National Foundation
for Educational Research

The Critical Thinking Advanced Extension Award Trial Examination

A report for the Qualifications and Curriculum Authority

Paul Newton, Chris Whetton, Ewan Adams
and Tandi Clausen-May

13 July 2001

Acknowledgements

The authors are very grateful to the many individuals who contributed to the project:

NFER Project Team

Christine Webster, Simon Rutt, Joanne Kaye, Liz Gibson, Margaret Parfitt

External Critical Thinking Consultants

John Butterworth, Simon Carver, Roy van den Brink Budgen, Dave Wells

Additional NFER staff

Debbie Banks, Bali Gill, John Hanson, Joan Howell, Pauline Benefield, Mohinder Rattu, John Kimber

QCA Project Team

Tina Isaacs, Patricia Bellas, Liz Francis, Julie Sohal, Kate Green

OCR Officers

Judith Cockerill, Tony Orgee

Members of the Critical Thinking Advisory Group

The schools, teachers and students that participated in the trial

Contents

Section 1 Introduction

Section 1 explains the rationale behind Advanced Extension Awards and discusses the specification for developing a trial examination in Critical Thinking.

1.1 The Advanced Extension Awards

Advanced Extension Awards (AEAs) are being introduced for 18-year-olds in England, Wales and Northern Ireland in Summer 2002 as part of the Government's response to the report *Excellence in Cities*. They are intended to complement the world class tests for 9- and 11-year-olds and to supersede Special papers.

AEAs should be designed so as to:

- stretch the most able A-level students by allowing them to demonstrate a deeper level of understanding than would be possible at A-level;
- be of a comparable standard to the most demanding examinations of other countries;
- be similarly accessible to students across a range of educational backgrounds (type of school, specifications studied, etc.);
- help differentiate between the most able candidates, thereby negating the need for universities to develop their own examinations to discriminate between candidates with indistinguishably high A-level grades.

AEAs are to be based on pre-existing GCE subject criteria (where these exist) but must be independent of individual specifications and require no additional teaching or resources. They are to assess, through external examination, both depth of understanding and the ability to think critically and creatively.

1.2 The Specification for the AEA in Critical Thinking

The AEA in Critical Thinking is intended to be introduced for first examination in Summer 2002. It is to be targeted at the top 10% of all A-level students, irrespective of institution of study or subjects studied. It must, therefore, be independent of all A-level specifications, including the AS in Critical Thinking.

1.2.1 The critical thinking construct

The *Specification for the Development of an Advanced Extension Award in Critical Thinking* set out the following six aims for the award, in terms of encouraging students to:

1. develop a critical capacity to evaluate their own and others' beliefs and knowledge claims in a variety of contexts;
2. understand the interpretative nature of knowledge including personal and ideological bias;
3. generate their own arguments and alternatives;
4. make interdisciplinary connections and synthesise arguments;
5. evaluate reasoning of different kinds, including recognising and evaluating assumptions;
6. develop an understanding of why critically examining knowledge is important.

In addition, the *Specification* required that the AEA should: "provide opportunities for students to demonstrate breadth of knowledge, depth of knowledge, an ability to transfer skills and make connections, integrate ideas and develop concepts, analyse the logical form of arguments, make judgements, assess the credibility of sources, evaluate evidence and examine questions from a broader standpoint than that of a single discipline."

Finally, it proposed the following assessment objective: "The AEA will assess candidates' abilities to apply and communicate effectively their knowledge, understanding and skills in critical thinking."

1.2.2 The nature of the examination

The *Specification* made clear that, to make it accessible to a wide range of students, the AEA in Critical Thinking should draw upon a range of verbal and numerical thinking skills from across the different traditions of critical thinking. This would include logical, numerical and scientific reasoning skills and the use of problem solving skills in novel situations.

While requiring that the examination should be external, written and timed the *Specification* allowed that different modes of assessment could be included; for example, pre-release materials; research; or extended essays. In fact, it specifically required that

different question types should be trialled, from open-ended essays to objective-test questions including multiple rating items.

Although the *Specification* suggested that optional questions might be employed, it was later agreed that – due to concerns for comparability – they would not be. There would, however, be two examination forms: Mode P (with pre-release material) and Mode Q (without pre-release material).

1.2.3 Additional concerns

Particularly as the AEA in Critical Thinking would have no A-level syllabus to ground it, the importance of developing detailed Guidance Material for trial centres was stressed.

Mark schemes were also to be developed which:

1. rewarded positive achievement;
2. maximised consistency of marking;
3. gave a general indication of the required learning outcomes;
4. provided details of the expected level of attainment characteristic of each grade.

The *Specification* indicated that there was no need to award a grade on candidates' work for the trial examination.

Finally, it was noted that the AEA in Critical Thinking must also be cost-effective in terms of both schools and colleges and awarding bodies.

1.2.4 The trial

The examination was to be trialled on a minimum of 200 candidates for each mode, from a range of centres. Each of these centres would receive feedback on candidates' results.

The schedule for the development and administration of the trial examination was acknowledged to be extremely tight, particularly bearing in mind:

1. the a-typical complexity of the construct specification process (see Section 2);
2. the absence of item writers with experience of writing for the target population (see Section 3);

3. the requirement to develop novel item formats that had never before been trialled in an examination context (see Section 2).

As such, the NFER proposed to extend the developmental period from two to four months and to reduce the time taken to process and report on the results. Even this revised timetable was felt to be (and proved to be) extremely demanding.

1.3 The Critical Thinking Advisory Group

A Critical Thinking Advisory Group (CTAG) was established by the QCA to oversee and steer the development of the trial examination. It was comprised of around 20 participants, from a range of backgrounds, selected to represent a variety of critical thinking interests. Amongst the groups represented were: academics specialising in critical thinking, problem solving and thinking skills; teachers and examiners of critical thinking; higher education institutions; high-achieving schools; awarding bodies; and others.

The CTAG contributed to the *Specification*, particularly to the definition of critical thinking and to the aims for the award. In addition, it met twice to consider progress that had been made by the NFER in developing the trial materials. The CTAG gave feedback on this progress, influencing the direction that the project took.

Section 2 Examining Critical Thinking

Section 2 describes the conceptual groundwork that enabled the AEA in Critical Thinking to be developed in accordance with the development specification provided by the QCA. The more practical details of the examination development process are discussed in Section 3.

2.1 Construct specification

The first task in designing any examination is to specify the characteristic to be assessed. This is relatively unproblematic for most traditional examinations: what is to be assessed is the level of proficiency obtained in a particular subject area, i.e., the level of knowledge, skill and understanding achieved through study of the syllabus. Construct specification for the Critical Thinking AEA was complicated owing to the lack of an associated syllabus. This meant that an independent definition of the construct would have to be agreed upon before the developmental work could begin.

2.1.1 Conflicting definitions

One of the major problems facing any attempt to assess critical thinking is the lack of a generally accepted definition. Presented below are a few potential candidates:

"the correct assessing of statements" Ennis (1962)

"reflective and reasonable thinking that is focused on deciding what to believe or do" Ennis (1985)

"disciplined, self-directed thinking which exemplifies the perfections of thinking appropriate to a particular mode or domain of thinking" Paul (1993)

"skilled, active interpretation and evaluation of observations, communications, information and argumentation" Fisher and Scriven (1997)

Each of these definitions attempts to highlight the essential characteristic(s) of a form of thinking that would allow one to call it 'critical thinking'. They tend to differ with respect to how broad a definition they adopt and in terms of the facet(s) of thinking that they emphasise.

An important distinction has been drawn by a number of academics between ability and disposition, i.e., it has been proposed that critical thinking should embrace not only thinking skills but also the disposition to think critically (e.g., Norris and Ennis, 1989). In

practice, though, the assessment of critical thinking has tended to focus predominantly upon skill rather than disposition (c.f., Norris, 1992) and it was decided that this should be true for the AEA.

One reason for the many definitions of critical thinking is that it has primarily been developed as an educational, rather than a psychological, construct. It has typically not been proposed that critical thinking represents a discrete, uni-dimensional, underlying, even innate ability (in the way that some have believed IQ to be). Instead it is typically assumed simply to be comprised of a set of skills that together facilitate good thinking and that may be enhanced through instruction and exercise.[1] In this sense, it is possible to see how the complement of skills that are taken to represent the paradigm of critical thinking might depend on the purpose to which it was to be directed. This is to present critical thinking as a social construct.

Understanding critical thinking in this way clarifies why its definition is not simply a matter of conceptual analysis but also of value judgement. This emphasises even more strongly the need for explicit prior construct specification. Indeed, for an AEA grade in Critical Thinking to be meaningful for users, the sense in which construct is to be understood and assessed ought to be specified at the same level of detail as a traditional examination syllabus. Hence the necessity of further unpacking the QCA's chosen definition:

"a form of reflective reasoning that uses a combination of skills, attitudes and information or knowledge, which facilitates good judgement and is sensitive to context"

2.1.2 Unpacking the chosen definition

Although an initial definition of critical thinking is a useful starting point, it is only once this is fleshed out that the construct to be assessed begins takes shape. Indeed, the above definition could be interpreted in any number of ways. This raised two major questions for the examination development process, one theoretical and one practical:

---

[1] Some, such as McPeck (1981), would take the multi-dimensionality argument a step further by suggesting that the skills of critical thinking must also be understood as domain-specific. However, contra McPeck, the assumption underlying the AEA is that critical thinking must be understood as a complement of thinking skills that are largely independent of content. It would be fair to say, though, that this debate has not yet been fully resolved.

1. what values ought to underlie the construction of critical thinking for the AEA?

2. how had critical thinking been constructed in previous, related contexts?

As indicated in the QCA's specification document, the primary intended purpose of the AEA in Critical Thinking was to: "help differentiate between the most able candidates, particularly in subjects with a high proportion of A grades at Advanced GCE, in order to obviate the need for universities to develop their own entry tests."

The implication, therefore, is that the values underlying the critical thinking construct ought to reflect the needs of university selectors in the UK. That is, the construct ought to reflect those aspects of critical thinking that are most effective in discriminating between candidates with the most and least 'potential' for success in Higher Education (HE). However, exactly what these aspects are is not obvious. One way of finding out might be simply to ask a large number of university selectors. This is the essence of the Delphi method which was adopted in the early 1990s in the USA to reach consensus over what critical thinking should be taken to mean (Facione, 1990). On the other hand, this approach assumes that the consensus of university selectors would be sufficient to specify those critical thinking components that are crucial to success in HE. This assumption is certainly questionable.

As it was not within the NFER's remit to conduct separate construct-specification studies, the only alternative was to reflect upon a range of potentially relevant formulations and to extract from them the components that appeared most appropriate. Figure 2.1 presents a selection of sources that were drawn upon and illustrates the considerable variation in the conceptions of critical thinking that they embraced (different ways of categorising critical thinking competence and different views concerning its core component skills).

1. NFER Critical Reasoning Tests

These were developed by the NFER for the selection of junior and middle grade managers (Smith and Whetton, 1992), although they are also suitable for school leavers or employees who are not currently in managerial positions. Figure 2.1 lists the three core components of critical thinking assessed through this test.

2. Watson-Glaser Critical Thinking Appraisal Test

This is one of the oldest tests of critical thinking. Figure 2.1 lists its five sub-test components.

NFER Critical Reasoning Tests	Watson-Glaser Critical Thinking Appraisal Test	Cornell Critical Thinking Tests (the Level Z tests)	OCR AS award in Critical Thinking	Delphi report	National Postsecondary Sourcebook on Assessment	Ennis (1985)	Fisher and Scriven (1997)
Analysis	Inference	Deduction	Identify the elements in a reasoned case	Analysis	Interpretation	Elementary classification	Critical interpretation
Evaluation	Recognition of assumptions	Meaning	Evaluate reasoning of different kinds	Evaluation	Analysis	Basic support	Critical communication
Planning	Deduction	Credibility	Recognise and evaluate assumptions	Inference	Evaluation	Inference	Critical knowledge
	Interpretation	Inductive inference (support)	Clarify expressions and ideas	Interpretation	Inference	Advanced clarification	Critical technique
	Evaluation of arguments	Inductive inference (prediction)	Present a reasoned case in a clear, logical and coherent way	Explanation	Presenting arguments	Strategy and tactic	
		Definition and unstated reasons		Self-awareness	Reflection		
		Assumption identification					

FIGURE 2.1. An illustration of the variety of core components into which critical thinking has been divided by a range of test developers, national panels and theorists.

3. Cornell Critical Thinking Tests

The Cornell Test is another with a long pedigree. It has two versions, the Level X and the Level Z, with the latter being aimed at a higher level. Figure 2.1 identifies the seven sections of the Level Z test.

4. AS in Critical Thinking

The UK's only public examination in critical thinking – OCR's Critical Thinking AS – has now been running for a few years. The five assessment objectives for the award (see OCR, 2000) are as listed in Figure 2.1.

5. Delphi report

Facione (1990) reported the results from a consultation with 46 experts, which aimed to reach a consensus on the definition of critical thinking and its role in education. Out of this process came an agreement (greater than 95 per cent of those involved) that critical thinking involved three factors: analysis, evaluation and inference. He also reported agreement amongst a high proportion of the experts consulted (greater than 87 per cent) on the importance of interpretation, explanation and self-awareness.

6. National Postsecondary Sourcebook on Assessment

The Sourcebook on Assessment (Erwin, 2000), sponsored in the USA by the National Centre for Educational Statistics, was produced with the intention of defining critical thinking, problem solving and writing and of developing a list of published tests in these areas analysed according to these definitions. Its definition of critical thinking was based upon studies by Jones et al (1995, 1997) who sought the opinions of policy makers, employers and educators. They proposed the six ability components presented in Figure 2.1.

7. Ennis

Robert Ennis has been at the forefront of critical thinking for many decades and has formulated his views on the essential components of critical thinking ability in a variety of publications (e.g., Ennis, 1985; Norris and Ennis, 1989). He identifies twelve sub-components of critical thinking skill within the five major components listed in Figure 2.1.

8. Fisher and Scriven

Finally, to indicate a wider emphasis upon the core competencies of critical thinking, Fisher and Scriven (1997) adopted the four groups presented in Figure 2.1.

2.1.3 Developing a framework for the AEA in Critical Thinking

It was not within the NFER's remit to conduct a full and systematic analytical review of critical thinking perspectives. However, sources such as those described above were considered in relation to the central aim of assessing a construct that would have direct relevance for university selectors. As such, the NFER was able to develop a construct specification in expansion of the general assessment objective (presented by the QCA in the development specification).[2] This was honed in collaboration with the QCA's Critical Thinking Advisory Group. It was reproduced in the Guidance Material for centres and students which was sent to schools taking part in the trial in advance of their examination date.

2.1.3.1 An expansion of the Critical Thinking assessment objective

The construct to be assessed through the Critical Thinking AEA was broken down into four components. These are described below as they appeared in the Guidance Material.

2.1.3.1.1 Assessment Objective Component 1

> *Demonstrate a critical capacity to evaluate different kinds of reasoning in a range of contexts, from different academic disciplines and everyday life.*

Two core skills are crucial to critical thinking: the ability to break down arguments into their component parts; and the ability to evaluate how well the component parts of an argument together constitute a valid argument form. As such, the evaluation of lines of reasoning is a core element of critical thinking. A further element of critical thinking involves the evaluation of beliefs and claims, drawing on more than the statements in isolation.

Thus, the first component incorporates the following areas:

---

[2] Technically, only one assessment objective is specified for the AEA in Critical Thinking: "The AEA will assess candidates' abilities to apply and communicate effectively their knowledge, understanding and skills in critical thinking."

Analysis – the ability to analyse arguments, and further break these down into claims, reasons or conclusions. Identifying missing information and underlying assumptions is an important part of analysing arguments effectively. Also important is the ability to determine whether particular evidence is relevant to the area under debate (regardless of the quality of the evidence in terms of its own logic and authenticity of source).

Inference – the ability to judge whether the reasons given in an argument are sufficient to justify a conclusion. This includes assessing the nature and strength of justifications offered in support of statements, including the evaluation of the use of evidence. Also covered in this area is the assessment of the credibility of different sources of information.

2.1.3.1.2 Assessment Objective Component 2

> *Demonstrate an understanding of the interpretative nature of knowledge including intentional and unintentional bias, and the techniques utilised in persuasive communication.*

Key areas within this component include:

Emotive language – the ability to detect the use of strong emotional language or imagery designed to trigger a response, for example, the ability to critique an advertisement or propaganda.

Misleading language – the ability to recognise the use of misleading language such as exaggeration or downplaying of an important topic.

Irrelevancies – the ability to detect use of irrelevancies introduced to divert attention from the issue at hand.

Bias – the detection of slanted definitions or comparisons which express bias.

2.1.3.1.3 Assessment Objective Component 3

> *Communicate arguments and alternatives clearly and accurately in a concise and logical manner.*

There are two aspects of this area: developing arguments and then communicating them.

Developing arguments – requires the generation of relevant considerations, other scenarios, examples and counter examples. This includes the ability to develop analogies

and other forms of comparisons to clarify meaning, or examples that help to explain something or remove troublesome ambiguities.

Communicating arguments – requires the presentation of arguments that are clear, logical and coherent. That is, the skills used to critique others should be applied to one's own reasoning. General criteria for good critical thinking and therefore students' written work include clarity, relevance, accuracy, fairness, completeness, precision, depth, breadth and adequacy.

2.1.3.1.4 Assessment Objective Component 4

> *Make connections and synthesise arguments, locating, selecting, categorising, comparing and integrating information.*

This component consists of the following areas:

Connecting and synthesising arguments – recognising how information from different areas or points of view can be combined, either to build a coherent argument or to represent differing points of view. Also identifying the main theme within an argument, and how the various lines of reasoning support this, if they do. In doing this, showing an awareness of the context and audience.

Locating and selecting information – identifying information and selecting the parts that are relevant to a specific argument and those which are irrelevant. This includes the ability to ask relevant and penetrating questions to clarify facts, concepts and relationships. Also identifying and seeking additional resources, such as resources in print, that can help clarify comprehension.

Comparing and categorising information – the ability to identify similarities and differences and to formulate and use categories, distinctions or frameworks in order to organise information and aid comprehension.

2.1.3.1.5 The Language of Critical Thinking

No questions specifically addressing the language of critical thinking will be included within the examination paper for the Advanced Extension Award in Critical Thinking. Instead, the extent to which students understand the language and processes of critical thinking should emerge from their responses to the examination.

Additionally, as part of the preparation for the examination, students will be given guidance information which describes technical terms relating to critical thinking. If these were explicitly tested, the guidelines could become a 'mini syllabus', rather than preparation materials. Because of this, evidence that students understand the language and processes of critical thinking will be examined through the assessment objective components set out above.

2.1.3.2 The status of the construct specification

The construct specification, as described in the above assessment objective components, should still be regarded as essentially Work in Progress. While it highlights qualities that appear to be relevant to HE courses of study, there are no grounds for believing that it is necessarily the most appropriate formulation. Components may have been included that would have been better left out (or vice versa) or components may have been appropriately included but unduly emphasised. There is further work to be done to refine and validate the construct specification. ***The importance of this task for any future developmental work should not be underestimated; it will require the investment of considerable time and effort.***

It should also be noted that the construct specification presented above would benefit from considerable elaboration. While it is coherent as a model of critical thinking (if not necessarily optimal) it lacks the extent of specification that is necessary to construct a strongly defensible examination. The unavoidable time restrictions of the trial examination schedule meant that question development had to precede largely in the absence of a full construct specification. Yet, without a full breakdown of sub-components it is not possible to determine the full range of skills and sub-skills to be assessed and it is not possible to develop an examination framework in which their relative weights are effectively distributed.

2.2 The sampling framework

Once a construct has been specified in sufficient detail it is possible to develop a framework within which to determine precisely *how* the examination will assess what it is intended to assess. This section will explain the logic of determining how and why the particular question constructs, contents and contexts were chosen.

2.2.1 Construct sampling

The construct specification should indicate the full complement of critical thinking skills embraced within the chosen definition. However, an examination construct is only fully specified when the 'weight' attached to each of the component skills is specified. For example, we might have chosen to weight the four assessment objective components differently, allocating 20% of the total examination marks to each of the first two components and 30% of the total marks to each of the second two components.[3] Furthermore, within each component, any independent sub-component skills would also need to be identified and weighted appropriately. In assigning weights we are effectively saying that, for the successful demonstration of competence in each component skill, a certain number of marks must be available.

When it can be assumed or demonstrated that a particular type of question is effective for assessing a particular component skill (and only that component skill) this weighting/sampling process is relatively straightforward. However, in the context of the Critical Thinking AEA, the situation was considerably more problematic. The two main problems were:

1. how to deal with questions that would inevitably assess an indeterminate combination of component skills;

2. how to deal with component skills that could not easily be translated into straightforward questions.

Notionally, an equal weighting of the four assessment objective components was assumed. In practice, though, it is not clear whether or not this weighting was achieved. Moreover, it should be noted that the decision to weight the components equally was somewhat arbitrary and should be a topic for debate during any future developmental work.[4]

---

[3] Ideally, then, the various components should be independent of each other, i.e., they ought to describe conceptually distinct skills.

[4] The ICAT Critical Thinking Essay Examination, for example, gives Analysis a weighting of 80% and Evaluation a weighting of only 20%.

2.2.2 Content sampling

It was assumed that the critical thinking construct embraces skills that are largely content-general – that critical thinking skills are equally applicable across a broad range of fields, from the academic to everyday life. However, this is not the same thing as assuming that the exercise of critical thinking skill is independent of content-specific knowledge. Thus, someone with a high level of content-specific knowledge would have to do less critical thinking to solve a problem in that field than someone with a lower level of content-specific knowledge (Fisher and Scriven, 1997). Therefore, the fact that students from a range of academic disciplines were to take the AEA presented a problem for the accurate assessment of critical thinking skill.

2.2.2.1 Approaches to content sampling

Three potential solutions were considered:

1. the content of **each question** should represent a **single A-level subject area** and, overall, the examination should represent the full range of A-level subject areas equally;

2. the content of **each question** should represent a **range of A-level subject areas** and, overall, the examination should represent the full range of A-level subject areas equally;

3. the content of **each question** should represent a **subject area (or areas) that would not be covered within any A-level** course of study.

The third option represents what Fisher and Scriven (1997) described as the 'homeless curriculum'. Relevant content areas might include: techniques for product and personnel evaluation; elementary decision theory; theory of terrorism; subliminal advertising and telemarketing; age discrimination; legislation of drugs; internet censorship; etc..

The 'homeless curriculum' approach to content sampling has much to recommend it, as it appears to offer the greatest prospect of eliminating subject-specialism bias. The 'single subject' approach has three main disadvantages. First, any individual question will clearly be biased in favour of students with a particular subject-specialism. Second, when some examination questions are worth many marks while others are worth only a few, this complicates the content sampling process significantly. Finally, when an examination contains a relatively small number of questions, the effective sampling of content areas is

prohibited. The 'cross-subject' approach is also problematic because it is hard to find questions that genuinely embrace a good range of disciplines with equivalent emphasis.

In fact, no decision was made between the three alternative approaches and questions reflecting each of them were included within the examination papers. In many ways this was a pragmatic choice owing to the lack of time in which to develop a more structured approach. Future development work should consider whether it would be better to adopt a single approach, in particular, the 'homeless curriculum' approach.

It should be noted that, even if a 'homeless curriculum' approach to content sampling was adopted, there would still be some questions that were more mathematical while others were more verbal while others were more scientific (etc.). Any future development work should bear this in mind and should attempt to develop a coherent formal content sampling framework.

2.2.2.2 The formal specification

The trial Guidance Material identified four general areas within which the examination questions would be framed. These were broad enough to embrace just about any kind of question content. Moreover, there was no formal statement concerning how these areas would be sampled within each examination. As such, the formal specification did not technically define or circumscribe the content of the AEA examination. The four areas were as follows:

Society, Politics and the Economy

The contexts included will require an understanding of: debates concerning the values of society; political processes; explanations of human behaviour in the past or present; the role of language; legal situations; and social and economic trends.

Science and Technology

The contexts included will require an understanding of the nature of scientific objectivity, and understanding of scientific methods and principles. The social, ethical and environmental implications of scientific discovery and technological development may provide the setting, including medical research and the implementation of treatments.

Logic, Mathematics and Numeracy

The contexts included will require an understanding of formal logical processes and of mathematical reasoning. They may include the interpretation of numerical information or the evaluation of probabilistic inferences.

Culture, Arts and Humanities

The context included will require an understanding of the nature and importance of culture, of aesthetic evaluation and of the processes of communication used by modern media.

2.2.3 Context sampling

There is an important distinction to be made between the content and the context of an examination question.[5] While content concerns what the presentation is about, context concerns how the presentation is presented. In its most general sense, critical thinking could be applied to presentations delivered in any of the five senses. As is traditional, though, it was decided to limit the AEA examination to visual presentations. Furthermore, as an examination of critical thinking, these visual presentations were restricted to those through which an argument could be conveyed. Among the many possible forms of visual presentation of argument were the following:

Verbal presentations

- constructed passages
- logical 'brainteasers'
- letters
- newspaper articles
- position statements
- press releases

[5] The choice of the terms 'content' and 'context' is somewhat arbitrary (in that other researchers might use them differently); however, the terms are used here simply to capture a key conceptual distinction.

- legal proceedings
- etc.

Numerical presentations

- graphs
- tables
- statistics
- probabilities
- till receipts
- insurance documents
- etc.

Pictorial presentations

- maps
- advertisements
- cartoons
- product labels
- etc.

Technically speaking, it was not obvious the extent to which formal context sampling would be necessary for the AEA in Critical Thinking (other than to the extent that it might overlap with concerns of content sampling[6]). In principle, the contexts employed ought simply to have been those most appropriate for discriminating between students with most and least 'potential' for success at HE (the basic technical criterion). In the absence of

---

[6] That is, if numerical presentations were more accessible to students with a mathematical training, or if verbal presentations were more accessible to students with a linguistic training, then the examination might generate biased results.

prior research into this question, it was decided simply to sample across a range of contexts. Future development work should consider whether more formal context sampling criteria ought to be formulated.

2.2.4 The sampling matrix

The preceding discussion would recommend the construction of a 'construct by content by context' sampling matrix which would guide the question development process. However, the considerable time restrictions of the trial limited the extent to which a formal matrix could be developed and utilised. At the root of this problem was the fact that question development had to take place in tandem with construct specification. There were also constraints on the application of a formal sampling matrix caused by decisions concerning examination format. In particular, the fact that essay questions were to be employed reduced the scope for content and context sampling and limited the potential for assessing component skills in isolation. Thus the sampling decisions were ultimately driven more by heuristic and pragmatic concerns than by formal specification. There is considerable work to be undertaken in respect of sampling that will depend upon the outcome of anticipated future construct development.

2.3 The assessment formats

In parallel with the development of a framework for sampling question contents and contexts, the type of questions to be developed must be decided upon. This section will explain the reasoning behind the choice of assessment formats for the Critical Thinking AEA.

2.3.1 The multiple choice format

Among the commercially available tests of critical thinking, the multiple choice assessment format has – by a long margin – been the most prevalent (e.g. the Watson-Glaser Critical Thinking Appraisal, California Critical Thinking Skills Test, Collegiate Assessment of Academic Proficiency Critical Thinking Test, Cornell Critical Thinking Test). There have been a number of reasons for relying upon the multiple choice format and the most important have tended to be:

1. ease of scoring (multiple choice tests do not require markers to be experts in critical thinking and tests may even be scored automatically – this makes the marking quicker, cheaper and, when expert markers are in short supply, manageable);

2. breadth of sampling (the smaller the assessment unit, the more items can be included and the more inclusive the construct, content and context sampling can be);

3. reliability (the more items can be included, the more reliable the aggregate score is likely to be).

On the other hand, the multiple choice format has been challenged, on validity grounds, for problems such as:

1. construct under-representation (for example, not being able to assess skills that require argument construction and communication);

2. process ambiguity (not providing evidence on how students arrive at answers).

The NFER was keen to ensure that the Critical Thinking AEA should both capitalise upon the strengths of the multiple choice format, but also incorporate alternative formats to provide crucial supplementary information. As such, the initial proposal for development was framed in terms of single mark multiple choice questions and multiple mark (short) constructed response questions. Included within this model were a number of questions with a particularly novel assessment format: multiple rating items.

2.3.2 The multiple rating format

The QCA was keen that the multiple rating format should be piloted as part of the Critical Thinking AEA trial. The idea of a multiple rating item is quite new (Fisher and Scriven note that it was probably first mentioned in print by Scriven in the 1991 *Evaluation Thesaurus*). Indeed, the trial was to be the first ever large-scale evaluation of this assessment format.

The basic structure of the multiple rating item is that of **stimulus passage** (e.g. a page of text expressing an argument), plus a number of **stimulus responses** (e.g. four summaries of the argument of varying quality), plus a **multiple rating scale** according to which each stimulus response is to be rated. The multiple rating scale extends from A to E and each rating is accompanied by a **descriptor**. In this model, A is typically the most positive end of the scale (with a descriptor that is, essentially, an elaboration of 'very good') and E is the worst (with a descriptor that is, essentially, an elaboration of 'very bad').

In requiring students to evaluate the quality or strength of stimulus responses (summaries, critical comments, etc.) to stimulus passages (arguments), the multiple rating format targets a crucial component of critical thinking. Moreover, due to a subtlety of scoring

that was originally intended to eliminate the possibility of payoff from guessing, it is well suited to critical thinking contexts where problems often do not have answers that are either black or white. Multiple rating items require students to select the rating that best describes the quality or strength of the stimulus response in relation to the stimulus passage. Therefore, it is clear that there ought to be a single rating grade that experts in critical thinking would generally agree to be the 'best'. However, as critical thinking typically deals with problems that do not have black and white solutions, and as the rating scale is ordinal and adjacent ratings will tend to differ by degree rather than qualitatively, it will often be the case that ratings adjacent to the 'best' will also be 'plausible best' answers.

The crucial point is that students are credited not simply for selecting the 'best' rating but also for choosing a 'plausible best' rating. In Fisher and Scriven's original specification, they proposed that: 1 mark should be awarded for choosing the 'best' rating; ½ mark should be awarded for choosing one of the two 'plausible best' ratings (either side of the 'best' rating); and 1 mark would be deducted for choosing a rating that diverged from the 'best' by more than one rating grade (thereby deterring students from random guessing).

While Fisher and Scriven noted that the deduction of marks is essential if the guessing payoff is to be eliminated (1997, p.191) this does not fit well with accepted UK examining practice. It was decided that the AEA mark scheme would employ neither negative marking nor the award of half marks. For the AEA in critical thinking, candidates would receive 2 marks for choosing the 'best' rating, 1 mark for a 'plausible best' and 0 marks for any other response.[7,8]

---

[7] Consider the situation in which the 'best' rating was B and both A and C were credited as 'plausible best' ratings. Under the Fisher and Scriven scoring model, 1 mark would be awarded for B, ½ mark for either A or C, and -1 mark for either D or E. The negative scoring means that a candidate who chose at random would, on average, tend to score zero marks per question (0.2x1+2x0.2x1/2+2x0.2x-1=0). This eliminates any payoff from guessing. Under the AEA model, a candidate who chose at random would, on average, tend to score 0.8 marks per question. This means that the guessing payoff is not eliminated.

[8] In fact, unless it is held that A and E may never be the 'best' rating, the Fisher and Scriven scoring model actually means that – when either A or E are the 'best' – a candidate who chose at random would, on average, tend to score -0.3 marks per question (thus introducing a net guessing penalty). Under the AEA scoring model, such a candidate would tend to score 0.6 marks per question.

2.3.3 Justification components

In one version of the multiple rating item, Fisher and Scriven proposed that candidates be given space to explain why they had chosen their rating grade for each stimulus response. Thus, in addition to marks for choosing the 'best' or 'plausible best' rating grade, they might achieve a mark for a strong justification of that choice. There are a number of potential benefits of employing response-justification:

1. it enables markers to credit the reasoning behind a selection;

2. it allows candidates who chose 'plausible best' ratings for good reasons to achieve additional marks, while candidates who chose 'best' ratings for no good reason would gain no further advantage;

3. it offers the opportunity to assess a different component of critical thinking – critical writing – a component that is not assessed through response-selection alone.

The NFER was initially hesitant about including response-justification for a number of reasons:

1. although a different component of critical thinking might be assessed, unless a candidate had achieved at least partial success on the response-selection task it is not clear how they could progress to success on the response-justification task (thus the justification mark would typically not be independent of the selection mark);

2. the justification would require marking which might introduce further problems (marking would take longer, be more expensive, decrease the reliability of the examination, etc.).

The potential for reducing the reliability of the examination – through marker inconsistency – was a particular concern. Response-justification is not a widely used assessment format and the extent to which (and the conditions under which) it can produce technically robust results is not well established.

The QCA's Critical Thinking Advisory Group (CTAG) was enthusiastic about the inclusion of a justification component within the multiple rating item format. In fact, at the final meeting of the CTAG it strongly recommended that even the multiple choice section should include a justification component. As such, response-justification was

included in both the multiple choice section and the multiple rating section of the trial examination.

2.3.4 The essay and short report format

Extended written responses have been employed only very infrequently in published tests of critical thinking (e.g. the Ennis-Weir Critical Thinking Essay Test, the Critical Thinking Assessment Battery). In the UK, the two key examples have been the MENO pre-university test (UCLES) and the AS in Critical Thinking (OCR).

While extended written responses are clearly desirable for assessing the ability to construct and communicate argument, the NFER was initially hesitant about the inclusion of essay questions:

1. essays require marking (consuming time, money and resources) and there are particular problems of marking reliability for extended written responses which are exacerbated by a severe deficit of markers with sufficient critical thinking expertise;

2. it is unclear the extent to which the capacity for critical writing is dependent upon a general capacity for writing (and there is a risk that candidates may end up being marked more for writing skill than for critical thinking skill);

3. if candidates are to be assessed on their ability to think and write critically about a specific subject area then they will need to be provided with sufficient information about that area to be able to construct a response (as there is no content specification) and this can lead to inappropriately long stimulus passages;

4. the ability to present a strongly reasoned argument concerning a specific subject area will be affected by prior knowledge of that area – while this is true of all critical thinking questions its impact will be magnified by the fact that a single essay will be worth many marks.

Despite concerns such as these, the QCA was keen that the AEA in critical thinking should require candidates to demonstrate their capacity for critical writing through extended written responses. This led to the development of two kinds of task: short reports and essays. The short reports were to be based upon stimulus passages that were presented during the examination: on the basis of information provided, candidates would present an argument in promotion of a particular perspective. The essays were to be based upon stimulus passages that were presented both prior to the examination (in pre-release material) and during the examination: on the basis of information provided,

candidates would present an argument for or against a particular perspective (the choice being theirs).

2.4 The examination structure

The AEA in Critical Thinking was to be a written examination with a duration of three hours. In common with other trial AEAs, it was to be piloted through two modes: one with pre-release material (including an extended essay); and one without pre-release material (including two short reports). Other than the short reports and essay, the two modes were to be identical. The two modes were, therefore, to have the following structure:

<table>
<tr><th>Mode P</th><th>Mode Q</th></tr>
<tr><td colspan="2">Section A
Multiple Choice Questions
10 questions without justification – 10 marks
5 questions with justification – 10 marks
Time Allowed (Notional) – 40 minutes</td></tr>
<tr><td colspan="2">Section B
Multiple Rating Questions
1 question without justification – 8 marks
1 question with justification – 15 marks
Time Allowed (Notional) – 50 minutes</td></tr>
<tr><td>Section C
Essay Question
1 question – 40 marks
Time Allowed (Notional) – 90 minutes</td><td>Section C
Short Report Questions
2 questions – 40 marks
Time Allowed (Notional) – 90 minutes</td></tr>
<tr><td colspan="2">Total Time Allowed – 3 Hours
Total Marks Available – 83</td></tr>
</table>

The AEA in Critical Thinking was intended to provide two grades of pass: Distinction and Merit, with Distinction being higher. Candidates who did not achieve the standard for Merit would be recorded as Ungraded.

Section 3 Development, Administration, Marking and Processing

Section 3 describes the methodology of the trial examination, from the development and administration of the question papers to the marking of scripts and analysis of results. It ends by discussing lessons that were learned along the way.

3.1 Examination development

3.1.1 Early construct specification and item development

The first task undertaken by the NFER critical thinking team was to develop the construct specification. As explained in Section 2, this task was particularly complex as critical thinking has tended to mean different things to different people for different reasons. While a construct specification framework was hinted at by the QCA's tender *Specification*, this needed considerable work to translate it into a sound basis for producing an examination.

As indicated in Section 2, this developmental work borrowed from many sources including:

1. published tests of critical thinking;
2. academic theses on critical thinking;
3. large-scale consultations on critical thinking.

An outline construct specification was prepared in expansion of the assessment objective and presented in the *First Submission of Critical Thinking Materials to QCA Critical Thinking Advisory Group* (the *First Submission*).

In addition to this outline, the *First Submission* also included examples of questions that indicated what the final examination might look like. These included:

1. multiple choice items;
2. multiple rating items;
3. short report questions;
4. essay question (with pre-release).

Each of the multiple choice questions consisted of a short stimulus passage, a question and five alternative answers, from which students were to select the correct one. The stimulus passages were predominantly text-based.

As discussed in Section 2, the basic structure of the multiple rating item was that of stimulus passage, plus a number of stimulus responses, plus a rating scale (from A to E) according to which each stimulus response was to be rated. Two types of stimulus response were devised: responses that **summarised** the argument of the passage; and responses that offered **critical comments** on the argument of the passage. Two distinct rating scales were, therefore, also utilised (one with quality of summary descriptors and the other with quality of criticism descriptors). The stimulus passages tended to be of around a side in length and the stimulus responses around one-third of a side.

The short report exemplar was based upon approximately three sides of abridged 'internet extracts' which were to form the evidence base from which students were to "take the role of a press officer..." and "... prepare a report for a reputable international magazine, putting forward the stance that there is nothing inexplicable in the phenomenon of lost aircraft and ships in the Bermuda Triangle." To ensure that students would present fair-minded argument, they were prompted to "be aware that your article will be published in tandem with that of a feature writer from 'Weird' magazine, who can be expected to document evidence and argue for the opposite viewpoint. The Navy will expect you to present a thorough and conclusive argument that addresses any points which this article may present."

In contrast to the short report, which contained all the necessary evidence within the question paper, the essay question was based upon material from an 11-page DETR document on climate change (presented as pre-release) and from a 5-page statement on global warming from the Heartland Institute (presented within the question paper). Students were asked "to produce a written response to the DETR's proposals, in light of recent information on global warming which is summarised below." They were also told that it "is important that you highlight both the positive and negative points in the DETR's document, and that the arguments you give are well reasoned and supported by the information below."

3.1.2 First Critical Thinking Advisory Group Meeting

The *First Submission* was presented to the Critical Thinking Advisory Group (CTAG) in early January 2001. The NFER explained the considerable challenges that this new

examination was presenting – particularly in the absence of an agreed construct specification – and the CTAG welcomed this acknowledgement.

CTAG also made a number of specific recommendations for future development, which QCA supported and instructed NFER to comply with:

1. the incorporation of more visual, scientific and mathematical material;
2. the incorporation of material that would require candidates to reflect on reasoning or generate their own reasoning;
3. the co-option of additional item-writers (to broaden the experience of the NFER team);
4. to reduce the reading load upon students;
5. to consider asking candidates to justify multiple choice selections (where questions may have more than one correct response);
6. to further clarify and distinguish the rating scale descriptors;
7. to tighten the level descriptors for the essay and short reports.

The NFER accepted and supported the CTAG's comments which, in fact, highlighted problems that went beyond the trial and that will continue to cause problems in years to come. These issues are discussed in detail in 3.6.

3.1.3 Secondary item development

In response to the CTAG's concern about broadening the item writing team, the NFER began the secondary development phase with the co-option of four consultants, each of whom had experience of working on one or more of the OCR, CIE or UCLES examinations mentioned above.[1]

The secondary development stage involved three main components:

1. further development of the construct specification;

[1] One of these consultants was subsequently forced to pull out of the item writing team.

2. honing existing items and writing new items;

3. preparation of the Guidance Material.

As noted in Section 2, this situation was less than optimal as, clearly, the construct specification ought to have been finalised before any item writing had begun. However, the necessity of piloting the trial in March meant that this was simply not feasible.

In addition to the revised construct specification, the *Second Submission of Critical Thinking Materials to QCA Critical Thinking Advisory Group* (the *Second Submission*) included a draft version of the Mode P Examination Paper, Answer Booklet and mark scheme, examination instructions and Guidance Material for Centres. The Guidance Material included: an introduction to AEAs; an introduction to critical thinking; examples of critical thinking questions; an explanation of the structure of the trial examination; a glossary of critical thinking terms; and a recommended reading list.

3.1.4 Second Critical Thinking Advisory Group Meeting

The *Second Submission* was presented to the Critical Thinking Advisory Group (CTAG) in early February 2001. In response to this submission, the CTAG made a number of general comments, which the QCA requested NFER to respond to. The most significant of these are presented below:

1. the assessment materials were still felt to be too long and too heavily text-based, resulting in too large a reading load;

2. there was felt to be too little opportunity for students to demonstrate their capacity for reflective reasoning;

3. in response to these two concerns, it was felt that the multiple choice items would benefit from including a justification component;

4. there was concern that the items might not be sufficiently different in terms of difficulty;

5. the necessity of carefully training the examiners to achieve marking reliability was stressed (such that substance rather than style be assessed).

3.1.5 Tertiary item development

In response to the CTAG's comments, a decision was made to reduce the reading load further still, wherever possible. In addition, the time allocation for each section was reconsidered, providing more time for answering the short report and essay questions. Finally, the multiple choice questions with the most ambiguous answers were given a space in which students could justify their selections. The resultant examination structure was as presented at the end of Section 2:

- 10 multiple choice questions (without justification);
- 5 multiple choice questions with justification;
- 1 summary multiple rating item (without justification);
- 1 critical comment multiple rating item with justification;
- 1 essay with pre-release (Mode P) ***OR*** 2 short reports (Mode Q)

3.1.6 Item ratification meeting

Owing to the inevitably ambiguous nature of many of the critical thinking questions, and to the fact that the item writers were working independently, it was necessary to convene an item ratification meeting before the draft examination papers could be approved. This involved the entire item writing team gathering to discuss the draft papers to reach agreement upon acceptable answers. Where agreement could not be reached, items would either be modified or rejected outright. The outcome of the item ratification meeting was to be a final draft of the Mode P and the Mode Q examination paper and the framework for their respective mark schemes.

3.1.6.1 Background to the meeting

The item writers were sent draft papers in advance of the ratification meeting. For each multiple choice and multiple rating question they were asked to indicate what they considered to be the 'best' answer/rating and to highlight any answers/ratings that they considered to be 'plausible best'.

It is important to note that 'plausible best' did not mean the same as a 'good distracter' – where a 'good distracter' is an answer that, while clearly wrong, might reflect a common, understandable mistake. A 'plausible best' answer means one for which a strong argument could be made that it is just as valid as the answer deemed to be the 'best'.

Generally speaking, a good multiple choice item should not have any 'plausible best' answers. Yet, in the context of a critical thinking examination, it can be hard to avoid this situation – indeed, embracing this situation might even be considered true to the construct itself.

For the justification components, the item writers were asked to consider what might count as a valid justification of a 'best' or 'plausible best' answer. Likewise, they were asked to consider the characteristics of a strong essay or short report.

3.1.6.2 Outcomes of the meeting

The most general comment on the item ratification meeting was that it took very much longer than anticipated. As a consequence, it was not possible to consider the short reports or essay question in depth. Many of the questions proved considerably more ambiguous than had previously been anticipated (even, in some cases, after their dissection during two CTAG meetings) and this led to extended discussion.

Although a number of the multiple choice questions were straightforward to ratify, having a simple stimulus structure and an unambiguous 'best' answer, others were not so accommodating. Even amongst the argument component identification items (familiar to the consultants through working on the Critical Thinking AS) there were numerous subtleties that caused problems. Disagreement sometimes focused on the nature of the intended conclusion of the argument contained within the stimulus passage. On other occasions, debate focused on the formal interpretation of the terminology used within a question, for example, whether a conclusion was actually stated rather than implied.

Generally speaking, it was possible to resolve ambiguities by changing the wording of the passage, question or answers. However, this process of clarification did tend to make questions easier than they had previously been. There was clearly a tension between reducing the ambiguity of the questions and reducing their demand. To the extent that an important component of critical thinking is the ability to 'see-through' ambiguity, this tension may not simply be of pragmatic significance.

By far the most problematic (and time-consuming) issue that the meeting addressed was the wording of the rating scales used within the multiple rating items. For the *First Submission*, the rating scale descriptors were taken directly from Fisher and Scriven (1997). For example, the rating scale that Fisher and Scriven (1997, p.180) described for argument plus critical comment multiple rating items was as follows:

A Correct and very powerful (very 'telling')

B Correct and with moderate force

C Correct and relevant but not very telling

D Correct but only marginally relevant

F Either incorrect, irrelevant or trivial

The Fisher and Scriven descriptors were subsequently revised in the light of concerns expressed during the first CTAG meeting. However, during the item ratification day, it became clear that a complete overhaul of the rating scale descriptors would be required if the questions were to function at all.

The meeting spent a considerable time in revision of the rating scales for both the summary and the critical comment multiple rating items. The main intention was to ensure that the five rating descriptors for each scale were as independent as possible and to ensure that they could encompass all possible scenarios. Although the resultant scales were not perfect – and perhaps, for reasons given in 3.6.8.1, could never be perfect – they were certainly a distinct improvement.

The final outcome of the meeting was satisfaction that each of the multiple choice questions had an answer that the item writers agreed was the 'best' (if not unanimously then by majority vote). They were also happy that the multiple rating scales were workable and that there were 'best' ratings for each multiple rating question. Finally, acceptable justifications for 'best' and 'plausible best' solutions were generated.

3.2 Examination administration

3.2.1 Participating centres

The intention was to trial the Mode P and Mode Q examinations on two separate samples of 200 students each. Assuming that each participating school would select around 10 of its best A-level students, this meant a sample of around 40 schools. QCA provided the NFER with a list of around 50 schools, each of which had expressed (via the QCA's website) an interest in the AEAs generally or in the Critical Thinking AEA specifically.

Each of these schools was invited to participate in the trial. To compensate for those that declined the invitation, a supplementary sample was selected by the NFER. These schools were drawn from a high-performing population and tended to over-represent

grammar and independent schools. In fact, despite an initial interest and agreement to participate, a number of schools subsequently dropped out of the trial. Unfortunately, these were largely from the comprehensive sector which meant that comprehensives were dramatically underrepresented in the final sample. The final centre breakdown is presented in Table 3.1.

TABLE 3.1. Number of schools that participated in the trial.

	Mode P	Mode Q
No. independent schools	6	7
No. grammar schools	8	5
No. comprehensive schools	1	2
No. other schools	3	4
TOTAL	**18**	~~18~~ 9

TABLE 3.2. Number of students that participated in the trial.

apparently one school we had no data on.

	Mode P		Mode Q	
	Male	**Female**	**Male**	**Female**
No. independent students	37	28	30	40
No. grammar students	58	26	27	41
No. comprehensive students	7	5	18	4
No. other students	9	16	29	24
TOTAL	**111**	**75**	**104**	**109**

When selecting students for the trial, participating centres were asked to bear in mind:

1. that students should be selected from the Year 13 cohort;

2. that students should be selected from amongst the most able in their cohort (or the most able in their subject area, see below);

3. that students should ideally be selected from across the main subject area groupings at A-level (i.e., sciences, social sciences, humanities, etc.).

In fact, a small number of centres also submitted a small number of students from Year 12 in addition to their Year 13 sample and these were included in the trial. Particularly as the AEA is intended to be content-free it was assumed that their inclusion would not distort the trial results significantly. The final student breakdown is presented in Table 3.2.

3.2.2 Pre-examination despatches

Two weeks in advance of the commencement of the trial, centres were sent Guidance Material. Two types were despatched:

1. Guidance Material for Students (one per student);

2. Guidance Material for Centres (one per centre).

The material for students was somewhat more informal and included less technical information on the AEAs and on critical thinking. Each of the above documents was printed in two formats, for Mode P and Mode Q, respectively. Students were informed that the Guidance Material was for information only and that they should not spend a great deal of time studying it and that there was no need to learn any of it.

In addition to the Guidance Material, students in the Mode P sample were also sent Pre-release Material. They were instructed to spend around an hour familiarising themselves with the information and to bring the document, unmarked, to the examination hall. They were told not to attempt to prepare an essay on the material in advance. Teachers were advised to present students with the Pre-release Material two weeks before the date of their examination.

3.2.3 The trial

The trial was scheduled to commence on 19 March and to run until 6 April. Schools were allowed to chose exactly when, during this period, they would administer the examination. The examination papers and associated documents were despatched to schools one week in advance of the commencement of the trial. The Mode P Examination Paper, Mode P Pre-release Material and Mode Q Examination Paper are attached as Appendices 3.1, 3.2 and 3.3.

3.2.3.1 The administration instructions

Centres were asked to administer the trial under standard examination conditions. The examination was to last three hours and the candidates were to answer all questions.

Answers to the first two sections were to be written directly in the Examination Paper, while answers to the last section were to be written in separate Answer Booklets.

After the examination, administrators were to distribute to each student a Data Form and Questionnaire. The Data Form collected background information on each student, such as sex, ethnicity, GCSE grades and predicted A-level grades. The Questionnaire posed a series of tick-box questions on impressions of the examination. Reactions to the examination were also collected from the administrator in a similar manner. The (Mode P) Data Form and Questionnaire and the (Mode Q) Administrator Data Form and Questionnaire are attached as Appendices 3.4 and 3.5.

3.2.3.2 The examination observations

Six schools were visited by an NFER researcher to observe the conduct of the examination in full and, subsequently, to conduct a focus group interview with the students. The purpose of the interview was to explore issues similar to those considered in the Questionnaires, but to do so in more qualitative depth. For each examination mode, one independent, one grammar and one comprehensive school was visited.

3.3 Examination marking

The marking team was comprised of three of the four item writing consultants and one NFER item writer. A Chief Examiner was not formally appointed and the team developed the mark scheme jointly, with the NFER Project Leader ultimately arbitrating any points of disagreement. Thus, the Project Leader fulfilled crucial aspects of the role of Chief Examiner, where necessary.

3.3.1 The mark scheme development

Although groundwork for the mark scheme had been conducted during the item ratification meeting, draft mark schemes had not been developed in full prior to the commencement of the trial. This situation was not ideal, as a full draft mark scheme should have been prepared in advance, yet the tight development schedule did not permit this. As such, the mark schemes were drafted during a two-day meeting of the item writers conducted mid-April. This process was informed by 10 completed scripts (5 Mode P and 5 Mode Q), drawn at random from the sample, photocopied and distributed to all participants in advance of the meeting.

3.3.1.1 Awkward decisions

Despite the fact that many of the questions had been scrutinised during CTAG meetings and during the item ratification meeting, the mark scheme development meeting still threw up issues that had not arisen before. In particular, a few questions that had appeared uncontroversial in previous discussions were now seen (by the same people) to be ambiguous after all. One multiple choice item (Question 15) and one multiple rating item (Question 16-2) was ruled to have two 'best' answers/ratings, both of which would be rewarded. For the remaining selection items, though, one and only one answer was ruled to be the 'best'.

Having said this, there was one item (Question 17-5) that simply divided the panel. Two of the item writers considered this to be a weak criticism (rating E) while two considered that it could, or should, be considered a strong criticism (rating A). Note that this was different from the above situation where general agreement existed that there was more than one 'best' answer. After a prolonged discussion that did not lead to a resolution of opposing views, the Project Leader ruled that one interpretation should over-ride the other and selected rating E as the 'best' answer. Note, further, that it would not have been possible to allow both A and E as 'best' because the rating scale was ordinal and A and E were at opposite poles. The only option would have been to score all selections as the 'best' thereby effectively eliminating the contribution of the selection component of this item.

3.3.1.2 Rewarding the justification components

For the selection component of each multiple choice and multiple rating item it was agreed that a mark should be awarded only for choosing the answer that the marking panel had deemed to be the 'best'. Even when there were alternative 'plausible best' answers there was still general agreement that a particular answer was the 'best' (cf. the above exceptions). Therefore, it seemed fair only to reward the selection of answers deemed to be the 'best' by the expert panel.

The justification components of the multiple choice and multiple rating items were more problematic. It might have been decided that marks should be awarded for valid justification only to those students who had also selected the 'best' answer. Yet this would have meant refusing to award justification marks to those students who had clearly understood the intention of the question – as shown by a valid justification – even though they had mistakenly selected the wrong alternative. Moreover, it would also have meant refusing to award justification marks to those students who had chosen a 'plausible best'

selection and given a strong justification of it. This would seem to defeat one of the main purposes for including a justification component.

It was therefore ruled that there should be three general situations that should lead to the award of a justification mark:

1. if the 'best' answer had been selected and a strong reason given for calling it the 'best' answer (hence, the justification would relate directly to the selection);

2. if a 'plausible best' answer had been selected and a strong reason given for calling it the 'best' answer (again, the justification would relate directly to the selection);

3. if a strong justification of the 'best' answer had been given but the 'best' answer had not been selected (thus, the justification would not relate to the selection).

In addition to these three general situations, it was ruled that there should be a further, exceptional, situation:

4. if an answer that had not previously been considered even to constitute a 'plausible best' answer had been selected and a strong reason given for calling it the 'best' answer (again, the justification would relate directly to the selection).

Situation 4 was ruled in recognition of the continuing reappraisal of answers that had occurred throughout the item development process. This had revealed that even experts had failed to spot certain plausible interpretations until the mark scheme development meeting and it was, therefore, quite possible that other plausible interpretations remained unnoticed. It was stressed that Situation 4 was unlikely to be a common occurrence and that, in a formal examination context, would require consultation with the chief examiner.

As a general point of principle, it was ruled that a positive marking approach would be adopted, such that a justification that included both a credible component *and* an incredible or inappropriate component would be awarded the justification mark. In addition, it was ruled that a justification that was based purely upon an elimination of alternatives would not be sufficient to be awarded a mark. That is, candidates should be expected to provide a positive explanation of their selection (although if they took this approach *and* eliminated alternatives they could be credited with a mark).

3.3.1.3 Marking the essay and short reports

As it was not possible to consider the essay and short reports in sufficient depth at the item ratification meeting, their marking had to be determined largely from scratch during the mark scheme development meeting. The Project Leader proposed that each of the short reports and the essay should be marked according to the same criteria, which might be referred to as the capacity for critical writing. Once again, the development process was hindered by the lack of a full construct specification, in particular, a previously agreed interpretation of critical writing. However, to facilitate a practical solution to this problem, the Project Leader circulated a list of criteria that had been abstracted from a range of sources (particularly published and unpublished critical thinking essay tests). The list highlighted seven sub-components within three main components.

The meeting generally agreed that the sub-components identified were appropriate criteria according to which the essay and short reports could be marked. However, there was some reticence in rewarding one of the sub-components – 'fairness' – if this was taken to mean giving equivalent emphasis to both sides of an argument; the point being that both the essay and short reports required students to argue for one perspective against another. Following discussion, the following performance components were proposed as marking criteria for the essay and short reports:

1. **Selection** (to identify and distinguish between central, peripheral and irrelevant issues; to address all main points; to provide support for claims and assumptions where necessary, targeting an argument at the appropriate level);

2. **Synthesis** (to reach conclusions through strong argument; to develop, explain and defend central issues and claims; to identify and integrate conflicting perspectives; to make logically sound inferences and to avoid bias; to generate new evidence and argument; to recognise when claims are tentative);

3. **Structure** (to formulate issues well; to present clearly, with effective 'signposting'; to develop argument in a manner that is easy to follow; to be consistent throughout).

In addition, it was ruled that marks would not be awarded simply for stylistic qualities that were independent of critical thinking. This led to the following guidelines:

1. marks should not be awarded simply for quality of English (spelling, punctuation, grammar or flair);

2. marks should not be awarded simply for quality of rhetoric (apparent persuasiveness – that stems from rhetorical style rather than strength of argument – should not be rewarded);

3. marks should not be awarded simply for quality of description, paraphrasing, summary or explanation (where this is not central to the development of the argument).

Each of the three answers within Section C of the Examination Paper (2 short reports or 1 essay) were to be marked out of 20 (marks for the Mode P essay question would subsequently be doubled). The marks were to be spread across the three components of critical writing in the manner presented in Table 3.3.

TABLE 3.3. Performance bands for marking each Section C question

	Band 3	Band 2	Band 1
Selection	0 – 2	3 – 4	5 – 6
Synthesis	0 – 3	4 – 6	7 – 9
Structure	0 – 1	2 – 3	4 – 5

In addition to generic performance descriptors for each performance band of each critical writing component, specific guidance relating to each question was presented.

The final mark scheme is attached as Appendix 3.6.

3.3.2 The co-ordination

It was initially intended that the two-day mark scheme development meeting should also incorporate an element of co-ordination and a sample of additional scripts were to be marked collectively during the second day. However, as with the item ratification day, the ambiguities of the trial examination led to an extended discussion which – while essential for clarification of the mark scheme – meant that little co-ordination could be undertaken. As such, a remote co-ordination exercise was conducted during the following week.

For this exercise, all four markers were sent a sample of 12 scripts, which had been randomly selected and photocopied. Half of these were from Mode P and half from Mode Q. All markers were to mark all questions for both modes according to the newly

developed mark scheme. Marks awarded to the 12 scripts by the 4 markers were collated and analysed for consistency.

Generally speaking, the consistency of marking was good for the justification components. It was slightly better for the multiple choice justifications that for the multiple rating justifications, but even for the latter the consistency was reasonable. Generally speaking, when there was any inconsistency of marking between the four markers, the tendency was for one marker to diverge from the other three (though not necessarily the same one). Markers were given full feedback on their performance in relation to the performance of the other markers (anonymously). Where they had diverged from the other three, they were asked to consider if this was due to a relative lenience or harshness that could be brought into line with the others for the marking of the live scripts. From comments that they had annotated on the mark sheets it looked likely that this was the case in at least a good number of instances.

For questions where more than the occasional script was marked inconsistently – and particularly when there were a number of scripts that had split the markers evenly – the Project Leader looked for clues as to why this inconsistency might have occurred (particularly from markers' mark sheet annotations). Question 17-4 was particularly problematic and this appeared to reflect an ambiguity that had been discussed at length during the mark scheme development meeting. The advice to the marking team concerned the interpretation of the main focus of the stimulus passage – a matter that the marking team had not fully agreed upon previously.

The marking of the essay and short report questions was generally satisfactory. Consistency was strongest between two of the markers who agreed to a considerable extent, both in terms of script ranking and absolute marks awarded, across all three questions. The remaining two markers tended to agree with the first two on one of the questions, but to diverge on two (and to diverge from each other as well). The Project Leader ruled that the marks awarded by the two most consistent markers should be reviewed by the other two who should amend their marking appropriately. Where the two 'discrepant' markers tended to be somewhat harsh, they were asked to award higher marks.

As a result of the co-ordination process, a small number of issues relating to the mark scheme were clarified:

1. [justification components] marks should be awarded for the specific act of justification, not simply for responding to the critical comment (or summary) in an intelligent or creative manner;

2. [essay and short reports] the opportunity for demonstration of competence should be taken into account when awarding marks for each question, that is, marks should be awarded in relation to the quality of performance that might reasonably be expected in response to a particular question;

3. [essay and short reports] where students' essays or short reports contained incorrect or inappropriate statements, claims, interpretations, inferences, etc., these should *not* be ignored under the principle of positive marking – students should be penalised for significant errors as they reflect a weakness of critical thinking.

Marks awarded to scripts during the co-ordination were not formally recorded and all 12 co-ordination scripts were marked formally during the subsequent live marking phase.

3.3.3 The marking and re-marking process

The marking was conducted during the last three weeks of May and consisted of two despatches. The first despatch consisted of the entire sample of scripts from all candidates. These were divided among the four markers. The second despatch consisted of a sample of 100 photocopied scripts – selected in a quasi-random fashion from the first despatch – which was to be re-marked by the four markers (25 scripts each).

3.3.3.1 Despatch A – the live marking

The markers were instructed to mark the scripts as they might for any other examination. As the data would not be entered directly from the Examination Papers, markers were required to record both student selections (letters from A to E for the multiple choice and multiple rating items) and student marks on individual mark sheets. (Incidentally, during the co-ordination, only a single clerical error of letter transfer was noted.)

The marking was to be conducted in red ink and markers were permitted to annotate scripts if they wished to. This was not required though. Once completed, scripts and mark sheets were returned to the NFER.

3.3.3.2 Despatch B – the re-marking

The intention of the re-marking exercise was to obtain an indication of the extent to which it was possible to achieve reliable marking. The design was such that each marker would re-mark a sample of 25 scripts – from either Mode P or Mode Q – that had previously been marked by another marker. This design, effectively, amounted to 4 mini-re-marking studies: M1 vs M4; M2 vs M3; M3 vs M1; M4 vs M2. Thus, in all, 100 scripts were marked twice.

The design ensured that:

1. the 100 photocopied re-mark scripts would not be marked until the very end of the marking period;

2. the 100 originals of the re-mark scripts would be marked at the end of the first despatch.

Thus all scripts within the re-marking study were marked towards the end of the marking period.

3.4 Examination processing

After scripts and mark sheets had been returned to the NFER, the mark sheet data were input (a double-entry process was used to eliminate input error). The data were analysed by the Statistics Research Analysis Group toward three main aims:

1. to produce feedback for centres on their students' results (in terms of student mark, student standard score and mean school performance);

2. to produce information on the functioning of the examination and of individual items;

3. to produce information on the reliability of marking.

3.5 Grade awarding

As there was no requirement for grades to be awarded to students in the trial, no formal grade awarding meeting was convened. However, grade performance descriptors were developed and indicative grade boundary recommendations were made on the basis of a statistical analysis of results.

3.5.1 The grade descriptions

The AEA in Critical Thinking is designed to provide two grades of pass: Distinction and Merit, with Distinction being higher. Candidates who do not achieve the standard for Merit will be recorded as Ungraded.

The following grade descriptions were developed to indicate the levels of attainment characteristic of Distinction and Merit. They give a general indication of the required outcomes at each grade. The grade awarded will depend on the extent to which the candidate has met the assessment objectives overall. Shortcomings in some aspects of the examination may be balanced by better performance in others. As such, the grade descriptions are to be understood in a 'best fit' sense.

1. **Distinction**. Candidates demonstrate, across a full range of issues and contexts, a high level of ability in critical thinking. They do this by critically evaluating statements and complex reasoning deriving from them, accurately assessing the nature and strength of justification, even when measures are taken to disguise this, forming an accurate view of the strength of the arguments presented. They can identify and discount the effects of subtle biases and disguised persuasive devices in written, diagrammatic and numerical sources. They can select and summarise arguments clearly and accurately and produce their own persuasive or critical pieces of successful argumentation. They can collect, organise and synthesise subtle arguments across a wide range of subject matter.

2. **Merit**. Candidates demonstrate, for a wide range of issues and contexts, the ability to evaluate the reasonableness of statements and a line of complex reasoning, assessing the nature and strength of justification and forming an accurate view as to the power of the argument presented. They can identify and discount the effects of biases and persuasive devices in written, diagrammatic and numerical sources. They can select and recount arguments and produce their own persuasive or critical pieces of argumentation. They can collect, organise and synthesise arguments across a wide range of subject matter. They can recount arguments and produce their own critical argumentation. They can collect, organise and synthesise arguments across a range of subject matter.

3.5.2 The boundary marks

From an analysis of the Mode P and Mode Q mark distributions, indicative grade boundary recommendations were made. These were based predominantly upon cohort-

referenced concerns, with the aim of providing an appropriate level of discrimination between students. On the assumption that the AEA trial represented a reasonable sample of the target population (in terms of ability range), it was decided that grade boundaries for Distinction and Merit should be placed at marks corresponding to a cumulative percentage of 33% and 67% of candidates, respectively, for both the Mode P and Mode Q samples. That is, Distinction would be awarded to the top 33% of students on Mode P and Mode Q, respectively; and Merit would be awarded to the following 34%. The remaining 33% of students on each mode would be classified as Ungraded. The marks that best facilitated these distributions from the trial data were as presented in Table 4.4.[2]

TABLE 4.4. Notional grade boundary marks and percentage of mark total (cum. percentage of students in parentheses)

	Mode P	Mode Q
Distinction	53 marks, 64% of max. mark (33.3%)	47 marks, 57% of max. mark (34.3%)
Merit	44 marks, 53% of max. mark (67.2%)	40 marks, 48% of max. mark (68.1%)

3.6 Issues arising from the development process

It would be misleading to present the quantitative analyses of Section 4, without first discussing the numerous qualitative issues that arose from the development process. Many of these concerned the validity of the examination, although there were also numerous practical lessons to be learned.

---

[2] There is, of course, an inevitable contradiction in simultaneously defining grade boundary marks both statistically and in terms of performance criteria. That is, there is no *a priori* reason to believe that students at statistically defined grade boundaries will necessarily display qualities recorded in pre-determined performance criteria. On the other hand, there are strong theoretical reasons for believing that performance criteria are insufficient for defining performance standards with any precision (as national curriculum experience has confirmed). The approach recommended in the present report would be for initial standard setting to be based principally upon statistical concerns and for grade descriptions (not criteria) to be honed through a subsequent analysis of student performance at grade boundary marks. (Of course, this is not to propose that standards should subsequently be *maintained*, from year to year, on a purely statistical basis.)

3.6.1 First things first

The most important practical lesson was that, for the development process to run smoothly, problems need to be solved sequentially rather than in parallel. This was most clear in relation to the construct specification. The practical imperative meant that examination development had to occur in parallel with the development of a detailed construct specification. Indeed, the construct specification should still be regarded as essentially Work in Progress. In many ways, this made much of the examination development work quite ambiguous and led to delays at later stages (i.e., item ratification and mark scheme development), as fundamental theoretical issues had to be disentangled before practical solutions could be arrived at. It also led to wasted effort as questions that failed to meet subsequently established criteria had to be rejected. Further development of the AEA in Critical Thinking should treat construct specification as its primary task. This is crucial since – in contrast to most examined subjects – there is no existing consensus to be drawn upon and the developmental work to date is incomplete.

3.6.2 The status of ambiguity

At the core of critical thinking is the notion of constructing and deconstructing complex everyday arguments, in which even the questions are often ambiguous, let alone the answers. As such, critical thinking tends to seek solutions that are more or less plausible rather than simply right or wrong. What, then, should be the role of question formats whose answers are generally supposed to be unambiguously correct or incorrect? During development of the AEA in Critical Thinking, this problem was faced most squarely in relation to the multiple choice questions.

One response would be to take steps towards making the plausible answer to each question very plausible (the 'correct' response) and the implausible answers very implausible (the 'incorrect' distracters). To some extent this was what happened during the development of multiple choice items. However, the greater the disparity between the plausibility of 'correct' and 'incorrect' answers, the easier a question often becomes. This is undoubtedly a major problem if the purpose of an examination is to discriminate between the best A-level candidates.

Another response would be to develop questions that were more clearly grounded in formal logic (or probability, etc.) and which, therefore, did have definite 'correct' and 'incorrect' answers. Once again, a small number of questions were subsequently developed in this manner. However, as critical thinking stems from the discipline of *informal* logic, skewing the examination too far in this direction would risk

misrepresenting the construct. The CTAG was divided on the extent to which formal logic 'brainteasers' were appropriate for inclusion in the AEA examination (apparently reflecting differing conceptions of the scope of critical thinking – which, again, emphasises the necessity of pre-specifying the construct).

The response recommended by the first CTAG meeting was to consider requiring students to justify their multiple choice selections. This would then allow them to generate and demonstrate their own reasoning. In addition, it would allow them to achieve marks for justifying responses that were not necessarily deemed to be the 'correct' or even the 'best'. Unfortunately, though, this does not entirely solve the problem of ambiguity. Instead, it tends to shift it from the student's response to the mark scheme and the marker's evaluation: exactly what kind of justification, and how much justification, should be required for a mark? This is to raise the spectre of marker reliability. For this reason, the justification of multiple choice answers was initially resisted by the NFER.

3.6.3 The problem of reading load

A crucial design requirement of the AEA in Critical Thinking was that it should not require teaching and that it should not assume any specific academic or general knowledge. On the other hand, critical thinking is largely concerned with reasoning about knowledge claims. If the trial examination was to assess how well students could apply their critical thinking skills to meaningful problem areas then it would have to ensure that they were given sufficient background information relating to those areas. This seemed particularly important for the essay and short report questions, but it was also a significant concern in developing the multiple rating and multiple choice questions. Unfortunately, though, providing background information in the body of each question tends to make the paper long-winded and leads to a considerable reading load.

While the load was reduced substantially during the development process, even the final question papers contained a significant amount of reading. Although there may be ways of circumventing this problem, certain question formats will always raise more problems than others. For example, the summary multiple rating items will generally require more reading than the critical comment multiple rating items. Likewise, if students are required to prepare their own argument from information provided, this will always necessitate the provision of a substantial reading load. Future development of the AEA in Critical Thinking will have to consider how this problem can best be dealt with.

3.6.4 The availability of experienced item writers, senior examiners and assistant examiners

Although the NFER critical thinking team had considerable experience of developing specific kinds of test items (e.g., items similar to those in the NFER's Critical Reasoning Tests) the first CTAG meeting felt that item writers with complementary experience should also be co-opted. This highlighted a particular problem for the development of an AEA in Critical Thinking: the serious lack of item writers with experience of developing critical thinking examinations.

At present, the UK's only public examination in critical thinking is OCR's recently developed AS in Critical Thinking. Under the banner of the CIE/UCLES, Cambridge has also developed tests of problem solving and (the nearest relative to the Critical Thinking AS) the MENO. However, none of these examinations involve a large number of experienced item writers. Moreover, there is very little tradition of teaching critical thinking in the UK, which means that there are very few critical thinking specialists in UK schools who might readily be trained as item writers. In fact, critical thinking has not even evolved into a coherent academic discipline within Higher Education and there are relatively few critical thinking courses on offer in UK universities meaning that few, if any, critical thinking specialists are presently being trained.

The key point, then, is that the pool of potential item writers for the AEA in Critical Thinking is very shallow indeed. Those with any experience of critical thinking item writing are generally already involved with the existing examinations. More importantly, the intention of the AEA was not only to explore new ways of assessing critical thinking, but to develop question papers aimed at a very select and narrow ability range. Thus, item writing experience gained in the context of, say, the AS in Critical Thinking might not transfer directly to the AEA in Critical Thinking.

This problem will have to be addressed in future years if the AEA in Critical Thinking is to develop successfully. Of course, it is not simply a problem for the item writing process, it is even more of a problem for the marking process.

3.6.5 To justify or not?

Justification components were introduced to overcome the possibility that candidates might select – for good reason – answers not deemed by the expert panel to be 'best'. Thus, while a candidate might loose a mark for the selection component s/he might still gain it for the justification component. While this appears to make the examination more

fair it also has the consequence of raising the stakes of each question (from one mark to two). While the mark scheme acknowledges that the justification mark is not dependent upon the selection mark, both marks are still dependent upon understanding the thrust of the specific question. This raises an interesting question: (even ignoring the problem of marker reliability) would a ten mark section with five independent selection plus justification items be more reliable – in the sense of test reliability – than a ten mark section with ten independent items and no justification? Expressing the point more directly, reducing the number of independent questions might reduce test reliability more than including a justification component could increase it. Allowing justification might, therefore, not make the examination more fair.

Of course, including justification components technically also changes the nature of the competence being assessed and this needs to be borne in mind. If this is the explicit intention, then all well and good. Whatever, it needs to be clear exactly what competence the justification is intending to assess. Generally speaking the mark scheme for the trial examination did not assess justification components for quality of critical writing, as marks were not awarded for how well the point was made, simply for whether the point was made. Indeed, marks were even awarded when incorrect or inappropriate points were made in addition to correct and appropriate ones. Thus, the intention of the justification component was essentially to assess the same competence as was required for the selection component – meaning that both marks were conceptually dependent upon the one competence.

An alternative approach might have been simply to award marks for a successful justification, regardless of the selection made (awarding no marks for selection). Whether this would lead to greater test reliability than awarding marks for selection only is worthy of investigation. Of course, in practice, it would be compounded by the impact of marking (un)reliability.

If the justification of multiple choice or multiple rating items is to be adopted in future examinations, then there is a need for further research into the characteristics of questions that are readily amenable to justification (and those that are not). Experience of the trial examination indicated that some answers are technically more amenable to justification than others. Certain questions (e.g., Question 11 and 12) were clearly amenable, requiring candidates to state the general logical principle underlying their choice. Other questions (e.g., Question 13 and 14) were considerably less amenable, almost requiring candidates simply to regurgitate sections of the stimulus passage.

Somewhat ironically, candidates were rewarded less for *justifying their selection* on the critical comment multiple rating items than simply for *providing a sound general evaluation of the force of the critical comment.* The point being that, in principle, a justification ought (first to cite the general evaluation but then) to focus primarily upon why one particular position on the rating scale was chosen rather than any other. Not only would this require an extended explanation, it would add an additional facet to the justification which would make it even more complex to mark. As such, candidates were rewarded simply for a sound general evaluation. The validity of such an approach is perhaps questionable.

3.6.6 To mark positively or not?

As noted earlier, in relation to the positive marking principle, the approach taken for the essay and short reports was different from the approach taken for the justifications. Positive marking meant that justifications that were poorly structured, or that contained incoherent or irrelevant information, were credited so long as the gist of the correct response was evident. However, the essay and short reports were marked specifically in terms of their structure and quality of selection and synthesis.

This may be defensible on the assumption that different critical thinking skills were being assessed by the different question formats. However, it is still true that positive marking resulted in some extremely poor justifications being credited – justifications that could hardly be said to reflect high quality critical thinking. This is (at least) an issue of face validity (the extent to which crediting such poor responses would appear plain wrong to markers, teachers and students alike).

3.6.7 The multiple choice questions

Although the multiple choice format is standard in many tests (and particularly tests of critical thinking) the trial faced a number of problems that are worth highlighting. The first problem stemmed from the fact that it was not possible to complete the construct specification in advance of the examination development. As the item writers were forced to develop items while the construct was still being honed this resulted in a lack of clarity concerning what the questions were primarily intended to assess. This meant that a range of question types were included, but without a full rationale for their inclusion. Exactly the same was true of the distracters for individual questions which were generally not constructed according to a formal template nor with specific errors of critical thinking in mind. Future development work will have to rectify this. On the other hand, for an examination aimed at such a select band of high achievers, it will be important that future

examinations are not too formulaic – as clear formulas are often easy to spot once their general principles have been learned or taught.

The second problem encountered was the difficulty of pitching critical thinking questions at a level that would be appropriate for discriminating between these high achievers. There was a general feeling, both amongst the item writers and the CTAG, that the multiple choice questions were not a great deal harder than those developed for the AS in Critical Thinking. Unfortunately, learning how to pitch questions at a higher level is likely to take time. Whereas a normal AEA might simply require knowledge of a greater depth and breadth than its corresponding A/AS, the nature of the difference between the Critical Thinking AEA standard and the standard of the AS in Critical Thinking is not quite so clear.

It is worth bearing in mind that critical thinking may simply turn out not to be a competence according to which such a select band of students can reliably be differentiated.

A final problem encountered in developing the multiple choice items applied also to the other assessment formats. This was the amount of time and effort that it took to write a strong item. Perhaps the most significant complicating factor is that the entire nature of a stimulus passage, question or answer can be changed radically through a subtle difference in the way that a single word or phrase is interpreted. If alternative readings are not noticed before the examination goes live then this can result in invalid items and compromised mark schemes. Critical thinking items require a lot of time and effort to develop – the higher the profile of the examination the more important this investment becomes.

3.6.8 The multiple rating items

The multiple rating item (which was included at the request of the QCA and the Critical Thinking Advisory Group) is a new assessment format that has never before been piloted in a large-scale examination. Perhaps, then, it is no surprise that it raised some significant issues during the trial. Among the key practical lessons to be learned was the importance of making clear to students exactly what they are expected to do. One weakness of the Guidance Material was that it did not make clear that positive explanations were required in justification of the critical comments and that elimination strategies ought not to be attempted. Another weakness was the suggestion present in the Guidance Material that students might begin by reading all of the summaries and ranking them in terms of quality before choosing the appropriate grade. This appeared to have been taken by certain

students to mean that no two summaries could be awarded the same grade, which was not the intention.

Another practical obstacle was that the elimination of negative marking from the mark scheme – in agreement with QCA principles but contra to Fisher and Scriven's advice – meant that the statistical properties of the items were not as intended. Specifically, candidates who selected answers at random would, by chance alone, gain more marks than candidates who declined to answer. Fisher and Scriven's intention was that there should be a penalty for guessing and that this should increase the further away from the 'correct' response a candidate's selection was.

3.6.8.1 The rating scales

Turning now to certain of the more conceptual problems, the first was the lack of precision inherent in the rating scale descriptors. Of some theoretical significance, the fundamental problems identified were precisely those that have rendered strong criterion referencing unworkable in large-scale educational assessments.

The first significant obstacle is that the language of descriptors is imprecise and typically does not admit of more than a few discrete levels (cf. the five levels from A to E). The second significant obstacle is that performance descriptors typically assume that variables described co-vary (when, in reality, they often do not). To exemplify this latter point, rating A might state 'identifies all major points with no errors of interpretation' while rating E states 'identifies few major points with substantial errors of interpretation'; but then what of a summary that 'identifies all major points with substantial errors of interpretation'? As identification does not co-vary with interpretation, the rating scale does not have a clear location for it.

The ultimate weakness of even the final versions of the rating scales was that the five descriptors were simply not sufficiently discrete. Consider, for example, the descriptors for A and B for the summary question:

A. Identifies all the main points of the passage, and presents them with no errors of interpretation.

B. Identifies all or most of the main points of the passage, and presents them with few, if any, significant errors of interpretation.

Now, while there is clearly a practical distinction between A and B, A is formally a subset of B. Therefore, technically speaking, any student who put B when either A or B was

deemed to the 'best' rating ought to be awarded full marks. Of course, this is not the practical implication of the scale, but it is still the formal implication.

In a similar vein, consider the descriptors for C and D of the critical comment question:

C. Raises a plausible criticism of the reasoning, but does not threaten the main conclusion of the argument.

D. Raises a very weak criticism of the reasoning, and does not threaten the main conclusion of the argument.

Now, while these are clearly different, the difference between a plausible and a very weak criticism (neither of which challenge the main conclusion) is undoubtedly one of degree rather than kind and will ultimately be a matter of judgement. Whether the consensus of a panel of expert critical thinking examiners is sufficient to 'objectivise' this judgement is certainly open to debate. If not, then penalising candidates for choosing the 'wrong' one would seem somewhat harsh.

3.6.8.2 The summary questions

The summary questions proved to be considerably more straightforward than the critical comment questions when it came to writing the marking key. However, that is not to say that there were no problems.

Close inspection of the stimulus passage revealed that the article had not one, but two, subtly different purposes:

1. to explain why ice ages come and go (see title and paragraph 1); versus

2. to explain why we cannot predict the next ice age (see paragraph 6).

If the main purpose was taken to be point 1, then parts of the passage about predicting the next ice age would be relatively incidental. If, on the other hand, the main purpose was taken to be point 2, then details concerning the causal mechanism would be relatively incidental. The decision between the two might well affect the rating grade chosen to represent a particular summary. As a practical compromise, it was ruled that both purposes were relevant. Yet a candidate would have good justification for choosing either 1, 2 or both.

This kind of problem could, presumably, be avoided through careful selection and scrutiny of the stimulus passage. A deeper problem, though, is the extent to which the

question is actually assessing a crucial critical thinking competence. While the ability to summarise a passage undoubtedly involves certain skills of critical thinking, the ability to précis is generally considered to be a linguistic skill. Making sure that the stimulus passage clearly took the form of a subtle argument would go some way towards making it more explicitly a critical thinking item. Yet whether it results in the examination being too heavily skewed toward more peripheral critical thinking competencies must still be considered.

3.6.8.3 The critical comment questions

One of the more surprising outcomes of the mark scheme development meeting was a major split between panel members concerning the 'best' rating for Question 17-5. After an extended discussion, the best explanation for this split appeared to be that the panel members had interpreted the argument of the stimulus passage in radically different ways. Although the passage (intentionally) did not reflect a simple formal structure, it was not anticipated that the subtle informal structure would be interpreted so differently. This proved to be a major obstacle to the marking and, effectively, rendered Question 17 invalid.

Perhaps, though, this indeterminacy of interpretation should not have been surprising. Critical thinking is grounded in the discipline of informal logic and a particular problem for this discipline has been to locate examples of everyday argument to which the principles of informal logic can effectively be applied (Antaki, 1994). Informal logic is ultimately unable to circumscribe everyday argument – even though informal logicians are explicitly concerned with the analysis of argument as it occurs in ordinary language.

The kind of evaluation of everyday argument that is required to answer critical comment multiple rating items involves going beyond what is actually said to what is meant. This is because everyday arguments typically do not conform to a straightforward logical structure where the premises and conclusion are stated in unambiguous terms. In everyday argument: crucial components of the argument may be left out; irrelevant components may be included; and the language in which it is couched may be interpretable in numerous ways. Of course, it is precisely the task of the critical thinker to be able to explore such ambiguities in an attempt to reduce what has been said to the bare bones of the argument that might have been intended. Crucially, though, there might well be a range of equally plausible interpretations of the argument that was intended. Moreover, without further interrogation of the source, it might not be possible to decide between these alternatives.

This was precisely the problem faced with Question 17: the marking team could not decide upon a single interpretation of the argument that was intended. In the absence of a single agreed interpretation it is hard to see how students could be asked to "rate the critical force of each comment as a challenge to the argument".

Yet the ambiguity was not confined to the interpretation of the passage but applied also to the interpretation of the comments. Here it becomes important to reflect upon the nature of the task being set. Are students being asked to rate the critical force of the 'stated' comment or of the 'intended' comment? The stated comment goes no further than the sentence itself, for instance: "Other adult hobbies, e.g., an interest in model railways, could be seen as equally childlike." In fact, in itself, this does not formally constitute a challenge to the argument of the passage at all. It only constitutes a challenge once its implications are drawn out, for example:

1. the argument of the passage is fundamentally premised on the assumption that collecting toys in adulthood leads to an over-sentimentalism in respect of the real objects that they represent;

2. adults who collect model trains are not over-sentimental about real trains;

3. this is, therefore, a powerful counter-example that undermines the fundamental premise of the argument.

Needless to say, this is not the only way in which the critical comment could have been interpreted. The point, then, is that a critical comment can only be rated for the *intended* meaning that it is taken to convey (because what is stated is insufficient to mount a full challenge). Moreover – and even more in relation to the critical comments than to the passage – this intended meaning can be read in a multiplicity of ways. *There can be no single correct interpretation of the critical comment.*

For the sake of argument, though, let us assume that there had been only one interpretation of the argument of the passage and only one interpretation of the critical comment and, further, that the critical comment (as developed above) had interpreted the argument of the passage correctly. Would we then have been able to pass judgement on the force of the critical comment? No. In the above example, all that we are now concerned with is to pass judgement on steps 2 and 3 (having assumed 1 to be correct). Even here though we get into trouble. If, for example, step 2 was true then this might support a rating towards one end of the rating scale; yet, if it was false, it might support a rating towards the other. Importantly, unless we could assume that the status of step 2 (as

a knowledge claim) was general knowledge, then it would not be fair to arbitrate the matter one way or the other.

It might be responded that this inherent ambiguity is precisely why justification is so crucial. Perhaps. If so, though, it would be more reasonable to provide candidates with a single critical comment and to give them at least an hour and a half to complete their justification as an essay. (This is not intended to be a flippant comment!)

The conclusion of this discussion is that the critical comment multiple rating item is not well suited to the simple selection-plus-short-justification format that was trialled. While this conclusion is, of course, based upon only one example, the problems that it raised would appear to be general. It is hard to see how the ambiguity of interpretation could be avoided without trivialising the exercise or reducing it to one of formal logic.

As a consequence of ambiguity of interpretation it may often be quite possible for both ends of a rating scale to be equally justifiable (as was observed most dramatically for Question 17-5). As such, the awarding of marks for the selection component is seriously problematic and it might be best simply to abandon this enterprise.

3.6.9 The essay and short reports

For students to demonstrate their capacity for critical writing, through extended constructed responses, this might be achieved in a number of ways. In developing the AEA, the two main types of item appeared to be:

1. argument evaluation questions;

2. argument construction questions.

The former generally involve the presentation of a relatively brief stimulus passage (containing the text of an everyday argument) and the instruction to write a critical evaluation of that passage (often encouraging students to introduce further evidence and argument as appropriate). In fact, this type of question is a key component of the AS in Critical Thinking.

In contrast to the argument evaluation format, an argument construction question would require students to draw upon the content of a stimulus passage (or passages) but would not require students directly to evaluate that passage (those passages). Thus, the instruction would not be to discuss the quality of argument presented in a specific text,

but to use the text (or collection of texts) as a resource for the construction of a new argument.

As many of the earlier (multiple choice and multiple rating) questions were intended to assess candidates' skills of evaluation, it was felt that the essay and short report questions should emphasise the capacity for synthesis. It was decided that these questions should, therefore, require students to reflect upon a range of evidence sources, using them as a resource for constructing their own arguments. It was further decided that the content of these arguments should be accessible to all and should relate to significant controversies selected from a range of everyday fields (e.g., political, socio-scientific, etc.).

As discussed in 3.6.3, the problem with presenting text as a resource for a question, rather than as the subject of a question, is that it tends to increase the reading load. This proved to be a significant problem for both the short reports and the essay question. In addition, the trial revealed a number of others problems presented by this assessment format.

3.6.9.1 The essay with pre-release (Mode P)

Having selected a key controversy and texts that an answer might be based upon, the major problem faced in developing the essay was exactly how the question should be phrased. While the general intention was for candidates to present a strong case for or against 'the issue' of the question this was made problematic by the fact that everyday controversies often tend not to be single issue matters. Indeed, real-world debate is often characterised by general discussion of costs and benefits without consideration of specific implications for alternative courses of action.

How, then, should each essay question frame its central controversy for students? To illustrate this concern in relation to the GM food question, students might have been asked to prepare an argument for or against either of the following propositions:

1. Government should place a moratorium on UK research into the genetic modification of food;

2. the potential risks associated with GM foods and crops should not dissuade Government from promoting this new technology.

The advantage of the first formulation would have been in tying down the issue to a very specific course of action, making the purpose of the argument quite clear. However, the disadvantage would have been the redundancy of a great deal of the general content contained in the pre-release material and stimulus passage. The advantage of the second

formulation was that it left the focus of the debate as wide as possible (the potential risks and benefits of GM crops and foods) meaning that very little of the content was clearly redundant. However, it did mean that the course of action was somewhat obscure (does 'promoting' mean: supporting the idea, if only in principle, during international debate; not introducing anti-GM legislation in the UK; providing major financial support for research and development in Wales; etc.). Not wanting to limit the scope of students' arguments, the second formulation was chosen over the first, with the consequence that the precise meaning of the course of action (to be argued for or against) remained somewhat ambiguous.

Unfortunately, though, there is a further problem with the general question format when used in conjunction with pre-release material. If students know that the question will be general rather than specific then – knowing the nature of the controversy from the pre-release – they will be able to prepare an essay in advance of the examination. Schools could therefore assist students through preparation lessons. These two obstacles raise significant concern over the essay with pre-release material format.

Both the short reports and the essay question were characterised by another potential limitation that is worth noting: they required candidates to present an argument in support of one side of a debate (and against the other). If an important element of critical thinking is the capacity for open-mindedness then this requirement may be thought to act against it (or, at least, against its assessment). On the other hand, if open-mindedness is interpreted as the fair representation of alternative perspectives then this challenge may be taken to have somewhat less force. The essay question went further by specifying that students could also choose which side of the debate to promote. If this opportunity were offered to candidates in the future, it would be important to establish that they were not penalised for the side that they chose to adopt (either through examiner bias or through differential demand characteristics of writing for or against a particular position).

3.6.9.2 The short reports (Mode Q)

An additional problem with providing a significant amount of content as a resource for students is the risk that students may end up simply regurgitating the material without demonstrating much depth of analysis. The tension, once again, lies in asking students to reason within a domain about which little, if any, specific content knowledge can be assumed. The rationale for the short report and essay questions was that specific content should be presented from a variety of perspectives and the students' task should, essentially, amount to the selection of key issues and the synthesis of evidence for or

against the position adopted. However, the more clear-cut the controversy was, the less of a critical contribution appeared to be required from the students. Thus, the markers reported concern that there was less opportunity to demonstrate the required skills of critical writing on the 'electoral reform' question than on the 'Bermuda triangle' question.

3.6.9.3 The marking of future examinations

Finally, the likely problems of scaling-up the marking of constructed response questions should be stressed once again. Schools in the UK generally do not employ teachers for their expertise in critical thinking and it is likely that there are few teachers with sufficient expertise to mark critical thinking essays reliably without a significant investment in training. Critical thinking essays, although ostensibly similar to essays in English or General Studies, require a significantly different approach to marking. Indeed, experience of marking for related subjects might even be a disadvantage for those who wish to enrol as examiners of critical thinking. Critical thinking examiners would, therefore, need very careful training in addition to a satisfactory grounding in the discipline itself.

3.6.10 Summary of issues arising from the development process

As the preceding sections have amply demonstrated, even before the results of the trial were apparent, numerous lessons were learned from the development process. For future examination, the three most general and significant concerns would appear to be:

1. to develop the construct specification fully;
2. to resolve the problem of item ambiguity (and its impact upon item difficulty, mark scheme validity and marking reliability);
3. to increase the pool of sufficiently competent item writers and examiners.

In addition, a number of more specific conclusions were reached:

1. multiple rating items are problematic and may well not be appropriate for high-stakes public examination;
2. the use of justification for multiple choice questions is in need of further research and development;
3. the 'argument construction' essay model should be reconsidered (in view of reading load and question ambiguity concerns);

4. the use of positive marking in critical thinking examinations should be reconsidered;

5. good critical thinking items take a very substantial amount of time to develop due, in particular, to a tension inherent in critical thinking between the theoretical necessity of logical precision and the practical ambiguity of its subject matter.[3]

---

[3] Incidentally, but importantly, the item writing team fully acknowledged that the questions developed for the trial examination were far from perfect. Had there been sufficient time available the questions would have been of a significantly higher quality. However, many of the general problems addressed within Section 3.6 would still have occurred, owing to the nature of the subject matter and to the type of questions that the team was required to develop.

Section 4 Evaluating the Trial

The presentation of results from the trial is broken down into three main sections, reflecting:

1. feedback from students and teachers;

2. general statistical functioning (of items and papers);

3. specific measurement analyses (of reliability and validity).

This is prefaced by a number of sample composition caveats that need to be borne in mind when reading the results.

4.1 Caveats regarding sample composition

As many of the analyses below present results for Mode P alongside results for Mode Q, it is important to stress from the outset that caution should be exercised when drawing comparisons. In fact, it would be unwise to infer directly from such comparisons as the samples of students allocated to Mode P and Q, respectively, were not directly comparable.

4.1.1 Sample selection

The samples of schools were largely provided for the NFER by the QCA and schools were allocated to a single mode of examination. Clearly, under these circumstances, it would not have been possible to construct precisely matched samples. Instead, the samples were constructed so as to be broadly similar (in terms of school type composition and gender balance). It should therefore not be assumed that the 'average critical thinking ability' of students within Modes P and Q was necessarily comparable, nor that the two groups were necessarily equivalent on any other dimension. As such, differences in performance or attitude between samples should not be over-interpreted.

It should also be recognised that many of the schools that participated in the trial had previously shown an interest in either Advanced Extension Awards or, more specifically, the Critical Thinking AEA. As such, we might expect these schools to be amongst the more positively disposed towards the examination. The feedback from teachers, in particular, should be interpreted in this light.

4.1.2 Sample characteristics

Tables 4.1 to 4.4 present data on the breakdown of various student characteristics by mode.

TABLE 4.1. Sample breakdown by ethnicity.

	Percentage of Students	
	Mode P	**Mode Q**
White	89 %	85 %
Mixed	3 %	2 %
Asian	5 %	9 %
Black	0 %	1 %
Chinese	2 %	1 %
No response	1 %	1 %

Table 4.1 represents the ethnic diversity of students in the trial. Clearly, there was not a great deal of diversity, with the vast majority of students considering themselves to be White.

TABLE 4.2. Sample breakdown by prior attainment.

Average GCSE score (A* = 8, A = 7, ... U = 0)	**Percentage of Students (Cumulative)**	
	Mode P	**Mode Q**
7.5 < x ≤ 8.0 (e.g., > 5 A*s & 5 As)	35 % (35 %)	30 % (30 %)
7.0 < x ≤ 7.5 (e.g., > 10 As)	37 % (73 %)	33 % (63 %)
6.5 < x ≤ 7.0 (e.g., > 5 As & 5 Bs)	19 % (91 %)	16 % (80 %)
6.0 < x ≤ 6.5 (e.g., > 10 Bs)	8 % (100 %)	12 % (91 %)

Virtually all students provided GCSE grades (and little else) as evidence of prior attainment. Table 4.2 displays the prior attainment of students in terms of average GCSE grade (including Short Course grade, but excluding grades from any other qualifications). As intended, the sample represented only the most able students.

TABLE 4.3. Sample breakdown by AS in Critical Thinking.

	Percentage of Students	
	Mode P	**Mode Q**
AS in Critical Thinking	13 %	12 %
No AS in CT	87 %	88 %

The data in Table 4.3 show that only a relatively small number of students had taken the AS in Critical Thinking. This limited the comparative analyses that could be conducted in relation to prior experience of critical thinking (as did the fact that these students also tended to cluster in particular centres).

TABLE 4.4. Sample breakdown by subject grouping.

	Percentage of Students	
	Mode P	**Mode Q**
No Specialism	15 %	16 %
Sciences	53 %	51 %
Humanities	22 %	23 %
Languages	10 %	10 %

Subject grouping was defined in terms of students' A-level choices. Where a majority of a student's subjects fell into one of the four categories listed in Table 4.4 this was classed as their specialism.[1]

4.2 Feedback from students and centres

Feedback from students and centres was obtained from site visits to a small number of schools and from questionnaires that were completed by all students and all teachers.[2]

---

[1] The specialisms were based upon an initial categorisation of each A-level into one of 16 subject area groups (based largely upon the NCER model). These groupings were as follows. SCIENCES: Sciences, Mathematics, ICTs, Technologies, Industries. HUMANITIES: Art, Geography, History, Humanities, General Studies, Critical Thinking, Music, Sport. LANGUAGES: English, Welsh, Languages.

4.2.1 Observations of the trial

Observations were undertaken in 6 participating centres: 3 Mode P and 3 Mode Q. Within each mode, 1 Independent, 1 Grammar and 1 Comprehensive school was visited. An NFER observer was present in each centre for the full duration of the examination and subsequently conducted a focus group interview with students involved in the trial. A full report on the six site visits is presented in Appendix 4.1.

4.2.1.1 Observation of procedures

The examination appeared to function well procedurally. The invigilator rubric was followed appropriately and students rarely felt the need to seek clarification of how they were required to approach the examination. Most candidates appeared to have sufficient time to complete the examination and, generally speaking, no more than two or three students completed it with more than half an hour to spare.

4.2.1.2 Focus group feedback

Feedback from the focus groups suggested that the subject matter chosen to examine critical thinking was generally perceived to be varied and interesting. However, many students were concerned that the 3 hour examination was too long. Specific comments on the questions of each section are presented below.

4.2.1.2.1 Section A – Multiple Choice

Students generally considered the multiple choice items to be interesting, valid and challenging. The multiple choice format was also considered to be effective for testing critical thinking. The justification component of this section was welcomed, although some students were not sure what was required for an effective justification. Others felt that there was not enough space in which to provide a full justification of why a particular response was chosen and why others were not.

4.2.1.2.2 Section B – Multiple Rating

The multiple rating section was received least well by students, although some students were happy with it and found it interesting and enjoyable. Many students found the rating

---

[2] The trial occasionally addressed the teachers as examination 'administrators', as this was one of their primary roles in the trialling process.

scale to be too subjective and vague. Likewise, many found the target comments or summaries too similar to distinguish between them.

Once again, although the justification components were welcomed as an opportunity to defend their responses, many students were unsure of the criteria for a response that would be credited. There was also some feeling that the justification component might be biased towards those with superior writing skill.

Other concerns with the items of this section included the layout, which was felt to require a lot of 'flicking backwards and forwards' between pages and, therefore, a substantial memory component. It was suggested that a 'tear-out page', on which the rating grades could be described, would overcome this problem. Some students were also confused as to whether the target summaries or comments were to be rated independently or, effectively, to be rank ordered without repetition of a rating grade.

4.2.1.2.3 Section C – Mode P Essay

Opinions differed as to the validity of the essay question. Some students considered the presentation of an argument to be an important skill of critical thinking, while others considered it biased in favour of those with superior English skills. There was a widespread feeling that the skills assessed were similar to those of English. A common criticism of this task was that it amounted to little more than copying or rephrasing material presented under different sub-headings. Similar to the concern over the assessment of English skills, those students who had studied Biology considered themselves to have been at an advantage in writing an essay on GM Food.

There was a feeling that the pre-release material was too much of a clue to the nature of the question that would be asked in the examination (meaning that students would be able to prepare an essay in advance). To overcome this, it was suggested that all of the reading material might be included in the examination booklet (assuming that sufficient reading time was also provided).

Other suggestions for improving the examination included the specification of a word limit and the writing of two separate arguments (defending each side of the debate respectively).

4.2.1.2.3 Section C – Mode Q Short Reports

Again, there was a feeling that the task demands favoured students with superior English skills and that the examination required more copying and rephrasing than critical

evaluation and analysis. On the other hand, some students felt that the debate aspect of the task was satisfying and that the identification of useful sources and the use of presentation skills made the Short Reports valid.

The first report (Bermuda Triangle) was generally considered to be more interesting and straightforward (although some students were concerned with its reading load). A number of students considered the material for the second report (Electoral Reform) to be somewhat odd, as the points 'against' the argument were presented in the notes of the protagonist arguing 'for' the case. Opinion differed as to whether this was confusing or integral to the task. The second report was felt to be less engaging than the first.

Some students felt that insufficient relevant information was provided for the production of a successful argument and that high marks would depend on prior knowledge. This was considered to be particularly likely for the second report.

4.2.2 Feedback from teachers

A questionnaire was provided for the teacher responsible for organising the Critical Thinking trial within each school.[3] This posed a number of questions regarding their impressions of the examination. A full breakdown of the responses to each question is presented in Appendix 4.2 and the key points are summarised below.

A series of 6 questions were posed and teachers were requested to respond with a tick corresponding to one of five agreement statements (Strongly agree, Agree, Unsure, Disagree, Strongly disagree). In fact, they were requested to provide 5 responses to each question; 1 for each section of the examination (multiple choice without justification, multiple choice with justification, multiple rating summary, multiple rating critical comment, short reports/essay). A final question concerned the value of the examination as a tool for university selectors.

The first six questions – posed as statements – were as follows:

4. The instructions are clear and intelligible.

5. Sufficient time is provided to answer the questions.

6. The questions are stimulating.

---

[3] This is referred to as the Administrator Questionnaire.

7. The questions are a fair test of critical thinking skill.

8. The questions are sufficiently challenging.

9. The question topics are interesting.

Teachers tended to respond very similarly between modes, across sections (within questions) and between questions. For no question, in relation to any section on either mode, did the percentage of teachers in agreement with each statement fall below 50%. That is, there was no statement for which fewer than 50% of teachers ticked either the 'Agree' or the 'Strongly agree' box (for any of the sections of the examination). This seems to represent an overwhelmingly positive impression of the trial examination.

However, the statement that received the least positive assessment was question 7. The percentage of teachers that agreed that the examination was a fair test of critical thinking skill ranged between half and two-thirds (but no higher). For both modes, there was more confidence in the multiple choice items than in both the multiple rating items and the essay/short report items. In fact, a trend across questions was for the essay/short report items to be associated with lower agreement ratings than any other item type.

The final question on the questionnaire posed the following statement:

10. This qualification would provide strong evidence to help universities select the best students.

This resulted in the least certainty and in a difference of perspective between modes. Whereas three-quarters of mode Q teachers agreed that the qualification would provide strong evidence, only two-fifths of mode P teachers agreed. Over half of the mode P teachers were uncertain, as were one-fifth of mode Q teachers.

4.2.3 Feedback from students

A very similar questionnaire was provided for students, who completed it immediately after having sat their examination. A full breakdown of responses is presented in Appendix 4.3 and the key points are summarised below.

The questions to which students responded – presented as statements – were as follows:

6. I understood the instructions.

7. I had enough time to answer the questions.

8. I enjoyed answering the questions.

9. The questions were a fair test of my critical thinking skills.

10. I found the questions challenging.

11. I found it hard to do what the questions asked me to do.

12. I found the question topics interesting.

While the students responded similarly between modes and across sections (within questions), they evidenced slightly more differential responding between questions than their teachers had. However, their responses were still positive, tending to range between majority agreement and overwhelming agreement.

Overwhelming agreement was noted in response to questions 6 and 7, which concerned clarity of instructions and the time needed to answer items. Across sections, no fewer than 70% of students agreed with these statements, with most sections receiving closer to 80% or 90% agreement. Most concern was expressed in relation to the Mode Q short reports, where 15% of students felt that they did not have sufficient time to provide answers.

Good levels of agreement were evident in relation to questions 9, 10 and 12 which concerned whether the examination was a fair test of critical thinking skill and whether it was challenging and interesting. As for teachers, there was most agreement amongst students that the multiple choice questions were a fair test of critical thinking skill and least agreement that the essays and short reports were. One-quarter of students felt that the Mode P essay was not a fair test and over one-fifth felt that the Mode Q short reports were not.

The lowest levels of satisfaction emerged from questions 8 and 11, which asked whether students enjoyed answering the questions and whether they found it hard to do what the instructions asked. Across both modes, students most enjoyed answering the sections that required the least written input from them (at least 72% agreement for the multiple choice items without justification and at least 49% agreement for the multiple rating item without justification). Similarly, in relation to the enjoyment statement, the sections for which most disagreement was recorded were the essay (38%) and the short reports (36%). When asked whether they had found it hard to do what the items asked them to do, roughly half

of the students tended to disagree and roughly one-third tended to agree. The multiple choice items were considered to cause the least problems (69% agreement for both modes).

Finally, students were presented with the following statement:

13. This qualification would provide strong evidence to help universities select the best students.

Their responses tended to be slightly less confident than those of their teachers, although there was still more confidence from Mode Q than from Mode P. Across both modes one-quarter of students disagreed with this statement. Similarly, one-quarter of Mode P students agreed with the statement, while two-fifths of Mode Q students did so.

4.3 General statistical functioning

Data arising from the Critical Thinking examination were subject to a number of statistical analyses in order to explore the technical quality of the examination as a test of critical thinking skill. The following presentation of results begins with an analysis of the general statistical functioning of the Mode P and Mode Q versions. Recall that Mode P and Mode Q differed only in terms of Section C. In fact, students performed very similarly between modes on the shared sections and, consequently, some of the data below have been aggregated.

4.3.1 Examination performance

Both modes of the Critical Thinking examination functioned reasonably well in terms of achieving a distribution of marks across the available range of 83. The minimum marks achieved were 17 (P) and 14 (Q) and the maximum marks were 69 (P) and 62 (Q). The mean marks achieved were 47 (P) and 43 (Q), with standard deviations of 9 for both modes.

Although the students were not formally matched across modes, they performed very similarly. Considering only performance on Sections A and B, the mean marks achieved for each mode were 22.7 (P) and 22.5 (Q). This means that the overall difference in performance between Mode P and Mode Q can, essentially, be put down to the contribution of Section C. As the response patterns for each item of Sections A and B were also very similar, the following presentation of results, for Sections A and B, represent data that have been aggregated across modes.

FIGURE 4.1. Mode P mark distribution.

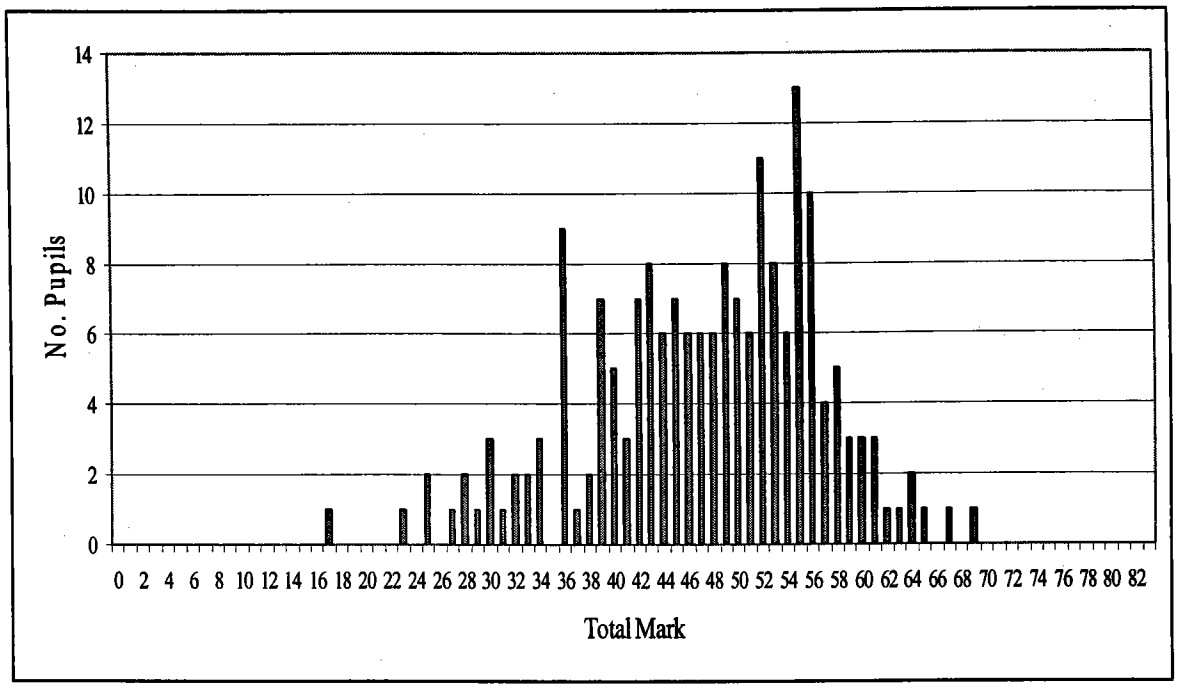
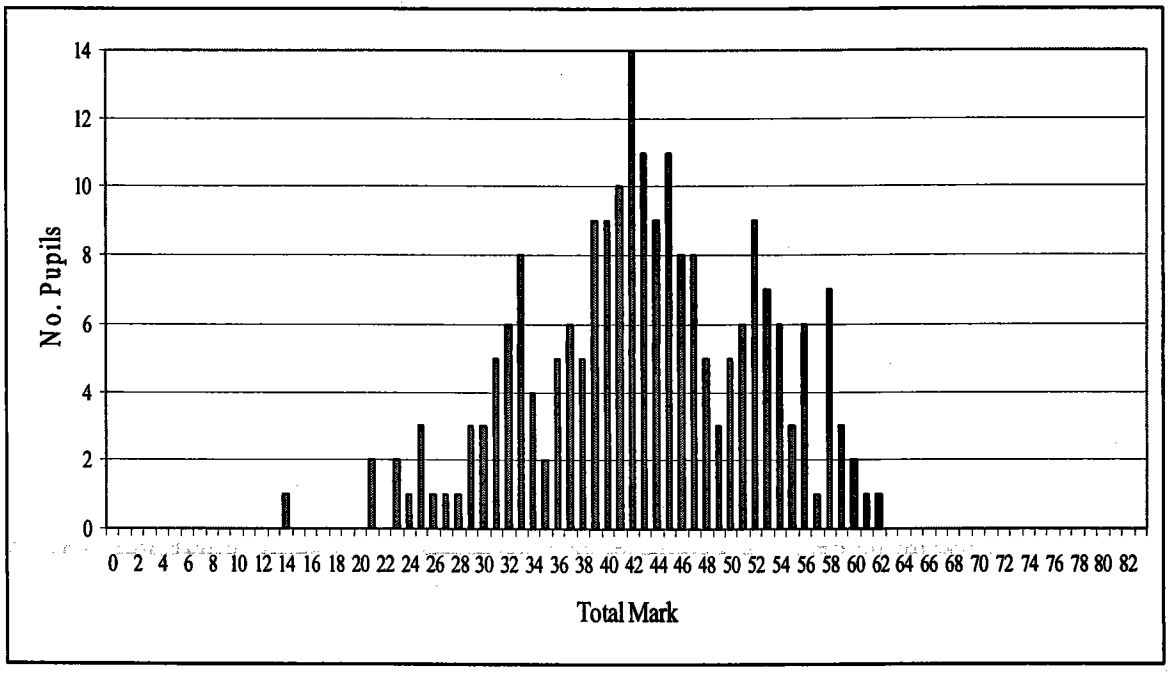


FIGURE 4.2. Mode Q mark distribution.



4.3.2 Multiple choice performance

4.3.2.1 The selection components (Q1-10 and 11a-15a)

The facility of a multiple choice item is defined, statistically, in terms of the percentage of students who select the correct response. As can be seen from Table 4.5, for items 1 to 15a, the facility ranged from 28% to 93%. Between these two extremes, facilities ranged from 39% to 76%.[4] This pattern was reasonable. However, as only four of the fifteen facility indices were below 50%, it would be fair to conclude that the multiple choice questions tended to be more easy than hard.

When evaluating the success of a multiple choice item it is important to consider not only the number of 'correct' selections, but also the number of students who chose each of the 'distracters'. As a general rule, when a distracter is selected by only a very few students then its utility is minimal (being too obviously a distracter). The greater the number of clear distracters an item has, the easier it will be to guess the best answer by chance alone.

On this criterion, it is clear that a number of the items suffered from multiple 'obvious' distracters (e.g., particularly items 1, 5, 6 and 15a). Because it is not always clear in advance exactly where such problems will occur it is good practice for multiple choice items to be pre-tested. It may then be possible for poor distracters to be replaced, or revised, to improve the functioning of a question.

4.3.2.2 The justification components (Q11b-15b)

The justification components of the multiple choice items functioned similarly to the selection components, with facilities ranging from 44% to 79% and with the majority being above 50%. For two of the items (11b and 12b) more students achieved a mark for the justification component than achieved a mark for the selection component. For the remaining three items the reverse was true.

The concordance between selection and justification components was also examined for each item. This revealed that the vast majority of students performed similarly on the selection and justification component (i.e., received a mark for both or received a mark for neither). The percentage of students who did not follow this pattern ranged from 8%

---

[4] Technically, the facility for 15a was the sum of percentage responses to both B and C (as there were both deemed to be 'best' answers) which gave a facility of 80%.

(15b) to 34% (11b). This suggests that both components were assessing a broadly similar competence.

TABLE 4.5. Section A & B performance data (Mode P and Q combined).[5]

Item	Percentage Students Selecting Each Response					
	BLANK	A	B	C	D	E
1	0	47	2	42	8	2
2	1	6	6	9	66	13
3	0	1	93	1	0	5
4	0	1	16	5	71	7
5	0	6	3	14	76	1
6	0	16	2	53	28	1
7	0	10	4	61	16	10
8	0	10	15	13	14	48
9	0	8	61	14	7	11
10	0	7	7	22	6	58
11a	1	21	20	39	2	18
12a	0	6	5	3	10	76
13a	1	6	56	21	2	14
14a	1	12	5	66	15	1
15a	0	20	56	24	0	0
16-1	0	26	45	23	5	1
16-2	0	4	18	38	32	8
16-3	0	37	35	17	9	3
16-4	0	9	22	32	26	11
17-1a	0	3	9	27	30	32
17-2a	0	13	46	27	11	3
17-3a	0	5	23	38	22	13
17-4a	1	19	32	16	21	12
17-5a	0	14	32	19	20	16

Item	% Students With Correct Justn.	
	0	1
11b	39	61
12b	21	79
13b	56	44
14b	41	59
15b	27	73

Item	% Students With Correct Justn.	
	0	1
17-1b	44	56
17-2b	57	43
17-3b	67	33
17-4b	63	37
17-5b	78	22

4.3.3 Multiple rating performance

4.3.3.1 The selection components (Q16 and Q17-1a to Q17-5a)

Multiple rating items differ from multiple choice questions through crediting up to three selections (one with 2 marks and up to two with 1 mark). They also differ in not having traditional distracters (as the responses represent a position along a continuous rating scale rather than independent alternatives). As such, the evaluation of multiple rating item functioning differs from that of multiple choice items. In particular, because the

[5] Note that the lighter shading indicates a response credited 1 mark (the 'correct' answer for questions 1 to 15a) and the darker shading indicates a response credited with 2 marks.

response alternatives are not independent, a traditional analysis of distracter functioning is inappropriate.

In terms of facility, the percentages of students that achieved one or two marks on each item are presented in Table 4.5. This table reveals that between 13% (17-3a) and 46% (17-2a) of students selected responses deemed by the expert panel to be 'best'.[6] The percentage of students that received no marks at all for an item ranged from 11% (16-3) to 66% (17-3a) with a mean of 40%.

Generally speaking, it would not be appropriate to claim that a single item was flawed just because a large number of students failed to chose the correct response or even tended to opt for a particular distracter. It may simply have been that they 'fell into the trap' laid by an item that was intended to assess whether students shared a common misconception. Indeed, the same might be said of multiple rating items; moreover, certain items were specifically designed with common errors of critical thinking in mind.

However, particularly bearing in mind the problems that the experts had encountered in agreeing 'best' ratings, it is worth considering patterns of error closely. Two items are most worthy of attention in this respect: 17-3a and 17-5a. In both of these cases the largest number of students chose a rating that was deemed by the expert panel to be worthy of no marks. For 17-3a, this rating was C, two rating grades away from the 'best' (E). For 17-5a, the rating was B, three rating grades away from the 'best' (E). Recall the problems that were encountered with 17-5a during the mark scheme development meeting (see 3.3.1.1) where the expert panel was split between rewarding A or E with 2 marks.

Responses to question 17-4a were also somewhat odd, with an almost bimodal distribution either side of the 'best' rating; thus, while only 16% of students opted for the 'best' rating (C), 32% and 21% chose the 'second best' ratings (B and D, respectively).

4.3.3.2 The justification components (Q17-1b to Q17-5b)

The students were less likely to be awarded marks for the justification components of Question 17 than for the components of Questions 11b to 15b. The percentage of students that received a justification mark ranged from 22% (17-5b) to 56% (17-1b).

---

[6] As, for item 16-2, both C and D were deemed to be 'best', the percentage achieving 2 marks on this question was actually 70%.

The concordance between selection and justification components was also examined for each item. The percentage of students that performed similarly on the selection and justification component (i.e., received at least one mark for both or received a mark for neither) was considerably less than observed for the multiple choice questions. The percentage of students who did not perform similarly ranged from 25% (17-2b & 17-5b) to 46% (17-4b). This suggests either that the selection and justification components were assessing somewhat dissimilar competencies or that the reliability of marking for the justification components was questionable (or both).

4.3.4 Essay performance (Mode P)

For the Mode P examination, students were required to write one essay which was marked out of 20 marks. These marks were divided between three sub-components of critical writing: selection; synthesis and structure. Table 4.6 represents the distribution of marks awarded for the Mode P essay.

TABLE 4.6. Section C performance data (Mode P).

Item	Percentage Students Awarded Each Mark									
	0	1	2	3	4	5	6	7	8	9
18a sel	2	3	10	23	24	39	1			
18b syn	3	1	5	10	12	18	27	15	9	0
18c str	2	5	13	31	41	8				

From Table 4.6, it is clear that the markers used a good range of the marks available to them. Least well used were the top marks for selection (6) and synthesis (9). The mean marks awarded for selection, synthesis and structure were 3.8, 5.2 and 3.3, respectively.

The marks available corresponded to notional performance bands. Students were most likely to be placed in the highest band for structure, then for selection and least likely for synthesis (and vice versa for likelihood of being placed in the lowest band). Just over half of all students were placed in the second band for synthesis (57%), while just under half were placed in the second band for selection (46%) and structure (45%).

4.3.5 Short report performance (Mode Q)

For the Mode Q examination, students were required to write two short reports which were each marked out of 20 marks. Exactly as before, these marks were divided between three sub-components of critical writing: selection; synthesis and structure. Table 4.7 represents the distribution of marks awarded for the Mode Q short reports.

TABLE 4.7. Section C performance data (Mode Q).

Item	Percentage Students Awarded Each Mark									
	0	1	2	3	4	5	6	7	8	9
18a sel	1	5	18	28	29	19	0			
18b syn	1	2	10	16	18	18	21	11	2	0
18c str	1	7	19	35	33	5				
19a sel	5	11	22	29	22	11	0			
19b syn	6	3	15	14	21	20	11	10	1	0
19c str	5	12	21	38	22	2				

Very similar patterns of results were observed for both Mode Q short reports as for the Mode Q essay. This included the markers using a good range of marks, although not using the top marks for selection or synthesis. The mean marks awarded for Q18 (Bermuda Triangle) were 3.8, 4.6 and 3.1, for selection, synthesis and structure, respectively. These were quite similar to the marks awarded for the Mode P essay. For Q19 (Electoral Reform) the mean marks were somewhat lower: 2.9, 4.0 and 2.7, respectively. This pattern was consistent with the co-ordination marking.

When the marks awarded were explored by band, very similar patterns were observed as for the Mode P essay. For both short reports, just over half of all students were assigned to the second band, regardless of sub-component being assessed. However, there was a greater likelihood of being assigned to the top band for structure than for selection and the likelihood was least for synthesis (once again, these patterns were reversed for placement in the lowest band).

4.4 Specific measurement analyses

In addition to the general item and score distribution analyses, more specific evidence was sought to determine whether marks from the trial examination would have supported reliable and valid inferences concerning critical thinking competence. Roughly speaking, the intention of these analyses was to establish the plausibility of the proposal that:

1. the marks awarded accurately represented a distinct underlying competence; and
2. the underlying competence represented by the marks was critical thinking.

A range of sources of evidence was considered in order to evaluate the plausibility of these claims. The results of these analyses are presented below under two broad headings: reliability and validity.

4.4.1 Reliability

4.4.1.1 Internal consistency

The range of marks attained on Mode P and Mode Q, respectively, was 52 (from 17 to 69) and 48 (from 14 to 62). There was also a reasonable dispersion of marks, with both modes having a standard deviation of 9 marks.

However, an important question remains: were those students who attained the higher marks generally more competent than those who attained the lower marks? A possible alternative might be that those who attained the higher marks were generally no more nor less competent than those who attained the lower marks and that the differences in marks achieved were largely due to 'chance'. Another alternative might be that the examination assessed a large number of relatively distinct competencies which did not ultimately cohere as a discrete construct.

Of course, implicit in these alternatives, is the assumption that critical thinking skill can genuinely be understood in terms of a single continuum of competence along which students can meaningfully be ranked. This assumption is, in itself, open to question. Indeed, there might conceivably be theoretical grounds for assuming that certain sub-skills of critical thinking are relatively independent of other sub-skills.

The assumption made in developing the Critical Thinking trial examination – an assumption that should equally apply across the range of public examinations – was that, while the specified component skills might be somewhat independent, there would still be sufficient dependence for the examination to cohere. That is, even if different items or sections assessed different aspects of critical thinking, reasonable performance correlations would still be observed between them. This would ensure that the aggregated mark total would have clear meaning as an index of critical thinking competence.

4.4.1.1.1 Reliability coefficients

To explore the reliability of marks arising from Sections A and B, the reliability coefficient, Cronbach's alpha, was computed. Coefficient alpha is a measure of inter-item consistency, i.e., the extent to which students perform consistently well (or poorly) across items. If each item of the examination had been designed to assess precisely the same skill then we would expect coefficient alpha to be high (close to +1). Although it was assumed that Section C might assess different skills from Sections A and B, it was assumed that the skills assessed within the latter would be more uniform.

From Table 4.8, it can be seen that the coefficients were generally low. They were highest for Section A alone and it is clear that the lack of inter-item consistency for Section B also brought the coefficients for A & B combined down. Even for Section A, though, the coefficients were not high. A further insight into these figures can be gained from Table 4.9, which displays the correlation coefficients between pupils' marks on each item and their total marks for the section. Again, these Item-Total correlation coefficients are very low.

For the results of the Critical Thinking AEA to be meaningful we would have hoped for somewhat higher coefficients, at least within sections. The coefficients for Section B are particularly worrying.[7]

TABLE 4.8. Coefficient alpha for Sections A and B.

Section	Mode	Justification inclusion	Coefficient alpha
A & B combined	P	With (34 items)	0.44
		Without (24 items)	0.10
	Q	With (34 items)	0.58
		Without (24 items)	0.38
A only	P	With (20 items)	0.53
		Without (15 items)	0.40
	Q	With (20 items)	0.63
		Without (15 items)	0.52
B only	P	With (14 items)	0.16
		Without (9 items)	-0.26
	Q	With (14 items)	0.26
		Without (9 items)	0.01

Also apparent from Table 4.8 was the fact that the coefficients were higher when justification components were included in the analyses. This should not be over-interpreted as, in many ways, the finding is not surprising. First, once the justification component is taken out, the section total is reduced which would thereby attenuate any subsequent correlation. Second, the justification components are not independent questions (merely an opportunity to obtain a second mark for an answer already provided)

[7] Had this been an A-level trial, yet piloted only on the 'top 10%' of A-level students, we would have been quick to point out that the low correlation coefficients were probably an artefact of trialling on a restricted ability range of students. However, as the AEA is intended specifically to discriminate between this target population, the problem of low correlation coefficients remains significant.

which should inevitably increase the apparent inter-item consistency. What this finding should *not* be taken to mean is that the examination is technically better with justification components included than without – that would be a spurious interpretation of the evidence caused by a confusion between consistency and reliability. No evidence has been provided in relation to the technical effectiveness, *per se*, of including justification components. Finally, a word should be said regarding the negative coefficient observed for Section B. In theory, this is not possible and it is likely to have arisen here as an artefact of range restriction.

TABLE 4.9. Section A Item-Total correlation coefficients (justification included).

Item	Mode P	Mode Q
1	0.11	0.03
2	0.24	0.31
3	0.05	0.31
4	0.21	0.13
5	0.06	0.18
6	0.13	0.22
7	0.27	0.25
8	0.14	0.29
9	-0.02	0.29
10	0.01	0.09
11a	0.29	0.33
11b	0.21	0.21
12a	0.31	0.28
12b	0.32	0.29
13a	0.19	0.09
13b	0.15	0.22
14a	0.14	0.18
14b	0.11	0.25
15a	0.19	0.26
15b	0.24	0.27

4.4.1.1.2 Correlation coefficients

Very similar information was gained by considering the correlation coefficients between sections of the trial examination.

The correlation coefficients presented in Table 4.10 and 4.11 are very low. Once again, this is disappointing. We would have hoped for higher figures, even assuming that the component skills assessed by the different sections were not exactly similar. Even the

correlation between the two short reports of Mode Q was no higher than 0.50 – and these were intended to be parallel questions assessed using the same criteria.

TABLE 4.10. Inter-section correlation coefficients (n.s. = not significant at P < 0.05).

		Section A	Section B	Section C
Mode P	Section A	-	0.17	0.19
	Section B	-	-	n.s.
	Section C	-	-	-
Mode Q	Section A	-	0.29	0.24
	Section B	-	-	n.s.
	Section C	-	-	-

TABLE 4.11. Inter-section correlation coefficients (n.s. = not significant at P < 0.05).

		A (1-10)	A (11-15)	B (16)	B (17)	C (18)	C (19)
Mode P	A (1-10)	-	0.16	n.s.	0.16	n.s.	-
	A (11-15)	-	-	n.s.	n.s.	0.16	-
	B (16)	-	-	-	n.s.	n.s.	-
	B (17)	-	-	-	-	n.s.	-
	C (18)	-	-	-	-	-	-
ModeQ	A (1-10)	-	0.26	0.26	n.s.	0.20	0.14
	A (11-15)	-	-		0.19	0.20	0.14
	B (16)	-	-	-	n.s.	0.18	n.s.
	B (17)	-	-	-	-	n.s.	n.s.
	C (18)	-	-	-	-	-	0.50

4.4.1.2 Marking reliability

The aim of the re-marking study was to compare the marking of two pairs of markers on Mode Q scripts and two pairs of markers on Mode P scripts, as presented in Table 4.12. Marking reliability was investigated both at the script total level and at the level of individual items, for items that required judgement (i.e., the justification components of Sections A and B and the essay/short reports of Section C).

The marking reliability analyses were computed separately for each of the four script samples. While this restricted each sample size to 25, which is not a large sample, it was considered that this would still provide reasonable evidence and that this approach would lead to far more meaningful conclusions.

TABLE 4.12. Design of marking reliability study.

Mode	No. scripts	Original marker (OM)	Re-marker (RM)
P	25	Marker 1 (M1)	Marker 4 (M4)
P	25	Marker 2 (M2)	Marker 3 (M3)
Q	25	Marker 3 (M3)	Marker 1 (M1)
Q	25	Marker 4 (M4)	Marker 2 (M2)

The script level analyses of marking reliability are presented below in Table 4.13.

TABLE 4.13. Script total comparisons for the four samples.

	No. scripts	Mode	Correlation coefficient		T-test		
			coefficient	sig.	mean diff.	t	2-tail sig.
M1 versus M4	25	P	0.89	0.000	2.00	3.074	0.005
M2 versus M3	25	P	0.83	0.000	4.32	6.727	0.000
M3 versus M1	25	Q	0.86	0.000	6.48	6.986	0.000
M4 versus M2	25	Q	0.88	0.000	5.44	5.809	0.000

The coefficients of correlation basically represent the extent of agreement in ranking between marks awarded by each marker. As is apparent from the four coefficients in Table 4.13, there was a promising level of agreement – bearing in mind the nature of the subject, the format of the questions and the fact that this was a trial examination. By way of comparison, a recent study of marking reliability for GCSE English revealed script total correlation coefficients that varied between 0.87 and 0.93 (Newton, 1996).

Yet inspection of the mean differences between markers is somewhat less encouraging. Even for the most consistent markers, there was an average disagreement of two marks.[8] For each of the marker pairs the mark differences were significant. The level of these

[8] Note that this was based on a straightforward average; we would, therefore, expect the mean of absolute mark differences to be higher still.

differences would undoubtedly impact upon the grades awarded to a considerable number of students.[9]

The item level analyses of marking reliability are presented in Appendix 4.4. For each of the ten justification questions, reliability data were presented in terms of: cross-tabulations; the percentage of item responses for which two markers agreed and disagreed; and a Kappa statistic – a formal coefficient of concordance.

Across each of the four marker pairs, the reliability of marking of justification components was promising. Although there was no question for which any of the four pairs agreed for all 25 item responses, the average mark agreement was 79% (i.e., agreement on 20 of the 25 item responses – where agreement means both markers awarding a 0 or both awarding a 1).

The Kappa statistic took the analysis of agreement a step further by considering the observed level of agreement in relation to the agreement that might be expected by chance. Clearly, if markers had been awarding marks at random then there would have been a 50% chance of agreement by chance alone (i.e., 00 and 11 versus 10 and 01). Most notable from the statistical analysis was the fact that question 17-5b caused serious problems across all four marking pairs and question 17-3b caused serious problems for three of the four. (It will be recalled from 4.3.3.1 that these two items were noteworthy due to the largest number of candidates selecting a rating grade that was not even worthy of a single mark.) Significant marking disagreements were only encountered for four other items and each of these items was only problematic for a single marking pair (although 17-1b tended towards being problematic for all four pairs).

Generally speaking, the marking of justification components for the multiple choice items was more reliable than the marking of justification components for the multiple rating items. Interestingly, the marking problems associated with 17-3b and 17-5b may also hint at a further conclusion: that the use of justification – to enable students to achieve marks for a good understanding that was not rewarded through their rating grade selection – may, at least to some extent, have failed. That is, it may be in precisely such circumstances that markers have the most problems in evaluating the worth of a

---

[9] Note that the mean mark differences for the Mode Q marking pairs were higher than those for the Mode P marking pairs. This is partly an artefact of computing the Mode P analyses from a total of 63 rather than 83 (i.e., not scaling the Mode P Section C marks by a factor of 2 for these analyses).

justification – to the extent that marking unreliability may undermine any benefit that might accrue from allowing students to justify their selections.

Turning to the marking of the essay questions (18 and, for Mode Q, 19), the tables in Appendix 4.4 reveal that the between-marker correlation coefficients were promising – once again, bearing in mind the nature of the subject and the fact that this was a trial examination. These ranged between 0.62 (M2 vs M3 on the Mode P essay) and 0.79 (M4 vs M2 on the Mode Q Electoral Reform short report). By way of comparison, a recent study of marking reliability for GCSE English revealed correlation coefficients for writing components that varied between 0.74 and 0.92 (Newton, 1996).

Somewhat more worrying, though, were the mean mark differences between markers. Although M1 and M4 were generally marking at a similar level, M2 and M3 tended to diverge both from M1 and M4 and from each other. Roughly speaking, M1, M2 and M4 were marking at a similar level for the Mode P essay, while M3 was tending to be consistently harsh. On the Mode Q short reports, M1 and M4 were once again marking at a similar level, while M3 was tending to be consistently harsh and M2 consistently lenient. In fact, similar patterns of response had been noted during the initial co-ordination, which means that steps taken to bring markers more into line had failed. Thus, if the essays or short reports are to carry forward into future examinations, it will be important to focus particularly on ways of enhancing the absolute agreement between markers (as evidence of relative agreement – agreement in ranking – was more encouraging).

4.4.1.3 Bias

A final issue linked to reliability is whether particular questions were responded to differently by different sub-groups of students. This involved computing 'differential item functioning' (*dif*) statistics for each item, broken down by different sub-groups of students. These analyses are generally considered to indicate items that may be inappropriately biased in favour of a particular sub-group (i.e., the sub-group may be performing differently on the item for reasons not to do with the construct we intend to assess). Importantly, these statistics do not indicate whether or not the test as a whole is biased; they simply consider whether particular items function in a different way from the rest. For example, if a comparison of male and female students revealed that females consistently out-performed males on all items except one – where the reverse was true – the *dif* statistic would highlight this item as potentially biased.

Differential item functioning analyses were conducted for items in Section A and B (combined across modes) in relation to three sub-groupings:

1. male vs female;

2. Critical Thinking AS vs no Critical Thinking AS;

3. Sciences vs Humanities vs Languages vs No specialism.

Only four items proved significant beyond $P < 0.01$. Items 6 and 7 favoured both male over female and Sciences over Humanities/No specialism. Items 14a and 14b favoured both girls over boys and No Critical Thinking AS over Critical Thinking AS.[10]

Items 6 and 7 were, perhaps, the most 'traditional': item 6 was, essentially, a formal logic problem; while item 7 was, essentially, a spatial mathematics problem. Item 14 required the identification of the implication of a passage.

4.4.2 Validity

The bottom line in any examination development exercise is whether the resultant scores accurately represent the extent to which individual students possess the knowledge, skill and understanding embraced by the to-be-assessed construct. Unfortunately, this statement does not lead to a straightforward evaluative question, let alone to a straightforward evaluative answer. Instead, establishing the extent to which scores from an examination may support valid inferences concerning students' capacities involves the construction of an argument based upon multiple sources of evidence.

Validity is traditionally explored from a range of different perspectives – using many sources of evidence – and we shall consider a number of these perspectives in the following sections. Ultimately, though, these different perspectives are unified through the notion of construct validity which, essentially, concerns the meaning of the examination scores (and whether they mean what we think they ought to mean).

---

[10] It should also be noted that there may have been some interaction between the sub-group *dif* statistics. For instance – in relation to 14 – the proportion of females in the No CT AS sub-group was considerably higher than the proportion of females in the CT AS sub-group and this could have contributed to the significant *dif* for this breakdown (in favour the No CT AS group).

4.4.2.1 Content evidence

Content evidence is sought to determine: whether the content of the examination is *relevant* to the content of the domain it is intended to assess; and whether the items of the examination sample the domain in a manner that *represents* it appropriately.

In fact, these were precisely the concerns of Section 2 of the present report. Unfortunately, the evidence for content relevance and representativeness was not overwhelming. It was concluded that the construct specification should still be regarded as essentially Work in Progress; moreover, the necessity of developing the specification in parallel with the test items limited the extent to which formal sampling techniques could be employed.

Once again, the importance of a full construct specification needs to be emphasised. Many traditional examinations rely heavily upon respected content specifications for their credibility and construct specifications are often (although inappropriately) overlooked. This is not possible for the AEA in Critical Thinking, however, as the award is to have no content specification. Consequently, not only must there be a well developed construct specification, there must also be strong evidence that there is actually a coherent construct of critical thinking to be assessed. Although construct validity ought to be at the centre of any validation exercise (Messick, 1989) this is clearly unavoidable in the case of critical thinking. It was noted in Section 2.1 that critical thinking has often been discussed as an educational construct rather than as a psychological construct, per se. However, if critical thinking is to form the basis of a high stakes assessment then it is important that the assessed construct should have psychological meaning and not simply educational import.

In this respect, the future development of an AEA in Critical Thinking would seem to recommend:

1. a thorough literature review relating to evidence of the construct validity of published critical thinking tests and examinations;[11]

---

[11] In fact, the evidence provided for published tests of critical thinking is not always overwhelming. For example, in the promotional literature for the ICAT Critical Thinking Essay Examination, it is proposed that the need for evidence on reliability and validity is not pressing (and it is not provided) because the examination is designed to be used for internal assessment. Instead, the publishers stress its 'face validity' (i.e., that it looks like a good test of an important skill).

2. further research into the psychometric properties of the AS in Critical Thinking.

4.4.2.2 Criterion evidence

Criterion evidence is sought to determine the extent to which scores on an examination are effective at predicting the criterion that the examination is intended to predict. In relation to the AEA in Critical Thinking, this would seem to be success in Higher Education. Clearly, no such evidence could be collected as part of the trial evaluation. However, serious consideration should be given to collecting this evidence once the examination goes live, if it is to do so.

4.4.2.3 Construct evidence

Construct evidence is sought to determine precisely what it is that the examination scores mean (and whether they mean what they were intended to). As already noted, construct validity subsumes all other perspectives upon validity. At the heart of construct validation is a theoretical rationale – a psychological (or curriculum) model of the construct that supports predictions concerning how the test items should function and how they should not function.

4.4.2.3.1 Convergent and discriminant evidence

One of the main sources of evidence pertinent to the meaning of scores from an examination comes from relating them to scores from other reliable and valid measures – both similar and dissimilar. When compared with scores from reliable and valid measures of a dissimilar construct we would predict low correlation coefficients (discriminant evidence); and when compared with scores from reliable and valid measures of a similar construct we would predict high correlation coefficients (convergent evidence).

During the present trial, information on students' prior GCSE scores and predicted A-level scores were obtained. The GCSE scores were averaged, and the predicted A-level scores totalled, to provide rough indices of general academic achievement. They were then correlated against performance on the trial examination.

It would seem reasonable to assume that critical thinking skill ought to be related to general academic achievement. After all, we would expect the skills of critical thinking to contribute directly to academic achievement. Yet it is not clear that critical thinking skill will necessarily have a large impact upon academic achievement as success at GCSE and A-level will also be dependent upon many other factors that are not part of the critical

thinking construct (for example, diligence and effort throughout an educational course as well as a capacity for learning).

TABLE 4.14. The relationship between performance on the trial examination and measures of general academic achievement.

		Section A Total	Section B Total	Section C Total	Section A&B Total	Exam Total
P	A-level Total	0.31	n.s.	n.s.	0.18	n.s
	GCSE Average	0.37	n.s.	n.s.	0.34	0.23
Q	A-level Total	0.26	0.18	0.16	0.28	0.28
	GCSE Average	0.19	n.s.	0.17	0.11	0.19

As such, we might predict correlation coefficients that were moderate, but not necessarily high. In fact, as Table 4.14 illustrates, the coefficients were low and, not infrequently, insignificant. The highest relationships were observed in relation to the multiple choice sections of the examination.

Unfortunately, by its very nature, the trial faced the technical problem of restriction of range in the GCSE and A-level scores, as all students were bunched at the higher grades. This inevitably attenuated the correlation coefficients, meaning that the correlation evidence was not as strong as it might have been.

Yet the coefficients of Table 4.14 may still convey some useful information. In particular, the fact that the coefficients for Sections B and C were lower than those for Section A might be taken as evidence that the multiple choice items functioned better than the other item types as predictors of academic performance.

4.4.2.3.2 Inter-item consistency

Inter-item consistency is generally discussed from the perspective of examination reliability. Yet it is also a crucial feature of construct validity. If there is no inter-item consistency then the examination is simply not assessing a coherent construct. Of course, it could still be that there is a coherent construct to assess even though the examination is failing to assess it. But it could equally be that the construct does not cohere; that is, the putative construct may simply not manifest itself as a psychological characteristic, or trait, in terms of which students differ in consistent ways.

From the inter-item and inter-section correlation coefficients of Section 4.4.1.1, the evidence supporting an underlying construct of critical thinking was not strong.

4.4.2.3.3 Performance of specified sub-groups

Another way of exploring the meaning of the assessed construct is to consider predictions concerning the likely performance of specified sub-groups. This involves an *a priori* evaluation of the construct in relation to how its manifestation might be expected to vary between sub-groups. Table 4.15 illustrates the sub-groups for which performance differences were investigated (the same as were investigated for item bias).

TABLE 4.15. Mean score/mark differences by sub-groups.

		Gender		A-level Subject Specialism				Critical Thinking AS	
		Male	Female	None	Sci.	Hum.	Lang.	No CT AS	CT AS
P	Average GCSE score	7.23	7.38	7.33	7.35	7.09	7.43	7.28	7.35
	Significance	0.052 (n.s.)		0.025				0.547 (n.s.)	
	Average CT exam mark	47.98	46.40	49.36	46.52	48.05	47.17	46.85	50.71
	Significance	0.262 (n.s.)		0.518 (n.s.)				0.060 (n.s.)	
Q	Average GCSE score	7.03	7.27	6.88	7.31	7.00	7.11	7.26	6.31
	Significance	0.068 (n.s.)		0.068 (n.s.)				0.000	
	Average CT exam mark	43.47	42.70	40.51	44.42	42.13	42.64	43.90	37.12
	Significance	0.540 (n.s.)		0.133 (n.s.)				0.000	

It was concluded that there was no reason to expect critical thinking skill to differ by gender or by subject specialism. However, the proposal was less clear-cut with respect to possession of an AS in Critical Thinking. On the one hand, the AEA was independent from the AS and was not designed to the same construct specification. On the other hand, the AEA was similar to the AS and the AS is generally associated with a formal course – a course that might well transfer to gains in the AEA examination. As such, it was predicted that, all things being equal, there would be some advantage of studying for the AS.

In fact, as shown in Table 4.15, all things were not equal. For one of the critical thinking sub-groups there was a difference in mean GCSE score which suggests that the groups were not well matched in terms of prior achievement. For this reason – and because there were only a few students, clustered within particular schools, in the CT AS group – the

results of the comparison of AEA performance between critical thinking sub-groups lack validity.

However, in the remaining sub-group comparisons, the predictions were supported. No significant differences in examination performance were observed either between male and female or between pupils with different subject specialisms.

4.4.2.4 Consequence evidence

In recent years, educational measurement professionals have realised that validity arguments need to consider the potential consequences of examination use, particularly any negative consequences that might not have been anticipated at the outset. This is explicitly not to propose that if the deployment of a new examination leads to negative consequences then it is necessarily invalid. However, wherever it is feasible that negative consequences might arise due to unanticipated construct validity problems, it is important to highlight these potential effects in advance and to watch out for them.

One possible consequence is that students, teachers or Higher Education selectors may reject the AEA in Critical Thinking – either because they do not consider it to be a fair test of critical thinking skill, or because they do not believe that it will provide useful information for university selection. While we have no evidence in relation to HE selectors, we do have evidence from the students and teachers that took part in the trial. As discussed in Section 4.2.2 and 4.2.3, both teachers and students generally agreed that the trial was a fair test of critical thinking skill. However, there was slightly less agreement that the examination would function well as a predictor of success at university (and there was less confidence amongst students than amongst teachers). Recall also that the schools participating in the trial were likely to have been more positively disposed to the AEA in Critical Thinking than other schools.

A second possible consequence is that students in well-resourced schools may receive considerable preparation for the AEA in Critical Thinking and, as a result, achieve higher marks. This is to highlight a potential source of confusion over the nature of the construct. We have assumed that, if critical thinking can be shown to cohere as an assessable construct, then it will be possible to offer it as an examination that does not necessitate formal instruction. However, this is not to say that performance on the examination will not be enhanced by formal instruction. If it was assumed that performance on the critical thinking examination could not be enhanced by formal instruction then evidence of enhanced performance from schools which did offer formal instruction would suggest problems of construct validity. Yet, as just explained, we have

made no such assumption and it remains to be seen the impact that instruction may actually have.

The final potential consequence identified by the present report is that it may not be possible to restrict entry to the target population. The AEA is presently being developed to discriminate between the 'most able' A-level students. This can be interpreted in a fairly straightforward manner for AEAs with associated A-levels. However, it is not clear how the 'most able' students will be selected for the AEA in Critical Thinking. If, for example, this criterion was interpreted as 'most ability in critical thinking' then how would students know their status in advance of the examination? Moreover, to the extent that there would be nothing to loose from entering the examination, how would it be possible to restrict the entry to the 'top 10%' (assuming that these students could actually be identified in advance)? The point is not that a large entry would necessarily be a bad thing; instead, what is to be cautioned against is the possibility that a far wider entry will lead to a dilution of standards due to the fact that so many would otherwise have to be failed. Alternatively, if the examination was adapted so as to be accessible to a wider range of students, then it would loose its power to discriminate between the most able.

4.5 Conclusions

From the evidence of the preceding sections, it should be clear that there was insufficient evidence to establish a strong validity argument in support of the trial examination. The possible reasons for this will now be considered. Four explanations stand out particularly. While each of these might explain the results of the trial independently, it is likely that the true explanation involves an interaction of problems from more than one, if not from all.

4.5.1 The specification may not have reflected a coherent psychological construct

According to this explanation, it is possible that:

1. the construct specification accurately defined a meaningful *educational* construct;
2. the items of the examination accurately assessed this construct;
3. there was insufficient relationship between performance on the various items – as the construct specification did not reflect a coherent *psychological* construct – and the aggregate marks were, therefore, rendered spurious.

As noted earlier, there is an important distinction between an educational construct and a psychological construct.[12] Thus, it is proposed that a construct can be educationally coherent while being psychologically incoherent. Educational coherence is a function of whether the full range of knowledge, skill and understanding associated with the construct has been embraced. Psychological coherence is a function of whether the separate components and sub-components of critical thinking tend to be correlated within individuals; that is, whether people who tend to be strong at certain aspects of critical thinking also tend to be strong in others (and vice versa). If there is no psychological coherence then we would expect different people to be competent in different aspects of critical thinking. There may, of course, be no principled reason to assume psychological coherence. However, if psychological coherence is limited, then mark aggregation becomes highly problematic and it becomes far more likely that any mark total differences will actually have been obtained by chance. In short, if the construct does not demonstrate at least a significant degree of psychological coherence, the critical thinking mark will loose its meaning.

Of course, it might also be possible that the specification did not even accurately define the educational construct. Indeed, it has been made clear within this report that the present construct specification should still be regarded as essentially Work in Progress.

4.5.2 The students may not have differed significantly in terms of critical thinking ability

According to a second explanation, it is possible that:

1. the construct specification accurately defined a construct that was coherent in both educational and psychological terms;

2. the items of the examination were well designed to assess this construct;

3. the target population was too homogeneous, in terms of critical thinking ability, for the examination to discriminate effectively between them.

The implication of this explanation is that – had we selected a much wider target population – we would have seen a much wider spread of marks; moreover, that this

---

[12] These terms are used for illustrating the argument and should not be assumed to reflect a distinction that all researchers would label in the same way.

spread would allow for coarser (yet valid) distinctions to be drawn between students in terms of their relative standing in relation to the construct. However, the students of the trial would still be bunched together (toward the top of the mark distribution range) and it would still not be possible to discriminate *between them* with confidence.

Although it is quite possible that there is some truth in this proposal, it is probably fair to conclude that it is unlikely to provide the sole explanation. The fact that students' marks were reasonably well spread out on the trial examination suggests that they were not simply too similar in terms of critical thinking ability. If they had been – and the examination had been a reliable and valid device for assessing a coherent psychological construct – then we would have expected the mark distributions to have been more tightly bunched.

4.5.3 The marking of the examination may not have been of a sufficiently high standard

A third explanation suggests the possibility that:

1. the construct specification accurately defined a construct that was coherent in both educational and psychological terms;
2. the items of the examination were well designed to assess this construct;
3. the target population was sufficiently diverse, in terms of critical thinking ability;
4. the markers were not capable of marking the questions reliably.

This explanation is, perhaps, the most simple to rule out as a single cause of the problems encountered. From the results presented earlier it is clear that there is still work to be done to enhance the reliability of marking. However, it is equally clear that the marking reliability statistics were promising. Moreover, it should be noted that problems were observed not simply for items that required a significant component of judgement, but also for items that required none. Items that required response selection were also associated with technical limitations (consider, for example, the moderate discrimination and internal consistency for Section A items).

4.5.4 The examination may simply not have been a good test of critical thinking

A final explanation suggests the possibility that:

1. the construct specification accurately defined a construct that was coherent in both educational and psychological terms;

2. the target population was sufficiently diverse, in terms of critical thinking ability;

3. the markers were capable of marking the questions reliably;

4. the items of the examination were not well designed to assess the construct.

The final explanation suggests that the project team simply failed to develop a sufficiently good examination (either for reasons beyond their control or for reasons within it). This is not a conclusion that the authors of the report would be happy reaching, as the Assessment and Measurement Department at the NFER is one of the most respected test developers in the UK, with significant experience of developing critical thinking tests. However, this possibility must be considered. Indeed, there were a number of factors (though largely beyond the project team's control) that may well have impacted upon the test development process with negative consequences. These included:

- a schedule for development that was extremely tight (particularly bearing in mind the requirement to work with item-writers who had only their 'spare time' to offer – a problem that was exacerbated by the shortage of item writers in the UK);

- the requirement to employ an item format that had never before been trialled in a large-scale public examination (the multiple rating item) – a requirement that will need to be reconsidered in the light of results from the trial;

- the relative inexperience of the item writing team in writing novel item types for a new, and highly specific, target population;

- the need to develop examination items in parallel with the construct specification.

For reasons such as these, it is likely that limitations in the examination development process contributed to some extent to problems that were apparent from the analysis of results. Such limitations will need to be resolved for future examinations.

Once again, though, it is important to stress that the development process was probably affected by each of the four explanations provided above. What is not clear is the extent to which the limitations of the final product can be attributed to their respective contributions. Importantly, it is unclear the extent to which the difficulties encountered

may be inherent in developing an AEA in relation to critical thinking. Further research and development will be necessary to shed light upon this issue.

4.6 Recommendations

The examination trial did not provide sufficient evidence to establish the feasibility of an AEA in Critical Thinking. Paramount amongst the problems encountered was a lack of evidence of construct validity. The inter-item consistency coefficients were low, as were the inter-section correlation coefficients. This interacted with a construct specification that is still to be regarded as Work in Progress. While the multiple choice items appeared to be promising, question marks remain over the utility of justification components. Likewise, while the evidence of marking reliability from the essays and short reports was promising, there is still scope for further research into the kind of multiple-mark constructed-response item that would be most suitable for the AEA. Most problematic was the lack of evidence in support of the reliability and validity of the multiple rating items: this assessment format may simply prove unworkable in the context of a large-scale high-stakes examination. A final threat to the feasibility of an AEA in Critical Thinking is the lack of people in the UK with expertise in critical thinking. This may well pose problems for the recruitment of a sufficiently skilled body of examiners.

The following recommendations are made:

1. The AEA in Critical Thinking should not be offered for examination in Summer 2002.

2. Further consultation and research should be conducted into the feasibility of an AEA in Critical Thinking. This should focus on both the conceptual and the practical problems that have been identified within the present report. As there are many significant issues to address and to resolve, this work should commence as soon as possible.

3. A particular emphasis should be placed upon the development of a construct specification that will cohere psychologically and will not simply possess educational significance. A thorough review of evidence for the construct validity of alternative published tests of critical thinking will be important in this respect.

4. Consideration should be given to whether further development of the multiple rating item format will be profitable.

5. Bearing in mind the complexity of the task – and the potential consequences of going live with a high-stakes examination that was sub-standard and that functioned

ineffectively – the development of an AEA in Critical Thinking should be treated as a medium-term, rather than a short-term, goal.

6. If – following further consultation and research – an AEA in Critical Thinking is deemed to be feasible, the examination should undergo a full re-trial.

Section 5 References

ANTAKI, C. (1994). *Explaining and Arguing: the Social Organisation of Accounts.* London: Sage.

ENNIS, R.H. (1962). 'A concept of critical thinking', *Harvard Educational Review*, **29**, 128–36.

ENNIS, R.H. (1985). 'A logic basis for measuring critical thinking skills', *Educational Leadership*, **43**, 2, 45–8.

ERWIN, T.D. (2000). *The NPEC Sourcebook on Assessment. Volume 1: Definitions and Assessment Methods for Critical Thinking, Problem Solving and Writing.* Washington, DC: US Government Printing Office.

FACIONE, P.A. (1990). *Critical Thinking: a Statement of Expert Consensus for Purposes of Educational Assessment and Instruction. Research Findings and Recommendations.* Newark, DE: American Philosophical Association.

FISHER, A. and SCRIVEN, M. (1997). *Critical Thinking: Its Definition and Assessment.* Norwich: Centre for Research in Critical Thinking.

JONES, E.A., DOUGHERTY, B.C., FANTASKE, P. and HOFFMAN, S. (1997). *Identifying College Graduates' Essential Skills in Reading and Problem-Solving: Perspectives of Faculty, Employers and Policymakers.* University Park, PA: US Department of Education.

JONES, E.A., HOFFMAN, S., MOORE, L.M., RATCLIFF, G., TIBBETTS, S. and CLICK, B.A. (1995). *National Assessment of College Student Learning: Identifying College Graduates' Essential Skills in Writing, Speech and Listening, and Critical Thinking* (NCES 95-001). Washington, DC: US Government Printing Office.

McPECK, J.E. (1981). *Critical Thinking and Education.* New York, NY: St. Martin's Press.

MESSICK, S. (1989). 'Validity.' In: LINN, R.L. (Ed) *Educational Measurement.* Third edn. New York, NY: MacMillan.

NEWTON, P.E. (1996). 'The reliability of marking of General Certificate of Secondary Education scripts: mathematics and English', *British Educational Research Journal*, **22**, 4, 405–20.

NORRIS, S.P. (1992). 'Testing for the disposition to think critically', *Informal Logic*, **14**, 2&3, 157–64.

NORRIS, S.P. and ENNIS, R.H. (1989). *Evaluating Critical Thinking*. New York, NY: Teachers College Press.

OXFORD CAMBRIDGE and RSA EXAMINATIONS (2000). *OCR Advanced Subsidiary GCE in Critical Thinking* (3821). Cambridge: OCR.

RICHARD, P. (1993). *Critical Thinking*. Santa Rosa, CA: Foundation for Critical Thinking.

SMITH, P. and WHETTON, C. (1992). *Critical Reasoning Tests*. Windsor: NFER-NELSON.

NFER HEAD OFFICE
National Foundation
for Educational Research
The Mere
Upton Park
Slough
Berks SL1 2DQ.
Tel: 01753 574123
Fax: 01753 691632
E-mail: enquiries@nfer.ac.uk
Web site:http: //www.nfer.ac.uk

NFER WELSH OFFICE
Chestnut House
Tawe Business Village
Phoenix Way
Enterprise Park
Swansea
SA7 9LA.
Tel: 01792 459800
Fax: 01792 797815
E-mail: scyanfer@abertawe.u-net.com

NFER NORTHERN OFFICE
Genesis 4
York Science Park
University Road
Heslington
York
YO10 5DG.
Tel: 01904 433435
Fax: 01904 433436
E-mail: jbh3@york.ac.uk