make achievement tests and assessments - 5th ed.* Massachusetts: Allyn and Bacon.

Grossman, H., and Grossman, S., (1993). *Gender issues in education.* Massachusetts: Allyn and Bacon.

Hambleton, R. K., and Murphy, E., (1992). A psychometric perspective on authentic measurement. *Applied Measurement in Education, 5(1), 1-16.*

Haney, W., and Madaus, G., (1992). Searching for alternatives to standardized tests: whys, whats, and whithers. In Burke, K., (ed), *Authentic assessment: a collection, 87 - 99.* Australia: Hawker Brownlow Education.

Hanna, G., (1994). Should girls and boys be taught differently? In Biehler, R., Scholz, R. W., Stråßer, R., and Winkelmann, B., (eds), *Didactics of mathematics as a scientific discipline, 303-314.* Dordrecht: Kluwer Academic Publishers.

Harmon, M., Smith, T. A., Martin, M. O., Kelly, D. L., Beaton, A. E., Mullis, I. V. S., Gonzalez, E. J., and Orpwood, G., (1997). *Performance Assessment in IEA's Third International Mathematics and Science Study.* Boston College: Center for the Study of Testing, Evaluation, and Educational Policy.

Harnisch, D. L., (1994). Performance assessment in review: new directions for assessing student understanding. *International Journal of Educational Research, 21(3), 317-339.*

Heuvel-Panhuizen, M. v.d., and Gravemeijer, K., (1993). Tests aren't all bad. An attempt to change the face of written tests in primary school mathematics instruction. In Webb, N. L., and Coxford, A. F., (eds), *Assessment in the mathematics classroom, 54 - 64.* Reston, Va: National Council of Teachers of Mathematics, Inc.

Khattri, N., and Sweet, D., (1996). Assessment Reform: Promises and Challenges. In Kane, M. B., and Mitchell, R., (eds), *Implementing Performance Assessment: promises, problems, and challenges, 1 - 21.* New Jersey: Lawrence Erlbaum Associates, Inc.

Knight, G., (1997). ... I do and I understand, and then I forget. The role of memory in mathematics education. In Biddulph, F., and Carr, K., (eds), *People in mathematics education, 1 - 5.* Waikato: The Mathematics Education Research Group of Australasia Inc.

Kulm, G., (1994). *Mathematics assessment: what works in the classroom.* San Francisco: Jossey-Bass Inc.

Lambdin, D.V., (1993). The NCTM's 1989 Evaluation Standards: recycled ideas whose time has come? In Webb, N. L., and Coxford, A. F., (eds), *Assessment in the mathematics classroom, 7 - 16.* Reston, Va: National Council of Teachers of Mathematics, Inc.

Lamon, S. J., and Lesh, R., (1992). Interpreting responses to problems with several levels and types of correct answers. In Lesh, R., and Lamon, S. J., (eds), *Assessment of authentic performance in school mathematics, 319 - 342.* Washington DC. : American Association for the Advancement of Science.

Leder, G., (1990). Gender and classroom practice. In Burton, L, (ed), *Gender and mathematics.* Cassell Educational Limited

Leder, G. C., Forgasz, H. J., and Solar, C., (1996). Research and intervention programs in mathematics education : A gendered issue. In *International handbook of mathematics education, 945 - 985.* Netherlands: Kluwer Academic Publishers.

Lesh, R., Lamon, S. J., Behr, M., and Lester, F., (1992). Future directions for mathematics assessment. In Lesh, R., and Lamon, S. J., (eds), *Assessment of authentic performance in school mathematics, 379 - 426.* Washington DC. : American Association for the Advancement of Science.

Lesh, R., and Lamon, S. J., (1992). Trends, goals, and priorities in mathematics assessment. In Lesh, R., and Lamon, S. J., (eds), *Assessment of authentic performance in school mathematics, 3 - 16.* Washington DC. : American Association for the Advancement of Science.

Linn, R., (1988). Dimensions of Thinking: Implications for Testing. *CSE Technical Report 282.* Los Angeles: UCLA.

Linn, R. L., and Baker, E. L., (1996). Can performance-based student assessments be psychometrically sound? In Boykoff Baron J., and Palmer Wolf D., (eds), *Performance-based student assessment: challenges and possibilities, 84 - 103.* Chicago: The National Society for the Study of Education.

Linn, R. L., Baker, E. L., Dunbar, S. B., (1991). Complex, performance-based assessment: expectations and validation criteria. *Educational Researcher, 20(8), 15-21.*

Linn, R., and Burton, E., (1994). Performance-Based Assessment: Implications of Task Specificity. *Educational Measurement: Issues and Practice 13(1), 5 - 8.*

Linn, R. L., and Dunbar, S. B., (1992). The nation's report card goes home: good news and bad about trends in achievement. In Burke, K., (ed), *Authentic assessment: a collection, 13 - 30.* Australia: Hawker Brownlow Education.

Magone, M., (1996). *A study of gender differences in responses to a mathematics performance assessment instrument consisting of extended constructed-response tasks.* An abstract for a presentation given at the American Educational Research Association conference, New York.

Magone, M. E., Cai, J., Silver, E. A., and Wang, N., (1994). Validating the cognitive complexity and content quality of a mathematics performance assessment. *International Journal of Educational Research, 21(3), 317-339.*

Masters, G. N., and Doig, B. A., (1992). Understanding children's mathematics: some assessment tools. In Leder, G. C., (ed), *Assessment and learning of mathematics, 249 - 268.* Victoria: The Australian Council for Educational Research Ltd.

Meyer, M. R., (1992). Gender differences in test taking: a review. In Romberg, T. A., (ed), *Mathematics assessment and evaluation: imperatives for mathematics educators, 169 - 183.* Albany: State University of New York Press.

Ministry of Education, (1990). *Tomorrow's standards. The report of the ministerial working party on assessment for better learning.* Wellington: Ministry of Education.

Ministry of Education, (1992). *Mathematics in the New Zealand curriculum.* Wellington: Learning Media.

Ministry of Education, (1993). *Education for the 21st century.* Wellington: Learning Media.

Ministry of Education, (1994). *Assessment policy to practice.* Wellington: Learning Media.

Monk, D. H., (1996). Conceptualizing the costs of large-scale pupil performance assessment. In Kane, M. B., and Mitchell, R., (eds), *Implementing Performance Assessment: promises, problems, and challenges, 119 - 137.* New Jersey: Lawrence Erlbaum Associates, Inc.

Nagasaki, E., and Becker, J. P., (1993). Classroom assessment in Japanese mathematics education. In Webb, N. L., and Coxford, A. F., (eds), *Assessment in the mathematics classroom, 40 - 53.* Reston, Va: National Council of Teachers of Mathematics, Inc.

Newmann, F. M., (1992). Linking restructuring to authentic student achievement. In Burke, K., (ed), *Authentic assessment: a collection, 133 - 147.* Australia: Hawker Brownlow Education.

Neyland, J., (1994). Problems with proposed unit standard framework for mathematics. *The New Zealand Mathematics Magazine, 31(1), 1-9.*

Niemi, D., (1996). Assessing conceptual understanding in mathematics: representations, problems solution, justifications, and explanations. *The Journal of Educational Research, 89(6), 351-363.*

Niss, M., (1993). *Investigations into assessment in mathematics education.* The Netherlands: Kluwer Academic Publishers.

Nuttal, D. L., (1992). Performance assessment: A message from England. In Burke, K., (ed), *Authentic assessment: a collection, 118 - 122.* Australia: Hawker Brownlow Education.

Parkes, J., (1996). *Optimal designs for performance assessments: the subject factor.* Paper presented at the NCME conference, New York.

Peressini, D., and Bassett, J., (1996). Mathematical communication in students' responses to a performance-assessment task. In Elliott, P. C., and Kenney, M. J., (eds), *Communication in mathematics, K-12 and beyond, 146 - 158.* Reston, Va: The National Council of Teachers of Mathematics.

Resnick, D. P., and Resnick, L. B., (1996). Performance assessment and the multiple functions of educational measurement. In Kane, M. B., and Mitchell, R., (eds), *Implementing Performance Assessment: promises, problems, and challenges, 23 - 38.* New Jersey: Lawrence Erlbaum Associates, Inc.

Robitaille, D. F., McKnight, C. C., Schmidt, W. H., Britton, E., Raizen, S., and Nicol, C., (1993). *TIMSS monograph no. 1: curriculum frameworks for mathematics and science.* Vancouver, B. C.: Pacific Educational Press.

Romberg, T., (1992). Concerns about mathematics assessment in the United States. In Stephens, W. M., and Izard, J. F., (eds), *Reshaping assessment practices: assessment in mathematical sciences under challenge, 35 - 45.* Australia: Australian Council for Educational Research Ltd.

Rowntree, D., (1987). *Assessing students: how shall we know them - 2nd ed* New York: Nichols Publishing Company.

Santos, M., Drisoll, M., and Briars, D., (1993). The Classroom Assessment in Mathematics network. In Webb, N. L., and Coxford, A. F., (eds), *Assessment in the mathematics classroom, 220 - 228.* Reston, Va: National Council of Teachers of Mathematics, Inc.

Shepard, L. A., (1992). Why we need better assessments. In Burke, K., (ed), *Authentic assessment: a collection, 37 - 47.* Australia: Hawker Brownlow Education.

Stephens, M., (1992). Forward. In Stephens, W. M., and Izard, J. F., (eds), *Reshaping assessment practices: assessment in mathematical sciences under challenge.* Australia: Australian Council for Educational Research Ltd.

Stephens, M., and Sullivan, P., (1997). Developing tasks to assess mathematical performance. In Biddulph, F., and Carr, K., (eds), *People in mathematics*

education, 470 - 476. Waikato: The Mathematics Education Research Group of Australasia Inc.

Stiggins, R., (1991). Facing the challenges of a new era of educational assessment. *Applied Measurement in Education 4(4), 263-273.*

Stipek, D., Salmon, J. M., Givvin, K. B., Kazemi, E., Saxe, G., and MacGyvers, V. L., (1998). The value (and convergence) of practices suggested by motivation research and promoted by mathematics education reformers. *Journal for Research in Mathematics Education, 29(4), 465-488.*

Suzuki, K., & Harnisch, D., (1996). *An investigation of the generalizability of performance-based assessment in mathematics.* Paper presented at American Educational Research Association conference, New York.

Swan, M., (1993). Assessing a wider range of students' abilities. In Webb, N. L., and Coxford, A. F., (eds), *Assessment in the mathematics classroom, 26 - 39.* Reston, Va: National Council of Teachers of Mathematics, Inc.

Szetela, W., (1993). Facilitating communication for assessing critical thinking in problem solving. In Webb, N. L., and Coxford, A. F., (eds), *Assessment in the mathematics classroom, 143 - 151.* Reston, Va: National Council of Teachers of Mathematics, Inc.

Tartre, L. A., (1990). Spatial skills, gender, and mathematics. In Fennema, E., and Leder, G. C., (eds), *Mathematics and gender : influences on teachers and students, 27-59.* New York: Teachers College, Columbia University.

The National Council of Teachers of Mathematics, (1991). *Mathematics assessment : myths, good questions, and practical suggestions.* Reston, Va: The National Council of Teachers of Mathematics.

TIMSS International Study Center, (1994a). *TIMSS Performance Assessment administration manual.* Boston College: Centre for the Study of Testing, Evaluation, and Educational Policy.

TIMSS International Study Center, (1994b). *Coding guide for Performance Assessment.* Boston College: Centre for the Study of Testing, Evaluation, and Educational Policy.

Wagemaker, H. (ed), (1993). *Achievement in Reading Literacy: New Zealand's performance in a national and international context.* Wellington: Research Section, Ministry of Education.

Wang, N., & Lane, S., (1996). Detection of gender-related differential item functioning in a mathematics performance assessment. *Applied Measurement in Education, 9(2), 175-199.*

Webb, N., and Romberg, T. A., (1992). Implications of the NCTM standards for mathematics assessment. In Romberg, T. A., (ed), *Mathematics assessment and evaluation: imperatives for mathematics educators, 37 - 60.* State University of New York Press: Albany.

Wiggins, G., (1992). Teaching to the (authentic) test. In Burke, K., (ed), *Authentic assessment: a collection, 69 - 82.* Australia: Hawker Brownlow Education.

Wiley, D. E., and Haertel, E. H., (1996). Extended assessment tasks: purposes, definitions, scoring, and accuracy. In Kane, M. B., and Mitchell, R., (eds),

Implementing Performance Assessment: promises, problems, and challenges, 61 - 89. New Jersey: Lawrence Erlbaum Associates, Inc.

Williams, A., (1993). Is a pass good enough in tertiary statistics? In Atweh, B., Kanes, C., Carss, M., and Booker, G. (eds), *Proceedings of the sixteenth annual conference of the Mathematics Education Research Group of Australasia (MERGA), 587 - 591.* Queensland University of Technology: Mathematics Education Research Group of Australasia.

Wilson, M., (1992a). Measuring levels of mathematical understanding. In Romberg, T. A., (ed), *Mathematics assessment and evaluation: imperatives for mathematics educators, 213 - 241.* State University of New York Press: Albany.

Wilson, M., (1992b). Measurement models for new forms of assessment in mathematics education. In Stephens, W. M., and Izard, J. F., (eds), *Reshaping assessment practices: assessment in mathematical sciences under challenge, 77 - 98.* Australia: Australian Council for Educational Research Ltd.

Wolf. R. M., (1994). Performance assessment in IEA studies. *International Journal of Educational Research, 21(3), 239-245.*

Zevenbergen, R., (1997). Do disadvantaged students fail mathematics or does mathematics fail disadvantaged students? In Biddulph, F., and Carr, K., (eds), *People in mathematics education, 23 - 37.* Waikato: The Mathematics Education Research Group of Australasia Inc.

Appendix A

Microsoft Excel was used to perform the chi-square test on selected data. The function CHITEST returns the test for independence. CHITEST returns the value from the chi-squared ($\chi 2$) distribution for the statistic and the appropriate degrees of freedom. The $\chi 2$ test first calculates a $\chi 2$ statistic and then sums the differences of actual values from the expected values. The equation for this function is CHITEST=p(X>$\chi 2$), where:

$$\chi = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}, \text{ and where } E_{ij} = \frac{n_i C_i}{N}$$

O_{ij} = actual frequency in the i-th row, j-th column

E_{ij} = expected frequency in the i-th row, j-th column

r = number or rows

c = number of columns

n_i = total row frequency

C_i = total column frequency.

For example:

	A	B	C
1	Observed		
2		Male	Female
3	0	58	35
4	1	11	25
5	2	10	23
6			
7	Expected		
8		Male	Female
9	0	45.35	47.65
10	1	17.56	18.44
11	2	16.09	16.91

The $\chi 2$ statistic for the data above is 16.16957 with 2 degrees of freedom.

CHITEST(B3:C5,B9:C11) equals 0.000308

Appendix B. Chi-test values for the dice task.

Gender

Question	Chi-test for Codes	Chi-test for Scores	Chi-test for Maximum	Chi-test for Missing	Number of students
std 3					
1	0.15	0.11	0.27	0.37	196
2	0.08	0.94	0.94	0.02	196
3	0.36	0.23	0.28	0.31	196
4	0.43	0.16	0.60	0.37	196
5a	0.40	0.17	0.17	0.25	196
5b	0.22	0.41	0.41	0.06	196
H8	0.63	0.13	0.13	0.52	161
J5	0.01	0.02	0.02	0.01	109
L13	0.59	0.18	0.18	0.46	98
U4	0.50	0.45	0.45	0.69	207
K4	0.46	0.79	0.79	0.97	109
L10	0.42	0.20	0.20	na	312
M3	0.63	0.84	0.84	0.15	98
Intersection gp					
H8	0.28	0.43	0.43	0.31	42
J5	0.05	0.34	0.34	0.01	42
L13	0.09	0.08	0.08	0.18	22
U4	0.70	0.24	0.24	0.80	80
K4	0.31	0.90	0.90	0.35	58
L10	0.16	0.39	0.39	na	81
M3	0.28	0.38	0.38	0.18	22
form 3					
1	0.55	0.85	0.67	0.34	258
2	0.12	0.59	0.59	0.71	258
3	0.73	0.72	0.48	na	258
4	0.10	0.07	0.17	0.08	258
5a	0.02	0.05	0.05	0.00	258
5b	0.29	0.98	0.98	0.07	258
L13	0.35	0.34	0.34	0.35	130
G1	0.01	0.82	0.82	0.01	336
L10	0.34	0.34	0.34	na	130
M3	0.51	0.48	0.48	na	134
N18	0.37	0.11	0.11	0.29	72
Intersection gp					
L13	0.32	0.31	0.31	na	39
G1	0.14	0.79	0.79	0.04	110
L10	0.97	0.98	0.98	na	39
M3	0.53	0.80	0.80	na	33
N18	0.16	0.16	0.16	0.16	17

Ethnicity

Question	Chi-test for Codes	Chi-test for Scores	Chi-test for Maximum	Chi-test for Missing	Number of students
std 3					
1	0.17	0.07	0.22	0.22	196
2	0.27	0.22	0.22	0.56	196
3	0.19	0.50	0.42	0.47	196
4	0.35	0.09	0.13	0.74	196
5a	0.28	0.55	0.55	0.69	196
5b	0.37	0.40	0.40	0.25	196
H8	0.01	0.00	0.00	0.26	158
J5	0.49	0.31	0.31	0.57	108
L13	0.60	0.53	0.53	0.27	98
U4	0.01	0.00	0.00	0.00	206
K4	0.44	0.69	0.69	0.37	108
L10	0.00	0.00	0.00	na	311
M3	0.21	0.20	0.20	0.52	98
Intersection gp					
H8	0.42	0.42	0.42	0.55	42
J5	0.93	0.60	0.60	0.93	42
L13	0.61	0.36	0.36	0.35	22
U4	0.00	0.00	0.00	0.00	80
K4	0.16	0.49	0.49	0.50	58
L10	0.28	0.33	0.33	na	81
M3	0.70	0.39	0.39	0.26	22
form 3					
1	0.32	0.49	0.68	0.97	256
2	0.40	0.17	0.17	0.45	256
3	0.60	0.41	0.19	na	256
4	0.78	0.98	0.86	0.91	256
5a	0.27	0.29	0.29	0.45	256
5b	0.03	0.11	0.11	0.10	256
L13	0.53	0.08	0.08	0.54	130
G1	0.19	0.59	0.59	0.21	334
L10	0.90	0.90	0.90	na	130
M3	0.17	0.11	0.11	na	133
N18	0.06	0.02	0.02	0.53	71
Intersection gp					
L13	0.40	0.40	0.40	na	39
G1	0.65	0.85	0.85	0.11	109
L10	0.79	0.79	0.79	na	39
M3	0.27	0.27	0.27	na	32
N18	0.57	0.57	0.57	0.57	17

Language

Question	Chi-test for Codes	Chi-test for Scores	Chi-test for Maximum	Chi-test for Missing	Number of students
std 3					
1	0.84	0.65	0.69	0.20	186
2	0.55	0.32	0.32	0.05	186
3	0.67	0.46	0.23	0.33	186
4	0.15	0.04	0.08	0.29	186
5a	0.52	0.39	0.37	0.93	186
5b	0.11	0.33	0.33	0.39	186
H8	0.00	0.00	0.00	0.28	156
J5	0.86	0.65	0.65	0.48	105
L13	0.41	0.13	0.13	0.86	90
U4	0.02	0.05	0.05	0.10	195
K4	0.59	0.16	0.16	0.70	105
L10	0.62	0.37	0.37	na	298
M3	0.19	0.11	0.11	0.71	90
Intersection gp					
H8	0.00	0.00	0.00	0.81	39
J5	0.73	0.66	0.66	0.27	39
L13	0.89	0.43	0.43	0.67	20
U4	0.27	0.08	0.08	0.15	77
K4	0.61	0.30	0.30	0.75	57
L10	0.51	0.13	0.13	na	76
M3	0.38	0.41	0.41	na	20
form 3					
1	0.72	0.51	0.38	0.27	254
2	0.48	0.19	0.19	0.60	254
3	0.98	0.92	0.90	na	254
4	0.11	0.41	0.20	0.92	254
5a	0.87	0.81	0.81	0.60	254
5b	0.58	0.86	0.86	0.71	254
L13	0.94	0.38	0.38	0.75	128
G1	0.17	0.06	0.06	0.07	331
L10	0.29	0.35	0.35	na	128
M3	0.00	0.00	0.00	na	134
N18	0.01	0.01	0.01	na	69
Intersection gp					
L13	0.66	0.66	0.66	na	38
G1	0.71	0.80	0.80	0.28	109
L10	0.53	0.53	0.53	na	38
M3	0.03	0.02	0.02	na	33
N18	1.00	1.00	1.00	na	16

Socio-economic status

Question	Chi-test for Codes	Chi-test for Scores	Chi-test for Maximum	Chi-test for Missing	Number of students
std 3					
1	0.52	0.33	0.13	0.89	177
2	0.31	0.04	0.04	0.50	177
3	0.88	0.68	0.36	0.95	177
4	0.45	0.77	0.83	0.56	177
5a	0.65	0.22	0.22	0.63	177
5b	0.51	0.39	0.39	0.76	177
H8	0.03	0.03	0.03	0.21	146
J5	0.11	0.67	0.67	0.01	98
L13	0.21	0.17	0.17	0.37	88
U4	0.46	0.27	0.27	0.09	185
K4	0.90	0.98	0.98	0.41	97
L10	0.07	0.00	0.00	na	279
M3	0.22	0.27	0.27	0.36	88
Intersection gp					
H8	0.22	0.14	0.14	0.09	40
J5	0.28	0.34	0.34	0.07	38
L13	0.55	0.53	0.53	0.29	20
U4	0.59	0.68	0.68	0.28	69
K4	0.91	0.66	0.66	na	49
L10	0.20	0.02	0.02	na	75
M3	0.17	0.02	0.02	na	20
form 3					
1	0.93	0.93	0.94	0.90	247
2	0.49	0.64	0.64	0.15	247
3	0.15	0.08	0.03	na	247
4	0.64	0.36	0.17	0.97	247
5a	0.02	0.04	0.04	0.63	247
5b	0.46	0.49	0.49	0.96	247
L13	0.83	0.91	0.91	0.34	125
G1	0.33	0.09	0.09	0.21	322
L10	0.50	0.50	0.50	na	125
M3	0.99	0.98	0.98	na	127
N18	0.03	0.14	0.14	0.25	67
Intersection gp					
L13	0.37	0.37	0.37	na	38
G1	0.17	0.49	0.49	0.08	105
L10	0.45	0.45	0.45	na	38
M3	0.47	0.47	0.47	na	30
N18	0.08	0.08	0.08	0.08	17

Do well

Question	Chi-test for Codes	Chi-test for Scores	Chi-test for Maximum	Chi-test for Missing	Number of students
form 3					
1	0.90	0.87	0.63	0.70	254
2	0.43	0.89	0.89	0.12	254
3	0.85	0.74	0.39	na	254
4	0.20	0.37	0.61	0.33	254
5a	0.22	0.39	0.39	0.06	254
5b	0.05	0.46	0.46	0.00	254
L13	0.36	0.09	0.09	na	128
G1	0.33	0.24	0.24	na	331
L10	0.19	0.19	0.19	na	128
M3	0.51	0.30	0.30	na	133
N18	0.03	0.03	0.03	na	69
Intersection gp					
L13	0.36	0.36	0.36	na	38
G1	0.35	0.18	0.18	na	107
L10	0.88	0.88	0.88	na	38
M3	0.54	0.44	0.44	na	32
N18	na	na	na	na	16

Life

Question	Chi-test for Codes	Chi-test for Scores	Chi-test for Maximum	Chi-test for Missing	Number of students
form 3					
1	0.78	0.63	0.34	0.77	255
2	0.49	0.61	0.61	0.39	255
3	0.83	0.71	0.60	na	255
4	0.65	0.83	0.52	0.80	255
5a	0.93	0.28	0.28	0.56	255
5b	0.33	0.45	0.45	0.38	255
L13	0.88	0.78	0.78	na	3
G1	0.58	0.56	0.56	na	331
L10	0.78	0.78	0.78	na	3
M3	0.95	0.86	0.86	na	131
N18	0.13	0.43	0.43	na	72
Intersection gp					
L13	0.77	0.77	0.77	na	39
G1	0.29	0.04	0.04	na	109
L10	0.59	0.59	0.59	na	39
M3	0.49	0.53	0.53	na	32
N18	0.06	0.06	0.06	na	17

Packaging with Multiple-choice tasks

Questions	Chi-test for Scores	Number of students	Questions	Chi-test for Scores	Number of students
Std 3			**Form 3**		
1 & H8	0.18	42	1 & L13	0.86	35
2 & H8	0.26	42	2 & L13	0.55	35
3 & H8	0.66	42	3 & L13	0.95	35
4 & H8	0.08	42	4 & L13	0.50	35
5a & H8	0.12	42	5a & L13	0.61	35
5b & H8	0.95	42	5b & L13	0.58	35
1 & J5	0.45	42	1 & G1	0.33	79
2 & J5	0.05	42	2 & G1	0.77	79
3 & J5	0.30	42	3 & G1	0.31	79
4 & J5	0.00	42	4 & G1	0.98	79
5a & J5	0.01	42	5a & G1	0.06	79
5b & J5	0.02	42	5b & G1	0.00	79
1 & L13	0.70	22	1 & L10	0.80	35
2 & L13	0.07	22	2 & L10	0.39	35
3 & L13	0.10	22	3 & L10	0.91	35
4 & L13	0.49	22	4 & L10	0.84	35
5a & L13	0.09	22	5a & L10	0.27	35
5b & L13	na	22	5b & L10	0.43	35
1 & U4	0.00	80	1 & M3	0.21	22
2 & U4	0.00	80	2 & M3	0.18	22
3 & U4	0.04	80	3 & M3	na	22
4 & U4	0.26	80	4 & M3	0.78	22
5a & U4	0.00	80	5a & M3	0.45	22
5b & U4	0.23	80	5b & M3	0.10	22
1 & K4	0.21	58	1 & N18	na	9
2 & K4	0.03	58	2 & N18	na	9
3 & K4	0.17	58	3 & N18	na	9
4 & K4	0.23	58	4 & N18	na	9
5a & K4	0.04	58	5a & N18	na	9
5b & K4	0.43	58	5b & N18	na	9
1 & L10	0.73	81			
2 & L10	0.13	81			
3 & L10	0.18	81			
4 & L10	0.51	81			
5 & L10	0.45	81			
6 & L10	0.27	81			
1 & M3	0.02	22			
2 & M3	0.66	22			
3 & M3	1.00	22			
4 & M3	0.28	22			
5a & M3	0.38	22			
5b & M3	na	22			

Appendix C. Chi-test values for the packaging task.

Gender

Question	Chi-test for Codes	Chi-test for Scores	Chi-test for Maximum	Chi-test for Missing	Number of students
std 3					
1	0.03	0.04	0.94	0.29	197
2	0.16	0.45	0.33	0.25	197
3	0.13	0.09	0.09	0.19	197
K3	0.89	0.82	0.82	0.33	161
L5	0.04	0.83	0.83	0.15	98
Intersection gp					
K3	0.45	0.25	0.25	0.30	74
L5	0.23	0.12	0.12	NA	24
form 3					
1	0.11	0.55	0.54	0.00	261
2	0.28	0.13	0.33	0.03	261
3	0.02	0.12	0.55	0.74	261
B11	0.80	0.88	0.88	0.93	305
K3	0.39	0.90	0.90	0.16	131
Intersection gp					
B11	0.77	0.53	0.53	0.94	103
K3	0.12	0.51	0.51	0.28	19

Ethnicity

Question	Chi-test for Codes	Chi-test for Scores	Chi-test for Maximum	Chi-test for Missing	Number of students
std 3					
1	0.72	0.97	0.84	0.36	195
2	0.41	0.16	0.26	0.11	195
3	0.09	0.53	0.08	0.26	195
K3	0.48	0.25	0.25	0.14	158
L5	0.80	0.37	0.37	0.44	98
Intersection gp					
K3	0.24	0.85	0.85	0.02	73
L5	0.23	0.93	0.93	NA	24
form 3					
1	0.12	0.11	0.05	0.01	260
2	0.01	0.00	0.00	0.09	260
3	0.02	0.01	0.27	0.21	261
B11	0.02	0.01	0.01	0.01	303
K3	0.54	0.29	0.29	0.78	131
Intersection gp					
B11	0.14	0.13	0.13	0.22	103
K3	0.12	0.86	0.86	0.21	19

Language

Question	Chi-test for Codes	Chi-test for Scores	Chi-test for Maximum	Chi-test for Missing	Number of students
std 3					
1	0.70	0.54	0.32	0.75	189
2	0.39	0.13	0.86	0.26	189
3	0.23	0.61	0.32	0.25	189
K3	0.91	0.40	0.40	0.70	156
L5	0.14	0.13	0.13	0.10	90
Intersection gp					
K3	0.15	0.01	0.01	0.70	71
L5	0.02	0.12	0.12	NA	22
form 3					
1	0.25	0.43	0.24	0.02	255
2	0.16	0.12	0.04	0.45	255
3	0.31	0.15	0.48	0.33	255
B11	0.43	0.98	0.98	0.65	302
K3	0.54	0.52	0.52	0.65	129
Intersection gp					
B11	0.77	0.53	0.53	0.66	101
K3	0.41	0.68	0.68	0.72	19

Socio-economic status

Question	Chi-test for Codes	Chi-test for Scores	Chi-test for Maximum	Chi-test for Missing	Number of students
std 3					
1	0.48	0.38	0.82	0.26	175
2	0.78	0.80	0.53	0.73	175
3	0.03	0.00	0.02	0.22	175
K3	0.23	0.39	0.39	0.41	146
L5	0.88	0.49	0.49	0.73	88
Intersection gp					
K3	0.82	0.91	0.91	0.41	69
L5	0.26	0.12	0.12	NA	22
form 3					
1	0.29	0.79	0.43	0.65	249
2	0.84	0.62	0.42	0.42	249
3	0.51	0.86	0.53	0.43	249
B11	0.70	0.62	0.62	0.52	291
K3	0.46	0.69	0.69	0.12	129
Intersection gp					
B11	0.51	0.98	0.98	0.60	98
K3	0.61	0.69	0.69	0.56	19

Do well

Question	Chi-test for Codes	Chi-test for Scores	Chi-test for Maximum	Chi-test for Missing	Number of students
form 3					
1	0.95	0.88	0.72	0.14	258
2	0.29	0.28	0.22	0.83	258
3	0.62	0.99	0.89	0.53	258
B11	0.86	0.88	0.88	0.92	304
K3	0.52	0.43	0.43	0.95	130
Intersection gp					
B11	0.41	0.88	0.88	0.84	103
K3	0.41	0.36	0.36	na	19

Life

Question	Chi-test for Codes	Chi-test for Scores	Chi-test for Maximum	Chi-test for Missing	Number of students
form 3					
1	0.25	0.43	0.07	0.36	259
2	0.45	0.24	0.13	0.09	259
3	0.00	0.01	0.49	0.46	259
B11	0.60	0.56	0.56	0.21	299
K3	0.85	0.45	0.45	0.89	129
Intersection gp					
B11	0.37	0.31	0.31	0.11	103
K3	0.47	0.73	0.73	na	18

Packaging with Multiple-choice tasks

Questions	Chi-test for Scores	Number of students
1 & K3	0.13	74.00
2 & K3	0.81	74.00
3 & K3	0.71	74.00
1 & L5	0.24	24.00
2 & L5	0.22	24.00
3 & L5	0.35	24.00
1 & B11	0.07	106.00
2 & B11	0.07	106.00
3 & B11	0.18	106.00
1 & K3	0.04	19.00
2 & K3	0.35	19.00
3 & K3	0.12	19.00