Policy and Developments in Mathematics Assessment in England

Sarah Maughan, National Foundation for Educational Research
Louise Cooper, National Foundation for Educational Research

1 Abstract

This paper sets the context by detailing the policy developments in mathematics assessment in England, highlighting the many changes that have occurred over recent years and the reasons for these. The reasons include: to encourage greater engagement with mathematics, which in turn could lead to greater participation rates beyond compulsory schooling; and to develop higher levels of deep understanding of mathematical concepts, allowing them to be used more effectively in new contexts, including in the work place.

The paper goes on to describe findings from on-going research that is being conducted at the National Foundation for Educational Research in England into the features of good mathematics assessment. Good mathematics assessment is defined as assessment that will support the teaching and learning of problem solving and of higher order skills in the mathematics classroom. Finally, the paper considers whether the current policy developments will encourage the use of our model of good assessment, and if not, what some of the constraints may be.

2 Introduction

Good quality mathematics education in England is seen as key to the success of the economy, despite on-going concerns that the students who leave school do not have all the skills required for the workplace. A landmark report highlighting this issue was Mathematical Skills in the Workplace (Hoyles *et al*., 2002) which concluded that there is an increasing need for mathematical literacy skills in the workplace, despite the fact that technology is used much more widely. In addition to this we have fewer students in England following a mathematics course past the end of compulsory schooling, with the numbers taking A Levels (the qualifications used for university entrance in England) in mathematics falling in the early 2000s (although increasing again now) and the numbers studying mathematics at university decreasing. These issues have led to a review of the mathematics curriculum at primary and secondary levels, and significant change in both the curriculum and the assessment of it in England.

In 2004, Professor Adrian Smith published his report into post-14 mathematics: *Making Mathematics Count* (Smith, 2004). The report highlighted the importance of mathematics for its own sake, for the economy and for individuals. It highlighted a number of issues with the provision of mathematics education including: the supply of appropriately qualified mathematics teachers, the need for a more coherent system of professional development for teachers of mathematics and problems with the testing and examination system. He went on to make a number of recommendations about systemic change which continue to have a major influence on mathematics education in England.

A few years later, Peter Williams was asked to review mathematics teaching in early years settings and primary schools (Williams, 2008), considering issues such as the most effective pedagogy. However, unlike the Smith Report, the Williams Report did not consider the most appropriate form of assessment of mathematics.

These two reviews have led to recommended changes to the mathematics curriculum, to the way in which it is assessed, and to the way in which teachers are trained to deliver it. The changes to the assessment system are the focus of this paper and are described in more detail below.

3 The Assessment of Mathematics in England

A national curriculum was introduced in England in 1988, and over the next few years, tests were introduced to assess the curriculum at the end of a number of key stages. The following table provides a brief (and simplified) overview of the assessment system.

Table 1: Simplified View of the Assessment System in England.

Assessment	**Age Group**	**Notes**
End of key stage 1 tests	Students in year 2 (age 7)	Tests were originally compulsory at the end of key stage 1 in English and mathematics. However, since 2005 the requirements have changed: the tests must still be used, but the results are used to inform teacher judgements which are then reported, rather than the test results being reported directly.
Optional tests	Students in years 3, 4 and 5	Towards the end of the 1990s tests were made available to primary schools following the same model as the end of key stage tests. The use of these tests has always been optional, although a large proportion of schools choose to use them (95% at the height).
End of key stage 2 tests	Students in year 6 (end of primary education, age 11)	Tests were originally compulsory in English, mathematics and science. Science was compulsory for the last time in 2009 although the English and mathematics tests continue.
Optional tests	Students in years 7 and 8	Early in the 2000s tests were developed for the first two years of secondary education. These are optional and are used much less than the primary optional tests (only about 33% of schools opt in).
End of key stage 3 tests	Students in year 9 (end of lower secondary education, age 14)	Compulsory tests were used in English, mathematics and science. All these tests were abolished in 2008.
GCSE	Students age 16 (end of compulsory schooling)	Students tend to select eight or more subjects to follow for 2 years, from the age of 14 to 16 years old. English, mathematics and science are compulsory.
AS/ A Level	Students age 18	After 16 students can choose to stay in education for further study leading towards university entrance. Students generally choose four subjects. No subjects are compulsory.

4 New Assessment Models and Pathways

A particular feature of the assessment context in England is the way that test results are used for accountability purposes. Over the last 20 years the Government has set up an accountability framework for all public services, in which targets have been set and then used to measure the performance of public service providers. For education, the targets were related to test and examination results – in primary schools the target related to the percentage of students achieving the expected level in the key stage 2 tests and in secondary schools the percentage of students achieving A*-C in five GCSEs (latterly this had to include English and mathematics). Progress against the targets has been published in the media in the form of league tables of schools. Failure to meet the targets leads to consequences for head teachers and teachers, initially in terms of support and guidance being provided but ultimately it could result in the school being closed. In May this year a new Government was elected in England. The new Government has pledged to review the system of accountability although they have stated that performance tables of schools will continue.

This use of test results for very high stakes purposes has led to an emphasis in schools on the skills and content needed to pass the tests, arguably at the expense of other subjects and of the aspects of the curriculum not covered in the tests (Mansell, 2007, Ofsted, 2008, Royal Society, 2010). In 2010 teacher unions called for a boycott of the key stage 2 tests and 25% of schools did not participate. The Government has announced that the tests will run again in 2011 in the same format, but that a review will be conducted. This will mean that a new model for the tests will be introduced from 2012 onwards.

In mathematics the nature of the tests, as well as the content, is said to determine the nature of teaching and learning and some argue that this inhibits the use of more effective or innovative modes of teaching (NCETM, 2008). In primary schools the teaching of mathematics centres around the key stage 2 tests, as targets are set for achievement in these, with much of the final year of schooling being spent on test preparation in some schools. In secondary schools the picture is more complex. The targets focus on success in the GCSEs in particular, although a number of changes are currently working their way through the system.

The Smith Report in 2004 introduced the idea of appropriate mathematics pathways. It suggested that no single model of mathematics assessment is right for all students, rather there should be '*a highly flexible set of interlinking pathways that provide motivation, challenge and worthwhile attainment across the whole spectrum of abilities and motivations'* (p. 8). The proposals include a new model for GCSE mathematics, including: the introduction of two linked GCSEs in mathematics as an alternative to one; the introduction of functional skills[1] in mathematics which a majority of students would enter alongside the GCSEs; the inclusion of more functional skills-type questions in the GCSE itself; and further changes to the examinations for post-compulsory mathematics.

The extent to which this high stakes accountability regime causes problems for the teaching and learning of mathematics depends somewhat on the nature of the tests. In 2008, Ofsted (the agency responsible for inspecting schools in England) found in its report *Mathematics: understanding the score*, that there was too much teaching to the test in mathematics lessons. This led one university mathematics expert to respond that teaching to the test is only a problem if it is '*the wrong test'*.

[1] Functional skills assessments are intended to assess the knowledge, skills and problem solving approaches that can be used in work, life and further learning.

The Ofsted Report made the point that test results had improved but understanding of mathematical concepts had not. This improvement had come about through familiarisation with test taking techniques, revision classes and intensive intervention. This suggests that success in the tests is possible without a deep understanding of the mathematical concepts involved.

In part as a result of this debate, NFER is currently conducting a small programme of research into existing mathematics provision and how tests should be designed to encourage the teaching of desirable skills. If it is possible to achieve well in the current tests without a deep level of understanding of the mathematics required, are there changes that could be made to the tests and examinations to prevent this, that is, what would 'the right test' look like? The first phase of this research was reported at the IAEA conference in Brisbane (Maughan, 2009) and is summarised in section 5 below. The following sections present results from later stages of the work and discuss the findings in light of the current policy context.

5 Phase 1 Research Results: Features of a good mathematics assessment

The early phase of the NFER research was made up of two distinct parts: a review of the question types and approaches used in existing mathematics tests and telephone interviews with mathematics experts about what a good mathematics question would be like.

A number of current maths assessments were reviewed, including national curriculum tests at key stage 2 and GCSE examination papers. The questions in each of these assessments were analysed in terms of the content and skills assessed and the types of items included. The review of content focused on the match with the programme of study for mathematics, with a particular focus being placed on categorising questions against the 'using and applying' strand for key stage 2. The skills were classified into different levels depending on the extent to which they required understanding or extended reasoning.

In terms of content, there was a high representation of calculating or number questions in the vast majority of assessments. Measuring questions usually appeared infrequently for younger students, but became more prominent (as geometry and measures) in assessments for older students.

Analysing questions with reference to the using and applying strand for key stage 2, revealed that the most popular types of questions used were those asking students to solve problems, or to use reasoning. However there was a low incidence of representing, enquiring and communicating questions.

One of the clear patterns that emerged with regards to the analysis of skills was that there are relatively few extended reasoning questions, and there are also few questions that were classified as assessing understanding. Instead, recall and computation questions seem to make up the bulk of the assessments.

The second part of this early research involved a number of telephone interviews. Researchers conducted interviews with mathematics experts, including teacher educators, maths test developers and university mathematicians. Interviewees were asked to describe a particular example of what they considered to be a good maths question. This example was then used to stimulate discussion about what they thought were the defining features of good maths questions. The following table provides those features which were given by more than one interviewee.

Table 2: Number of times different features of 'good' maths tests were mentioned.

Feature	**Number of Mentions**
Open-ended	6
Using interesting or unusual scenarios	5
Connections made across different areas of maths	5
Promotes good practice	3
Students should learn something	3
Using real life situations	3
No time limit/ extended time limit eg a week	2
Multi-step questions	2
Interesting	2
Tests what it is that you want to know about the students (ie is valid)	2
Should stimulate ideas for the classroom	2

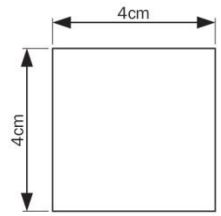
A number of the features that came up in response to this question relate more to a good maths test rather than a good maths question. These include:

- a variety in the types or genres of questions;
- a good flow between questions;
- good content/ curriculum coverage;
- start with easy questions and progress to more difficult ones;
- a range of 'lengths of reasoning' (time needed to answer a question).

The features of a good maths question were grouped into a number of recurring themes and discussed in Maughan, 2009. One example of a good maths question given in the interviews is provided below.

Example 1: an example of a 'good' maths question.[2]

Q1.



The square above is an **equitable shape**, that is its area is the same as its perimeter.

a) Draw **two** more equitable shapes

b) When does a shape have the same area and perimeter?

This example was provided as a good example of an open-ended question that could be answered in different ways by different students with a range of abilities. It involves generalising from a given example to further examples and then constructing a rule, this requires a high degree of understanding. The mark scheme could be in the form of level descriptors, similar to an English writing mark scheme.

6 Phase 2 Research: Evidence of 'good' features in mathematics assessments

Following the interviews with maths experts, NFER researchers reviewed a number of existing maths assessments to find examples of good maths questions (with the features highlighted in the telephone interviews). As part of the review we looked at assessments of mathematics from different contexts, including national and international surveys (including the Assessment of Performance Unit (APU) and the Trends in International Mathematics and Science Study (TIMSS)), classroom assessment materials (a range of materials developed by publishers and government agencies) and high stakes tests and examinations (including the ATM-SEG GCSE).

6.1 Examples of open-ended questioning and extended tasks

The strict administration and marking arrangements for most national or international surveys mean that in many cases, open-ended tasks are largely excluded in favour of short answer or closed questions that are relatively straightforward to mark. However, this is not always the case. Given its international context and high profile, the TIMSS assessment requires administration and marking reliability that are as high as possible, yet even in this context it is possible to produce some limited extended questions. The review considered three such examples from the 2007 paper: 'Class Trip',

[2] This example has not been worked up for use in a real test, but is presented as it was described in the interview.

'Triathlon' and 'Marytown'. In order to address the multi-step nature of these questions, students can be awarded 'follow through' marks in situations where they make a mistake in an earlier stage.

Another, much more radical, example of extended work being used for high stakes assessment is in the ATM-SEG GCSE which ran from 1986-1994. By using a moderation model, participating teachers were able to assess a sample of on-going investigative work carried out in the classroom at any time during the two years of the course. The assessed work varied, depending on the school, and gave students the opportunity to investigate an area of personal interest, as well as giving enough time for students to explore a topic in depth, often working on a project for several weeks. This allowed teachers and students a high degree of flexibility, however the use of the moderation model is potentially costly to run, and it might raise questions about public trust and reliability, especially in today's context in which the results of GCSEs are used for accountability purposes for secondary schools.

Examples of extended questioning being used in low stakes assessment are more prevalent. The reasons for this are likely to be twofold. Firstly, the low stakes assessments do not have the same time restrictions as many high stakes assessments and tasks can be carried out over hours or even weeks. Secondly, the fact that the assessment is low stakes means that there is less pressure for the results to be reliable on a national scale, and less pressure for the assessment to be marked within a certain time frame. Therefore there is more scope for using extended questioning which can be more difficult to mark reliably and often takes more time to mark.

The Year 5 Optional Skills Assessments in Wales are a good example of innovative assessment which aims to capture the student's thinking process as and when it happens. Groups of students work through a range of extended tasks, some of which may take several lessons to complete, and choose their own strategy for doing so. Whilst the class is working on a task, the teacher focuses on one or two students at a time and assesses how well they can demonstrate skills described in the Skills Framework. Teachers are provided with a list of prompting questions, which are interspersed within the task, and which aim to get students to talk about what steps they have taken and to explain their thoughts and reasons. Teachers are also encouraged to give students assistance and advice in situations where there is a perceived impasse in the flow of the task.

The Mathematics Assessment Resource Service (MARS) at Nottingham University has developed a range of innovative, optional open-ended tasks, from 5 to 45 minutes in length, administered and marked by the class teacher. They are designed to give summative and formative feedback to guide student development and also to aid the development of teaching skills. The 'scoring guide' credits several different methods of arriving at an answer, using the depth of a student's answer to differentiate between final results.

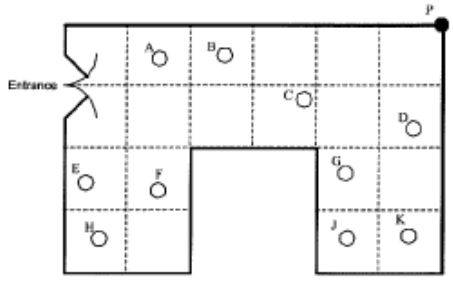
Two examples of MARS questions are provided below.

Example 2: MARS Security Camera.

SECURITY CAMERA

A shop owner wants to prevent shoplifting.
He decides to install a security camera on the ceiling of his shop.
The camera can turn right round through 360°.
The shop owner places the camera at point P, in the corner of the shop.

Plan view of the shop



1. The plan shows ten people are standing in the shop.
   These are labelled A, B, C, D, E, F, G, H, J, K.
   Which people cannot be seen by the camera at P ?
2. The shopkeeper says that "15% of the shop is hidden from the camera"
   Show clearly that he is right.
3. (a) Show the best place for the camera, so that the it can see as much of the shop as possible.

   (b) Explain how you know that this is the best place for the camera.

Example 3: MARS Consecutive Sums.

CONSECUTIVE SUMS

The number 15 can be written as the sum of consecutive whole numbers in exactly three different ways:

$$15 = 7 + 8$$
$$15 = 1 + 2 + 3 + 4 + 5$$
$$15 = 4 + 5 + 6$$

The number 9 can be written as the sum of consecutive whole numbers in two ways:

$$9 = 2 + 3 + 4$$
$$9 = 4 + 5$$

The number 16 cannot be written as a consecutive sum.

Now look at other numbers and find out all you can about writing them as sums of consecutive whole numbers.

Write an account of your investigation. If you find any patterns in your results, then try to explain them fully.

Which numbers can not be written as the sum of consecutive whole numbers?

6.2 Barriers to Using Open-ended Questions and Extended Tasks

Getting started/getting lost

In an open-ended problem students may be required to find their own way 'into the problem' and work their way through it. This process will inevitably require some time spent trying out different approaches and deciding on a particular route. In some maths assessments, assessors interact with the students at least some of the time (e.g. Year 5 Optional Skills Assessments). The help that is offered to students needs to be taken into account when drawing conclusions from the results of the assessment. Therefore the assessor will need to decide carefully when to intervene and what type of support to give each student and the marking process must enable the teacher to provide information about what support has been given.

Time taken to answer

Traditional mathematics tests are usually constructed in part with the aim of minimising the time students have to spend on each question. This allows a larger number of questions to be included and a better coverage of the curriculum. In the test development process, unusual questions are sometimes rejected by review groups and expert panels due to the concern that some students would spend too much time puzzling over the question and then run out of time for answering other questions which would have been within their capabilities. Rejecting unusual questions can, however, lead to unhelpful restrictions in task types. Some argue that students should be expected to direct their own behaviour and possess such levels of meta-cognition that they know when to move on to the next task or try a different approach.

Context

Open-ended questions are often set in a context both to assess how students solve real-life problems and to help them think their way through the problem. Vappula and Clausen-May (2006) compared students' performance, and chosen approaches, in similar questions set in and out of context. They found that *'regardless of the additional elements of reading and translating a context into something that produces a numerical answer, in none of the comparisons did contextualisation hinder a greater number of pupils than abstract presentation did.'* Also, *'in division questions, contextualisation encouraged pupils to attempt the question and to use informal or drawn methods.'*

The level of contextualisation varies from one assessment task to another. Test questions typically use only a few sentences to set the scene in order to minimise the time and effort required for reading. Extended tasks make more use of building up the problem context. Optional assessments have more flexibility than compulsory assessments in the time required for getting to grips with the context and the potential need for teacher help.

It is important that the context is accessible to all students. Students from different backgrounds may vary in their familiarity or interpretation of different contexts. Task contexts that engage students could result in better performance than those that do not generate much interest. Sometimes when trying to meet all the different demands from a suitable task context, there is a risk that the scenario becomes contrived or unrealistic.

Communication

Communicating mathematics is an area that features in several assessments and is important in bringing mathematics into a real life scenario. Communication can include written, oral and symbolic

communication. It can take place in a group situation where the student must argue their point of view persuasively or as a presentation where the student is, in effect, communicating to a passive audience. In each case the student must put their mathematical ideas into a format that can be understood by the listeners. This requires a good understanding of the mathematics involved and also of the audience's knowledge-base. One example of mathematical communication was given by MARS where the student is required to give instructions over the telephone to another student on how to make a specific T-shirt design.

7 Discussion

As mentioned in the introduction above one criticism of the current tests is that they narrow the curriculum. Students are taught only those sections of the curriculum that are assessed by the tests and the tests include few questions that assess deep understanding or extended reasoning. It is not likely to be possible to assess all aspects of the curriculum using short response questions. Therefore, the use of a mixture of different types of assessment types to suit the different demands of what is being measured may be more appropriate. A simple example would be to use closed item types for assessing the knowledge of simple facts and to use extended tasks or group work for assessing communication and problem solving skills. This may be considered as part of the proposed review of the key stage 2 mathematics tests. It has been announced that the current model of testing will continue in May 2011 but that a new model is being discussed for later years. This provides the opportunity for more open-ended tasks to be included in a new model.

At the secondary level a new pair of GCSEs is currently in development. The two GCSEs will each have a slightly different focus with one assessing Applications of Mathematics and the other assessing Methods in Mathematics. The new GCSEs are '*designed to encourage learners to develop problem solving skills in mathematics*' (OCR, www.ocr.org.uk/qualifications/type/gcse_2010/maths/app_of_maths_pilot). In addition to the new linked pair of GCSEs, all existing GCSEs in mathematics have been revised to include an aspect of functional skills assessment within them. The assessments tend to be made up of more open-ended tasks than in GCSE and be set in real world contexts. Both of these recent changes may lead to the requirement for students to develop a deeper understanding of mathematics to be able to respond.

The issue with time taken to answer the more open tasks is likely to be alleviated in these new tests and examinations by the use of a variety of assessment types. It is possible for a small number of open tasks to be used to assess deeper understanding while a number of more objective questions could also continue to be used to assess knowledge and to ensure curriculum coverage.

An alternative approach to the time issue is to modify the purpose of the tests. If the purpose of the assessment is not to give student- or school-level feedback, a sample of students can be used rather than the whole cohort. TIMSS makes use of a model where different samples of students sit different tests with some common link items. This allows a wide range of topics to be assessed without presenting too many test questions to any one student. A further related benefit of using a sample rather than all of the target population was seen in TIMSS performance assessment and in Assessment of Performance Unit assessments. These surveys made extensive use of practical tasks which would have been time consuming, costly and difficult to manage if the whole cohort was assessed. The policy context in England is such that the newly elected Government has stated that the end of key stage 2 tests and GCSE examinations will remain and will continue to be used for accountability purposes (although the exact targets and the way they are used are being reviewed). However, this does leave the opportunity for a sample approach to be adopted at other stages, and this is being

discussed for the end of key stage 3 where the tests have been abolished, and for the key stage 2 science tests. If this approach is adopted and proves to be successful, it may pave the way for similar changes at other stages.

The extent to which communication ought to be a compulsory aspect of the mathematics curriculum continues to be debated. Some argue that reading and writing demands in a mathematics test affect the validity of the test – that it is not the student's mathematics skill that is being assessed. Others argue that for mathematics skills to be useful, for example in the workplace, a student must be able to communicate their findings and explain their approach. The latter argument seems to be gaining support and students will be assessed on the quality of their written communication in the new GCSEs.

It is apparent that a number of the features of 'good' mathematics questions that we found in our earlier research are being incorporated within the new qualifications for secondary schools and the opportunity remains for them to be incorporated for primary schools through the proposed review of the key stage 2 tests. However, some difficulties remain with including all of the highlighted features. For very unusual or open-ended tasks it remains the case that students may not always know how to start. In other subjects with open tasks, such as English, this problem does not arise in the same way. It is unlikely that any of the new models will allow a test or exam invigilator to provide guidance that is then taken into account in the marking process. For this reason, the extent of the open-endedness or unfamiliarity may be constrained. An alternative approach may be to provide a choice of questions but this raises other complications in terms of comparability of the tasks.

Another difficulty may be the marking of the tests. The current tests can be marked relatively reliably due to the dominance of objective or short, constructed response questions. If a wider range of question types were to be included, such as the open-ended or extended tasks described in this paper, then there will be an impact on marking reliability. Those arguing for the use of tests and examinations that encourage good teaching and learning may well be happy to accept the lower levels of reliability as a necessary compromise. There is no reason to assume that the marking reliability would be any lower than that for existing high stakes tests and examinations that use open-ended tasks, such as in English, art or history. However, the policy context in which the results are used for accountability purposes may mean that arguments in favour of keeping the tests as reliable as possible win out.

Although this paper focuses on innovative ways of assessing mathematics, it is useful to remember that traditional tests have some advantages when used appropriately. The relatively simple presentation of problems means that it is easy to untangle the specific skills and procedures needed to solve them. In this sense the data tends to be 'cleaner' for analysis than in more complex assessment situations where the many links between cognitive processes and other factors produce a lot of 'noise'. Tests are also straightforward to administer which can lead to more consistent data being collected. So, for some purposes, it can be useful to use traditional tests. The extent to which the same tests can be used to encourage and support the effective teaching of deep understanding in mathematics continues to be debated.

9 References

Hoyles, C., Wolf, A., Molyneux-Hodgson, S. and Kent, P. (2002). *Mathematical Skills in the Workplace. Final Report to the Science, Technology and Mathematics Council*. London: University of London, Institute of Education.

Mansell, W. (2007), *Education by Numbers: the Tyranny of Testing*. London: Politico.

Maughan, S. (2009). *What is a good maths assessment?* Presented at the International Association for Educational Assessment conference, Brisbane, September 2009.

National Centre for Excellence in the Teaching of Mathematics (2008). *Mathematics Matters: Final Report* [online]. Available: https://www.ncetm.org.uk/files/309231/Mathematics+Matters+Final+Report.pdf [1 March, 2010].

OfSted (2008). *Mathematics: Understanding the score - Messages from the inspection service*. London: Crown.

Smith, A. (2004). *Making Mathematics Count: the Report of Professor Adrian Smith's Inquiry into Post-14 Mathematics Education*. London: The Stationery Office [online]. Available: http://www.mathsinquiry.org.uk/report/MathsInquiryFinalReport.pdf [1 March, 2010].

The Royal Society (2010). *Science and mathematics education, 5 – 14: 'state of the nation' report.* London: The Royal Society. ISBN: 978-0-85403-826-8, July 2010.

Williams, P. (2008). *Independent Review of Mathematics Teaching in Early Years Settings and Primary Schools: Final Report.* London: DCSF [online]. Available: http://publications.teachernet.gov.uk/eOrderingDownload/Williams%20Mathematics.pdf [1 March, 2010].

Vappula, H. and Clausen-May, T. (2006) *Context in Maths Test Questions – Does It Make A Difference?* Research in Mathematics Education, Volume 8 Issue 1, pp 99-155.