ISSUES IN CONSTRUCTING FORMATIVE TESTS IN MATHEMATICS

Paper presented at the International Association for Educational Assessment Conference,
Hong Kong, September 1–6, 2002

W.A. Neill
Senior Researcher
New Zealand Council for Educational Research
Wellington
New Zealand

ABSTRACT

A series of mathematics tests called Achievement in Mathematics (AIM) is being developed for New Zealand schools. These tests aim to give detailed information about student performance in mathematics. The emphasis of the tests is on formative assessment that informs teaching and learning. The tests assess the five strands of the mathematics curriculum: Number, Measurement, Geometry, Algebra, and Statistics. Each test assesses all the objectives at a given level of an individual strand. To span all the objectives in this way, it is required that students answer as much of a test as possible. Some strategies for this are discussed. These include a discussion of cycling the level of question difficulty throughout a test, and increasing the level of difficulty through the parts of each question. Three measures of accessibility are proposed, based on different omission rates. Again, it is suggested that the accessibility of questions should cycle throughout the test. A brief discussion of some other criteria used in the construction of these tests is included.

1. Introduction

Many countries have gone down the track of compulsory testing of their students at a selection of year levels. New Zealand does not have a compulsory testing regime in the primary school sector. High-stakes national assessment takes place only in the final three years of secondary education. For the first ten years of students' education, New Zealand has a range of assessment initiatives aimed at gaining the assessment information needed at student, classroom, school, and system-wide levels. These give quality information to all the stakeholders in the education system, from students, parents, and teachers to policy makers and politicians. Each assessment initiative has its own distinctive niche within these wide assessment needs. This has the benefit of giving teachers a meaningful choice of assessment tools that they can use within their classrooms and schools. This helps them match the distinctive nature of their own local communities, as well as the learning needs of the students within the school.

The major assessment initiatives are:

- Progressive Achievement Tests (PAT)
- Essential Skills (ESA)
- The National Education Monitoring Project (NEMP)
- The Assessment Resource Banks (ARB)
- Assessment Tools for Teaching and Learning (asTTLe)
- National exemplars.

(See Croft, 1999, 2001; Chamberlain, 2001; Flockton, 1999; Neill, 1998; Reid, 1993; and the asTTle website.)

New Zealand is also involved in many international studies of student achievement, such as TIMMS, PISA and PERLS, so the performances of its students, schools and system can be seen in a wider international perspective.

Work has begun on a new series of mathematics tests that focus on specific areas of the mathematics curriculum. These tests are currently called Achievement in Mathematics (AIM). In time these may supersede the PAT Mathematics tests (Reid, 1993).

The AIM tests are strongly formative in nature. They attempt to assess all the objectives of specific strands and levels of Mathematics in the New Zealand curriculum (Ministry of Education, 1992).

1.1 Mathematics in the New Zealand curriculum

The mathematics curriculum is composed of six strands. Five of these are in different mathematical content areas. The sixth is Mathematical Processes, which impact on all the other five strands. Each context strand is broken down into two or three major achievement objectives. Each achievement objective is in turn divided into specific objectives. The five content strands are listed below, along with their achievement objectives.

Strand	**Achievement objectives**
Number	Exploring number Exploring computation and estimation
Measurement	Estimating and measuring Developing concepts of time, rate, and change
Geometry	Exploring shape and space Exploring symmetry and transformation
Algebra	Exploring patterns and relationships Exploring equations and expressions
Statistics	Statistical investigations Interpreting statistical reports Exploring probability

Each of these strands, except Number, define eight levels describing the development of the strand from Year 1 to Year 13. The achievement objectives are specific to each level, so that, for example, the objectives at Level 2 for Exploring Number are different from those at Level 3.

Each AIM test is targeted at assessing all the objectives of a given strand at a specified level (e.g. Number, Level 2) that are amenable to pencil-and-paper testing. The tests will cover Levels 2 to 5 of the curriculum. By getting some students to sit tests at two levels at standardisation, we will be able to provide feedback on the level at which each student is performing.

2. Designing mathematics tests

In designing the mathematics material, a number of principles have emerged. These principles relate to producing the intact tests and are based on the experience gained in the AIM project. Some of the principles are also based on the development of resources on the ARB project. A selection of these are mentioned below. These principles, of course, supplement the traditional measures of validity, reliability and so on that must always accompany good test design.

2.1 Formative versus summative assessment

The two ends of the assessment spectrum are formative (assessment *for* learning) and summative (assessment *of* learning). In AIM, the main thrust is formative assessment. There are three main ways of achieving this. Firstly, the tests are designed to look at all objectives within an individual strand. Secondly, the analysis of the results will provide information broken down in such a way that it can identify the individual strengths and weaknesses of students. This can then be fed back into teaching and learning at the level of the individual student, a group of students, the whole class, or the whole school (depending on the level of aggregation or disaggregation of the data). Thirdly, each question will have diagnostic information provided. This will be principally aimed at common wrong answers and the underlying misconceptions that lead to them. There may be some questions which also look at the alternative correct answers that students produce. The ARBs in mathematics and science have already adopted this diagnostic information approach, and it is well regarded by the teachers who use the banks (Doig, 1990; Gilbert and Neill, 1999; Marsden and Croft, 1999; Neill, 2001). As discussed in section 2.4, constructed responses have a stronger formative dimension than do selected responses, and hence they predominate in the AIM tests.

While the principal focus is on formative assessment, both AIM and the ARBs can also be used for summative purposes. With AIM, both normative information and leveling information is envisaged. Leveling information does have some formative value, but it does not reveal a student's specific areas of strengths and weaknesses. Both kinds of information can summarise where a student is at. In fact, the same assessment material can be used in either a formative or a summative way. This indicates that the divide between formative and summative assessment is not so much between the style of assessment, but in the purpose of testing and the uses to which the results are put.

2.2 Spanning the curriculum

In Achievement in Mathematics (AIM), each test is aimed at assessing all the objectives of a given strand of the mathematics curriculum that are amenable to pencil-and-paper testing. At the moment, the four tests that are undergoing trialing are at Levels 2 to 5 of the Number strand. To achieve this, a map of these levels was made. Each objective that was identified had test items written to assess the student's level of understanding of that concept. A separate question was written to test each specific objective, and each question was made up of several parts.

To ensure that students are assessed on all the curriculum objectives, it is important to ensure that they complete as much of the test as possible. This means that the tests emphasise power rather than speed. The great majority of students have sufficient time to complete the test. The following design principles were adopted to try to maximise the proportion of the test that a student answers. These principles are outlined below.

2.2.1 Controlling question difficulty

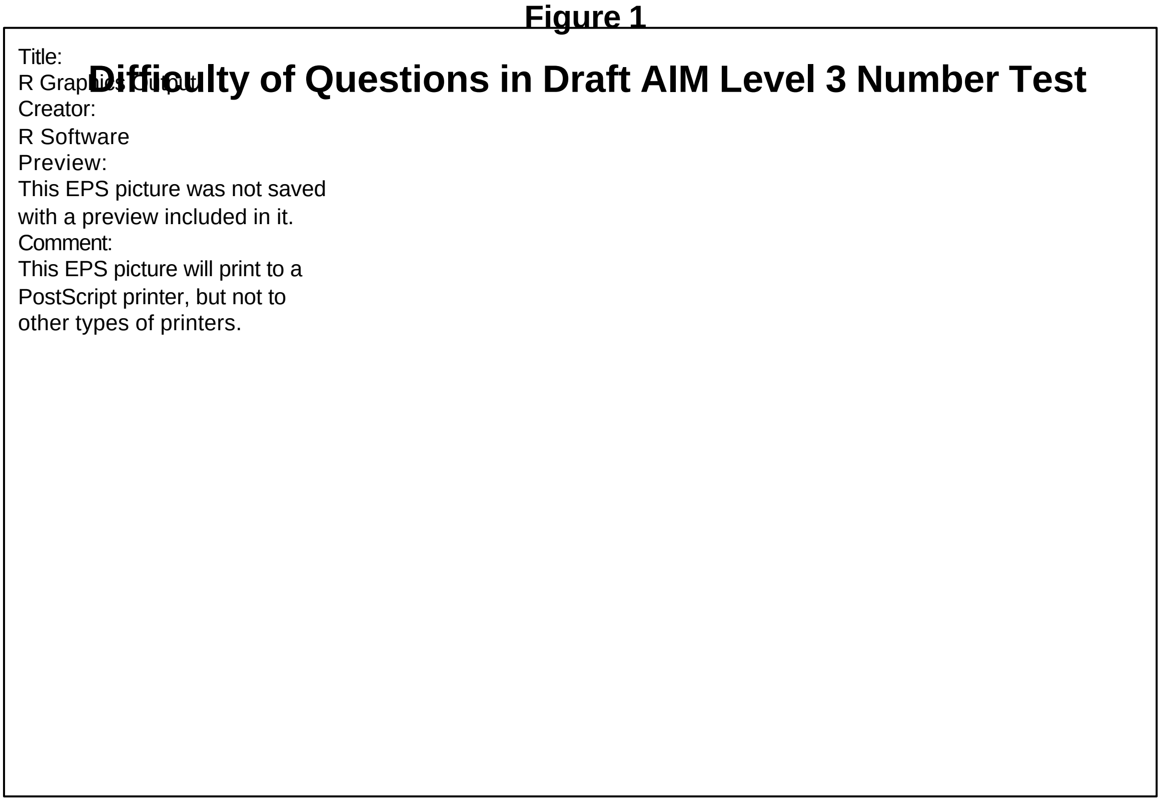
To maximise the proportion of a test that students answer, the difficulty of questions needs to be controlled. Two separate aspects of controlling difficulty are cycling the average difficulty of a question, and ensuring that the component parts of each question have increasing levels of difficulty.

2.2.1.1 Cycling the average difficulty of a question

In designing the AIM tests, the average difficulty of individual questions throughout the test varies in a cyclical manner. The easy questions are spread throughout the test, along with the hard ones. The tests preferably begin and end with relatively easy questions. Our experience in both ARB and AIM trialing strongly suggests that this encourages the student to persevere with the test.

The line in Figure 1 shows the relationship between question number and average question difficulty. For example, it can be seen that Question 15 of the Level 3 Number draft test is hard, because only about 20% of students answer it successfully. It seems preferable that questions like this be placed somewhat earlier in the test. However, if a question is relatively time consuming, then there is a case for placing it last (irrespective of difficulty), to prevent students from spending too long on it, and so omitting some of the other questions. While there is already relatively good cycling of difficulty levels in this test, some improvements can be made.

Figure 1


Difficulty of Questions in Draft AIM Level 3 Number Test
Title:
R Graphics Output
Creator:
R Software
Preview:
This EPS picture was not saved
with a preview included in it.
Comment:
This EPS picture will print to a
PostScript printer, but not to
other types of printers.

This alternation of average question difficulties between easy and hard contrasts with many other tests, where the questions typically get more difficult as the test progresses (for example, the Burt Word Reading Test – New Zealand Revision (Gilmore, Croft and Reid, 1981) and the PAT Mathematics test (Reid, 1993)). This is because the primary aim of these tests is to ascertain what level the student is performing at overall. In AIM we are assessing how well the student is performing against all of a range of objectives. AIM is primarily concerned with the power of the test, rather than with speed. This is consistent with the formative emphasis. The intent is to obtain formative information across all objectives of a curriculum level and strand of mathematics (e.g. all objectives in Level 3, Number).

2.2.1.2 Increasing difficulty within questions

Each question in the test focuses on a particular objective, and contains a number of parts looking at this objective. The questions are constructed so that the parts become increasingly difficult. The first part of the question will typically be conceptually easier than later parts. Early parts may be amenable to a wider range of simpler or street-wise strategies, but the final part often requires an understanding of the general or deeper mathematical principles involved. A good example comes from the Level 5 test question about calculating percentages.

Example 1

Write each of these marks as a ***percentage****.*

a) Kirsten got 30 out of 40 in a test. __________ *%*

b) Nick got 21 out of 30 for his project. __________ *%*

c) Natasha got 4 out of 9 for her homework. __________ *% (Answer to 1 d.p.)*

- Part a) requires recognising 30/40 as ¾ and knowing that this equals 75%. For this part, 51% of students were correct.
- Part b) requires either understanding how to calculate a percentage as 21/30 × 100/1, or recognising that 21/30 is equivalent to 7/10, which is 70%. Only 26% of students were correct.
- Part c) requires using the general formula 4/9 × 100/1 = 44.4%. Only 11% of students were correct.

The percentages of students who were correct for each part of each question (referred to as Question Part Difficulties) are used to measure the level of difficulty. In Figure 1, the levels of difficulty for each question part have been plotted. The individual letters (*a*, *b*, etc.) show the difficulty for parts *a, b*, etc of each test question. Notice how, for each question, the first part (*a*) is easiest, with subsequent parts becoming progressively harder. In some questions (e.g. Question 4) there is a big increase in the difficulty of subsequent parts. For other questions (e.g. Question 10) there is only a little increase in the difficulty of subsequent parts. The reasons behind the different levels of difficulty for each item would be useful to explore. This would be one useful dimension in a differential item analysis.

2.2.2 Accessibility of questions

The accessibility of questions I define here as the percentage of students who actually answer a particular question or part of a question (as opposed to the percentage who answer correctly). Accessibility differs from question to question. It is important to alternate the more accessible questions with the less accessible ones. Preferably, the initial question and the final question should be highly accessible, to encourage students to do as much of the test as possible. The more accessible a question, the fewer students omit it. Students omit questions for three broad reasons.

- They do not complete any questions past a certain point in the test because they run out of time, or for a variety of other reasons. The percentage of students who do this I define as the Overall Omission Percentage (OOP).

- They skip an entire question, but attempt at least one further question in the test. The percentage of students who do this I define as the Entire Question Omission Percentage (EQOP).
- They do not answer an individual part of a question, but do answer at least some more questions in the test. The percentage of students who do this I defined as the Within Question Omission Percentage (WQOP).

These three patterns are discussed below.

2.2.2.1 Overall Omission Percentage (OOP)

Students sometimes do not complete the test past a certain point either because they run out of time or for a variety of other reasons. The percentage of students who do not answer a particular question or any subsequent questions in the test I call the Overall Omission Percentage (OOP). In Figure 2, the OOP has been plotted for the Level 3 Number draft test. Notice that the rate increases throughout the test, as more and more students fail to answer any questions past that point in the test. For the final question, however, there is a sudden jump in the numbers who did not answer. It is unclear what the OOP for the final question should be. The students who did not answer could simply have run out of time, but it is unlikely that there would be a sudden jump in the number in this category. Alternatively, it could be that students who did not do the final question looked at it and decided not to answer it. This is closer to EQOP, as defined below. I have taken the approach that the OOP for the last question should be extrapolated from the OOPs from the previous questions (see dotted line on the graph).

The formula for OOP is as follows:

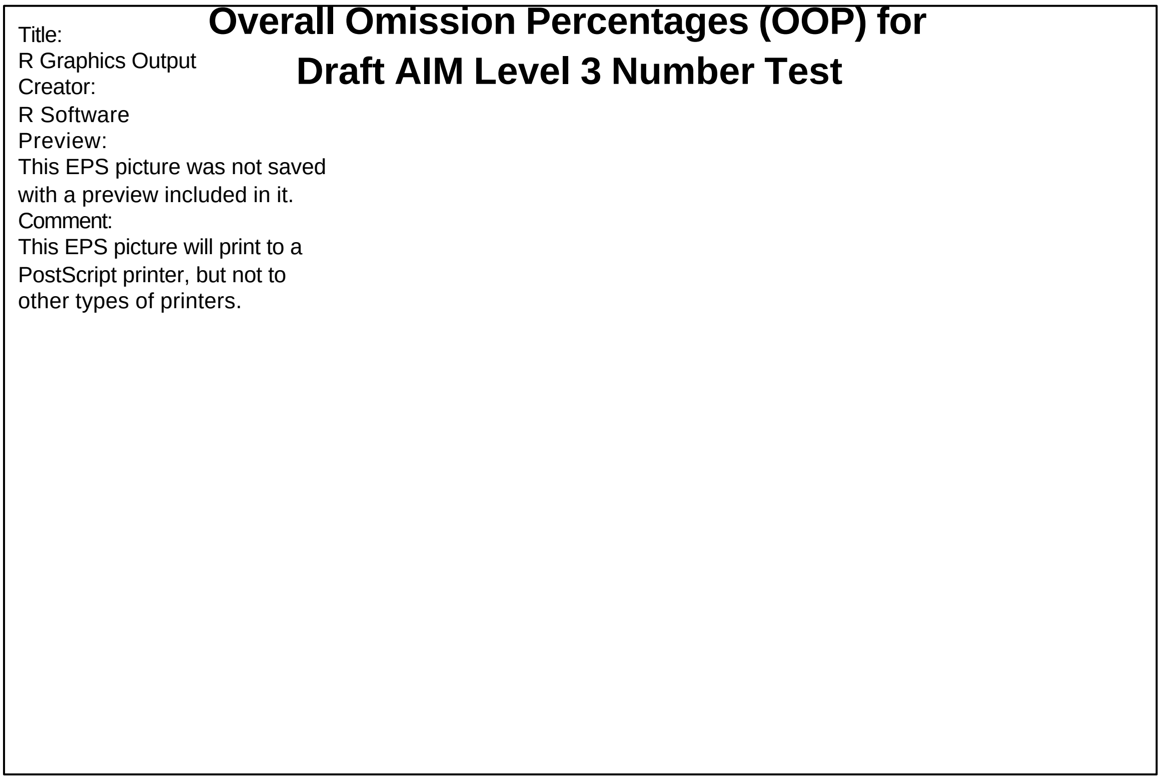
$$\mathrm{OOP}_\mathrm{i} = \mathrm{X}_\mathrm{i} / \mathrm{N} \times 100 \quad .\ .\ .\ . \quad (1)$$

Where: i = The question number.

X_i = Number of students who answered neither question i or any further questions.

N = Total number of students.

Figure 2



2.2.2.2 Entire Question Omission Percentage (EQOP)

This I define as the percentage of students who skip an entire question, but attempt at least one further question in the test (remember that the test is made up of questions with multiple parts). The EQOP gives a measure of how accessible a whole question is. As discussed above, it is unclear if the omission of the very last question in a test is due to that particular question being inaccessible, or to the student simply failing to answer any more questions in the test, as defined for the OOP. The extrapolated line in Figure 2 shows the likely OOP for the final question. The balance of omissions for the final question is then taken to be the EQOP for the final question. Students who have stopped answering further questions in the test as defined in the OOP are omitted from the EQOP as it cannot be ascertained if they would have attempted that question if they had the time or inclination to continue on with the test past the point at which they stopped.

The formula for EQOP is as follows:

$$\text{EQOP}_i = E_i / (\text{N-X}_i) \times 100 \quad . . . . \quad (2)$$

Where: E_i = Number who attempted no part of question *i*, but answered some subsequent questions (except for the last question where X_i is extrapolated from the previous values of X).

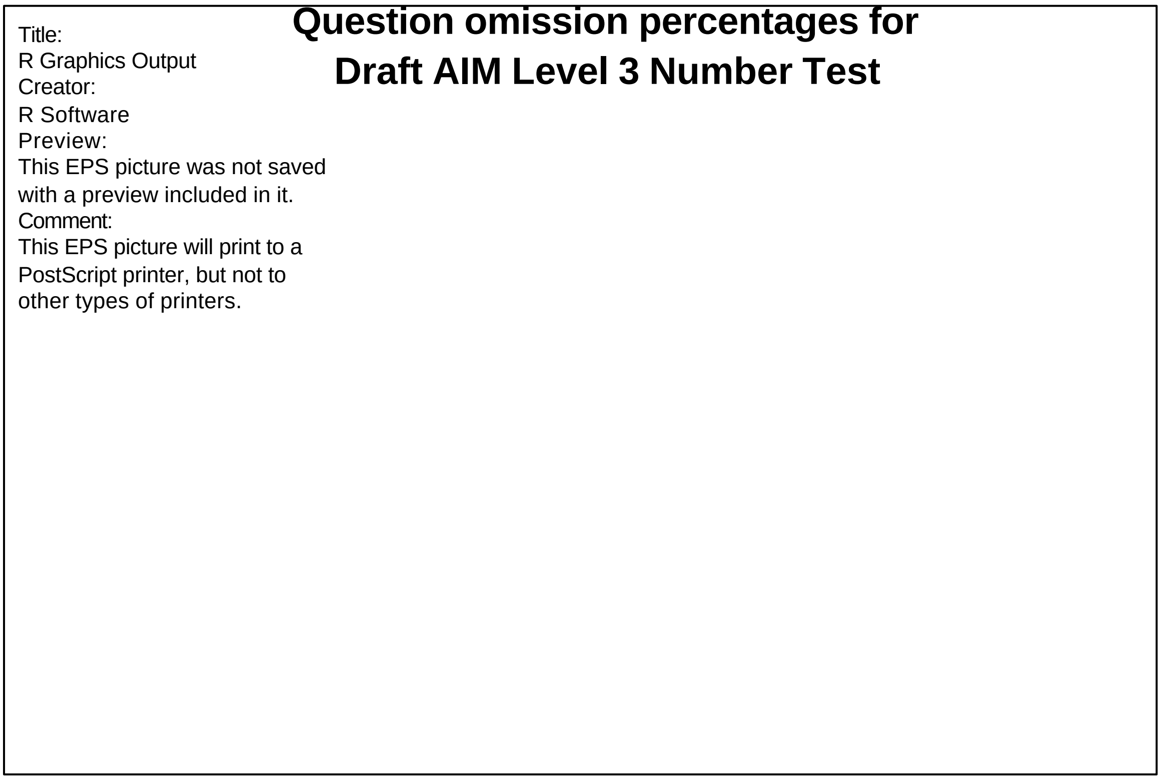
i, X_i, N as defined in (1) above.

The EQOP gives a measure of how accessible a whole question is. A high EQOP shows that many students ignore that question, while a low EQOP shows that most students attempt at least part of that question.

To ensure that students answer as much of the test as possible, the less accessible questions can be interleaved with the more accessible questions. The solid line in Figure 3 shows the EQOP for the Level 3 Number draft test. Notice how the EQOP varies from question to question. Our experience in producing ARB and AIM tests shows that it is advisable to mix the more accessible questions with less accessible ones to encourage students to continue answering the test. However, the last question has a high EQOP. It would be better to ask this question earlier in the test, or to rewrite it. A high EQOP does not necessarily mean that the question is poor, only that students are reluctant to answer it. Conversely, a low EQOP does not necessarily denote a good question. Students may be attracted to attempt a question merely because it seems worth a guess.

A useful exploration could be made of the reasons why students avoid some questions more than others. It could be the difficulty of the question, its content, its style and language, or the type of question (selected response or constructed response). The relationship between difficulty and EQOP is explored in section 2.2.3.

Figure 3


Question omission percentages for
Draft AIM Level 3 Number Test
Title:
R Graphics Output
Creator:
R Software
Preview:
This EPS picture was not saved
with a preview included in it.
Comment:
This EPS picture will print to a
PostScript printer, but not to
other types of printers.

2.2.2.3 Within Question Omission Percentage (WQOP)

The WQOP concerns the percentages of students who do not answer an individual part of a question, but do answer at least some more questions in the test. It therefore measures the within question accessibility of the component parts of the question. Again, the students who omit the remainder of the test from a given point onwards are disregarded.

The formula for the WQOPis as follows:

$$\text{WQOP}_{ij} = M_{ij} / (N\text{-}X_i) \times 100 \quad .\ .\ .\ . \quad (3)$$

Where: j = Part number within question i.

M_{ij} = Number of students who omit part j of question i, but answered some subsequent questions (except for the last question where X_i is extrapolated from the previous values of X).
e.g. $M_{5,3}$ is the number of students who omit part 3 of Question 5.

i, $X_{i,}$ N as defined in (1) above.

In Figure 3, the WQOPs for each question in the Level 3 Number draft test are displayed as the individual letters (*a, b* etc) For almost all questions, the WQOP increases throughout the question as fewer students answer the later parts of a question than answer the earlier parts. Remember that the parts of the question increase in difficulty.

Each question has its own variation of WQOP between its individual parts. Notice how the WQOP stays very constant for questions 3, 4, 5, and 6. This means that students answer the later parts of these questions roughly as often as the earlier parts. This means that these questions are more internally accessible than those with bigger gaps between their WQOPs. It would be ideal to spread questions of this type throughout the test, with the last question having only a small change in WQOP, to encourage students to complete the test.

By contrast, Questions 7 and 11 have very big differences in WQOP between their component parts. This means that far fewer students are willing to attempt the later parts of these questions, compared with the earlier parts. Again, questions with a big difference in accessibility between their parts should be split up, to avoid students stopping answering any more questions.

Inevitably, the final few questions will have bigger WQOPs as student fatigue sets in. Our experience suggests that the last question should have a relatively small variation in WQOP to maximise the chances of test completion. Question 15 should hence be placed elsewhere in the test.

It is of interest to note that very few students skip the first part of a question, but then attempt subsequent parts of it. This is probably because the parts become successively more difficult. Thus the WQOP for the first part of a question is close to the EQOP for that question. Written mathematically, this is $\text{EQOP}_i \approx \text{WQOP}_{i1}$.

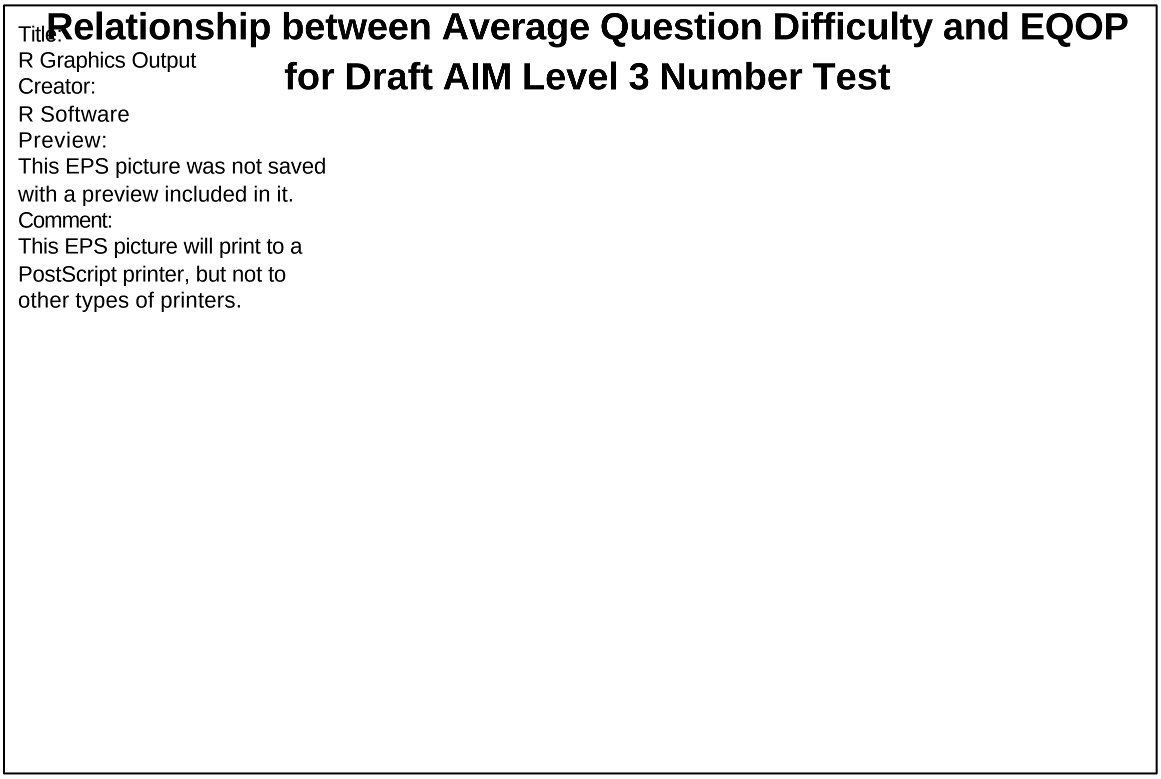
A further research question would be to investigate how the WQOP profiles differ for able and less able students. One would hypothesise that less able students would have bigger differences of WQOP within a question, while more able students would have smaller differences. This would mean that fewer able students stop answering as they proceed to the later and more difficult parts of a question. It could be that this is counterbalanced by the fact that less able students are less likely to attempt the question at all, and so have both a higher “threshold” (WQOP_{i1}), and subsequently a smaller difference between WQOPs than anticipated. It may be that an alternative measure to WQOP could be used that looks only at students who answered at least a part of the question of interest. It would be useful to investigate the differential item functioning for whole questions (EQOP) and for individual parts of questions.

2.2.3 Relationship between item difficulty and omission percentages

There is only a weak link overall between the difficulty of questions and the rates at which students answer them. This relationship differs markedly from question to question.

This is shown in Figure 4, which plots the average difficulty of a question and the Entire Question Omission Percentage (EQOP). If a relationship existed, one would have expected the points to be roughly linear. That is, one would expect the correlation coefficient between these two variables to be strongly negative, meaning that harder questions have fewer students attempting them. However, the value of the correlation coefficient is only 0.021, which is clearly non-significant at the 5% level, suggesting that no relationship exists between difficulty and EQOP.

Figure 4



Relationship between Average Question Difficulty and EQOP for Draft AIM Level 3 Number Test

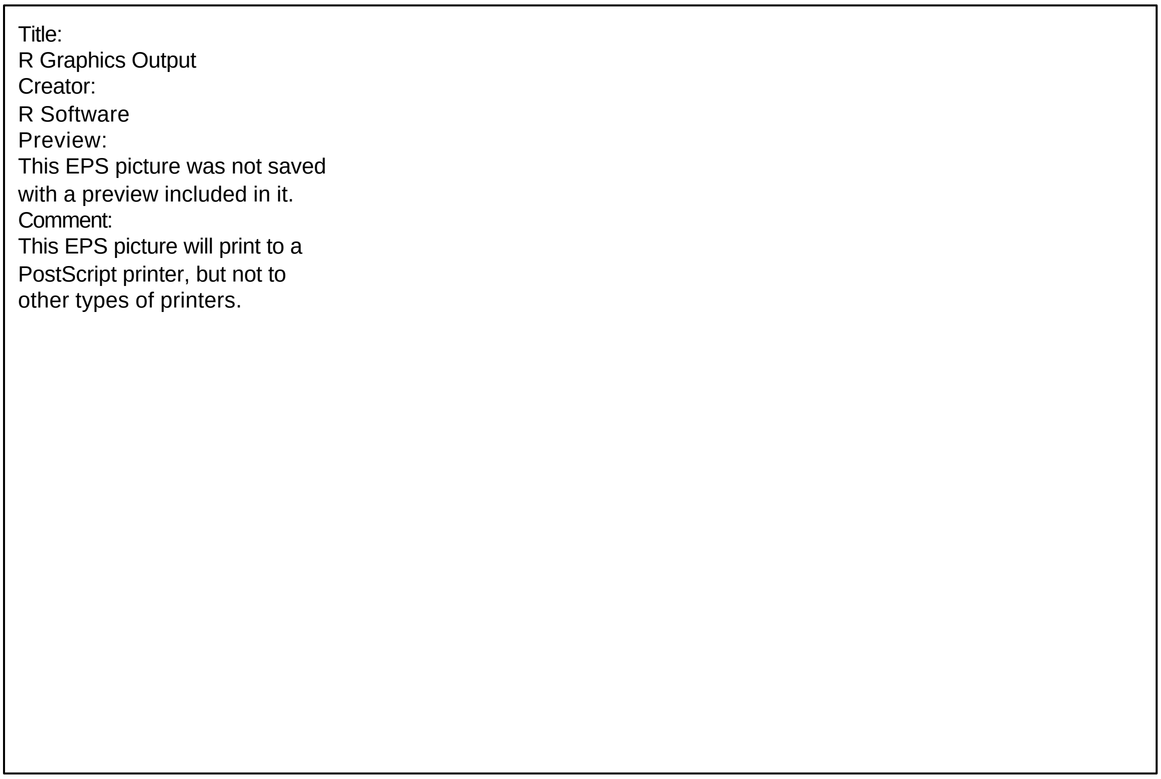
Title:
R Graphics Output
Creator:
R Software
Preview:
This EPS picture was not saved with a preview included in it.
Comment:
This EPS picture will print to a PostScript printer, but not to other types of printers.

A graph similar to Figure 4 could be plotted of WQOP by difficulty for each part of a question. An analysis of the correlation coefficient for the relationship between WQOP and the item difficulty of individual parts of a question shows only a weak relationship. The overall correlation coefficient is –0.486, which is significant at the 1% level. This is, however, influenced by the two high WQOPs in Question 11 and the high WQOPs in Question 15. By eliminating these five points, the correlation coefficient drops to –0.354, which is non-significant at the 5% level, again indicating that item difficulty is only a minor factor in omission rates for this test.

For individual questions, the relationship between the difficulty of different parts of the question and the number of students who do not answer individual parts (as measured by WQOPs) varies. For example, in Question 4, the second part is far harder that the first part (75.5% vs 14.0%); yet students are equally likely to answer each part of the question. By contrast, in Question 11, the four parts increase in difficulty by only a modest amount, but the percentages of students not answering specific parts of the question increase from 3.6% to 17.0%, by far the largest difference in the Level 3 Number draft test. This is demonstrated in Figure 5. This underlines the fact that the relationship between item difficulty and the percentage of students omitting a part of a question is a function of the specific question. It would be useful to explore the reasons behind the different relationships between item difficulty and omission rates within questions.

Figure 5

Relationship between Difficulty and WQOP for Draft AIM Level 3 Number Test

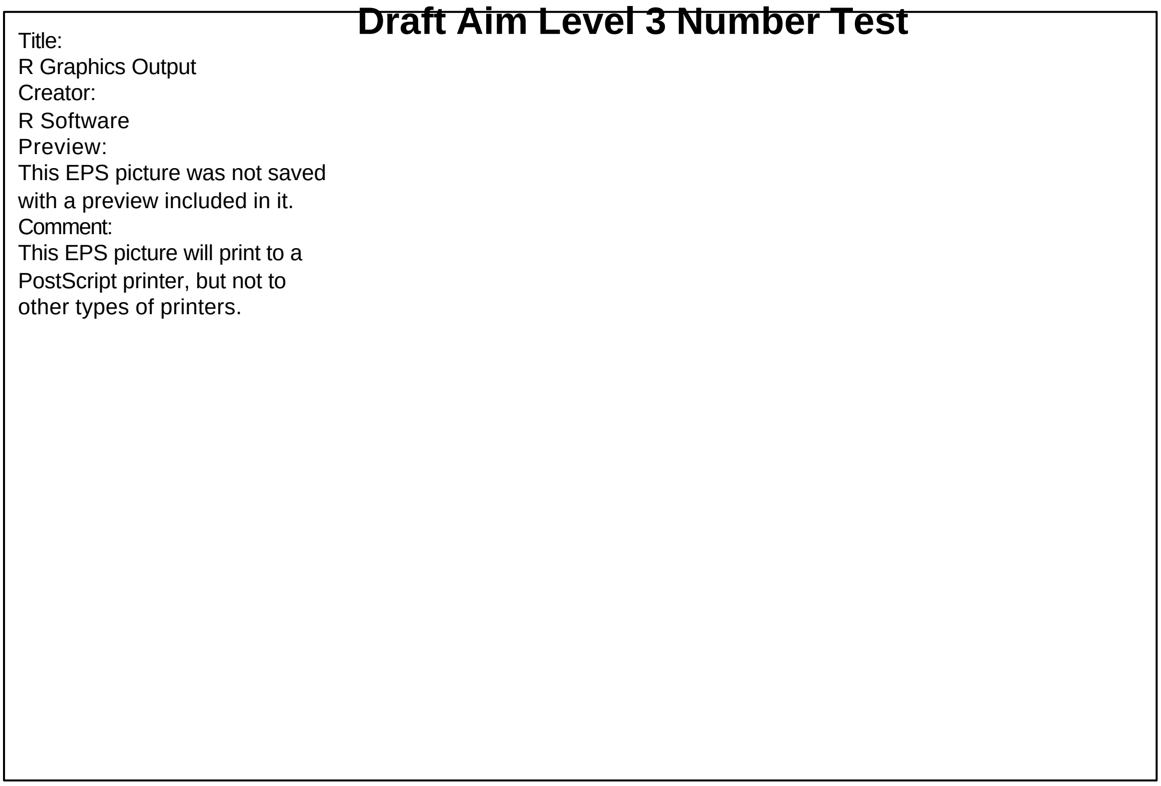


Title:
R Graphics Output
Creator:
R Software
Preview:
This EPS picture was not saved
with a preview included in it.
Comment:
This EPS picture will print to a
PostScript printer, but not to
other types of printers.

2.3 Growth in achievement of students

The tests were trialed on students at two different year levels (Years 6 and 7 for the Level 3 Number draft test). It was expected that there would be significantly more Year 7 students getting a question correct than the Year 6 students. This is shown graphically in Figure 6. Each question, other than Question 2, shows an appreciable growth in achievement between Year 6 and Year 7 students. The growth varies from 0.6% for Question 2 up to 24.6% for Questions 7 and 10. The overall growth from Year 6 to Year 7 was 13.3%. If more that an occasional question did not show growth, then reasons for this should be sought, or the test redesigned.

Figure 6



Draft Aim Level 3 Number Test

Title:
R Graphics Output
Creator:
R Software
Preview:
This EPS picture was not saved with a preview included in it.
Comment:
This EPS picture will print to a PostScript printer, but not to other types of printers.

2.4 Constructed versus selected response

In both the AIM tests and in the ARBs there is a mix of selected response questions (multiple choice, matching responses, ordering responses etc) and constructed responses which range from a simple number to a longer response from the student. In formative assessment, constructed responses are generally far preferable although selected response can have some role to play. The preference for constructed response is especially strong when the test is formative (Croft, Dunn and Brown, 2001). There are a number of reasons for this preference.

- A constructed response clearly implies the method by which the student reaches their answer. This therefore has a strong diagnostic dimension. The incorrect answer often indicates the misconception a student has. With a multiple choice distractor, the student may have chosen an incorrect response for a variety of reasons. We can tell that they do not understand the principle being tested, but we have only weak clues as to why. They may well have guessed. The formative-diagnostic potential is hence eroded, as the choice of a distractor does not clearly imply the reason behind that choice.
- Conversely, the selecting of the correct response does not clearly imply that the student understands the particular concept of a question. They may have got it correct by chance or by test-wise strategies. Cumulatively a high score implies strong skills overall, but does not imply the possession of each individual skill. Again this suggests that multiple choice is better suited to summative rather than formative assessment.
- Figure 5 shows that students are just as likely to answer subsequent parts of some questions even though they are much harder that the earlier parts. Questions 4, 5 and 6 on this graph are the most noticeable examples, and each one is a selected response. This constitutes a prima facie case that students are happy to guess in selected response question. The low EQOPs for these questions (shown in Figure 3) also support the case for guessing.

2.5 Reporting of results

It is anticipated that the results will be available in a range of ways to teachers and schools. The primary motivation is for the results to inform teaching and learning. This means that specifying the particular areas that a student has strengths and weaknesses in is important. Tools such as kid maps of achievement will be available as well as “class maps” of strengths and weaknesses at a class or school level. These formative tools will be supplemented by the more formal measures such as norms in the form of stanines. There will also be leveling information as to what level of the curriculum a student is performing at. This later measure summarises the overall level of achievement, but does not of itself show a students strengths and weaknesses. It will describe what a typical student at that given level has or has not achieved, but this does not necessarily pertain to the individual student. It can therefore be seen as more a summative than formative measure.

3. Conclusion

As discussed in the introduction of the paper, the AIM tests are designed to complement the other assessment tools that are available to New Zealand schools and teachers. By having a range of assessment resources, teachers can make meaningful choices to help get the best fit for the particular needs of their students and schools. Reliable, valid, and nationally comparable data can be obtained without going down the more prescriptive route of compulsory national testing.

The AIM tests have the distinctive feature of looking at specific strands of the curriculum in depth. A primary focus is on formative and diagnostic power, with an emphasis on constructed responses with relatively few selected responses. The second main principle is to give full coverage to the specific strand of the curriculum being assessed. To achieve this the tests are designed to maximise the number of questions that are attempted by students. Three measures of the rates at which students omit questions are proposed: the percentage of students who stop doing the test past a certain point (OOP), the percentage who omit a whole question (EQOP), and the percentage who omit individual parts of questions (WQOP). By taking account of these, plus alternating item difficulty, high response rates are achieved.

Bibliography

Croft, C. (1999). *School-wide assessment. Using the Assessment Resource Banks*. Wellington: New Zealand Council for Educational Research.

Croft, C., Dunn, K., & Brown, G. (2001). *Essential Skills Assessments—Information Skills. Teachers Manual.* Wellington: New Zealand Council for Educational Research.

Chamberlain, M. (2001). *The Development of exemplars in New Zealand. Background and rationale*. www.tki.org.nz/r/assessment/two/research1_e.php .

Doig, B. (1990). *Diagnostic mathematical profiles*. Melbourne: Australian Council for Educational Research.

Flockton, L. (1999). *School-wide assessment. National Education Monitoring Project*. Wellington: New Zealand Council for Educational Research.

Gilbert, A., & Neill, A. (1999). *Assessment Resource Banks in mathematics and their diagnostic potential*. Paper presented at NZAMT Conference, Dunedin, New Zealand.

Gilmore, A., Croft, C., & Reid, N. (1981). *Burt Word Reading Test – New Zealand Revision*. Wellington: New Zealand Council for Educational Research.

Marsden, C., & Croft, C. (1999). What do students know in science? Analysis of data from the Assessment Resource Banks. ***set:*** *Research information for teachers, No. 2.* Wellington: New Zealand Council for Educational Research.

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

Neill, A. (2001). An introduction to the Assessment Resource Banks (ARBs and their diagnostic potential. *The New Zealand Mathematics Magazine, 38* (1).

Neill, A. (1998). Assessment Resource Banks in mathematics: How can we help teaching and learning? Mathematics teaching: Where are we at? Seminar proceedings. Wellington: New Zealand Council for Educational Research.

Reid, N. (1993). *Progressive Achievement Test of mathematics*. Wellington: New Zealand Council for Educational Research

The asTTle package. www.tki.org.nz/r/asstle/deliverable_e.php .