National Research and Development Centre
for adult literacy and numeracy

Research Report

Assessing adult literacy and numeracy: a review of assessment instruments

Greg Brooks, Karen Heath and Alison Pollard
University of Sheffield

February 2005

INSTITUTE OF
EDUCATION
UNIVERSITY OF LONDON

EUROPEAN UNION
European Social Fund

Published by the National Research and Development Centre
for Adult Literacy and Numeracy

This report may be downloaded as a PDF document from
the NRDC website at **www.nrdc.org.uk**

We welcome feedback on the content and accessibility
of this publication. This should be sent to:
Publications
NRDC
Institute of Education
20 Bedford Way
London WC1H 0AL.

Telephone: +44 (0)20 7612 6476
Fax: +44 (0)20 7612 6671
email: publications@nrdc.org.uk

ISBN 0 954727436

NRDC is a consortium of partners led by the Institute of Education, University of London (see back cover for a list of members) and is part of the Bedford Group for Lifecourse and Statistical Studies at the IoE. **www.ioe.ac.uk/bedfordgroup**

Design: Chapman Design Limited
Photography: Phillip Meech
Print: Bradell

Assessing adult literacy and numeracy: a review of assessment instruments

Greg Brooks, Karen Heath and Alison Pollard

CONTENTS

LIST OF TABLES

This report is funded by the Department for Education and Skills as part of ***Skills for Life***: the national strategy for improving adult literacy and numeracy skills. The views expressed are those of the author(s) and do not necessarily reflect those of the department.

Project team

Professor Greg Brooks, project director, University of Sheffield.
Karen Heath and Alison Pollard, research officers, University of Sheffield.
Jacquie Gillott, project secretary, University of Sheffield.

Authorship

Karen Heath was principally responsible for the development of the analytic framework, trawled the relevant section of the Institute of Education, University of London's library, including the Basic Skills Agency archive within it, and wrote appendices B and E, section A.1, the notes on the analytic framework in appendix C, all but one of the reviews of literacy instruments, and the first draft of the main text. She also made the key presentation on the review at a seminar at the Institute of Education, University of London, in October 2002, and in Autumn 2004 supplied numerous corrections of detail for the final version. Alison Pollard contributed to the development of the analytic framework and wrote all the reviews of numeracy instruments except the instrument adapted from the ***Skills for Life*** survey. Greg Brooks directed the project, wrote section A.2, the reviews of the numeracy instrument adapted from the ***Skills for Life*** survey and of the literacy assessment instrument developed for NRDC by NFER, and the final draft of the main text, and desk-edited the entire report.

Independent peer review

The report was read and peer-reviewed by Andrew Barbour, Tim Deignan, Pat Dreyer, Marjorie Hallsworth and Isabella Jobson.

Authors' acknowledgments

We wish to express our gratitude to:

- the ***Skills for Life*** Strategy Unit (formerly known as the Adult Basic Skills Strategy Unit) within the Department for Education and Skills, for funding the project;
- all those individuals and organisations who provided instruments for review;
- Felicity Rees, who began the collection of numeracy instruments and contributed to the initial discussions on the analytic framework;
- our peer-reviewers; and
- Gaye Houghton, who drew our attention to Charnley and Jones (1979), Good and Holmes (1978) and the *Progress Profile*, engaged in a lively email discussion about the ideas in the review, and carried out a thorough first edit of the main text and appendix D during August–September 2004.

Summary

The profession needs valid, reliable and manageable instruments for assessing adult literacy and numeracy, and the National Research and Development Centre for Adult Literacy and Numeracy (NRDC) in particular needs such instruments for its own research programme. This report provides a review of existing assessment instruments.

There had been no previous thorough review of adult literacy and numeracy assessment instruments used in Britain. A review was particularly opportune in 2002, when this project began, because several new instruments had appeared recently (the national tests of key/basic skills at Levels 1 and 2, the Basic Skills Agency's *Initial Assessment* pack (2nd edition), Cambridge Training and Development's *Target Skills*, the tests for the ***Skills for Life*** survey of adult basic skills needs being conducted in 2002/03).

A total of 15 quantitative, summative instruments used to assess adult literacy and/or numeracy in Britain in the period 1991–2002 were identified, obtained and analysed. The analysis was carried out against a checklist and framework derived from theory, previous analyses, and the research team's experience.

- The major criteria for useful instruments were that they should be secure (unpublished, or not readily available), be aligned to the new QCA National Standards, and (for use in research projects), have parallel forms.
- No wholly suitable instruments meeting these criteria were found.
- For the 2004 sweep of the British Cohort Study 1970 it was recommended that the instruments used in the early 1990s be used with some modification.
- For NRDC's research projects it was recommended that new literacy and numeracy instruments be commissioned.

The report also provides a brief history of quantitative and qualitative assessment instruments used in Britain in the period 1972–2004, a brief review of some United States (US) instruments, and criteria that should be met by good tests.

During 2003 a new literacy assessment instrument meeting all NRDC's requirements was developed for NRDC by the National Foundation for Educational Research (NFER), and a review of this is also included. However, for NRDC's numeracy projects a less than fully satisfactory adaptation of items from the 2002/03 ***Skills for Life*** survey of needs was developed in 2003. A review of this is included, with a recommendation that a better instrument be developed.

1. Context and aims

1.1 Context

As part of the government's ***Skills for Life*** strategy for improving adult literacy, numeracy and ESOL (English for Speakers of Other Languages) in England, in February 2002 the Adult Basic Skills Strategy Unit (ABSSU) within the Department for Education and Skills (DfES) established NRDC. Two major strands of NRDC's work are:

- research relating to productivity and social inclusion; and
- research relating to best practice in teaching and learning.

Within the first strand, two of the major projects which began in 2002 were an enhancement of a birth cohort study sweep to be conducted in 2004 by the Centre for Longitudinal Studies at the Institute of Education, University of London, and a project aimed at improving the literacy and numeracy of young offenders and disaffected young people. The former was intended to add a whole-cohort assessment of literacy and numeracy levels to the British Cohort Study 1970 (BCS70) sweep already planned for 2004, and the latter was a controlled trial investigating whether improvement in literacy can reduce offending. For both projects the best available assessment instruments would be needed.

Within the second strand mentioned above, it was envisaged that many of the lines of research that began with reviews and explorations of the field would lead rather quickly to empirical studies, such as investigations of factors and techniques thought to influence learners' progress; these too would need valid and reliable assessment instruments for judging the amount of progress made by learners.

More widely, practitioners and researchers would benefit from an up-to-date review of available instruments for assessing adult literacy and numeracy, since

> "Assessment is a necessary part of successful teaching, and the appropriate and effective use of assessment for formative, diagnostic and summative purposes is the trademark of a skilled, professional teacher."
>
> (F. Rees in Brooks et al., 2001a, p.107)

A review was particularly opportune in 2002 because several new instruments had appeared recently (e.g. the national tests of key/basic skills at Levels 1 and 2, Cambridge Training and Development's *Target Skills*), while others had been revised (e.g. the BSA's *Initial Assessment* pack) or developed (e.g. the tests for the ***Skills for Life*** survey of adult basic skills needs which was to be conducted in 2002/03). Several of these developments were driven by the new *National Standards for adult literacy and numeracy* (QCA, 2000) – the level boundaries in these Standards had been adjusted upwards somewhat, relative to the previous BSA standards, and sub-levels within entry level were defined for the first time.

This project was therefore commissioned by NRDC in 2002 to review available and recently produced instruments for the quantitative, summative assessment of adult literacy and numeracy. This is the report of that project; it is primarily addressed to those who already have some knowledge of the technicalities of assessment, and therefore takes as read acquaintance with terms such as formative, summative, diagnostic, screening, parallel forms, floor effect, facility value, etc.

1.2 Aims

The aims of this review were to:

- analyse selected British adult literacy and numeracy assessment instruments that are quantitative in form and summative in purpose;
- evaluate the potential of those instruments for use in the cohort study and NRDC research projects, and their suitability in their own terms; and
- make recommendations for the use of existing instruments and/or the development of new ones.

1.3 Exclusions

This section discusses only the main categories of instrument which were excluded – a fuller list is given in appendix A.

The *Adult Literacy Core Curriculum* (Basic Skills Agency, 2001a) and the literacy section of the National Standards cover speaking and listening as well as reading and writing. However, there appear to be no available British instruments for directly assessing the oracy skills of adults who are native speakers of English, and therefore all the literacy instruments reviewed were tests of reading and (aspects of) writing, with one partial exception. This was the learndirect/Ufi instrument *Word Skills Check*, which contained items supposedly assessing speaking and listening. However, no speaking was involved, and the answer formats for the listening items (which would not be accessible at a computer with no sound board) were rather artificial. This aspect of this instrument was therefore not reviewed.

English for Speakers of Other Languages (ESOL) instruments were not a concern of this review. Many such instruments are produced in Britain, but mainly for the English as a Foreign Language (EFL) field, and it is doubtful if any instruments are available which are produced in Britain and intended for assessing the abilities in English of ESOL adults with basic skills needs. Indeed, NRDC's Effective practice in ESOL study has adapted an EFL oracy test for measuring attainment and progress. (It is, of course, recognised that many ESOL learners attend mainstream provision – Brooks et al., 2001b found 18 per cent in their initial sample of over 2,100 learners.)

No exclusions parallel to those just mentioned for literacy were necessary in numeracy.

However, for both the cohort study and intervention studies, NRDC needed a review of *summative* assessment instruments, that is principally those designed to measure learners' attainment at the end of a course of study or in a survey. Such instruments can also be used at the beginning of a course of study in order to measure, through re-administering them at the end, how much progress learners make during it. Therefore, this review covers only instruments that have been, or can be, used summatively. Despite their titles, this includes such instruments as the Basic Skills Agency's *Initial Assessment*, which can be used at both the beginning and the end of a course.

This focus on summative instruments has meant the exclusion of the following further whole categories of both literacy and numeracy instruments:

- screening instruments, in particular the Basic Skills Agency's *Fast Track* (1st edition, Basic Skills Agency, 2000);
- diagnostic materials, in particular those developed by the BSA and partners and published by the DfES in 2003 (for an account of the development of these, see Sewell, 2004); and
- formative assessment systems, since these are inherently multifarious and often specific to a single institution or even teacher.

Four other restrictions on the coverage of this review need to be stated here. First, only quantitative instruments were included; any that were purely qualitative were deemed unsuitable for the cohort study and research projects.

Secondly, it was decided not to review instruments used before 1991 (see section A.2 for a partial history and a justification of this cut-off date).

Thirdly, a decision had to be taken on the lower age limit of learners for whom the instruments were intended. On inspection, the tests devised for the assessment of 15-year-olds in PISA 2000 (Programme of International Student Assessment) and Level 4 of the *Edinburgh Reading Test* (which is standardised up to age 16) clearly had lessons for the adult sector, while those used with 13-year-olds in TIMSS (the Third International Mathematics and Science Study) in 1994 did not. The lower age limit was accordingly set at 15.

Finally, instruments developed and used in the US were not included in the main review. For the reasons for this, and for a brief review of a few US instruments, see appendix B.

The instruments which were chosen for review are listed in section 3.

2. Previous research

No previous comprehensive review of adult basic skills assessment instruments used in Britain was found (and the present review is itself not comprehensive – see again the list of exclusions). However, MacKillop (1997) gave a full account of the earlier assessment instruments developed by Adult Literacy and Basic Skills Unit (ALBSU), and her discussion of various methods of assessment available to adult literacy tutors proved very useful. Her thesis was particularly interesting because she had taught in England as well as North America, and therefore had practical experience, as well as academic knowledge, of both systems. However, she discussed only a selection of the literacy assessment materials used in the United Kingdom (UK). Neither did her work cover assessment in England since the introduction of the National Standards and Core Curricula, as it was completed before these were introduced; and her work was not concerned with numeracy.

One of MacKillop's main conclusions was that the use of multiple-choice tests in the United States (US) (which dominated, and dominate, adult basic education assessment practice there), prevented adult literacy schemes from achieving their goals. In contrast with US practice she characterised the approach to assessment in Britain then as 'authentic', which may now seem somewhat ironic, given the increasing move towards multiple choice here.

A recent and very detailed review of US instruments was provided by Kruidenier (2002), and this is drawn on extensively in appendix B.

One previous review of assessment *techniques* in literacy and numeracy was located (University of Sheffield Division of Education, 1989), but proved to be rather out of date. Two previous reviews of specific instruments were also located. Lindsay, Weinberger and Hannon (1992) provided a preliminary evaluation of the ALBSU Basic Skills Assessment materials published in 1992 (ALBSU, 1992). When located at the Institute of Education, University of London's library, the materials proved to have been a forerunner of the first edition of *Initial Assessment*; they have been out of use for some years. Finlay (1997) based a critique of ALBSU's (1993) first screening test for reading on using both it and informal methods with *one* learner. The screening test itself would not have been relevant to this review.

Finlay and Harrison (1992) looked at issues around the use of specific tasks – such as reading maps or writing letters – as the basis for teaching, assessment and programme evaluation in adult literacy. This approach has clear potential in terms of making courses relevant to learners. However, as Finlay and Harrison acknowledged, assessing some tasks is more difficult than assessing others. They suggested that approaches to assessment could be placed on a continuum from formal to informal, and that it would be interesting to investigate the degree to which the amount of progress claimed is linked with the position on this continuum of the assessment approach used – but no such research appears to have been carried out. Finlay and Harrison also made the interesting point that a learner did not need to make progress to achieve a *Wordpower* certificate, only to produce evidence of current attainment. This suggests that the achievement of a certificate cannot necessarily be relied upon as an indicator of progress.

Whiteman (1998) investigated adult literacy tutors' views on measuring learners' success, and Ward and Edwards (2002) went further by investigating adult literacy learners' own views on their progress in their 'learning journeys'. There is important work still to be done in this area, but none of it seems yet to have influenced the current wave of summative assessment instruments.

The review of research on assessment and accreditation in the adult basic skills field in Brooks et al., (2001a, chapter 9, pp.107–10) was very brief, and limited in scope to 'systems of assessment providing results to be reported to students …; instruments used only in research studies [we]re not considered' (p.107). Brooks et al.,'s main source of information was a Further Education Funding Council Inspectorate report (FEFC, 1998) – this covered many of the major awards for literacy, numeracy and ESOL at BSA Entry Level and Level 1 available to and used for students in the Further Education sector (though not those provided through the various Open College networks). The inspectors described the situation as 'a plethora of different awards, supposedly at the same level, but in reality requiring very different skills, competences or understanding'. The inspectors judged that levels differed not only between providers of qualifications but even between different colleges giving the same award, and there were significantly different achievement rates for different certificates supposedly accrediting the same level of the same framework.

The FEFC report did not make recommendations, but its closing description of effective awards could be read as such:

> "Awards in numeracy, literacy and ESOL at Entry Level and Level 1 are most effective when they:
>
> - provide an appropriate curriculum framework and structure for organising learning;
> - define standards accurately and align them to national levels;
> - specify learning outcomes and performance criteria;
> - have assessment linked to the standards;
> - ensure consistent application of standards through rigorous verification;
> - provide opportunities for dual accreditation of key skills and basic skills within the same scheme;
> - encourage the development of both functional and creative language, and both functional and conceptual numeracy; and
> - allow some flexibility for teachers to match content, context and assessment mode to the learning needs of their students."
>
> (FEFC, 1998, p.43)

Conceptual frameworks of this sort were useful in informing the analysis conducted in this project. However, none of the literacy and numeracy instruments used for the awards surveyed by the FEFC Inspectorate were selected for review here because they were all abolished in 2002.

The FEFC report covered ESOL as well as literacy and numeracy. Although (as stated in section 1.2) ESOL instruments were not a concern of this review, useful categories and suggestions for our framework and approach were found in Weir (1993), including the reviewers' practice of trying out instruments on themselves.

The FEFC report was concerned only with systems for assessing accredited learning. Recent work by Grief and Windsor (2002) for the Learning and Skills Development Agency on the recording of outcomes in non-accredited provision was also taken into account.

Though there appears never to have been a comprehensive British review of instruments for assessing adults' basic skills, there have been several at school level. Vincent and his colleagues (Vincent and Cresswell, 1976; Vincent et al., 1983), Pumfrey (1976, 1985) and Hurry et al., (1996) all provided reviews of reading tests and their use in the school sector. The rapid pace of change in the school curriculum means that all these texts except Hurry et al., (1996) are now outdated. The same applies to Ridgway's (1987) review of school mathematics tests. However, Vincent's (1976) work remains a useful introduction to some of the statistical techniques used in test development, and all the sources mentioned in this paragraph provided useful classifications and analyses of assessment instruments and their purposes.

The lack of relevant research on assessment noted by Brooks et al., (2001a) was also noted in a Learning and Skills Research Council (LSRC) report published some time after the work reported here was carried out. Torrance and Coultas (2004) set out to investigate the question 'Do summative assessment and testing have a positive or negative effect on post-16 learners' motivation for learning in the learning and skill sector?' (this is the main title of their review) – 'but did not identify any material of direct relevance to the review question' (summary). Instead, they 'effectively review[ed] "the field of assessment in post-compulsory education in the UK" ' (summary). Two of their main conclusions were that 'across the compulsory and post-compulsory sectors as a whole it would appear that summative assessment and testing do more harm than good' (p.35) and that 'many [learners] fear testing and ... there is evidence that this can precipitate drop-out and deter progression' (p.35) – a partial echo of MacKillop's finding and salutary warnings for all in the field.

The outcomes of this sketch of previous research can be summed up by saying that there was very little such research, but that the analytical frameworks used in some of it were useful for this review.

3. Method

Two half-time research officers were recruited to work on the review, one to cover literacy (KH), the other numeracy (AP). Both had teaching qualifications, a higher degree, and recent and relevant teaching experience, as well as an academic interest in the area and backgrounds in consultancy. Their brief was to collect and analyse a range of relevant British assessment instruments, using an analytic framework which was guided in part by previous work, drafted in advance, and modified after trials with the first few instruments.

3.1 Collecting examples

Tests and other instruments were collected using various sources of information, including specialists in the field, internet sites, and publishers' catalogues. The instruments collected included the most recent ones produced or sponsored by the Basic Skills Agency. Given the wide range of assessment materials in use, it was impossible to review every one. A list of those which were reviewed is given in table 1, with an indication of whether they were reviewed for literacy, numeracy or both.

All the instruments listed in table 1 have been used in England, and all but two (IALS, PISA) were developed specifically for use in Britain.

3.2 Designing an analytic framework

An analytic framework was compiled using concepts from the sources mentioned above, and in the light of the researchers' professional knowledge and experience. From the framework a summary chart was developed and trialled on which the main points about each instrument could be summarised on an 'at a glance' basis, together with additional detail where appropriate. The initial version of this chart was modified as a result of experience with the first few instruments and project team discussions, and the final version is reproduced in appendix C, together with notes on its development and sources.

The questions in the framework were intended to:

- support the process of review;
- log standard details about every instrument;
- ensure that issues of validity, reliability, manageability and utility were addressed in as much detail as was reasonable;
- allow notes to be made about whether the assessments reviewed would meet the criteria for the contemplated research purposes; and
- allow the general features of, and important points about, each assessment to be summarised.

A narrative commentary was also written for each instrument.

Table 1: list of instruments reviewed

	Reviewed for literacy?	Reviewed for numeracy?
A. Instruments used in previous studies		
1. Tests used in previous lifetime cohort study sweeps: BCS70 * (1991/92), NCDS5 ** (1995) (Ekinsmyth and Bynner, 1994; Bynner and Parsons, 1997)	Y	Y
2. International Adult Literacy Survey (IALS, 1994/98) (Carey, Low and Hansbro, 1997)	Y	Y
3. *Progress in Adult Literacy* study (1998/99) (Brooks et al., 2001b)	Y	n/a
4. Programme for International Student Assessment (PISA 2000) (OECD, 2001)	Y	Y
B. Other paper-based instruments		
5. *Assessing Progress in Basic Skills: Literacy.* (Basic Skills Agency, 1997a)	Y	n/a
6. *Initial Assessment* (1st edition) (Basic Skills Agency, 1997b)	Y	Y
7. *Initial Assessment (2nd edition)**** (Basic Skills Agency, 2002)	Y	Y
8. *Edinburgh Reading Test, Level 4 (3rd Edition)* (Educational Assessment Unit, University of Edinburgh, 2002)	Y	n/a
9. *Entry Level Certificate in Adult Numeracy.* (AQA, 2002)	n/a	Y
10. *Skillscape.* (Smith and Whetton, 2000)	Y	Y
C. Computer-based instruments		
11. *Target Skills: Initial Assessment.* (Version 1.1) (CTAD, 2001)	Y	Y
12. *Number Skills Check.* (Release 5) (**learndirect**/Ufi, 2001a)	n/a	Y (briefly) ****
13. *Word Skills Check.* (Release 2) (**learndirect**/Ufi, 2001b)	Y	n/a
14. *National Tests in Adult Literacy. Practice Tests. Literacy Level 1 and Level 2.* ***** (D*f*ES, 2002)	Y	n/a
15. Tests for ***Skills for Life*** Survey of Needs (2002) (CDELL, 2002)	Y	Y
D. New instruments		
16. Numeracy assessment instrument adapted by NRDC from ***Skills for Life*** survey (2003)	n/a	Y
17. *Go! Literacy Assessment Instrument.* Developed for NRDC by NFER (NRDC, 2004)	Y	n/a

Notes to Table 1:

* Instrument also used in *Older and Younger* study (1993–94) (Basic Skills Agency, 1995)

** Instrument also used in *Literacy and Numeracy Skills in Wales* study (1995) (Basic Skills Agency, 1997c)

*** Paper-based but designed for ICT conversion

**** Technical problems were encountered using the version supplied with some versions of Windows.

***** Only Level 1 reviewed; remarks generally apply also to Level 2.

3.3 Reviewing the instruments

The instruments, or as much of each as was needed to reach a judgment on it, were tried out by the relevant reviewer, and points arising noted. Any available documentation was read. Most attention was paid to the more recent instruments, on the assumption that these would come to be more frequently used than older ones. However, as centres may have stocks of older instruments, or be using similar materials, some older instruments were also reviewed. A review was completed for each instrument, consisting of a completed framework and a narrative commentary. All the detailed reviews are in appendix D, where instruments are listed in the same order as in table 1.

While it is hoped that this information will be found useful, it is emphasised that the detailed reviews are very much working documents. An attempt has been made to avoid repetition, so that when an issue has been dealt with in some detail in the course of one review, less detail appears in subsequent ones. The reviews should be read as a whole.

In assessing the instruments' suitability in their own terms, each was judged against the purposes stated for it in the manual and/or deducible from the nature of the instrument itself. Except in the completed frameworks very little is said about the instruments' suitability in their own terms. A little general commentary on this aspect is included in appendix E, along with a list of desirable characteristics of tests, but in the main text we concentrate on evaluating the instruments' potential for use in the cohort study and intervention studies.

3.4 Relating the framework to the aims of the project

For the purposes of *both the cohort study and research projects*, literacy and numeracy assessment instruments were required that would be:

- quantitative;
- summative;
- acceptably valid;
- acceptably reliable;
- reasonably manageable (quick for testers to administer and score);
- acceptable and appropriate to testees (not taking too long, not overfacing those with poorer skills by being well beyond their current level of attainment, culturally suitable and free from bias);
- based on generic rather than context-specific content (so that in principle they apply everywhere and are not restricted to particular circumstances); and
- cost-effective.

In addition to the above common requirements, the two types of study had some that were markedly distinct. Instruments for the *cohort study* needed to:

- cover both literacy and numeracy in a similar fashion, for ease of administration in the same project;
- be suitable for use by non-teaching personnel (in the household survey format of the cohort study sweep); and
- take account of tasks used in previous sweeps (to allow for comparability over time, if required by the cohort study enhancement project).

But we did not need to have parallel forms, while those for *research projects* needed to:

- be suitable for use by practitioner-researchers and fieldworkers involved in the intervention studies; and
- have parallel forms (so that different forms can be used at pre- and post-tests, in order to avoid practice effects).

Instruments for research reports could use quite different forms of instrument for literacy and numeracy.

There were, moreover, two features where the requirements of the cohort study and the research projects were similar but not identical.

First, for the cohort study the instruments needed to be unpublished, or at least not widely known and used so far, at the date of the sweep in 2004. This was so that familiarity with the tests on the part of any adults who might have encountered them in basic skills provision would not distort the results. For research projects, however, the instruments needed to be reserved for research purposes, that is kept unpublished and therefore secure, at least until the end of the last research project arising from NRDC's work – and possibly beyond then so that researchers could continue to use them. This implied keeping the research project instruments secure well beyond the date required for the cohort study instruments.

Secondly, for both purposes the instruments needed to be aligned with the *Adult Literacy and Numeracy Core Curricula* (Basic Skills Agency, 2001a, b) and especially with the Standards, since from 2002 programmes were expected to work within them, and progress towards the government's targets was defined in terms of them:

- For the cohort study it seemed that it would be sufficient for the results to estimate the proportions of the cohort achieving at each of six 'levels', namely Below Entry 2 – At Entry 2 – At Entry 3 – At Level 1 – At Level 2 – Above Level 2. There seemed to be no policy or research interest in differentiating between Levels 3 and 4, or between Entry 1 and Pre-Entry. In the latter case it would probably have been impractical anyway, given the requirement in the curricula that learners at these levels tackle only items that are personal and familiar. Any instrument or system which attempted to assess on these terms would contradict the requirement for teaching and therefore assessment to be based on generic rather than context-specific content.
- For research projects, on the other hand, while measures of progress would need to take account of the specifics of the Standards and Curricula, instruments would need to provide much finer discrimination, especially at lower levels. Progress from one level to another is far too blunt a measure for pedagogical innovations to be assessed against, and statistically and educationally significant gains might be missed if a finer scale were not used. Also, learners at lower levels may well make only small steps of progress which nevertheless are educationally and personally significant for them and which therefore deserve to be recognised and acknowledged.

The summary chart contained questions relating to all these features, and others, and a chart was completed for each of the tests mentioned in table 1.

4. Principal conclusions

4.1 Principal criteria

In this section we once again concentrate on the instruments' potential for use in the cohort study and research projects.

While we were reviewing the instruments that existed in 2002 (the special status of the last two instruments listed – the numeracy assessment instrument developed by NRDC in 2003 from the ***Skills for Life*** survey materials and the literacy assessment instrument developed for NRDC by NFER – is discussed in section 4.7), it became apparent that, in addition to technical matters such as reliability and validity, two criteria were crucial for deciding whether an instrument could be used in the cohort study and research projects – whether it was secure; whether it was aligned to the new QCA Standards – and that another criterion was crucial for research projects (but not for the cohort study), namely whether an instrument had parallel forms. Our judgements on the instruments against these important criteria are listed in table 2, with some comments.

A particular feature of the existing literacy instruments was that only a few made any attempt to assess writing (*Progress in Adult Literacy* study, *Assessing Progress in Basic Skills: Literacy*, *Initial Assessment* (both editions), *Skillscape*), and even some of these assessed only surface features (both editions of *Initial Assessment*).

Table 2: list of instruments with main criteria and comments

	Literacy/ numeracy?	Secure?	Aligned with new QCA Standards?	Parallel forms?	Comments
1. Tests used in previous lifetime cohort study sweeps	Both	N published but not well known	N old ALBSU/BSA Standards	N	Some of the literacy items include numeracy. Number of items low for good reliability. Some potential for including items in 2004 to provide comparisons over time.
2. International Adult Literacy Survey	Both (numeracy in form of 'quantitative literacy')	N not formally published but available to enquirers	N not aligned originally, because international, but quasi-aligned via *Progress in Adult Literacy* study	N	Definitional issues: prose, document and quantitative literacy scales used. Though the new Standards are supposed to correspond with IALS levels, the differences in the conceptualisation of the domain/ achievement in levels mean that ultimately decisions about correspondences would have to be matters of judgement, probably using item-by-item comparison. Judged not to be appropriate to assess progress in programmes intended to cover the new curricula.

	Literacy/ numeracy?	Secure?	Aligned with new QCA Standards?	Parallel forms?	Comments
3. Progress in *Adult Literacy* study	Literacy	N most items were derived from earlier studies and some of those had been published	Y quasi-aligned after the event	Y	Interesting use of items from IALS prose and document scales. Though Brooks et al., report that no floor effect was apparent in their reading results, they found that there were too few items that learners at lower levels could complete successfully, so not suitable for assessing progress within entry level. Also not suitable for assessing progress on programmes intended to cover the new curriculum. Attempt made to assess writing.
4. Programme for International Student Assessment	Both	N some items published in report	N international	N	Not designed as a measure of adult literacy or numeracy.
B. Other paper-based instruments					
5. *Assessing Progress in Basic Skills: Literacy*	Literacy	N	N old ALBSU/ BSA Standards	Y	Meant as progress measure to sit alongside Initial Assessment 1st edition as placement measure – but Initial Assessment was often used as progress measure. Includes writing.
6. *Initial Assessment* (1st edition.)	Both	N	N old ALBSU/ BSA Standards	Y	Replaced by 2nd edition
7. *Initial Assessment* (2nd edition)	Both	N	Y	Y	Very heavy weighting given to spelling in literacy marking scheme – potentially discriminatory against dyslexics. Question marks over suitability of multiple-choice format and the effects of this on validity.
8. Edinburgh Reading Test, level 4 (3rd edition)	Literacy	N	N earlier concept	N	Not designed as a measure of adult literacy.
9. Entry Level Certificate in Adult Numeracy	Numeracy	N items made available after each round of testing	Y	Y, in sense that new form used at each round of testing	As title states, covers only entry level
10. Skillscape	Both	N published but perhaps not well known	N developers' concept of domain	N	Standardised. Time judged by reviewer to be particularly important factor in results. The assessment of writing is interesting.

	Literacy/ numeracy?	Secure?	Aligned with new QCA Standards?	Parallel forms?	Comments
C. Computer-based instruments					
11. *Target Skills: Initial Assessment*	Both	N	Y	Y adaptive	Neither reviewer felt the instrument met the criteria for a good test. Problematic validity, especially tests of writing which didn't require any.
12. *Number Skills Check*	Numeracy	N	Y	Y adaptive	Not recommended for use in research projects: problems with quality of items. Did not meet criteria developed in course of project for good test, especially for learners at lower levels of achievement. Bewildering range of task types/formats, etc.
13. *Word Skills Check*	Literacy	N	Y	Y adaptive	As *Number Skills Check* plus problems with validity of results for writing – no writing was required.
14. National Literacy Test Level 1 (Practice Version)	Literacy	N	Y	Y adaptive	Though only Level 1 seen, test is known to cover only Levels 1 and 2 in both literacy and numeracy, so range not suited to either proposed purpose.
15. Tests for ***Skills for Life*** Survey of Needs	Both	Y	Y but very broad	Y adaptive	Alignment with standards insufficiently precise for research projects, but might be acceptable for cohort study.
D. New instruments					
16. Numeracy assessment instrument adapted by NRDC from ***Skills for Life*** survey	Numeracy	Y	Y	N	The 20 most reliable numeracy items were selected, modified slightly, and adapted back from computer-based to paper form. Too short to be very reliable as a whole, and has no parallel form.
17. *Go! Literacy Assessment Instrument.* Developed for NRDC by NFER	Literacy	Y	Y	Y	Specifically developed for use in NRDC's research projects, but not suitable and too late for the 2004 cohort study sweep.

4.2 Candidate instruments for research projects

From table 2 it is immediately apparent that the only instruments which existed in 2002 with even a qualified Yes in all three criteria columns were the literacy and numeracy tests which were being used in the ***Skills for Life*** needs survey, but, as the detailed review in appendix D shows, those tests had other substantial drawbacks (see also Heath, 2003).

Consideration was therefore given to an instrument that met the alignment and parallel forms criteria but which, while not fully secure, was perhaps not well enough known for the availability of some of its items to be a problem, namely the *Progress in Adult Literacy* study tests. However, even though partly constructed from what were supposed to be very simple items from previous studies, these tests were found to have very few items that learners within Entry Level could consistently manage. This would have meant that it would have been very difficult to detect small steps of progress made by Entry Level learners using this instrument.

Therefore, we concluded that none of the instruments existing in 2002 were fully suitable for use in NRDC's research projects. Further implications for NRDC's research projects are discussed below.

4.3 Candidate instruments for the cohort study

For the cohort study, the only semi-strong candidates were again the tests used in the *Progress in Adult Literacy* study and the ***Skills for Life*** needs survey.

However, the *Progress in Adult Literacy* study did not cover numeracy; it did not even cover quantitative literacy, since no IALS items of that type were used. For the cohort study it was important that literacy and numeracy be assessed in parallel formats.

This might seem to leave the tests developed for the 2002/3 national ***Skills for Life*** survey as the least unsuitable candidate – but computer-delivered tests seem to us to have severe problems, which need to be discussed before our overall judgements for the two types of study can be stated.

4.4 Advantages and disadvantages of computer-based tests

A computer-based test that is simply a paper-based test transferred to another format is an expensive waste of time, effort and resources. Of the instruments reviewed, only the impending computerised version of the new edition of *Initial Assessment* was thought to be just a paper-based test transferred to computer, and it may be that adaptations were due to be made in the process. The five computer-based instruments listed above were all developed specifically for computer administration and avoided being just computerised paper-based tests.

Moreover, they are all adaptive, and this feature of computer-based tests requires comment. 'Adaptive' here means that a computer-based test does not, like a paper test, have to present the same items across a range of difficulty in the same order to all test-takers. ('In the same order' here refers to the practicalities of printed matter, not necessarily to test-takers' strategies.) For a reliable estimate of a learner's level of attainment using a paper-based instrument, statisticians consider 25 items to be the absolute minimum, and prefer at least 30. Instead, an adaptive test presents one item at a time, chooses each item on the basis of the learner's performance on previous items, and can 'zero in' on an estimate of the learner's level of attainment using significantly fewer items, and much quicker, than a paper-based test. In theory, this is the strongest advantage of adaptive computer-based tests.

For such a system to work, however, there are very strong design requirements. The items must be very closely tied to the domain specification. There must be a large number of items, so that learners can re-take the test at short intervals without encountering the same items and therefore gaining a higher score simply through remembering those items (practice effect). The items must be very finely graded in difficulty, if fine gradations of score are required. Whatever degree of discrimination is required (even if only the six levels 'Below Entry 2 – Entry 2 – Entry 3 – Level 1 –

Level 2 – Above Level 2'), all the items must have been extensively trialled so that their statistical properties are firmly established (e.g. facility value, reliably indicating attainment within a particular level, etc.). Given the very short time within which the items for the ***Skills for Life*** survey were developed, it always seemed unlikely that they had received adequate trialling, and the report on the survey (Williams et al., 2003, especially pp.14–15, 223–7 and 233) shows that they did not.

Moreover, the reviewers felt that on the whole all the ICT-based test materials they saw:

a) required additional skills over and above those being assessed;
b) tended not to meet the criteria for good tests in terms of internal consistency, clarity of instructions, etc.;
c) tended to skew the domain in line with how assessments targeted at various bullet points in the Standards could be devised within the limitations of the medium; and
d) were biased against those without ICT experience.

In so far as they used multiple-choice tasks, the ICT-based test materials also shared the limitations of multiple-choice testing, which are discussed in the next section. We therefore could not unreservedly recommend the use of any of the existing ICT-based test materials.

An alternative view of at least the national basic and key skills tests may be found in Kingdon (2003). He summarised technical issues arising from the evaluation of the first on-demand and on-screen presentations of those tests in January 2002, preparatory to the national roll-out of these formats later that year. Though he located problems with the technical delivery of the tests, he seems to have taken a broadly positive view of them as the way to go. He did list some outstanding questions, such as:

> "whether presentation of tests on-screen will change significantly the psychometric properties of the test items… [whether] the different ways in which the participating awarding bodies present the same questions … also make items easier or more difficult.
> (Kingdon, 2003, p.98)"

But he did not discuss broader questions of validity, alignment with the Standards, backwash, and suitability for ESOL learners. For Kingdon these questions were less important than 'some of the practical issues that awarding bodies and centres will face when implementing the online on-screen delivery models for basic and key skills tests' (p.99).

4.5 Item formats

Consideration was given to the use of multiple-choice items, which feature frequently in ICT-based instruments. This technique has attractions in terms of manageability and objectivity of scoring, and well-established techniques for item development and design exist, if not banks of unpublished items with potential for use. On the other hand, this assessment practice is not a reliably embedded part of UK test culture, and though the new national tests at Levels 1 and 2 use such methods, those tests have been the subject of criticism already. Multiple-choice assessments of writing are of particularly doubtful validity.

Cloze passages (and sentences) are popular as tests of reading, not least because machine-scorable or ICT-based forms of some cloze-type items can be produced. In requiring learners to fill gaps, it requires literacy (including quantitative literacy) skills, knowledge and understanding to be applied in a manner and for purposes that rarely occur in real life. Therefore, the use of cloze in this context cannot be recommended.

We therefore concluded that none of the instruments existing in 2002 were fully suitable for use in the 2004 cohort sweep either.

4.6 General conclusions

Cohort Study 2004

No wholly suitable secure instrument aligned to the new Standards was found for either literacy or numeracy.

Research projects

No wholly suitable secure instrument aligned to the new Standards and with parallel forms was found for either literacy or numeracy.

4.7 Recommendations

Given this bleak assessment (which was made known to NRDC in early Autumn 2002), how was NRDC to proceed?

Cohort Study 2004

On balance, it was decided that the instruments for use in the 2004 sweep would, in order to maintain continuity, use some of the items used in the sweeps in the 1990s, but supplemented by newly-devised items, with all items aligned to the new Standards. This decision was implemented for the 2004 cohort study sweep – see Parsons et al., (2005).

Research projects

For literacy, the conclusion was reached that the progress made by learners working towards Entry Level 1 is likely to be so personal and individual that external assessment instruments are unlikely to reflect gains made, and other methods of assessing progress, including soft measures, should be considered. This judgement was based upon a close reading of sections of the relevant curricula, but might not be shared by the developers of the new *Initial Assessment*, which does include non-context-specific materials for assessing Entry Level 1 literacy.

It was, therefore, suggested that a set of literacy assessment materials, linked through the judgment of experienced tutors with the new Standards, should be developed for the assessment of reading in research projects, and that these tasks should use mainly open-ended (supply) response types. The materials should, as far as possible, meet the criteria set out in our recommendations for the design of good tests (see appendix E).

It was strongly recommended that these tasks, including and especially those at lower levels, should include continuous prose in order to encourage good practice in the teaching of strategies for initial reading.

Moreover, we recommended that writing should also definitely be assessed, and suggested that the assessment of writing adequacy in the *Skillscape* instrument might be a suitable model. One attractive feature of this assessment tool is that it looks at what people can do, not at their weaknesses. A second possibility was the use of a pictorial stimulus of some kind. In interpreting any results, the conditions under which the writing was produced should be taken into account, especially if respondents were not given the chance to check and re-write. We also recommended that a new instrument be developed for numeracy.

It would not always be possible for research projects to incorporate computer-based assessment, which will still be unrepresentative of teaching and learning circumstances in most cases in the foreseeable future anyway. We therefore recommended that computer-based assessment should not be used in NRDC's research projects.

4.8 Implementation

At the date of finalising this paragraph (September 2004) the recommendation that a numeracy instrument should be developed for NRDC's research programme has been shelved indefinitely. An alternative approach has been taken to the need for a numeracy instrument, namely adapting the one used in the ***Skills for Life*** needs survey. This decision was taken in full awareness of the limitations of that instrument. The 20 most statistically reliable items from it were selected to cover the major areas of numeracy at levels from Entry 1 to Level 2 and used to create one test form (not enough were available to create parallel forms). The 20 items were converted back from the computer-administered form in which they had been used in the ***Skills for Life*** survey to the paper-based form in which they had originally been developed. Some items were also further adapted in the light of experience of using them in NRDC's numeracy projects in academic year 2003/04, and those versions are in use in academic year 2004/05. A review of the latter version of this instrument is included in appendix D.

However, most of the recommendations for a literacy assessment instrument were implemented between the presentation of the first draft of this review to NRDC in October 2002 (by Karen Heath, with samples from several of the instruments) and the production of the final draft in the summer of 2004. The literacy instrument was commissioned from NFER in late 2002 and developed during 2003, a pilot version was available for research use in the Autumn of 2003, and the finalised instrument was delivered on schedule in January 2004. It covers both reading and writing, at levels from 'Entry 2 or below' to 'Level 2 or above'. There are two parallel forms, statistically equated. Each form of the writing test covers all the relevant levels of the Standards. However, within each form of the reading test there are two versions, an easier one intended for learners at Entry Level, and a less easy one for learners at Level 1 or above; and in case of doubt over which version to ask a particular learner to attempt there is a very short Locator booklet. Both the reading and the writing tests yield approximate National Standards levels, and the reading tests also yield scaled scores (on a 0–100 scale with a mean of 50 and a standard deviation of 10) derived from a standardisation exercise carried out by NFER as part of the piloting. All forms of the tests are based on two specially-written 'magazines', and in this respect they follow the pattern of the literacy sections of *Skillscape* and of Level 2 of the second edition of *Initial Assessment*.

In 2004 the new instrument was in use in at least four NRDC projects, with more to follow, and it was being kept secure. However, a little user reaction can be noted: several tutors whose classes were observed in 2003/04 for the Effective practice in reading study (NRDC project PG3.9ESF, directed from the University of Sheffield) reported that some learners liked the 'magazines' so much that they asked if they could take them home. In order to show that this instrument meets NRDC's requirements in full, a review of it is included in appendix D. To follow up this successful development, we again recommend very strongly that an assessment instrument for numeracy be developed.

References

ACACE (1982). **Adults' Mathematical Ability and Performance.** Leicester: Advisory Council for Adult and Continuing Education.

ALBSU (1987). ***Literacy, Numeracy and Adults: Evidence from the National Child Development Study.*** London: Adult Literacy and Basic Skills Unit.

ALBSU (1994). **Tasks and Showcards for Assessing Reading, Writing, Oral Communication Skills and Basic Maths**. London: Adult Literacy and Basic Skills Unit.

ALBSU (1992). **Basic Skills Assessment Materials.** London: Adult Literacy and Basic Skills Unit.

ALBSU (1993). **Assessment Reading and Maths. A screening test.** London: Adult Literacy and Basic Skills Unit.

AQA (2002). **Entry Level Certificate in Adult Numeracy.** Guildford: Assessment and Qualifications Alliance.

Basic Skills Agency (1997a). **Assessing Progress in Basic Skills: Literacy.** London: BSA.

Basic Skills Agency (1997b). **Initial Assessment**. London: BSA.

Basic Skills Agency (1997c). **Literacy and Numeracy Skills in Wales**. London: BSA.

Basic Skills Agency (1997d). **International Numeracy Survey: a comparison of the basic numeracy skills of adults 16–60 in seven countries**. London: BSA.

Basic Skills Agency (2000). **Fast Track.** London: BSA.

Basic Skills Agency (2001a). **Adult Literacy Core Curriculum**. London: BSA.

Basic Skills Agency (2001b). **Adult Numeracy Core Curriculum.** London: BSA.

Basic Skills Agency (2002). **Initial Assessment: An assessment of literacy and numeracy level.** (2nd Edition) London: BSA.

Benn, R. (1997). **Adults Count Too: mathematics for empowerment.** Leicester: NIACE.

Brooks, G. (1998). 'New emphasis on old principles. The need for clarity of purpose and of assessment method in national testing and for national monitoring.' In C. Harrison and T. Salinger (Eds) **Assessing Reading 1. Theory and Practice.** London and New York: Routledge, 110–21.

Brooks, G. (2001a). "Progress in adult literacy: putting the record straight." **Adults Learning**, 12(10), 15–16.

Brooks, G. (2001b). 'Appendix 1. Progress in adult literacy: putting the record straight. A reply to the NIACE commentary on the NFER research report Progress in Adult Literacy.' In NATFHE (Comp.) 25 Years of Basic Skills Work: Failed opportunities, new start? Report of joint NATFHE/NIACE Conference, 9 May 2001, 14–16.

Brooks, G., Giles, K., Harman, J., Kendall, S., Rees, F. and Whittaker, S. (2001a). **Assembling the Fragments: a review of research on adult basic skills.** London**:** Department for Education and Employment Research Report no.220.

Brooks, G., Davies, R., Duckett, L., Hutchison, D., Kendall, S. and Wilkin, A. (2001b). **Progress in Adult Literacy: do learners learn?** London: Basic Skills Agency.

Brooks, G., Gorman, T.P., Harman, J., Hutchison, D. and Wilkin, A. (1996). **Family Literacy Works: The NFER evaluation of the Basic Skills Agency's Family Literacy Demonstration Programmes**. London: Basic Skills Agency.

Bynner, J. and Parsons, S. (1997). **It Doesn't Get Any Better.** London: Basic Skills Agency.

Bynner, J. and Steedman, J. (1995). **Difficulties with Basic Skills.** London: Basic Skills Agency.

Carey, S., Low, S. and Hansbro, J. (1997). **Adult Literacy in Great Britain: a survey of adults aged 16–65 in Great Britain carried out by Social Survey Division of ONS as part of the International Adult Literacy Survey (IALS).** London: The Stationery Office.

Carr-Hill, R., Passingham, S. and Wolf, A. with Kent, N. (1996). **Lost Opportunities: the language skills of linguistic minorities in England and Wales.** London: Basic Skills Agency.

Charnley, A.H. and Jones, H.A. (1978). **The Concept of Success in Adult Literacy.** Cambridge: Huntington Publishers.

CDELL (2002). **Tests for *Skills for Life* Survey of Needs.** Nottingham: University of Nottingham.

Coben, D., O'Donoghue, J. and FitzSimons, G.E. (Eds.) (2000). **Perspectives on Adults Learning Mathematics: research and practice.** Dordrecht, The Netherlands: Kluwer.

Cohen, L., Manion, L. and Morrison, L. (2000). **Research Methods in Education***,* 5th edition. London: RoutledgeFalmer.

Coffield, F., Moseley, D., Hall, E. and Ecclestone, K. (2004) **Should we be using learning styles? What research has to say to practice.** London: Learning and Skills Research Centre.

CTAD (2001). **Target Skills: Initial Assessment.** Cambridge: Cambridge Training and Development.

D*f*ES (2002). **National Tests in Adult Literacy. Practice Tests. Literacy Level 1 and Level 2.** London: Department for Education and Skills http://www.qca.org.uk/nq/ks/level_1_2_tests.asp Accessed 1/5/02.

Educational Assessment Unit, University of Edinburgh (2002). **Edinburgh Reading Test.** 3rd Edition. London: Hodder and Stoughton.

Educational Testing Services (1990). **ETS Tests of Applied Language Skills.** New York: Simon & Schuster.

Ekinsmyth, C. and Bynner, J. (1994). **The Basic Skills of Young Adults.** London: Basic Skills Agency.

Elley, W. (1992). **How in the World do Students Read?** Hamburg: IEA.

FEFC (1998). **Numeracy, Literacy and ESOL: evaluation of Entry and Level 1 Awards.** *National Report from the Inspectorate.* Coventry: Further Education Funding Council.

Finlay, A. (1997). "Informal measures challenge the suitability of the Basic Skills Agency's reading test." **Reading**, 31(2), 29–34.

Finlay, A. (2000). Talk about reading: students tell what they know. In E. Millard (Ed.) **Enquiring into Literacy: Papers from the Literacy Research Centre** [pp.116–29].

Sheffield: Department of Educational Studies, University of Sheffield.

Good, M. and Holmes, J. (1978). **How's It Going? An alternative to testing students in adult literacy**. London: Adult Literacy Unit. 2nd edition, London: Adult Literacy and Basic Skills Unit (1982).

Gorman, T.P. (1981). "A survey of attainment and progress of learners in adult literacy schemes." **Educational Research**, 23(3), 190–8.

Gorman, T. and Moss, N. (1979). **Survey of Attainment and Progress in Adult Literacy Schemes.** Slough: National Foundation for Educational Research.

Grief, S. and Windsor, V. (2002). **Recognising and Validating Outcomes of Non-accredited Basic Skills and ESOL.** London: Learning and Skills Development Agency.

Hamilton, M. (2001). **Commentary on the NFER Research Report Progress in Adult Literacy.** Leicester: NIACE.

Hamilton, M. and Barton, D. (2000). "The International Adult Literacy Survey: what does it really measure?" **International Review of Education**, 46(5), 377–89.

Heath, K. (2003). "From Moser's dream to multiple choice." **RaPAL Journal**, no. 50, 12–16.

Heath, K. (2004). "Assessment in adult literacy: the origins of my research interest." **RaPAL Journal**, no.53 (Spring), 13–18.

Holland, D. and Street, B. (1994). Assessing adult literacy in the United Kingdom: the *Progress Profile*. In C. Hill and K. Parry (Eds.) **From Testing to Assessment: English as an international language** [pp.229–49]. London and New York: Longman.

Hurry, J. and Sylva, K. with Lee, L.F. and Mirrelman, H. (1996). **Standardised Literacy Tests in Primary Schools: their use and usefulness**. London: Schools Curriculum and Assessment Authority.

Jones, H.A. and Charnley, A.H. (1978). **Adult Literacy: a study of its impact**. Leicester: National Institute of Adult Education, University of London.

Kingdon, M. (2003). Computer-based assessment of basic and key skills: quality and compliance. In C. Richardson (Ed.) **Whither Assessment?** London: Qualifications and Curriculum Authority.

Kruidenier, J. (2002). Literacy assessment in adult basic education. In J. Comings, B. Garner and C. Smith (Eds.) **Annual Review of Adult Learning and Literacy**, vol.3 [pp. 84–151]. San Francisco, CA: Jossey-Bass.

La Marca, P.M. (2001). “Alignment of standards and assessments as an accountability criterion.” **Practical Assessment, Research and Evaluation**, 7(21).

learndirect/Ufi. (2001). **Number Skills Check**. Release 5. CD Rom. Sheffield: Ufi.

learndirect/Ufi. (2001). **Word Skills Check**. Release 2. CD Rom. Sheffield: Ufi.

Lewandowski, L.J. and Martens, B.K. (1990). “Selecting and evaluating standardized reading tests.” **Journal of Reading**, 33, 384–8.

Lindsay, G., Weinberger, J. and Hannon, P. (1992). **A Preliminary Evaluation of the ALBSU Basic Skills Assessment Materials.** Sheffield: University of Sheffield Division of Education Working Paper 9/92.

NRDC (2004). **Go! Literacy Assessment Instrument**. London: National Research and Development Centre for Adult Literacy and Numeracy.

OECD (2001). **Knowledge and *Skills for Life*: first results from PISA 2000.** Paris: Organisation for Economic Cooperation and Development.

Parsons, S. and Bynner, J. with Foudouli, V. (2005). **Measuring basic skills for longitudinal study: the design and development of instruments for use with cohort members in the age 34 follow-up in the 1970 British Cohort Study (BCS70)**. London: National Research and Development Centre for Adult Literacy and Numeracy.

Psychological Corporation (1991). **Wechsler Individual Achievement Tests: Reading Comprehension.** San Antonio, TX: Psychological Corporation.

Pumfrey, P.D. (1976). **Reading: tests and assessment techniques.** London: Hodder and Stoughton for UK Reading Association.

Pumfrey, P.D. (1985). **Reading: tests and assessment techniques.** 2nd edition. London: Hodder and Stoughton for UK Reading Association.

QCA (2000). **National Standards for Adult Literacy and Numeracy**. London: Qualifications and Curriculum Authority.

Ridgway, J. (1987). **A Review of Mathematics Tests.** Windsor: NFER-Nelson.

Rodgers, B. (1986). "Change in the reading attainment of adults: a longitudinal study." **British Journal of Developmental Psychology**, 4(1), 1–17.

Sewell, J. (2004). 'Diagnostic assessment within the ***Skills for Life*** strategy.' Paper presented at the conference of the International Association for Educational Assessment, Philadelphia, June 2004. Available online at http://www.nfer.ac.uk/research/papers/DiagnosticAssess.doc Accessed 29 November 2004.

Smith, P. and Whetton, C. (2000). **Skillscape**. Windsor: NFER-Nelson.

Social Surveys (Gallup Poll) Ltd (1981). **Adult Numeracy Study**. Leicester: Advisory Council for Adult and Continuing Education.

Strucker, J. (1997). "What silent reading tests alone can't tell you: two case studies in adult reading differences." **Focus on Basics**, 1(B), 13–16.

Test of Adult Basic Education (TABE) (1987). Monterey, CA: CTB/McGraw-Hill.

Torrance, H. and Coultas, J. (2004). **Do summative assessment and testing have a positive or a negative effect on post-16 learners' motivation for learning in the learning and skills sector? A review of the research literature on assessment in post-compulsory education in the UK**. London: Learning and Skills Research Council.

Tout, D. and Schmitt, M.J. (2002). The inclusion of numeracy in adult basic education. In J. Comings, B. Garner and C. Smith (Eds.) **Annual Review of Adult Learning and Literacy,** vol. 3 [pp.152–202]. San Francisco, CA: Jossey-Bass.

University of Sheffield Division of Education (1989). **Initial Assessment in Literacy and Numeracy – a review of techniques for Employment Training and YTS**. Sheffield: Training Agency.

Venezky, R.L. (1997). Literacy assessment in the service of literacy policy. In A.C. Tuijnman, I.S. Kirsch and D.A. Wagner (Eds.) **Adult Basic Skills. Innovations in measurement and policy analysis.** Creskill, NJ: Hampton Press Inc.

Venezky, R.L., Bristow, P.S. and Sabatini, J.P. (1994). **Measuring Gain in Adult Literacy Programs. NCAL Technical Report TR93-12.** Philadephia, PA: National Center on Adult Literacy.

Vincent, D. and Cresswell, M. (1976). **Reading tests in the classroom.** Windsor: National Foundation for Educational Research.

Vincent, D., Broderick, C., and Cresswell, M. (1983). **A Review of Reading Tests: a critical review of reading tests and assessment procedures available for use in British schools.** Windsor: NFER-Nelson.

Ward, J. with Edwards, J. (2002). **Learning Journeys: Learners' Voices. Learners' views on progress and achievement in literacy and numeracy.** London: Learning and Skills Development Agency.

Weir, C. (1993). **Understanding and Developing Language Tests.** Hemel Hempstead: Prentice Hall.

Whiteman A (1998). **Tutors' view on measuring success: the assessment and monitoring of students' progress in 'Ambridgeshire's' Adult Literacy Classes. Unpublished MEd dissertation, University of Sheffield.**

Wilkinson, G.S. (1993). **Wide Range Achievement Test – Revised (WRAT-3).** Chicago: Riverside Publishing Co.

Woodcock, R. (1991). **Woodcock Language Proficiency Battery – Revised.** Chicago: Riverside Publishing.

Williams, J. with Clemens, S., Olenikova, K. and Tarvin, K. (2003). **The *Skills for Life* Survey: a national needs and impact survey of literacy, numeracy and ICT skills.** London: Department for Education and Skills.

Internet sites accessed include:

http://www.nifl.gov/nifl/eff.html Accessed 30/9/04
www.ncal.org
http://www.ctb.com/products/product_detail.jsp
http://www.casas.org/02TrainingProfDev/PEPFinalNDN.pdf
http://www.dfes.gov.uk/readwriteplus/learning
http://www.qca.org.uk/nq/ks/level_1_2_tests.asp Accessed 01/05/2002
http://ww.nifl.gov/lincs/collections/eff/standards/how_they_work.html Accessed 10/07/2002
Practice Versions of the National Tests in Adult Literacy.
http://dfes.gov.uk/readwriteplus/learning Accessed 29/05/2002
www.pisa.oecd.org

Appendix A. Assessments not covered by this review

A.1 List of exclusions

Several kinds of literacy assessment were excluded from consideration. Reasons for this included limitations of time, the specialised nature of some of the tests, and lack of relevance to the purpose of the review. Not covered within this review are instruments in the following categories:

- Assessments of reading practices or habits or of attitudes to reading or motivation. The focus of the review was comprehension, and in any case there appear to be no such instruments for adult learners in Britain.
- Word recognition tests – the focus of the review was text level.
- Tools used to assess learning styles – for a detailed review of these see Coffield et al., (2004).
- Tools used to diagnose dyslexia.
- Other diagnostic assessment tools, in particular that developed by the Basic Skills Agency and published in 2003 (see Sewell, 2004).
- Screening instruments, e.g. *Fast Track* (Basic Skills Agency, 2000).
- Instruments designed for use with those aged less than 15, including those which report results in terms of reading age or spelling age.
- Tests purely of, or based upon, spelling, whether or not these are standardised using samples of adults.
- Aptitude and personality tests.
- Tests of or including assessments of broader key or core skills, e.g. *Basic and Key Skills Builder* (West Nottinghamshire College).
- Task-based screening proformas and similar profiling tools.
- External accreditation schedules and programme specifications (for an analysis of one such scheme see Heath, 2004).
- Award-bearing tests in use up to, but abolished in, 2002.
- Early Adult Literacy Unit/Adult Literacy and Basic Skills Unit/Basic Skills Agency assessment tools (e.g. *How's It Going?* (Good and Holmes, 1978) – see below – and the Basic Skills Assessment materials published in 1992 (ALBSU, 1992)), and early versions of the BSA screening test.
- Formative assessment instruments.
- Tools used in the contexts of particular projects, such as family literacy schemes (e.g. those reproduced in Brooks et al., 1996).
- Assessments specifically designed for use with ESOL learners.
- Assessments designed for learners who use Braille, Sign, or some other mode of communication.
- Assessments in terms of wider benefits of learning, such as improved self-confidence, including self-assessment instruments.
- Assessments in use before 1991 – but see below.

The exclusion of these assessment types should not be taken as a denial of their merits. Rather, it reflects the need to focus and prioritise within various constraints.

A.2 A partial history

Partial *adj*. (1) incomplete; (2) not impartial.

The first national survey of adult basic skills in Britain was the literacy survey carried out in 1972 (Rodgers, 1986). The sample was some 3000+ of the 17,000 people involved in the oldest lifetime cohort study, the National Survey of Health and Development. The cohort members had all been born in a week of March 1946, and were therefore aged 26 at the time of the survey. The test used was the Watts-Vernon, a 35-item five-option multiple-choice sentence-completion test which had been devised around 1938. The cohort members had also taken this test in 1961 at the age of 15. The 1972 results were interpreted as showing 'an illiteracy rate [sic] of less than 1 per cent' (the criterion being a score of less than 11 out of 35), and the average score had improved significantly since 1961. However, the test was already both aging and out of tune with current concepts of reading, and the quantitative approach taken was probably out of tune with prevailing opinion.

In the very early days of the mass adult basic skills movement in Britain (the mid 1970s), most assessment appears to have been informal and individualised, and formative in intention. When summative scores were needed, school-age tests were often used (Jan Walker, personal communication, December 2004), of the sort which yield reading ages – a meaningless concept when applied to adults. Very much in the qualitative (informal, individualised, formative) spirit were two small-scale surveys of the views of organisers, teachers and learners carried out in 1975–78 by Jones and Charnley (1978; Charnley and Jones, 1979). No exact figures were given for the numbers of people interviewed in the first survey, which was carried out in 1975–77, but the number of research areas was six (Jones and Charnley, 1978, p. 9) and at a later point (p. 73) the numbers of learners interviewed in two of these areas appear to be given as 41 and 39; perhaps about 240 learners took part in all. No attempt was made to measure learners' progress directly. However, Jones and Charnley (1978, pp. 93–95) provided nearly three pages of quotations from interviews with learners testifying to their progress, and then (p. 96) gave this summary:

> "[R]elatively few students were found who were ready to exercise their skills in public. They might sign a cheque in public with confidence, but only after the details had been filled in at home. This reluctance was most noticeable in writing. Students generally believed that they had made more progress with reading than with writing; but that most progress of all had been in feeling better about themselves. And this was seen more frequently in those who had attended groups than among those under tuition at home."

Later (p. 110) they made this judgement based on all the evidence in the survey:

> "The main body of those who persevered in tuition made progress in their skills that was slow but steady, and there were a very few examples of astonishingly rapid progress towards reading and writing."

The survey reported in Charnley and Jones (1979) appears to have been a spin-off from the one just summarised. They derived a set of putative indicators of success in adult literacy from the literature, trialled them with one group of teachers, and then used a revised set of criteria with a larger group of teachers (49) over a period of 18 months in 1976/77. They interviewed each of these teachers twice and discussed with them the profiles of 68 of their learners; and also interviewed 35 of the 68 learners once each. (The other 33 had left provision before they could be interviewed.) From all the evidence, Charnley and Jones derived a large set of criteria of success, grouped into five categories (the names of these and the examples are quoted verbatim from Charnley and Jones, 1979, pp. 176–7):

- Affective personal achievements (e.g. an improvement in self-reliance).
- Affective social achievements (e.g. better relationships with all members of the family).
- Socio-economic achievements (e.g. increasing the capacity of re-entering the employment market).
- Cognitive achievements (e.g. reading achievements displaying an increase in comprehension skills; writing achievements displaying an improvement in the ability to write letters).
- Enactive achievements (e.g. regular attendance for tuition; an ability to evaluate the reliability of various communications media).

For present purposes the salient feature of both of these reports is their total reliance on interview evidence rather than on directly recording learners' attainment or progress, by whatever method or instrument.

There has been some further research recently in the same tradition which needs to be logged here even though out of chronological sequence. Finlay (2000) interviewed a limited number of learners to find out their views on success in reading; Ward and Edwards (2002) directed a project which asked a sample of adult literacy learners and their tutors how they knew they (the learners) were making progress; and Eldred (2002) gathered many stakeholders' views and found that confidence is the most frequently noted marker of progress.

Even while the two Jones and Charnley projects were in progress, a need was clearly already being felt for more structure, and a framework for letting learners see and record the progress they were making. This was the avowed purpose of what appears to have been the first adult literacy assessment instrument published in Britain, *How's It Going?* (Good and Holmes, 1978, especially p. 3). This provided charts for logging learners' progress in reading and writing. The chart for reading contained 20 items, the chart for writing 19. Some examples of the items were 'Reading is communication', 'Key words' and 'Speed' for reading; 'Writing is communication', 'Dictionary' and 'Writing for speech' for writing. Each item was classified as one or more of attitudes, skills and knowledge, using the acronym ASK. Each chart then provided boxes for logging each learner's progress from Beginning through Not Bad to With Ease, three boxes at each of these levels; each entry was dated and recorded as '¸ = knows it, aware of it at this level', or '0 = is working at it, within this level', or 'X = starting off at this level'. Teacher and learner were to discuss the learner's progress and agree the entries. Explanations are given of all the various levels of each item. An exemplar of the Reading chart shows entries dated from January 1977 to May 1978.

The importance of this instrument is that it represented a determined and thought-through attempt to produce a learner-friendly and non-quantitative but systematic formative record. The subtitle, *An alternative to testing students in adult literacy*, is revealing. In use, the booklet must have been a handful for teachers to take in, let alone learners – it runs to 77 B5 sides – but after a time may have become sufficiently familiar, to teachers at least, for recording to have become impressionistic rather than based on constant look-up of details.

At about the same time, 1977–79, the first national survey of adult literacy learners' progress in England was carried out by a team at NFER led by Tom Gorman (Gorman, 1981; Gorman and Moss, 1979). The target sample was set at 2,000 students; pre-test materials were despatched in December 1977 to 1,831 students; by March 1978, 1,238 sets of materials (68 per cent of those distributed) had been returned. Post-testing was undertaken in June 1979. Scheme organisers were asked to distribute the post-tests to those learners from the pre-test sample who were still receiving tuition. A total of 1,158 learners (94 per cent of the pre-test sample) were traced; 378 (33 per cent of those traced) were no longer receiving tuition; and 194 (17 per cent of those traced) failed to respond; so that post-tests were received from 586 learners (51 per cent of those traced). However, some learners returned incomplete sets of post-test materials, so that the post-test sample was only about 40 per cent of the pre-test sample.

The tests were administered at both stages by basic skills teachers, and consisted of tests of reading, writing and spelling (details of which tests were used do not seem to be available). Thus this was a heavily quantitative and overtly summative project, quite different from the Good and Holmes approach. Yet because it was a survey and involved only a small percentage of learners it probably did not significantly affect practice. The only other study of adult literacy learners' *progress* (as opposed to attainment at one point in time) in Britain is the *Progress in Adult Literacy* study of 1998–99 (item 3 in the main text and appendix D).

Two national surveys were carried out in 1981. One involved members of the second lifetime cohort study, the National Child Development Study (ALBSU, 1987). These people had all been born in a week of April 1958 and were therefore aged 23 at the time of the survey. They were asked to complete self-report questionnaires containing questions about their abilities in reading, writing/spelling, and numeracy. Between 4 and 9 per cent reported problems in each skill; 13 per cent reported difficulty in at least one. This survey demonstrates continued reliance on self-report – but was also one of the first to include numeracy.

The other 1981 survey was carried out by Social Surveys (Gallup Poll) Ltd for the Advisory Committee on Adult Continuing Education, and attempted to measure the numeracy attainment of a nationally representative sample of adults selected according to ITV regions (ACACE, 1982; Social Surveys (Gallup Poll) Ltd, 1981). The instrument used was clearly inadequate – just 11 very simple items. For what it is worth, 21 per cent of the sample scored less than 6 out of 11.

Meanwhile, in the classroom 'the tradition of learner-centred teaching and assessment', as Holland and Street (1994, p. 231) put it, was maintained, and their own work continued it. In 1987 ALBSU funded Holland to research the assessment of adult literacy, and in 1989 she produced the loose-leaf, ring-bound *Progress Profile*. This bore a family resemblance to *How's It Going?* Assessment had both formative and summative aspects. The formative aspect was based on five questions (Holland and Street, 1994, p. 239):

> "Where do I want to go?
> What do I need to learn?
> How am I going to get there?
> How far have I got?
> Where to next?"

The first question was used at the beginning of a course, to help learners decide their general aims individually with their teacher, and the others as appropriate. Prompt Cards gave examples of aims in various areas of literacy and at various levels. 'The *Progress Profile* also asks students and teachers at the end of a specified period (we suggest 40 hours, but different programmes may adjust this to their own needs) to fill in a form entitled the *Progress Review*' (Holland and Street, 1994, p. 241). Here learners entered their aims and elements contributing to them, and recorded their progress by shading in, as they saw fit, between one and four boxes below each element. The boxes were defined as meaning (Holland and Street, 1994, p. 243):

> 'I've started, but I've still got a lot to do.
> I'm about half way there.
> I'm almost confident. I'm nearly there.
> I'm confident about this now.'

At the bottom of the sheet was a further box for answering the question 'How have you used what you've learnt?' Thus both aspects of the *Progress Profile* were non-quantitative but systematic. It seems as learner-friendly as *How's It Going?*, and perhaps even more teacher-friendly since the materials were less bulky and flexibility in use was emphasised throughout, including teachers being able, almost required, to add their own items to it.

Between October 1989 and April 1990 about 1,300 people received training from Holland in the use of the *Progress Profile* during 40 one-day sessions across England and Wales. Yet Holland and Street (1994, p. 244) report that 'One of the difficulties that arose most frequently was teachers' resistance to the very idea of assessment', despite even the summative aspect of the *Progress Profile* being almost the antithesis of traditional test-based assessment. It is not clear how widely it went into use, or for how long.

Also developed in the late 1980s were ALBSU's *Wordpower* and *Numberpower* systems. Hamilton and Merrifield (2000, p. 268) describe these as 'national accreditation for adult basic education learners' – therefore presumably the first to have this status. They were, Hamilton and Merrifield continue, 'based on a national set of standards that identifie[d] competencies... organised by level.' Since *Wordpower* and *Numberpower* will have been extremely familiar to many in the field, no attempt will be made to illustrate them here. The accreditation was achieved gradually and piecemeal, by building up a portfolio of evidence, none of it (apparently) quantitatively assessed but all of it ticked off in personal achievement records. The approach seems to have been intended to marry the qualitative, formative tradition with the need to have some more evidence-based method for justifying the award of a summative accreditation.

None of the instruments summarised above feature in the main parts of this review. Most were either not summative or not quantitative (or neither); those that were both summative and quantitative were out of date or otherwise no longer suitable. This is the justification for setting the cut-off date for instruments to be reviewed at 1991.

By the time Holland and Street's chapter was published, the model of assessment of adult basic skills in Britain was changing, as they themselves noted (p. 233); one very noticeable change was the overcoming, or at least submerging, of resistance to quantitative assessment. The change in the 1990s was dramatic: in the period 1991–96 in Britain there was one national survey of adult literacy alone, one of adult numeracy alone, and six which covered both. Though self-report questionnaires about difficulties featured in several, none used purely qualitative methods. All used summative, quantitative attainment tests, and the results of these have since been the main focus of attention, e.g. from the Moser Committee. The instruments used in five of these surveys are covered as items 1 and 2 in the main text and appendix D (the others were either not available or not suitable).

The frequency of national surveys has diminished markedly since 1996 – the only examples have been the *Progress in Adult Literacy* study of 1998–99, PISA at age 15 in 2000, and the ***Skills for Life*** needs survey of 2002–03 (items 3, 4 and 15 in the main text and Appendix D). However, in the period 1997–2002 a plethora of assessment instruments were developed and published. Ten such instruments are reviewed here (items 5–14).

And the emphasis on attainment measures in NRDC's research projects, and the development for NRDC by NFER of the literacy assessment instrument reviewed here as item 17, show that the emphasis on quantitative measures of attainment continues. It is certain that learner-centred, qualitative, formative assessment has not disappeared from professional practice in classrooms, but it is currently much less visible and receives much less attention.

Appendix B. Brief review of some US instruments

Preliminaries

Just as in Britain so in the United States the state of adult basic skills exercises policy makers. A recent initiative there is *Equipped For the Future* (EFF – see http://www.nifl.gov/nifl/eff.html), which can be seen as a parallel development to the production of the National Standards and Core Curricula here. This hugely expensive initiative sets out a meta-framework for assessment, based on 'analysis of what adults do in their roles as workers, citizens, and members of families and communities'.

Within these standards a communication skills category appears, in which the following skills are listed: *Read with Understanding; Convey Ideas in Writing; Speak So Others Can Understand; Listen Actively; Observe Critically.* Tout and Schmitt (2002) report that only one of the 16 EFF standards specifically mentions numeracy or mathematics: *Use Math to Solve Problems and Communicate.* While seeing this as positive in terms of 'adults using a range of purposeful skills to participate effectively in society', Tout and Schmitt regret that the EFF standards do not go further and make explicit the links between maths and the other standards, and that no examples of applying the maths standard are provided.

However, EFF is not an assessment instrument itself, but a framework for defining and devising such instruments. Moreover, up to the time of writing no new instruments seemed to have appeared based on it, and this appendix therefore concentrates on instruments which pre-dated it.

When considering US adult literacy and numeracy assessment instruments, it is helpful to remember that, in the United States, the word 'literacy' is often used in a sense that incorporates some of what we might refer to as 'mathematics' or as 'numeracy'. It is also important to bear in mind, when considering materials from the United States, that standardised testing is often mandatory in that country, and that assessment is often high stakes – with funding implications for the basic skills schemes involved. In effect, the requirement for standardised testing has led to the widespread use of multiple-choice tests throughout the educational system.

The following account is based largely upon a summary and review by Kruidenier (2002) of some of the tests most commonly used in the United States. Much less complete and less up to date is NCAL Technical Report TR95-03, 48 pages, $8.00. A summary is available from the National Center for Adult Literacy web site (www.ncal.org) not to be confused with the NCSALL site.

The National Institute for Linguistics website www.nifl.org says that examples of US tests are available online, but little detail is available without paying money. The whole tests – or banks of standardised items from which tests can be compiled – are secure, and only examples are available. The reviewers therefore relied on such examples as they saw, together with US-based reviews.

The following additional websites were accessed:

1) For (very little) information online about the Test of Adult Basic Education (TABE):
←http://www.ctb.com/products/product_detail.jsp?pageNumberCurrent=4&FOLDER%3C%3Efolder_id=36223&bmUID=1033157629100→
2) For information online about Equipped For the Future (EFF):
http://www.nifl.gov/lincs/collections/eff/ or http://www.nifl.gov/nifl/eff.html
3) For the Comprehensive Adult Student Assessment System (CASAS)
http://www.casas.org/02TrainingProfDev/PEPFinalNDN.pdf

Commentary

As part of the current project, we considered the question 'Are there instruments in use in the US that could be used in NRDC's research?' If there were, at least two benefits would accrue: first, the cost and difficulty of developing new British instruments might be avoided; secondly, direct statistical comparisons between US and British programmes in terms of learners' progress might be possible. However, it very rapidly became apparent that no US test would be suitable for NRDC's purposes, and this is the principal reason for not including any US instruments in the main part of this review. For example, a booklet was examined from the Test of Adult Literacy Skills (TALS) published by Educational Testing Services (1990). The language, or dialect, used in the text, and some of the background or cultural detail, were judged to be too different from those of England to allow the tests to be used in England in their current form, especially with adults for whom the familiarity of a text, or at least its cultural assumptions, is an important factor in its readability. These problems seemed likely to recur with many materials deriving from the US. For example, the adjective 'provisional' is not used when speaking of driving licences. The formats of everyday documents, which English learners would be familiar with British versions of, were different, which seems highly likely to affect the reliability and validity of the test if used on a different population

The adult basic skills test most widely used in the United States is the TABE (Test of Adult Basic Education). This test was normed on 'ABE Adults' (Strucker, 1997, p.15) and is based on short passages followed by multiple-choice tests. This description is taken from Venezky et al., (1994):

> The TABE ... [includes tests of] ... vocabulary, reading comprehension, language, spelling, and mathematics abilities. All of these tests were constructed to measure basic skills using skill models for the areas involved.
>
> The TABE is a battery of norm-referenced tests that require multiple-choice responses and is the most frequently used commercial test in adult literacy programs... Each test has four graduated but overlapping levels (Easy, Medium, Difficult, Advanced) with alternate forms available for each. Also available is a locator test for determining the appropriate level for full-scale testing. This locator test includes 25 multiple-choice vocabulary items and 25 multiple-choice arithmetic items and requires 37 minutes for administration.

It may be seen from this account that writing is not assessed in the TABE. This is perhaps the most obvious problem with multiple-choice assessments of basic skills. Among the issues here is the risk that if a skill is not assessed, or is not assessed in sufficient depth, it will either not be taught at all, or be taught inadequately.

Tout and Schmitt (2002) explain that the TABE includes two math sections, *Computation* and *Concepts/Applications* and that to some extent it diagnoses skills.

The TABE, which is based upon a skills model of literacy, can be contrasted with the TALS, also norm-referenced, which uses items stated to be based on functional literacy tasks of a kind likely to be met with in the course of everyday life. As in the International Adult Literacy Survey (IALS), the TALS items are grouped into three types, said to represent prose, document and quantitative literacy. Therefore, the TALS must be subject to the sort of questions about validity and utility for teaching and learning purposes raised about the IALS scales in appendix D.

A third assessment popularly used in the US, which also uses multiple-choice, together with some supply items, is the CASAS (Comprehensive Adult Student Assessment System) – see http://www.casas.org/02TrainingProfDev/PEPFinalNDN.pdf. The CASAS system includes the Beginning Literacy Reading Assessment portion of a Life Skills Assessment, and the Reading portion of the Basic Skills for Assessment in Employability. Students are assessed on their ability to comprehend reading materials such as medical forms, labels, and passages about legal issues and community services. The CASAS items also include what, in other frameworks, would be called quantitative items.

Whereas the skills measured by the TABE were 'obtained by examining ABE curriculum guides' and other materials, and include comprehension and silent vocabulary, the CASAS is based upon competencies. The word 'competency' has several meanings; here it seems to mean 'life-like tasks', a list of more than 300 of which was compiled by the US Department of Labor. These tasks involve both literacy and numeracy.

The multiple-choice and other tests are stated by Kruidenier (2002) to be keyed to these competencies, with suggested instructional material linked with the specified tasks being available. An example competency is *'interpret advertisements, labels, or charts to select goods and services'*. Given the emphasis often placed on transparency of standards, i.e. on the need to stipulate precisely what is required to pass, or, in the behaviourist language often used in connection with criterion-referencing and competence-based assessment, to demonstrate mastery, it is interesting to note how very far from precise this competency is.

Given the assessment methods used in the CASAS system, though the texts which candidates have to read may be drawn from life, or be realistic, thus giving an appearance of validity, the actual activities performed by those taking the tests are not realistic, and will, further, require a range of test-taking skills which, though presumably part of US culture, are relative alien to the culture of the UK.

Kruidenier's analysis is interesting, not least because he explains that all the assessments discussed include aspects of **both** norm- and criterion-referencing. Though these two types of referencing are often contrasted, in fact each form of assessment always implicitly includes an element of the other.

There are questions about how far such functional assessments provide useful formative information for teachers and learners. Kruidenier states that little research has been done in the US into the kinds of informal assessment carried out by adult literacy tutors in the course of teaching, or into links between assessments of the kind listed above and teaching. Tout and Schmitt (2002) report that the situation is similar in respect of numeracy, with research being minimal. Further, the sector seems to share with the English sector the problem that, due to lack of funding, some tutors have only a rudimentary training, and may not be fully aware of the strengths and limitations of various assessment regimes and tests.

As Venezky (1997) has suggested in the US context, some distinctions drawn between functional literacy and 'most other forms of literacy' derive from 'differing interests in basic and applied abilities'. Translated into the English context, these discussions would link into discussions about the relationship between basic skills and key skills, and about how best (and even whether) to ensure that adults in English basic skills classes, whatever the context, are taught using methods most likely to result in transferable, and therefore flexible, basic skills.

The kinds of assessment discussed by Venezky may be contrasted with the kind of performance assessment of the application of basic skills in real life, which schemes like Wordpower have traditionally supported in the UK, in which, in theory, witness statements about reading and writing activities in the workplace may be presented for external accreditation, rather than actual examples of written work.

Tout and Schmitt (2002) mention that, in the US, 'the main drivers of the math curriculum in ABE are the GED [General Educational Development] exam and commercially published workbooks'. The 1998 GED maths test consisted largely of multiple-choice word problems presented in adult contexts, classified as 50 per cent arithmetic, 30 per cent algebra, and 20 per cent geometry. Clearly this test would not be appropriate for the English context. Interestingly, Tout and Schmitt mention work in the United Kingdom, and especially that of Coben, O'Donoghue and FitzSimons (2000), and of Benn (1997), with interest. Like Brooks et al., (2001a), Tout and Schmitt look with interest to developments in Australia.

Mention should also be made of a group of sophisticated norm-referenced tests closely associated both with intelligence testing and with the assessment of dyslexia by psychologists. Sub-sections of some of these tests have been used in studies of progress in adult basic education in the US:

- Wide Range Achievement Test – Revised (WRAT-3) (includes word recognition and spelling) (Wilkinson, 1993).
- The Wechsler Individual Achievement (and other) Tests (Psychological Corporation, 1991).
- The Woodcock Reading Mastery Tests (Rev.) and Woodcock Language Proficiency Battery – Revised (Woodcock, 1991).

Versions of some of these tests adapted for use in Britain are available. None were reviewed here, partly because they are commercially available and therefore not secure, partly because their purposes were too specific.

Conclusions

- Though it is probable that useful insights might be gained from further study of US assessment materials and frameworks, it cannot be recommended that such materials be used for NRDC's research purposes in studies in England.
- If used here, tests from the US should be anglicised, and re-standardised upon relevant English populations.

Appendix C. Review framework

Some notes on the development of the framework are at the end of this appendix.

1 Title of instrument.
Publisher.
Date of publication.
Edition no.

2 Cost of instrument per set/copy.

3 Is the instrument 'secure' (unpublished)?

4 Stated purpose of instrument.

5 Are there parallel forms? If so, how many?

6 Are there any obvious manageability problems?

7 What, if any, stipulations about the training of administrators are made?

8 Time needed/allowed, if known.

9 Is the test available in different media? If so, explain briefly.

10 Is there a clear statement of the population for which the instrument is intended? If so, explain briefly.

11 a) How is the test 'tailored' for the range of achievement, if at all?
b) Is the test pitched at appropriate levels of ease/difficulty for the target learners?

12 If the test is stated to be suitable for ESOL as well as for other learners, what evidence is there of differentiation, including in the reporting of results?

13 Do the materials seem free from bias? List any difficulties.

14 Are the materials suitable for adults?

15 Is there an element of self-assessment in the instrument? If so, explain briefly.

16 Is extraneous knowledge required (e.g. in terms of literacy levels required to comprehend problems/questions, etc./ numeracy/IT)? Comment briefly as appropriate.

17 Does the test report results for:

- ■ reading
- ■ writing
- ■ spelling
- ■ punctuation
- ■ numeracy?

18 a) At what levels are results reported, if reported in levels?

b) If results are reported in different terms, explain briefly.

c) Has the test been explicitly aligned or re-aligned to the new QCA standards?

19 Assessment approach and alignment to the Adult Basic Skills Core Curricula and Standards. Discuss, if and as appropriate, as follows:

Is alignment-

a) broad;

b) covering a range of, or all, elements;

c) suitably balanced in terms of emphasis; and/or

d) appropriate in terms of required depth of skill, knowledge and understanding?

20 Describe the item types used. Comment if appropriate.

21 Do the materials/items appear to test what they are intended to test?

22 What, if any, technical information is available about the test, e.g. pilots, test-retest reliability, reliability between parallel forms, with different assessors, as 'split halves', relationship with results on other instruments, etc. Comment on any likely difficulties in these areas.

23 Is a degree of uncertainty in results, or of limitation in the instrument, recognised, reported, evaluated?

24 Are there any other points that might usefully be made about the instrument, e.g. minor proofing errors, tasks that seem particularly difficult?

25 Rate the usefulness of the reported results to
a) learners;
b) tutors; and
c) the wider world (high, medium or low).

26 Sum up likely positive and negative effects of this instrument in terms of 'backwash'/teaching to the test.

27 Form a judgement about the suitability of the instrument:
(a) for its stated purpose;
(b) for NRDC purposes in
- cohort studies
- intervention studies
(high, medium or low? Comment in text if needed).

The framework is based on the assumption that all assessment should be valid, reliable, manageable and useful, requirements that are, however, 'notoriously impossible to reconcile' (Brooks, 1998). While recognising that he was giving a 'counsel of perfection', Brooks set out six principles for assessment:

- The purpose that the assessment is intended to serve should be clear.
- The instruments should be valid in relation to the intended domain.
- The instruments should be as unbiased as possible.
- The system of assessment should be clear, and only as complicated as needed for the intended purpose.
- Outcome measures should be reliable.
- The experience of taking the assessment should be useful to those assessed, and the outcome measures should be useful both to the subjects of the assessment and to the wider society.

The questions in the framework cover these issues, plus all those in the checklist suggested by Lewandowski and Martens (1990), while also taking into account the adult literacy and numeracy standards and curricula. Most of the questions require no comment; only those for which the justification may not be obvious are commented on here.

Question 9 requires the existence of alternate forms of an instrument in different media to be noted. When different versions were available, question 22 would enable a reader to discover whether studies of the concurrent validity of the different versions had been made.

In theory, adult literacy and numeracy classes in England have often aimed to help learners to meet self-defined needs. Question 15 allows for consideration of whether the instrument provides for the learner to 'self assess'.

La Marca (2001) is the source for question 19. The word 'elements' is used here in a general sense, not in the specific sense it has in the adult literacy and numeracy curricula. Operationalising standards is not straightforward. The sections in this question support analysis of how each instrument has operationalised the standards.

Question 20 is about the type of assessment task used, not about the text types used as stimuli. Multiple-choice, cloze procedure and sentence-completion tasks are examples of types of item.

Face validity is addressed by question 21. If a test does not look to a learner as if it is testing something relevant, his or her performance may be affected. Potential users of the results may also lack confidence in the instrument.

Question 22 covers a number of technical points, including reliability. Lewandowski and Martens (1990) explain that a standardised test should be 'consistent with itself (internal consistency reliability) and consistent over time (test-retest reliability)'. Psychometric theory asserts that a test that lacks reliability cannot be valid.

Generally speaking, two statistical approaches have been used in the assessment of literacy and numeracy: classic test theory and item response models. It is sometimes claimed that item response models (also referred to as item response theories) have the advantage of providing difficulty levels for items that are not tied to particular standardisation populations or to particular collections of test items. (See Cohen et al., 2000, p.325.) Item response models are increasingly used, together with multiple-choice items, in computer-based adaptive language testing in the ESOL field. The IALS also used item response models. The use of either of these methods of measurement would be indicated in the response to question 22.

All assessments include some degree of error. An observed test score is taken to be the true score plus a measurement error. Error has various causes, which may be linked with the instrument, the raters, or the candidates, or with more than one of these. It is not limited to norm-referenced or standardised tests. Question 23 focuses on this aspect of assessment.

The framework is intended to be detailed enough to support a thorough review of any instrument. Inevitably, some of the categories in the framework overlap. Question 24 provides the opportunity to record observations that do not fall neatly into any particular category.

Appendix D. Detailed reviews of assessment instruments

This appendix presents a completed analytical framework and a narrative commentary for each of the instruments listed in table 1 in the main text.

A. Instruments used in previous studies

1. Instruments used in previous lifetime cohort study sweeps

1	Title of instrument. Publisher. Date of publication. Edition no.	Two similar instruments are discussed together. Both have been used in previous cohort studies: Ekinsmyth, C. and Bynner, J. (1994) *The Basic Skills of Young Adults*. London: BSA. Bynner, J. and Parsons, S. (1997) *It Doesn't Get Any Better*. London: BSA.
2	Cost of instrument per set/copy.	N/a
3	Is the instrument 'secure' (unpublished)?	No
4	Stated purpose of instrument.	Assessing skills of two samples of adults as part of longitudinal lifetime cohort studies, with a view to drawing links with other aspects of life.
5	Are there parallel forms? If so, how many?	No
6	Are there any obvious manageability problems?	No
7	What, if any, stipulations about training of administrators are made?	N/a. Accounts of training of administrators are given
8	Time needed/allowed, if known.	About 35 minutes.
9	Is the test available in different media? If so, explain briefly.	No.
10	Is there a clear statement of the population for which the instrument is intended? If so, explain briefly.	See 4 above.
11	a) How is the test 'tailored' for the range of achievement, if at all? b) Is the test pitched at appropriate levels of ease/difficulty for the target learners?	a) See below. b) Not targeted at learners but at cohort samples. Items covered most of the range of difficulty except that there were very few aimed at people above the median of the population.

12	If the test is stated to be suitable for ESOL as well as for other learners, what evidence is there of differentiation, including in the reporting of results?	N/a.
13	Do the materials seem free from bias? List any difficulties.	Broadly. The materials could better reflect the multi-racial nature of society.
14	Are the materials suitable for adults?	Yes.
15	Is there an element of self-assessment in the instrument? If so, explain briefly.	No (measures of self-reported problems are given, but none of the performance measures involve self-assessment).
16	Is extraneous knowledge required (e.g. in terms of literacy levels required to comprehend problems/questions, etc./ numeracy/IT)? Comment briefly as appropriate.	General cultural and background knowledge would be an advantage.
17	Does the test report results for ■ reading ■ writing ■ spelling ■ punctuation ■ numeracy?	Only for reading and numeracy as defined in the (old) ALBSU/BSA Standards. However, some of the literacy tasks in Bynner and Parsons (1997) require ability to extract numerical information from charts or text, and some of the numeracy tasks involve reading text or instructions.
18	a) At what levels are results reported, if reported in levels? b) If results are reported in different terms, explain briefly. c) Has the test been explicitly aligned or re-aligned to the new QCA Standards?	a) & b) Results are analysed and reported in various ways, e.g. numerically and by level based upon analysis of results. c) No
19	Assessment approach and alignment to the Adult Literacy and Numeracy Core Curricula and Standards. Discuss, if and as appropriate, as follows: Is alignment a) broad b) covering a range of, or all, elements c) suitably balanced in terms of emphasis and/or d) appropriate in terms of required depth of skill, knowledge and understanding?	N/a – these instruments were devised and used before the (new) Standards and Curricula existed.

20	Describe the item types used. Comment if appropriate.	Show cards using mainly 'realia' with understanding of ability to scan for stated facts or information tested. Mainly right/wrong answers, with some alternatives provided where appropriate. Some true/false questions. See below.
21	Do the materials/items appear to test what they are intended to test?	Broadly.
22	What, if any, technical information is available about the test, e.g. pilots, test-retest reliability, reliability between parallel forms, with different assessors, as 'split halves', relationship with results on other instruments, etc. Comment on any likely difficulties in these areas.	Pre-testing and piloting was carried out for both studies, but other technical information seems to be unpublished. See below.
23	Is a degree of uncertainty in results, or of limitation in the instrument, recognised, reported, evaluated?	Yes. In particular, it is stated that the assessments may under-represent the level of difficulties. However, it may be more appropriate, and valuable, given the context, to investigate empirically.
24	Are there any other points which might usefully be made about the instrument, e.g. minor proofing errors, tasks that seem particularly difficult?	Some of the tasks now appear out of date. See below.
25	Rate the usefulness of the reported results to a) learners; b) tutors; and c) the wider world (high, medium or low).	a) Low. b) Low. c) Medium in terms of purposes of survey.
26	Sum up likely positive and negative effects of this instrument in terms of backwash/teaching to the test.	N/a.
27	Form a judgement about the suitability of the instrument: (a) for its stated purpose (b) for NRDC purposes in - cohort studies - research projects (high, medium or low? Comment below if needed).	a) Stated purpose: medium b) - Cohort studies: see below. - Research projects: not suitable.

It cannot be recommended that either set of tasks be re-used in their entirety.

The main problems with the re-use of instruments from these studies are:

- the instrument from the earlier study was apparently rejected by those involved in the later survey;
- a number of the items now appear dated; and
- these studies both aimed to report results in line with the BSA Standards, which have been replaced.

The following summary of the cohort studies is based largely on that in Bynner and Parsons (1997).

In the 1990s, the Adult Literacy and Basic Skills Unit (ALBSU), which in 1995 became the Basic Skills Agency (BSA), commissioned several basic skills surveys. Four of the surveys were carried out by the Social Statistics Research Unit at City University, which then moved to the Institute of Education, University of London, and became the Consortium for Longitudinal Studies. Two of the surveys were based on birth cohorts, one born in a week in 1958 (the National Child Development Study, NCDS) – this group was studied at age 37 in 1995, and the second born in a week in 1970 (the British Cohort Study 1970, BCS70) – this group was studied at age 21 in early 1992. In each of these cohorts the full sample consisted originally of over 17,000 individuals. A different set of literacy and numeracy tasks was used in each study. The set used in the BCS70 sweep of 1992 was also used in the cross-sectional *Older and Younger* survey in 1993–94 (BSA, 1995), and the set used in the NCDS5 sweep in 1995 was also used in a cross-sectional survey in Wales in 1995 (BSA, 1997c).

The value of cohort studies, as opposed to studies of adults in provision, is that as well as providing insights into possible relationships between basic skills levels and other aspects of life, they allow a window onto the skills of a broader range of adults; those in adult basic skills tuition may not be typical of those whose skills are broadly within the same range. However, cohort studies by definition study people of just one age in any sweep, and are therefore themselves usefully supplemented by cross-sectional surveys.

Ekinsmyth and Bynner (1994) and Bynner and Steedman (1995) reported on a survey of a 10% sample of the BCS70 cohort members at the age of 21 carried out in early 1992. It included a half-hour assessment of basic skills designed by Cambridge Training and Development Ltd.

The tasks were designed 'specially for the survey to tap the different levels of the ALBSU (now BSA) Standards – Wordpower and Numberpower... For Wordpower these comprise Foundation Level and *three* higher levels. For Numberpower there is a Foundation Level' [later called Entry level] 'and *two* higher levels. The crucial distinction for this report is between Foundation Level signifying only a minimal grasp of the basic skill, and people who were reasonably competent in it. For the former group we go further in identifying, for literacy and numeracy respectively, people whose performance in the assessment was *very poor*, ie. they had barely any literacy or numeracy at all' (Bynner and Parsons, 1997, pp.12–13).

The assessment comprised a series of tasks, some of which required more than one answer, and which were linked to showcards. The materials were presented to respondents by trained administrators working with a script.

The set of nine literacy tasks focused upon reading and spanned four levels of difficulty. These tasks used a range of text types, including a map, an advertisement, graphs, a page from a video manual, and argumentative and fiction writing. The fiction item, understanding of depended heavily upon knowledge of, or ability to guess at, the meaning of the word 'rout'.

The ten numeracy tasks spanned three levels of difficulty, asking cohort members, for example, to calculate change from £20 for a purchase costing £17.89 in the first task, and to compare the cost of a jacket originally costing £200 with a 12 per cent discount with the cost of one originally costing £250 with 1/3 off the price. Cohort members were asked to do these calculations without a calculator if possible, but if they chose to use one, simply the fact was recorded. One of these numeracy tasks seems unfortunately worded; the script requires the interviewer to ask 'how many coins will you hand over to the shopkeeper?' but the required answer is '£14', not a number of coins.

The basic skills assessment might be terminated either at the request of the cohort member, or on the suggestion of the interviewer if the cohort member became uncomfortable, or in the event of a major interruption – the assessments were often carried out during home visits.

The scoring for the materials was objective and manageable. All tasks had answers that could be coded simply as correct or incorrect. Answers were given verbally, and the interviewers recorded them by ticking the appropriate boxes on the interview proformas. Samples of writing were obtained, but were not analysed or scored.

Ekinsmyth and Bynner (1994) explain that the development of the tasks was a key part of the study. Page 12 of their report emphasises that, though the tasks were pre-tested and piloted, there was insufficient time to develop a standardised test, and that while frequent use is made in their analyses of overall aggregate scores, these scores should be seen as 'no more than a means of summarising diverse data'. The drop in mean scores as the levels got more difficult suggested that the operationalisation of the ALBSU/BSA Basic Skills Standards had been broadly successful. Pilot testing showed that most respondents could do all the literacy tasks, though a significant minority had problems with all of them, and a few items discriminated within the group of good readers. The numeracy tasks discriminated across the whole sample. The authors acknowledge (p.15) that 'further work would usefully be done on the development of the instrument', but state that its diagnostic value had been considerable. Perhaps because of the reservations, a new instrument was used for the NCDS age 37 survey.

A survey of a 10 per cent sample of NCDS cohort members was carried out at age 37 in 1995 (Bynner and Parsons, 1997). This survey had two aims: first to test and replicate findings about the disadvantages associated with basic skills problems from the BCS70 age 21 study, and secondly to look at the effect of basic skills difficulties over a more extended period of the life cycle. The survey instruments included a new basic skills assessment using a set of functional literacy and numeracy tasks.

Each task consisted of a show card, or visual stimulus, with a number of questions to be answered, or some exercises to be performed. The tasks were, as in the previous study, designed to correspond to the BSA Wordpower and Numberpower standards, with a reasonable spread of tasks at each level. They were based on a set of 'Tasks and Showcards' developed for ALBSU by the National Foundation for Educational Research in 1991–93 (ALBSU, 1994) and used since then in several basic skills research projects. The general level of difficulty of the items increased through the test. After two pilots, and consultation with the BSA, the number of tasks was reduced to eight for literacy and nine for numeracy. Reference is made in appendix 5 of the report to a writing assessment, but this is not discussed in the report. In fact, though the writing samples were collected, they were never analysed (John Bynner, personal communication, 2000).

Many of the tasks are based upon realistic-looking texts, perhaps to give a sense of familiarity or relevance, or in an attempt to contextualise skills in a realistic way. Thus, for example, the first literacy question is based upon what is intended to look like a leaflet or poster advertising a concert, but the lack of any graphics in the text makes it less than fully realistic. This particular item was also used in the previous study, and was the only item to be retained. The most difficult 'literacy' item includes graphs and charts, and, therefore, arguably involves both literacy and numeracy.

Some of the numeracy tasks attempt to contextualise skills in terms of scenarios in which comparatively large sums of money are spent. It could be argued that, if cohort member respondents are not well off, the use of texts in which people are supposed or invited to imagine spending money on luxurious items is unfortunate.

Several items used were broadly topical at the time, and mention dates, which makes them less suitable than others for re-use without modification. It is reasonable to assume that subjects spending a considerable amount of time answering questions and doing assessments will put more into the task if it has a relevant feel, and the items are interesting. Items that seem out of date are not motivating.

A noticeable difference from the previous study is that no literary passage was used to assess reading.

In this survey, each question was coded as correctly answered, incorrectly answered, or not attempted. This is an improvement in practice over the previous study. When a respondent failed to answer three consecutive questions correctly, the assessment was judged to have been completed and no further questions were asked. While this practice could be justified on the grounds that it spared a respondent who found the tasks difficult an unpleasant experience, it might have influenced the results. The data presented in the report show that some respondents who got answers on easier tasks incorrect succeeded with some tasks at higher levels. However, only three respondents on the literacy tasks, and 12 on the numeracy tasks, fell in this category.

For the other cohort members, total scores were calculated separately for literacy and numeracy. The scoring procedures, as well as the methods for terminating the assessment, differed, therefore, from those in the earlier study. For literacy tasks, the maximum score was 23; for literacy, 18. Scores were re-scaled to fall within the range 0–10, and then grouped into four ability categories: very low, low, average, and good. 'Very low' was 'generally below the Wordpower and Numberpower Foundation Levels' and 'low' is 'at or barely scraping above, Foundation Level' (Bynner and Parsons, 1997). Thus, though the materials aimed to be criterion-referenced to Wordpower and Numberpower levels, there were, as in the previous study, elements of norm-referencing and of researcher judgment in the scoring and reporting procedures.

While it cannot be recommended that either instrument is used in its entirety in future cohort study sweeps, the items seem to have succeeded in discriminating between cohort members quite successfully for the purposes of the projects, and the instruments and procedures seem to form a reasonable basis for the development of an updated version for use in future sweeps.

2. International Adult Literacy Survey

N.B. In this survey, numeracy was assessed in the form of 'quantitative literacy', i.e. arithmetical problem-solving questions based on the same texts as the literacy items (see box 17 below).

1	Title of instrument. Publisher. Date of publication. Edition no.	International Adult Literacy Survey (IALS, 1994–98) in Carey, S. Low, S. and Hansbro, J. (1997) *Adult Literacy in Great Britain*. London: The Stationery Office.
2	Cost of instrument per set/copy.	N/a.
3	Is the instrument 'secure' (unpublished)?	Not published but available from The Data Archive to bona fide researchers wishing to analyse or use the items (see, e.g., Brooks et al., 2001b, next entry).
4	Stated purpose of instrument.	Estimate literacy and numeracy levels of adult population sample.
5	Are there parallel forms? If so, how many?	Not in the usual sense – see next box.
6	Are there any obvious manageability problems?	Although there were far too many items for any one participant to attempt, the items were arranged in seven blocks, of which each participant attempted only three. Also, a Core block of six very simple introductory items was used to screen out people who would not be able to tackle any of the main blocks. However, quite a few of the items had a high ratio of text to questions, thus requiring well-developed scanning abilities.
7	What, if any, stipulations about training of administrators are made?	Very systematic training was given to the market research fieldworkers employed as administrators.
8	Time needed/allowed, if known.	30–45 mins.
9	Is the test available in different media? If so, explain briefly.	No.
10	Is there a clear statement of the population for which the instrument is intended? If so, explain briefly.	Yes – adults aged 16–65 living in Britain.

11	a) How is the test 'tailored' for the range of achievement, if at all?	a) Apart from the Core block (see 4 above), not at all – participants were expected and encouraged to attempt every item in the blocks they were presented with, even the most difficult.
	b) Is the test pitched at appropriate levels of ease/difficulty for the target learners?	b) Not targeted at learners but at a population sample. Items covered most of the range of difficulty except that there were very few that those with very poor skills could consistently get right.
12	If the test is stated to be suitable for ESOL as well as for other learners, what evidence is there of differentiation, including in the reporting of results?	It seems to have been assumed that ESOL participants, being few in number, would have to tackle the items on the same basis as monolinguals.
13	Do the materials seem free from bias? List any difficulties.	All items were based on real-life texts, translated from the source languages into all the others needed. As a result, many items are presented in formats that are not quite the familiar ones – but in an international survey items that replicated too closely the conventions of one country would be unsuitable in all others, so this was an attempt to use texts that were both authentic and neutral.
14	Are the materials suitable for adults?	Yes.
15	Is there an element of self-assessment in the instrument? If so, explain briefly.	No.
16	Is extraneous knowledge required (e.g. in terms of literacy levels required to comprehend problems/questions, etc./numeracy/IT)? Comment briefly as appropriate.	See next box.
17	Does the test report results for ■ reading ■ writing ■ spelling ■ punctuation ■ numeracy?	Only for reading and numeracy. The items were classified as assessing 'prose literacy' (non-arithmetical questions based on continuous texts), 'document literacy' (non-arithmetical questions based on non-continuous texts, e.g. maps, timetables), and 'quantitative literacy' (arithmetical questions based on the same texts as the other categories, e.g. filling in an order form for theatre tickets, including calculating the total cost, given the required details in an

		advertisement). 'Quantitative literacy' can be taken as a proxy for numeracy where that takes the form of arithmetical problems embedded in text, but it does not cover pure calculation.
18	a) At what levels are results reported, if reported in levels? b) If results are reported in different terms, explain briefly. c) Has the test been explicitly aligned or re-aligned to the new QCA Standards?	a) On a five-level scale pragmatically derived in earlier US surveys of adult literacy and numeracy, from 1 (low) to 5 (high). Because so few people scored at level 5, levels 4 and 5 were usually merged for reporting purposes. b) Also reported on a 0–500 scale – see below. c) Broadly, since IALS level 1 corresponds to Entry level of England's new National Standards, IALS level 2 to England's level 2, etc. (see Brooks et al., 2001b, esp. pp.121–22).
19	Assessment approach and alignment to the Adult Basic Skills Core Curricula and Standards. Discuss, if and as appropriate, as follows: Is alignment – a) broad; b) covering a range of, or all, elements; c) suitably balanced in terms of emphasis; and/or d) appropriate in terms of required depth of skill, knowledge and understanding?	For the reasons given immediately above, these questions are not strictly relevant. However, approximate answers would be: a) Yes. b) A wide range but not pure calculation. c) Yes. d) Yes.
20	Describe the item types used. Comment if appropriate.	No multiple choice – all items were supply/constructed response type, requiring only short answers and therefore minimal writing skills.
21	Do the materials/items appear to test what they are intended to test?	Broadly, yes.
22	What, if any, technical information is available about the test, e.g. pilots, test-retest reliability, reliability between parallel forms, with different assessors, as 'split halves', relationship with results on other instruments etc. Comment on any likely difficulties in these areas	Considerable piloting was done. Reliability as envisaged in this question was not an issue for this one-off instrument. Some of the items had been used in previous US surveys but no results are reported on this. Some of the items were re-used in the *Progress in Adult Literacy* study – see next entry.

23	Is a degree of uncertainty in results, or of limitation in the instrument, recognised, reported, evaluated?	Confidence intervals are routinely reported where appropriate.
24	Are there any other points which might usefully be made about the instrument, e.g. minor proofing errors, tasks that seem particularly difficult?	No.
25	Rate the usefulness of the reported results to a) learners; b) tutors; and c) the wider world (high, medium or low).	 a) Low. b) Low. c) High – much use has been made of these results for policy-making, not least in the Moser Report.
26	Sum up likely positive and negative effects of this instrument in terms of backwash/ teaching to the test.	N/a.
27	Form a judgement about the suitability of the instrument: a) for its stated purpose; b) for NRDC purposes in – cohort studies – research projects (high, medium or low? Comment below if needed).	 a) High. b) – for cohort studies: Low – for research projects: Low Cohort studies need items that are more clearly UK-based, and the need for individual administration, and the lack of parallel forms, make the tests unsuitable for research projects.

The International Adult Literacy Survey (IALS), an OECD-supported initiative at first chiefly centred on Canada and the United States, took place in three phases, in 1994, 1996 and 1998. It was the first international survey of adult literacy and one of the first two international surveys of adult numeracy. (In 1996, simultaneously all four parts of Britain participated in the second phase of IALS and the three parts of Britain participated in the International Numeracy Survey – for the latter see BSA, 1997d. The International Numeracy Survey instrument is not covered in this report because it consisted of very few items and was probably not very reliable. For what it is worth, the British sample had the lowest average score of the seven nations participating.)

In IALS, within each of the domains (prose, document, quantitative) each item and each participant was attributed a score on a 0–500 scale, using Item Response Modelling. For items, this represented an estimate of ease/difficulty; for participants, an estimate of attainment. The definitions were reciprocal: for example, a score of 300 for an item meant that participants with that score had an 80 per cent probability of getting it right, and for participants with that score that they had an 80 per cent probability of getting items with that score right. Once participants and items had been attributed scaled scores, the scale was divided into the five levels derived from earlier US surveys (see above). In theory, an individual could be at one level on one IALS scale, and a different level on another scale.

The extent to which the IALS domains have been unequivocally established is controversial – see especially Hamilton and Barton (2000). The statistical techniques used assume the existence of unidimensional traits; they do not demonstrate their existence. The fact that items which do not fit the model are excluded from scales, rather than being used as a basis for further investigation, perhaps in search of other scales, if a psychometric approach is taken, or in terms of what a student has failed to learn if the interest is educational, is a criticism that has frequently been made of the use of Item Response Models in educational assessment.

Despite these problems with the IALS scales, there is evidence that the idea that 'document literacy' exists is being accepted and elaborated by some adult basic skills practitioners. For example, Macrae (1999) describes prose, document and quantitative literacies as 'ways of coding' meanings, and as 'codes'.

The IALS system is not suitable for either of NRDC's proposed research purposes. The items are easily available, which makes them unsuitable for research projects. The approach to the literacy domains taken in the survey is so different from that in previous cohort studies that it would be difficult to maintain continuity with those studies.

3. Progress in Adult Literacy study

1	Title of instrument. Publisher. Date of publication. Edition no.	Brooks, G. et al., (2001b) *Progress in Adult Literacy: Do Adults Learn?* London: Basic Skills Agency. N.B. The text cited is a report, not a set of assessment materials. The writing tasks used in the study are reproduced in the report. The reading tests used are not but were available to the reviewer (KH) because the director of this review project (GB) also directed the *Progress in Adult Literacy* study and had a full archive set of the tests, thus allowing completion of all details here.
2	Cost of instrument per set/copy.	N/a.
3	Is the instrument 'secure' (unpublished)?	Writing tasks: no – published in report. Reading tests: only partly. A few items were newly developed for the research task and are therefore fully secure, but most were borrowed from previous studies: – IEA 1991 survey of reading at age 9 (Elley, 1992); not published but available to researchers – BSA/NFER test item bank (ALBSU, 1994) (see above under previous cohort study sweeps); secure except that some items were used in the following study. – *Lost Opportunities* survey of linguistic minority adults (1994–95) (Carr-Hill et al.,

		1996); published, but not well known. – IALS (see above); not published but available to researchers. Only the 25 simplest IALS items were used. The five newly written items were all based on the IALS texts used.
4	Stated purpose of instrument.	Summative assessment for survey purposes.
5	Are there parallel forms? If so, how many?	Writing: no. Reading: two forms, each with two opening sections, one 'simple' and one 'very simple', and a common main section.
6	Are there any obvious manageability problems?	Yes – the reading tests were administered individually and some learners took over an hour to complete them.
7	What, if any, stipulations about training of administrators are made?	None. The reading tests were administered by fieldworkers, the writing tasks by the adult literacy learners' tutors.
8	Time needed/allowed, if known.	Untimed; see also 6 above.
9	Is the test available in different media (paper and electronic)? If so, explain briefly.	No.
10	Is there a clear statement of the population for which the instrument is intended? If so, explain briefly.	Adult basic skills literacy learners.
11	a) How is the test 'tailored' for the range of achievement, if at all? b) Is the test pitched at appropriate levels of ease/difficulty for the target learners?	a) Writing: not tailored – all learners tackled same task. Reading: Learners were assigned to levels of the opening section using tutors' recommendations. b) Yes, except that there were very few reading items that the poorest (Entry Level) readers could tackle.
12	If the test is stated to be suitable for ESOL as well as for other learners, what evidence is there of differentiation, including in the reporting of results?	The survey was based on mainstream adult literacy provision, not ESOL provision. Despite this, 18 per cent of the participants had English as an additional language – they were assessed in exactly the same way as the native speakers.
13	Do the materials seem free from bias? List any difficulties.	Some of the texts used in the reading tests seem less than fully suited to British learners – see below.

14	Are the materials suitable for adults?	Some of the texts used in the reading tests seem less than fully suited to adult learners – see below.
15	Is there an element of self-assessment in the instrument? If so, explain briefly.	No.
16	Is extraneous knowledge required (e.g. in terms of literacy levels required to comprehend problems/questions, etc./ numeracy/IT)? Comment briefly as appropriate.	General cultural and background knowledge would be an advantage for the reading tests.
17	Does the test report results for ■ reading ■ writing ■ spelling ■ punctuation ■ numeracy?	Reading and writing.
18	a) At what levels are results reported, if reported in levels?	a) Writing: Not reported by levels. Reading: reported by levels from New Standards Entry Level 1 (or below) to Level 3 (criterion-referenced).
	b) If results are reported in different terms, explain briefly.	b) Writing: only by length of text written, in words, and quality of handwriting. For reading, norm-referenced results were also reported. The known British facility values of the IALS items were used to scale all items to the IALS scale.
	c) Has the test been explicitly aligned or re-aligned to the new QCA Standards?	c) Writing: no. Reading: yes. The new Standards were made available to the researchers part way through their project, and it was known that the New Standards Entry Level/Level 1 boundary had been deliberately set to correspond to the IALS Level 1/Level 2 boundary. By extrapolation from the score corresponding to that boundary, bands of scores on the IALS scale were imputed to the new Standards levels stated in a) above.

19	Assessment approach and alignment to the Adult Basic Skills Core Curricula and Standards. Discuss, if and as appropriate, as follows: Is alignment – a) broad; b) covering a range of, or all, elements; c) suitably balanced in terms of emphasis; and/or d) appropriate in terms of required depth of skill, knowledge and understanding?	1) Writing: n/a – the tasks were designed before the new Standards and Curriculum were devised. 2) Reading: a) broad b) covering a range of elements. c) See below. d) Yes.
20	Describe the item types used. Comment if appropriate.	Writing: Single-sentence prompts. Reading: A mixture of open-ended and multiple-choice items. Text types mostly 'prose' and 'document'. One narrative text was used. See below.
21	Do the materials/items appear to test what they are intended to test?	See below.
23	Is a degree of uncertainty in results, or of limitation in the instrument, recognised, reported, evaluated?	Yes. See e.g. Brooks et al., (2001b), p.95.
24	Are there any other points which might usefully be made about the instrument, e.g. minor proofing errors, tasks that seem particularly difficult?	See below.
25	Rate the usefulness of the reported results to a) learners; b) tutors; and c) the wider world (high, medium or low).	a) Low. b) Medium. c) Medium.
26	Sum up likely positive and negative effects of this instrument in terms of backwash/ teaching to the test.	N/a since instrument not released.
27	Form a judgement about the suitability of the instrument– a) for its stated purpose; b) for NRDC purposes in – cohort studies – research projects (high, medium or low? Comment below if needed).	a) Medium. b) – cohort studies: low. – research projects: low. Cohort studies need items that are more clearly UK-based, and the need for individual administration, and the lack of parallel forms, make the tests unsuitable for research projects.

This report is an account of a study of the progress made in reading and writing by adults in adult basic skills literacy classes. Appendix B of the report (p. 95) sets out the reporting requirements for the reading part of the study, which were that the tests should provide both norm- and criterion-referenced information of particular kinds. Parallel forms were required for an AB/BA design, to avoid any improvements in scores being attributable either to one version of the test being easier than the other or to practice effects. At initial consultation meetings it was further agreed that each version of the test 'should have two entry points, one simple and the other very simple, so that adult literacy tutors could indicate which level of test each of their students should attempt' (p. 95).

> "It was concluded very early in the project that no existing instrument could meet even one of the [project's design] requirements, let alone all [of them]. There was also insufficient time to develop new tests. It was therefore decided to devise a set of tests by borrowing existing items." (Brooks et al., 2001b, p. 95)

For the sources of the existing items see box 3 above. Comments on the reading tests in draft were invited from experts in the field (Brooks et al., 2001b, Acknowledgments).

The use of items from these particular studies seems to have been determined partly by the requirement to provide both norm- and criterion- referenced information, and statistical techniques were used to translate all the results into a common scale. Therefore, the final results depend heavily upon statistical norm-referencing techniques.

The final structure of the tests was as follows:

	Form A		Form B	
	Form AL	Form AH	Form BL	Form BH
Opening Section	10 very simple items	10 simple items	10 very simple items	10 simple items
	↓	↓	↓	↓
Main Section	16 IALS(-based) items		14 IALS(-based) items	

Four modes of response were used: circling part of a text, giving a verbal answer to the tester, ticking of multiple-choice items, and writing open-ended responses in answer to factual comprehension questions. 'Item' in the above chart generally means 'question'. Some items had more than one part, so it was necessary to find more than one piece of information to complete the item successfully. Some of the items had had levels ascribed to them in the course of the IALS study. The items based upon a particular text vary in level. Thus, the item requiring interpretation of a line graph about the number of injuries in firework accidents is IALS Level 2 (UK Level 1), but the question based upon a chart in the same text showing the value, in UK pounds, of fireworks sold in the Netherlands in 1991 is IALS Level 3 (UK Level 2).

In Form AH, the first five items are factual comprehension questions about a set of instructions for planting seed sticks, and the next five are multiple-choice questions based on a story called 'The Bird and the Elephant'. The remaining items are based upon six texts used in IALS: a *unicef* advert, a scrambled egg recipe, an employment application form, charts about victims of firework accidents in the Netherlands, a prose article about the first person to swim round Manhattan three times, and the label from a Canadian bottle of aspirin.

Form BH opens with five questions based on a gallons-to-litres conversion table, followed by questions based on a letter from one neighbour to another that were also used in one set of the cohort study materials reviewed above. The remaining items in form BH are based upon five texts used in IALS: a notice of union council election results, an advert for *Les Misérables,* a chart about nuclear waste, a chart comparing the proportion of women teachers in the Netherlands with the proportions in other countries, and an article about preventing and fighting fires.

An attempt was made to ensure similarity in response modes between the parallel forms, but the forms were not fully balanced in terms of number of items, item type or text types.

Also, some of the texts seemed less than fully suited to adult British learners. In some cases this resulted from their being translations from other languages (e.g. both articles from the Netherlands); other items, chosen because they were known to be at simple, or very simple, levels, had very much the feel of children's material through being derived from a school-level survey (e.g. the tasks based upon a map and upon a story).

The method of using two opening sections per version, one simple and one very simple, was negotiated with tutors at initial meetings and is another factor rendering the parallelism of the forms problematic. The point of this was to ensure that learners working towards or within Entry Level 1 had some items that they could complete, an aim that the authors admit (p. 65) was not wholly achieved. It could be seen as an alternative to giving students at different levels wholly different tests based either upon an initial screening instrument or tutor judgement. Despite the test being tailored in this way, it seems that most students persevered with items at higher levels. Given that the difficulty of items did not increase in a straightforward way through the test, this was reasonable. However, it raises questions about how to tailor a test in such a way as to avoid putting learners who find a subject difficult through a potentially distressing and off-putting experience.

The choices, leaving aside computerised algorithmic 'adaptive' shaping of tests for the moment, seem to be broadly as follows:

- give differentiated papers based upon either tutor suggestion or the results of an initial screen;
- give a common paper in which the general level of difficulty increases and stop the session after a set number of consecutive incorrect responses (this was essentially the procedure used in the cohort studies discussed above);
- give a common paper in which the general level of difficulty increases and allow the student to decide when to stop;
- give a paper in which the level of difficulty of items is deliberately varied, explain this to the students, and ask the students to do as much as they can, leaving to them the decision about when to terminate the session (this was the approach used in IALS).

Much will depend upon what is known about the difficulty of items, and upon the nature of the alignment of the test with the Standards and Curriculum.

Greg Brooks is on record as saying that the reading tests were 'cobbled together' and therefore 'ramshackle', and nothing in this analysis need modify those opinions.

Nevertheless, the tests appear to have served the stated purpose adequately.

The writing prompts used were very basic (pre-test: 'Please say a bit about what you hope to learn here'; post-test: 'Please say a bit about what you have learnt here'). Because the writing tasks had to be applicable in adult literacy classes all over England and Wales, and be standard for all participants, they could neither offer alternatives nor take account of local interests or teaching. As a result they provided beginning writers with little support. For many learners the problem of thinking of something to say is as big a problem as having the courage to put pen to paper to say it. Fear of making mistakes is a powerful motive for writing nothing, and some of the learners in the study did in fact write little or nothing. The use of pictorial and visual stimuli, perhaps based on everyday scenes familiar to most students, such as town centres, might have provided more support for the learner whose mind empties itself at the first sight of a piece of white paper.

The main point in favour of the assessment of writing in this research is that, in contrast with many other materials reviewed here, an actual sample of writing was obtained.

The marking scheme, though detailed, was not quite straightforward, and focused on technical aspects of the writing to the exclusion of the communication of thoughts or ideas. In future research, consideration should be given to finding more supportive ways of assessing writing, and to making the criteria clear to tutors and students: thus any 'teaching to the test' would be of a positive rather than a negative kind, and students would be encouraged to write. Length of script (the main statistically significant outcome in the writing section of this study) might be a more useful criterion in this context.

In the school sector, one of the arguments put forward for the use of coursework was that in examinations candidates have no chance to plan, draft and edit their work. Those candidates who are good at producing high-quality first drafts are at an advantage. There are questions here that should be thought through if and when new assessment materials are produced. If candidates are not given the chance to polish their writing, this must be made clear in any report.

After the publication of the research report, NIACE commissioned Professor Mary Hamilton to write a critique of it, which was published, first on the internet, and then in hard copy (Hamilton, 2001). Greg Brooks' response was also published in both forms (Brooks, 2001a, b). Since the reading tests have not been published and were not available to her, Hamilton was able to comment on them only in general terms. She found 'no problem in principle with combining items from different sources', but thought that items might not be as discriminating when used with different populations. She saw 'changing the criteria against which some of the items are referenced mid-way through the study' as an additional difficulty. This is a complicated area, complicated further by the difficult terms of reference given to the original researchers. However, it is not correct to say that the criteria against which 'some' of the items were referenced were changed part way through the study. **One** of the sets of criteria against which **all** the items had to be 'referenced' was changed: the Standards (from the old, ALBSU set to the new, QCA set). The Levels within the new Standards were also supposedly aligned 'as a design feature' with the Levels of a quite different set of scales – the IALS scales. A problem here, as explained in the *Adult Literacy in Britain* report on IALS, is how to reconcile scales of literacy based upon differing views of the domain. 'Without undertaking a detailed analysis of tasks representative of each level for the different standards it is not possible to establish the congruence of the different hierarchies' (Carey et al., 1997, p. 16).

The authors of the report themselves acknowledge weaknesses of the instruments, for example that 'there were ... too few reading items which the very weakest students could manage' (p. 65), and that 'Students within Entry Levels 1 and 2 of the new Basic Skills Standards (29 per cent of the full sample) could not generally manage any of the test items' (p. 66). The question of whether an externally devised test could be expected to register the progress made by adults at or working towards Entry Level 1 of the new Standards is dealt with elsewhere in this report.

For the reasons given in box 27, the reading tests and writing tasks used in the *Progress in Adult Literacy* study cannot be considered suitable for use in NRDC's projects.

4. Programme for International Student Assessment

N.B. In this survey, numeracy was assessed in the form of 'mathematical literacy', i.e. arithmetical problem-solving questions based on the texts (see Box 17 below).

1	Title of instrument. Publisher. Date of publication. Edition no.	PISA in *Knowledge and* ***Skills for Life****: first results from PISA 2000* (OECD, 2001)
2	Cost of instrument per set/copy.	N/a.
3	Is the instrument 'secure' (unpublished)?	Mostly – some samples of the tasks are in the report and more at www.pisa.oecd.org
4	Stated purpose of instrument.	Assessment of attainment in literacy (mainly) and maths (science was also assessed) in an international sample of 15-year-olds 'to measure how well young adults ... are prepared to meet the challenges of today's knowledge societies' (p.14).
5	Are there parallel forms? If so, how many?	Not in the usual sense – see next box.
6	Are there any obvious manageability problems?	Although there were far too many items for any one participant to attempt, the items were arranged in several blocks, of which each participant attempted only 2 or 3. However, quite a few of the items had a high ratio of text to questions, thus requiring well-developed scanning abilities.
7	What, if any, stipulations about training of administrators are made?	Very systematic training was given to the market research fieldworkers employed as administrators.
8	Time needed/allowed, if known.	2 hours.

9	Is the test available in different media? If so, explain briefly.	No.
10	Is there a clear statement of the population for which the instrument is intended? If so, explain briefly.	15-year-olds.
11	a) How is the test 'tailored' for the range of achievement, if at all? b) Is the test pitched at appropriate levels of ease/difficulty for the target learners?	a) Difficulty of items covers full range. b) Yes.
12	If the test is stated to be suitable for ESOL as well as for other learners, what evidence is there of differentiation, including in the reporting of results?	Not so stated – students were mainly tested in the national language of the various countries and speakers of other languages were tested in exactly the same way as native speakers.
13	Do the materials seem free from bias? List any difficulties.	Yes – extensively trialled in all languages and checked for cultural acceptability.
14	Are the materials suitable for adults?	Many would be, though designed for 15-year-olds.
15	Is there an element of self-assessment in the instrument? If so, explain briefly.	No.
16	Is extraneous knowledge required (e.g. in terms of literacy levels required to comprehend problems/questions, etc./numeracy/IT)? Comment briefly as appropriate.	No.
17	Does the test report results for ■ reading ■ writing ■ spelling ■ punctuation ■ numeracy?	Within literacy, only for reading 'Mathematical literacy' was treated as a proxy for numeracy.
18	a) At what levels are results reported, if reported in levels? b) If results are reported in different terms, explain briefly. c) Has the test been explicitly aligned or re-aligned to the new QCA Standards?	a) At 6 IALS-like levels, from below Level 1 to Level 5, based on score ranges on a scale with a mean of 500 and standard deviation of 100. b) N/a. c) N/a because this was an international survey.

19	Assessment approach and alignment to the Adult Basic Skills Core Curricula and Standards. Discuss, if and as appropriate, as follows: Is alignment – a) broad; b) covering a range of, or all, elements; c) suitably balanced in terms of emphasis; and/or d) appropriate in terms of required depth of skill, knowledge and understanding?	N/a because this was an international survey.
20	Describe the item types used. Comment if appropriate.	Some multiple-choice but mainly supply (open-ended).
21	Do the materials/items appear to test what they are intended to test?	Yes.
22	What, if any, technical information is available about the test, e.g. pilots, test-retest reliability, reliability between parallel forms, with different assessors, as 'split halves', relationship with results on other instruments etc,. Comment on any likely difficulties in these areas.	Very detailed and professional – see p. 220 of *Knowledge and **Skills for Life*** and the *PISA 2000 Technical Report.*
23	Is a degree of uncertainty in results, or of limitation in the instrument, recognised, reported, evaluated?	Yes.
24	Are there any other points which might usefully be made about the instrument, e.g. minor proofing errors, tasks that seem particularly difficult?	No.
25	Rate the usefulness of the reported results to a) learners; b) tutors; and c) the wider world (high, medium or low).	a) Low. b) High. c) High.
26	Sum up likely positive and negative effects of this instrument in terms of backwash/ teaching to the test.	N/a since tests mainly unpublished.
27	Form a judgement about the suitability of the instrument: a) for its stated purpose; b) for NRDC purposes in – cohort studies – research projects	a) High. b) – cohort studies: low – research projects: low

(high, medium or low? Comment in text if needed).	Cohort studies need items that are more clearly UK-based, and the need for individual administration, and the lack of parallel forms, make the tests unsuitable for research projects.

The OECD is the sponsor of the Programme for International Student Assessment (PISA). PISA is intended to mount a survey of the skills of 15-year-olds every three years. The first survey was carried out in 2000, when more than quarter of a million students took part in 32 countries, including 9,340 in England and Scotland (Wales and Northern Ireland did not participate). Another large cohort (the exact number seems not to have been reported yet) took part in a further 13 countries in 2002.

PISA covers attainment in reading, maths and science (labelled respectively as 'reading literacy', 'mathematical literacy' and 'scientific literacy'). In each round, one of these areas will be the principal focus of assessment and the others subsidiary: in 2000 (and 2002) the main focus was reading, in 2003 it was maths, in 2006 it will be science, and in 2009 reading again. So far, details are available only on the 2000 round. Science will not be mentioned again in this review.

> "Reading literacy is defined in PISA as the ability to understand, use and reflect on written texts in order to achieve one's goals, to develop one's knowledge and potential and to participate effectively in society. This definition goes beyond the notion that reading literacy means decoding written material and literal comprehension... PISA 2000 employed about 140 items representing the kinds of reading literacy that 15-year-olds would require in the future."
>
> (OECD, 2001, p.21)

Later (p. 34), the report states that:

> "The concept of reading literacy in PISA has three dimensions ...: the type of reading task, the form and structure of the reading material, and the use for which the text was constructed. Personal competence is best understood in terms of the first of these. The other two are properties of the task materials that were helpful in ensuring that a range of diverse tasks were included..."

The text types used are defined on p. 22 as continuous prose passages (such as narration, exposition and argumentation) and non-continuous texts, including lists, forms, graphs and diagrams.

Three scales were used to measure students' performance on the reading tasks. These covered (briefly) retrieving information; constructing meaning and making inferences; and relating text to their knowledge, ideas and experiences.

The report defines mathematical literacy as:

> "the capacity to identify, understand and engage in mathematics, and to make well-founded judgements about the role that mathematics plays in an individual's current and future private life, occupational life, social life with peers and relatives, and life as a constructive, concerned and reflective citizen ... also implies the ability to pose and

> solve mathematical problems in a variety of situations, as well as the inclination to do so, which often relies on personal traits such as self-confidence and curiosity ...”

The areas of mathematical literacy for assessment – content and process in any situation in which mathematics is used – led to the choice of tasks in order of closeness to the student – private life, personal and school life, work and sports, local community and society, scientific – authentic tasks, even if fictional. A combination of question types was used. A number of units presented a situation or a problem on which the student would be set several tasks. Different combinations of diagrams and written information introduced each unit. Two-thirds of these could be marked right or wrong. Students demonstrated proficiency by answering problems correctly and showing whether they understood the underlying mathematical principles involved in the task.

Overall, the detail and effort put into PISA seem very impressive, and the results reveal a great deal about international comparisons. In particular, adult literacy and numeracy could benefit from domain specifications and test development as rigorous as those applied here. But the tests, however well designed for their own purpose, would not be suitable as they stand for either cohort study sweeps or research projects.

B. Other paper-based instruments

5. Assessing Progress in Basic Skills: Literacy

1	Title of instrument. Publisher. Date of publication. Edition no.	Basic Skills Agency (1997). *Assessing Progress in Basic Skills: Literacy*. London: Basic Skills Agency.
2	Cost of instrument per set/copy.	Assessment pack £25.00.
3	Is the instrument ‘secure’ (unpublished)?	No.
4	Stated purpose of instrument.	To track small gains in progress made by adults in basic skills classes.
5	Are there parallel forms? If so, how many?	Yes. Three at each level.
6	Are there any obvious manageability problems?	No.
7	What, if any, stipulations about training of administrators are made?	None, but the pack is intended for use by tutors, rather than as a test.
8	Time needed/allowed, if known.	Around 40 minutes.
9	Is the test available in different media? If so, explain briefly.	No.
10	Is there a clear statement of the population for which the instrument is intended? If so, explain briefly.	‘It is designed primarily for use with adults at relatively low levels of literacy’. (Guidance Booklet, p. 4).

11	a) How is the test 'tailored' for the range of achievement, if at all? b) Is the test pitched at appropriate levels of ease/difficulty for the target learners?	a) An initial 'diagnostic' or 'placement' test is given. This involves graded reading passages and, for those who complete the first reading exercise, a form-filling exercise. b) Yes.
12	If the test is stated to be suitable for ESOL as well as for other learners, what evidence is there of differentiation, including in the reporting of results?	N/a.
13	Do the materials seem free from bias? List any difficulties.	The test might present particular difficulties for some ESOL learners in terms of cultural background knowledge required.
14	Are the materials suitable for adults?	Yes.
15	Is there an element of self-assessment in the instrument? If so, explain briefly.	No, but it is stated that answers should be reviewed with the student.
16	Is extraneous knowledge required (e.g. in terms of literacy levels required to comprehend problems/questions, etc./numeracy/IT)? Comment briefly as appropriate.	No.
17	Does the test report results for ■ reading ■ writing ■ spelling ■ punctuation ■ numeracy?	Reading and writing Assessment of writing reported in terms of style and language, grammar, punctuation, spelling, legibility, vocabulary, layout.
18	a) At what levels are results reported, if reported in levels? b) If results are reported in different terms, explain briefly. c) Has the test been explicitly aligned or re-aligned to the new QCA Standards?	a) At 4 levels, A-D, where D was intended to be roughly equivalent to the old Wordpower Level 1, and A-C might be considered equivalent to subdivisions of the old Foundation (later Entry) level. b) N/a. c) N/a – instrument devised before the new Standards.

19	Assessment approach and alignment to the Adult Basic Skills Core Curricula and Standards. Discuss, if and as appropriate, as follows: Is alignment a) broad b) covering a range of, or all, elements c) suitably balanced in terms of emphasis and/or d) appropriate in terms of required depth of skill, knowledge and understanding?	Not aligned to new standards. See below.
20	Describe the item types used. Comment if appropriate.	A mixture of objective tasks and open-ended writing tasks requiring more general evaluation.
21	Do the materials/items appear to test what they are intended to test?	The materials are a good mixture of continuous prose and other reading materials.
22	What, if any, technical information is available about the test, e.g. pilots, test-retest reliability, reliability between parallel forms, with different assessors, as 'split halves', relationship with results on other instruments, etc. Comment on any likely difficulties in these areas.	Pilots were conducted, but no information about these was available to the reviewer. In particular, there was no information about whether the instrument was administered, as the guidance booklet suggests, at certain intervals of time, and successfully reflected progress within those time intervals. N/a.
23	Is a degree of uncertainty in results, or of limitation in the instrument, recognised, reported, evaluated?	Yes.
24	Are there any other points which might usefully be made about the instrument, e.g. minor proofing errors, tasks that seem particularly difficult?	The assessment grids used for writing were found a little difficult to use. See below.
25	Rate the usefulness of the reported results to a) learners; b) tutors; and c) the wider world (high, medium or low).	a) Medium. Might be discouraging for those making slow progress. b) Medium. c) Low.
26	Sum up likely positive and negative effects of this instrument in terms of backwash/ teaching to the test.	Positive: tutors and learners are encouraged to do some continuous writing. Potential negative effect: emphasis on technical 'competence' at lower levels might be discouraging for learners at those levels.

27 Form a judgement about the suitability of the instrument:	
a) for its stated purpose;	a) Medium.
b) for NRDC purposes in	b)
– cohort studies	– cohort studies: not suitable. Upper end of range not covered.
– research projects	– research projects: not suitable. Instrument is not secure.
(high, medium or low? Comment below if needed).	

This set of assessment materials is designed to help to track the progress made by learners in adult basic skills classes. It has four levels, A-D, with three versions at each level. The highest level is intended to be roughly at the old Wordpower Level 1.

This instrument was not found to be appropriate for either of the proposed research purposes. The range is inappropriate for cohort study purposes, and it is published, which makes it unsuitable for research projects.

The instrument aims to assess a student's grasp of the following areas:

- Reading comprehension;
- Factual/business writing skills (including letter writing); imaginative writing;
- Spelling and Punctuation, including proof reading; and
- Reading and interpretation of forms, tables and other information displays. Completion of forms.

Grammar and handwriting are assessed through the writing task. Grammar is also assessed through proof-reading exercises in which learners are asked to re-write sentences correctly.

If the aim of adult basic skills classes to help adults to make broad progress in their skills, then a positive aspect of this instrument is that a variety of items is used, including both prose passages and other real-life documents. On the other hand, if classes are organised on a task-specific basis, with the aim of helping learners cope with some pressing real-life literacy task, then the instrument would be less suitable.

Generally, the reading tasks focus on ability to locate factual information within the text. There is no requirement for inference, analysis, or evaluation of texts, or for the interpretation of images or other graphical information.

Assessment is of two kinds: marks based upon objective comprehension and other tasks, and completion of assessment grids for written work. The grids have the following categories listed down the left hand side*: task completed, appropriate style and language, grammar, punctuation, spelling, legibility, vocabulary* and *layout*. Three headings appear at the top of the grid: *Needs Attention, Needs More Work, and Competent*. Each heading is sub-divided into three subsections, resulting in nine squares across each category. These squares are shaded to show the achievement of the learner. Thus, four shaded squares on the vocabulary row would indicate that vocabulary needed 'more work'. This appears to be a development of the assessment grids in *How's It Going?* and the *Progress Profile* (see section A.2).

The level is determined in two stages: learners move up a level if their score on the initial items reaches a certain level **and** 'most items on the grid are in the 'competent' column and none are in the 'needs attention' one.' The problem here is that levels of competence at each level are not clear. An assumption has been made, supported by some versions of the BSA Standards, that for adults in basic skills classes progress consists of being able to do simple things perfectly, then more complex things perfectly and so on. Though this view seems logical, and is tempting for that reason, it is not in accordance with the actual ways in which writers make progress in writing.

The guidance booklet provides examples of student work that have been assessed as being at various levels, together with completed assessment grids. The reviewer was not always clear why particular pieces of work had been assessed as being at particular levels. For example, a short piece assessed as being at Level D appeared to have less merit than a longer piece assessed as being at Level C.

It seems possible that the grading method, with its combination of objective marks for reading and exercises and more impressionistic marking of writing, needs to be re-examined. It might be better, for example, to consider reporting levels separately for writing and for reading.

Assessment of writing is a difficult and complex matter, but it could be argued that in this case the use of exactly the same categories for the assessment grid at all levels has been unhelpful. In the absence of more information about the pilots, it seems reasonable to suppose that final decisions about levels may have been influenced by performance on other aspects of the test and by more holistic judgments made on the basis of professional experience of working with adult learners and the Wordpower Levels within which the assessments were intended to operate.

This assessment pack is not aligned to the new Adult Literacy Standards and curriculum. Revision seems unlikely in view of the availability of new tests. Though it has certain attractive features, such as its use of a variety of text types, and the welcome validity of its attempt to assess some of the of sub-skills of writing through an actual writing task, the time needed for administration and marking of this instrument would appear to render it an unlikely choice for use by centres who will be revising their record-keeping to conform with the requirements of the new accountability system and/or who use external accreditation which has its own requirements in terms of record keeping and assessment criteria.

6. Initial Assessment (1st edition)

1	Title of instrument. Publisher. Date of publication. Edition no.	Basic Skills Agency (1997). *Initial Assessment* London: Basic Skills Agency.
2	Cost of instrument per set/copy.	£9.00 plus postage and packing per ring binder at time of publication.
3	Is the instrument 'secure' (unpublished)?	No.
4	Stated purpose of instrument.	Initial screening tool for use in range of contexts.
5	Are there parallel forms? If so, how many?	Three.

6	Are there any obvious manageability problems?	No. Suitable for both individual and group use.
7	What, if any, stipulations about training of administrators are made?	Designed for general use. See page 2 of documentation.
8	Time needed/allowed, if known.	35 minutes, plus time to introduce the assessment.
9	Is the test available in different media? If so, explain briefly.	No.
10	Is there a clear statement of the population for which the instrument is intended? If so, explain briefly.	Intended for general use in a range of contexts, including prisons, business and industry, and further education colleges.
11	a) How is the test 'tailored' for the range of achievement, if at all?	a) Each 'element' (reading, spelling, punctuation and numeracy) has two graded parts, aligned with Entry Level and Level 1 of the old ALBSU/BSA Standards.
	b) Is the test pitched at appropriate levels of ease/difficulty for the target learners?	b) Very roughly. The reading passages seem very difficult for learners at Entry Level.
12	If the test is stated to be suitable for ESOL as well as for other learners, what evidence is there of differentiation, including in the reporting of results?	None.
13	Do the materials seem free from bias? List any difficulties.	The test might present particular difficulties for some ESOL learners in terms of cultural background knowledge required.
14	Are the materials suitable for adults?	Yes.
15	Is there an element of self-assessment in the instrument? If so, explain briefly.	No.
16	Is extraneous knowledge required (e.g. in terms of literacy levels required to comprehend problems/questions, etc./ numeracy/IT)? Comment briefly as appropriate.	N/a.
17	Does the test report results for ■ reading ■ writing ■ spelling ■ punctuation ■ numeracy?	Reading, spelling, punctuation, numeracy.

18	a) At what levels are results reported, if reported in levels? b) If results are reported in different terms, explain briefly. c) Has the test been explicitly aligned or re-aligned to the new QCA Standards?	a) Below Entry Level, at Entry Level, at Level 1 above Level 1 of the old BSA Standards, a summary version of which is provided in the documentation. N.B. For comment on a widespread confusion about this, see below. b) N/a. c) No.
19	Assessment approach and alignment to the Adult Basic Skills Core Curricula and Standards. Discuss, if and as appropriate, as follows: Is alignment – a) broad; b) covering a range of, or all, elements; c) suitably balanced in terms of emphasis; and/or d) appropriate in terms of required depth of skill, knowledge and understanding?	N/a.
20	Describe the item types used. Comment if appropriate.	Reading: two cloze procedure passages per assessment. Spelling: incorrect spellings to be re-written correctly. Punctuation: correcting given sentences. Numeracy: several items per level. All are objective in the sense that each has a single right answer.
21	Do the materials/items appear to test what they are intended to test?	See below. Assessment of spelling seems arbitrary. Punctuation assessed via 'exercises' – a method that is lacking in validity.
22	What, if any, technical information is available about the test, e.g. pilots, test-retest reliability, reliability between parallel forms, with different assessors, as 'split halves', relationship with results on other instruments, etc. Comment on any likely difficulties in these areas.	No information available to reviewer.
23	Is a degree of uncertainty in results, or of limitation in the instrument, recognised, reported, evaluated?	Yes. See page 2 of the documentation.

24	Are there any other points which might usefully be made about the instrument, e.g. minor proofing errors, tasks that seem particularly difficult?	The scoring procedures were unclear and caused great confusion in the field.
25	Rate the usefulness of the reported results to a) learners; b) tutors; and c) the wider world (high, medium or low).	a) Low. b) Medium. c) Low.
26	Sum up likely positive and negative effects of this instrument in terms of backwash/teaching to the test.	N/a. This is not a high stakes assessment.
27	Form a judgement about the suitability of the instrument: a) for its stated purpose; b) for NRDC purposes in – cohort studies – research projects (high, medium or low? Comment below if needed).	a) Medium, if used in accordance with the caveats in the introductory materials. b) – cohort studies: not suitable. – research projects: not suitable.

This package was designed to be used in a range of contexts as a screening instrument, and was very widely used for this purpose, including in prisons (see Brooks et al., 2001a, p. 22). But it had three parallel forms and therefore also came to be quite widely used as a summative assessment, that is, one form would be used at the beginning of a course and another at the end to measure progress, hence the instrument's being reviewed here. The package has sections for reading, writing (spelling and punctuation) and numeracy. Three parallel literacy assessments are provided, each consisting of two cloze procedure passages, two sections on spelling, and two sections on punctuation. Similarly, three sets of parallel items are provided for numeracy. Within each version of the instrument for each skill area the second part is more difficult than the first.

The format of the instrument was straightforward, but the scoring and reporting were not.

Marks were awarded for each part of the instrument for each skill area, and then converted, using the assessment grid provided, to levels. Each part was shown as corresponding to three levels, as follows:

From Part A
Below Entry Level
At Entry Level
Above Entry Level

From Part B
Below Level 1
At Level 1
Above Level 1.

If learners scored below or at Entry Level on Section A, the result was clear. Those whose score on Section A of a skill placed them above Entry level were asked to attempt Section B. If they then scored at or above Level 1, the result was also clear. But what of those who scored above Entry level on Section A but below Level 1 on Section B? The documentation in the package gave no guidance on this, and this lacuna had an unfortunate consequence: many in the field thought this meant that students could be 'Above Entry Level but below Level 1'. However, this was a conceptual error – in a criterion-referenced system, there is no justification for such a 'limbo' status – and in the second edition of *Initial Assessment* this non-existent category has been abandoned and clearer guidance given.

In some people's eyes, cloze procedure, the name of which derives from Gestalt perceptual psychology's concept of 'closure', should never be used for assessment, because it represents virtually no real-life task, though it has a place in teaching as a valid text discussion task (see, for example, Finlay, 1997). To others, cloze procedure is an ideal reading assessment because it allows a respondent to draw upon what might now be referred to as text- and sentence-level knowledge, as well as knowledge about the world, though there is some evidence that in practice not all respondents make use of such cues when completing these tasks. Previous experience of cloze procedure passages would be an advantage. Cloze tests have the additional benefits of being easy to give and quick to mark, being adaptable for IT-based use as well as for paper versions.

Though the theoretical basis of the tests has been challenged – for instance, Finlay (1997) suggests that the BSA cloze tests underestimated learners' abilities to understand texts on the subjects they were studying – they have been the subject of much research, and it appears that cloze tests can be devised that correlate highly with other measures of reading ability.

The tests of spelling and punctuation have problems of validity, in that actual samples of writing are not required, and the focus is on technical accuracy rather than on the ability to convey meaning. It is to be hoped that colleges and other centres take steps to obtain examples of learners' own writing either at the same time as administering any such test, or soon afterwards.

A slight advantage that these materials have over some IT or CD-Rom based assessments is that direct evidence of candidates' performance, however slight, is generated and could be kept in a learner's file, where tutors can see at a glance exactly what the learner did. This might be more useful than a print-out or record of results.

This instrument was not found to be appropriate for either of the proposed research purposes. It does not cover a suitable range for cohort studies, and as it is published it would be inappropriate for use in research projects.

7. Initial Assessment. 2nd edition

1	Title of instrument. Publisher. Date of publication. Edition no.	Basic Skills Agency (1997). *Initial Assessment. An assessment of literacy and numeracy level.* (2nd edition) London: Basic Skills Agency.
2	Cost of instrument per set/copy.	£17.50 for ring binder. Contents photocopiable for educational use.

3	Is the instrument 'secure' (unpublished)?	No.
4	Stated purpose of instrument.	'An indication of an individual's overall level ... to place learners in appropriate learning programmes ... not a profile of skills within those levels ...'
5	Are there parallel forms? If so, how many?	Three versions for each skill, two for any post-16 audience, one 'developed for use with adults enrolled on family programmes.' See below.
6	Are there any obvious manageability problems?	No.
7	What, if any, stipulations about training of administrators are made?	None. 'The assessment has been designed to ... be administered and marked by non-specialists.'
8	Time needed/allowed, if known.	20 minutes per skill, plus time for practice and giving instructions to students.
9	Is the test available in different media? If so, explain briefly.	No.
10	Is there a clear statement of the population for which the instrument is intended? If so, explain briefly.	See question 5.
11	a) How is the test 'tailored' for the range of achievement, if at all?	a) Items generally increase in difficulty. 'Candidates can stop at any time. If they do finish early, they can check their work ... Reassure any candidates who cannot complete many questions and offer them something else to do.'
	b) Is the test pitched at appropriate levels of ease/difficulty for the target learners?	b) Broadly. The judgement of the literacy reviewer, based upon close reading of the *Adult Literacy Core Curriculum* and of the relevant sections of the glossary of the document, was that the items in Entry level 1 literacy were not sufficiently based upon that which is personal and familiar to the learner. However, this is a matter of personal judgement.
12	If the test is stated to be suitable for ESOL as well as for other learners, what evidence is there of differentiation, including in the reporting of results?	N/a.

13	Do the materials seem free from bias? List any difficulties.	Broadly. It appears that statistical checks were made in the course of the development process.
14	Are the materials suitable for adults?	Yes.
15	Is there an element of self-assessment in the instrument? If so, explain briefly.	No.
16	Is extraneous knowledge required (e.g. in terms of literacy levels required to comprehend problems/questions, etc./ numeracy/IT)? Comment briefly as appropriate.	General cultural and background knowledge would be an advantage. An effort has clearly been made to keep the language requirements of the numeracy questions to a minimum. However this sometimes means that students have to work out what the problem is, which for some learners may be more difficult than the item itself.
17	Does the test report results for ■ reading ■ writing ■ spelling ■ punctuation ■ numeracy?	Results for numeracy and literacy are reported in levels, arrived at by checking off total marks attained against a chart. Literacy means reading and spelling; writing is not assessed.
18	a) At what levels are results reported, if reported in levels? b) If results are reported in different terms, explain briefly. c) Has the test been explicitly aligned or re-aligned to the new QCA Standards?	a) Below Entry 1; Entry 1; Entry 2; Entry 3; Level 1. b) N/a. c) Yes.
19	Assessment approach and alignment to the Adult Basic Skills Core Curricula and Standards. Discuss, if and as appropriate, as follows: Is alignment – a) broad; b) covering a range of, or all, elements; c) suitably balanced in terms of emphasis; and/or d) appropriate in terms of required depth of skill, knowledge and understanding?	a) These materials have been developed using a mixture of norm- and criterion-referencing techniques. See question 22. Alignment is best described as very broad. b) Not all elements are assessed; in particular, writing is not assessed. c) The literacy materials are unbalanced because of the high proportion of marks given to spelling – half the 1-mark-per-question items are tests of spelling. The nature of the tests limits to some extent the validity of the assessment materials. d) Within the limitations just mentioned, yes.

20	Describe the item types used. Comment if appropriate.	See below.
21	Do the materials/items appear to test what they are intended to test?	Very broadly for both skills. The main weakness is in the emphasis placed upon spelling, which counts for 50 per cent of the marks in the literacy test. The time allowed, which is relatively short, will influence results.
22	What, if any, technical information is available about the test, e.g. pilots, test-retest reliability, reliability between parallel forms, with different assessors, as 'split halves', relationship with results on other instruments, etc. Comment on any likely difficulties in these areas.	A little. The materials 'have undergone two rounds of trial, with participation from over 2000 learners and two hundred tutors working in the following settings: prisons, adult community colleges, colleges of further education, family literacy and numeracy classes, training providers, voluntary organisations, sixth form colleges.' The results of the trials were 'analysed statistically to indicate the difficulty level and potential bias of the individual questions. The reliability of the marking scheme was also investigated. The results from the analysis and tutor judgements were used to make a final selection of questions and decisions regarding the score boundaries for each level from Entry 1 to Level 1.' No information is given about, for example, how far GCSE, Wordpower, or other qualifications held by subjects match with the results on this test.
23	Is a degree of uncertainty in results, or of limitation in the instrument, recognised, reported, evaluated?	Yes, but not perhaps as thoroughly as it could have been. The test does not claim to assess writing. However, if the test is interpreted sensibly and used only for very initial screening purposes, this need not be a problem. Charts in the introductory materials setting out what learners at a particular level can do are misleading if intended to suggest that this instrument provides such information.
24	Are there any other points which might usefully be made about the instrument, e.g. minor proofing errors, tasks that seem particularly difficult?	None.
25	Rate the usefulness of the reported results to a) learners; b) tutors; and c) the wider world (high, medium or low).	a) Low. b) Low/Medium. c) Low.

26	Sum up likely positive and negative effects of this instrument in terms of backwash/ teaching to the test.	Since this is an initial test, this should not be an issue. Obviously, no test is perfect, and this is intended only to be an initial test, but it could be argued that the emphasis on accurate spelling sends unhelpful messages both to the learners themselves and to those administering the test. Also, if this version (because of its parallel forms) becomes as widely used as the 1st edition as a progress measure it might have backwash effects.
27	Form a judgement about the suitability of the instrument: a) for its stated purpose; b) for NRDC purposes in – cohort studies – research projects (high, medium or low? Comment below if needed).	a) Any allocation to a group or judgement about the level 'at which a student should begin work' made on the basis of this test should, and presumably will in practice, be provisional and subject to revision. b) – cohort studies: not suitable. – research projects: not suitable.

These materials are intended to be used by non-specialist staff, and to provide 'an indication of' an individual's overall level' with a view to placing learners in 'appropriate learning programmes'. They comprise three parallel sets of items for numeracy and for literacy, one of each set being stated to have been developed for use with adults on family programmes. There are 72 items for literacy and 50 for numeracy in each set. The items include multiple-choice and open-ended questions. All answers are either right or wrong. A chart shows score bands for Entry Levels 1, 2 and 3 and Level 1.

This instrument was not found to be suitable for either of the proposed research purposes. The range is too limited for cohort studies, and the instrument is not sufficiently valid for use in research projects. The materials were judged to have some value for their stated purposes, subject to any subsequent placement being regarded as provisional, pending further, more valid, assessment.

The instructions provided for administrators contain advice designed to help them to minimise any anxiety experienced by learners. No information is provided about the manner in which results are reported to learners. Presumably each centre using the materials will deal with this in its own way.

For each skill there is a short practice test, designed to familiarise students with the techniques of recording their answers. There are four practice items for literacy, two for numeracy. For both literacy and numeracy, a line in the question text indicates that students must circle one of a set of four possible answers, and where a box is provided, the open-ended answer should be written in it. The suggested script for administrators repeats this information. Rough work on the question paper is permitted. Instructions for candidates may be given in community languages if necessary.

Despite these precautions, it was felt that the assessment methods might present as much difficulty for some learners, especially those at initial levels, as the actual content of the test. This was felt to be particularly so in respect of numeracy. Whereas in literacy, similar items tended to be grouped together, in numeracy there was much more unpredictable variation in item format from question to question.

The grouping of items for literacy is interesting. There are four groups of questions in each version, each group beginning with a short text. This grouping suggests that originally it was intended that the items in each group should be aligned, with some degree of precision, with the Adult Literacy Core Curriculum.

It is clear that more items were piloted than appear in the final assessment materials. It would be extremely interesting to see how far the facility values found for the items in the course of the pilot reflected their supposed relative difficulty as represented by the original plan of alignment. Any serious degree of mismatch might be taken as evidence that the apparent precision of some of the progression ladders in the curriculum, and indeed, of the whole performance management and assessment systems constructed upon it, are spurious.

Students completing these materials do as many of the items as they can, missing out any they get stuck on, though these can be returned to later. There is an implicit acknowledgement here of the fact that learners will sometimes get an answer at a lower level wrong, yet be able to answer items from higher levels. Also acknowledged is the fact that a candidate who spends too much time thinking about one or more individual items might run out of time, achieving a lower score than one who adopts a different strategy. The statement in the introduction that the materials do not provide a profile of learner achievement is, therefore, reasonable.

For both literacy and numeracy, parallel forms have been created by using very similar items in all three versions. There are slight differences in the literacy materials between the numbers of items in each category of spelling item presented in each set. Version one, for example, begins with four multiple-choice items, whereas versions two and three have only two such items. This balances out in the next group of items: version one has four items in this category, versions two and three have six.

Literacy

For literacy, reading and spelling are assessed, though the results are reported in terms of literacy levels. Half the possible marks are for accurate spelling. This weighting is inappropriate and must make taking the test an unpleasant experience for anyone with specific spelling difficulties. It is tempting to conclude that the emphasis on spelling, which is assessed in three different ways, reflects the limitations of the assessment methods rather than the spelling needs of the learners.

It could also be argued that presenting lists of incorrect spellings together with the correct one is potentially confusing, and not, as the introduction suggests, supportive, and that this affects the validity of the spelling items. The obvious way to test spelling is to read out a list of words. Again, it is tempting to conclude that the possibilities suggested by the assessment techniques available have had an undue influence on the form of the assessment and, therefore, upon the validity of the results.

It is hard to decide which, if any, elements or descriptors of the core curriculum many of the reading items are designed to be aligned with: this reflects the facts that reading is a complex activity which, in practice, is not easily broken down into discrete Elements and Descriptors, and that learners have to read the items before deciding which alternative answer is correct. Few of these items require any depth of understanding. Some are apparently designed to assess understanding of the purposes of texts, and are liable to the same sort of criticism as the items in the CDELL materials (see instrument no.15, below): the depth of knowledge and understanding required to answer the items correctly is less than that suggested in the curriculum.

There is very little difference between the generic materials and those designed for use in family programmes. There were more references to children and to school in the latter. To the extent that these became repetitive, it could be argued that the range of vocabulary used in the assessment materials for family programmes was more limited than that used in the other forms. If the aim of family literacy and numeracy programmes is to enable parents who lack literacy and numeracy skills to help their children with homework, then greater use of words likely to appear in children's reading and maths materials, or that children might want to write, would be more appropriate.

The language in both the generic and the family learning materials tends to be stilted and unnatural, a classic problem in texts designed to be read using limited sight vocabulary and simple phonics, compounded here by the constant use of full forms when abbreviations would sound more natural. The sentiments expressed sometimes verge on the patronising.

Don't go that way or you will be ________ a big mistake.

I need hlep (SIC) to finish the patio in the garden.

He had to drive slowley (SIC) to work because of the icy conditions.

These days, there are lots of opportunities to do a variety of ________ to gain national qualifications.

I hope the school trip will not be on the same days that I am at ________

Bedtime is usually a good time to read ________ stories.

As a parent I am sometimes worried that my son is not doing what I ________ he can do.

N.B. The Level 2 literacy assessments have separate reading and answer booklets. This format contributed to the decision to adopt a similar format for all the levels of the literacy assessment instrument developed for NRDC by NFER in 2003 – see item 16.

Numeracy
A range of the knowledge and skills involved in number, measures, shape and space and handling data is assessed. No calculators are allowed; ability to use a calculator is not assessed.

Some numeracy items appear to draw on more than one aspect of the Standards. Without asking a learner who had answered a question incorrectly, it is not clear which aspect of the Standard they had failed at.

Some vertically set out addition questions have no space in the conventional place under the line for answers to be written, thus somehow subverting the usefulness of setting such sums out in this way.

An effort has clearly been made to use as few written instructions as possible, with the result that, for some items, students have to work out what is required. For example, in a question requiring understanding of millimetres the only indication that this is what is required is the appearance of 'mm' to the right of the answer box (version 2, question 36).

Other questions also rely heavily on students' knowledge of, or ability to work out, the test rules: a picture of a clock showing quarter past five, with 5:15 am written in a box next to it, is shown side by side with a picture of a clock showing five past seven with an empty box next to it. The student who is struggling to tell the time might be more challenged by the problem of understanding the question than by the numeracy task itself.

A few questions make a gesture towards the use of context. The wording in these is often unhelpful. They seem to fall between two stools, being neither the full texts of the IALS model, nor simple pencil and paper tests. There is, combined with this, a conflict between real-world and test knowledge. For example, in each version, a question based upon a map asks the distance between two towns. A key on the map states that the scale is 1cm to 10km. Thus far, the item seems to be realistic. However, a straight line is drawn between the two towns, and labelled in cm. This is not realistic. Learners who did not notice that 'km' appears to the right of the answer box might not realise that this item requires them to use the scale to convert the cm to km.

Though the different numeracy tests seem broadly parallel, when the individual items apparently intended to be parallel are examined in detail, differences in difficulty reveal themselves. For example, three parallel items require the daily cost of hiring a bicycle to be halved. The sums of money are £6.50, £7.50 and £8.50. It is harder to divide the second of these by 2 than it is to divide either of the others because 7 is an odd number.

Question 6 in Version 2 is as follows

10p – 2p – 2p – ________ = 5p

Question 6 in version 3 is

10p – ________ – 2p – 1p = 5p

The question in version 3 is clearly more difficult than the question in version 2. This is because the gap is in a different position in the line of the sum. It would be interesting to discover whether these two items had the same or different facility values in the trials.

In contrast, Question 21 on version 2 and Question 27 on version 3 are identical:

21, 24, 27, 33

It should be noted that no instructions at all are provided for this question; the learner has to work out what is required.

Finally, the figures on some items, such as those requiring learners to state how many ml of liquid a pictured flask contains, were too small.

Conclusions
Given that this test does not claim to be strictly aligned to the standards, and the obvious elements of norm-referencing and judgment in the development of the scoring system, it could be argued that it is very broadly suitable for its intended purposes if interpreted with a well-informed view of its limitations.

8. Edinburgh Reading Test, Level 4, 3rd edition

1	Title of instrument. Publisher. Date of publication. Edition no.	*Edinburgh Reading Test*. Hodder and Stoughton. 2002. 3rd edition. The Test has four levels, corresponding with ages. Only the fourth, highest, level is reviewed here. London: Basic Skills Agency.
2	Cost of instrument per set/copy.	£10.99 per loose-leaf ring binder set.
3	Is the instrument 'secure' (unpublished)?	No.
4	Stated purpose of instrument.	Various including target-setting, diagnostic, predictive.
5	Are there parallel forms? If so, how many?	No.
6	Are there any obvious manageability problems?	No.
7	What, if any, stipulations about training of administrators are made?	None. It seems to be assumed that they will be qualified teachers. Guidance about marking and about the interpretation and use of results is given.
8	Time needed/allowed, if known.	45 minutes, plus practice time.
9	Is the test available in different media? If so, explain briefly.	No.
10	Is there a clear statement of the population for which the instrument is intended? If so, explain briefly.	School ages. ERT 4 is for ages 12:0 to 16:6.
11	a) How is the test 'tailored' for the range of achievement, if at all? b) Is the test pitched at appropriate levels of ease/difficulty for the target learners?	a) N/a. b) Broadly. 'Quotients' below 70 are not reported, and a recommendation is made that further specialist assessment should take place in such cases.

12	If the test is stated to be suitable for ESOL as well as for other learners, what evidence is there of differentiation, including in the reporting of results?	N/a.
13	Do the materials seem free from bias? List any difficulties.	Yes. One item is based around a football match, another around a recipe. This may represent an attempt to provide gender balance.
14	Are the materials suitable for adults?	Broadly.
15	Is there an element of self-assessment in the instrument? If so, explain briefly.	No.
16	Is extraneous knowledge required (e.g. in terms of literacy levels required to comprehend problems/questions, etc./ numeracy/IT)? Comment briefly as appropriate.	General cultural and background knowledge would be an advantage.
17	Does the test report results for ■ reading ■ writing ■ spelling ■ punctuation ■ numeracy?	There are five subtests: skimming, vocabulary, reading for facts, points of view, inferential comprehension. Results are reported for these and for the test as a whole. Part of the possible value of the test lies in its allowing users to see whether respondents' scores on a subtest are within the range expected.
18	a) At what levels are results reported, if reported in levels?	a) Not reported in levels.
	b) If results are reported in different terms, explain briefly.	b) Raw scores obtained by adding correct responses are converted to normally distributed quotients. Age adjusted percentiles for each quotient are given. Reading ages based upon samples of English and Stirling schoolchildren are given, ranging from 11:7 to 16:6 based on the mid score of children in particular age groups. Tables show expected subtest scores for particular whole test results.
	c) Has the test been explicitly aligned or re-aligned to the new QCA Standards?	c) N/a.

19	Assessment approach and alignment to the Adult Basic Skills Core Curricula and Standards. Discuss, if and as appropriate, as follows: Is alignment a) broad b) covering a range of, or all, elements c) suitably balanced in terms of emphasis and/or d) appropriate in terms of required depth of skill, knowledge and understanding?	Not explicitly aligned. Would not cover the lower levels.
20	Describe the item types used. Comment if appropriate.	All multiple-choice.
21	Do the materials/items appear to test what they are intended to test?	Broadly.
22	What, if any, technical information is available about the test, e.g. pilots, test-retest reliability, reliability between parallel forms, with different assessors, as 'split halves', relationship with results on other instruments, etc. Comment on any likely difficulties in these areas.	Tables and details of standardisation samples are provided in the manual.
23	Is a degree of uncertainty in results, or of limitation in the instrument, recognised, reported, evaluated?	Yes. 'The scores from all tests are liable to error.' Manual, page 21.
24	Are there any other points which might usefully be made about the instrument, e.g. minor proofing errors, tasks that seem particularly difficult?	No. The vocabulary items have an arbitrary feel.
25	Rate the usefulness of the reported results to a) learners; b) tutors; and c) the wider world (high, medium or low);	a) Medium. b) Medium. c) Low.
26	Sum up likely positive and negative effects of this instrument in terms of backwash/ teaching to the test.	Issue unlikely to arise. Broad curriculum assumed, as the nature of the sub-tests and discussion in the manual make clear. Not a high stakes instrument.

27	Form a judgement about the suitability of the instrument a) for its stated purpose b) for NRDC purposes in – cohort studies – research projects (high, medium or low? Comment below if needed)	a) Medium. Any allocation to a group for literacy or numeracy or indeed for other subjects should be highly provisional and subject to revision. b) – cohort studies: not suitable – range of levels not appropriate. – research projects: not suitable – instrument not secure.

These tests were not suitable for either of the research purposes. Their range of coverage was inappropriate. The tests were found to be quite suitable for their stated purposes, subject to those making use of them having read and understood the manual. There is little to add to the general points made about this instrument on the summary framework.

Since examples may be found of adults in adult basic skills classes being given reading and spelling ages, it is worth making the point that such practices are illogical and seem to reflect lack of understanding of the tests involved. Briefly, such tests are norm-referenced in respect of particular samples (age groups) and therefore are not applicable in terms of their own theory to other population groups. Such measures have little or no information value for learner or tutor, and the reporting of results to adults in terms of children's ages is demeaning.

9. Entry Level Certificate in Adult Numeracy

1	Title of instrument. Publisher. Date of publication. Edition no.	Entry Level Certificate in Adult Numeracy. AQA. September 2002.
2	Cost of instrument per set/copy.	Pack of 10 tests £25.00. Pack of 50 tests £100. Pack of 100 tests £185. Marking guides £4.00 each.
3	Is the instrument 'secure' (unpublished)?	No.
4	Stated purpose of instrument	Not stated in so many words, but intended as statement of achievement at Entry Level.
5	Are there parallel forms? If so, how many?	Yes – three versions.
6	Are there any obvious manageability problems?	No.
7	What, if any, stipulations about training of administrators are made?	None required.

8	Time needed/allowed, if known.	Entry 1: 20 minutes for Calculator Paper + 40 minutes for Non-Calculator Paper Entry 2 and 3: 30 minutes for Calculator Paper + 60 minutes for Non-Calculator Paper
9	Is the test available in different media? If so, explain briefly.	No.
10	Is there a clear statement of the population for which the instrument is intended? If so, explain briefly.	Adults.
11	a) How is the test 'tailored' for the range of achievement, if at all? b) Is the test pitched at appropriate levels of ease/difficulty for the target learners?	a) Described as graduated tests, 3 repeated blocks of ten skills, at Entry 1, 2 and 3. b) Difficult to say (see text), but matches Adult Numeracy Core Curriculum elements.
12	If the test is stated to be suitable for ESOL as well as for other learners, what evidence is there of differentiation, including in the reporting of results?	Not mentioned, though stated to avoid bias.
13	Do the materials seem free from bias? List any difficulties.	Not much of a cultural mix in choice of names. Tasks relate to everyday material.
14	Are the materials suitable for adults?	Some childish-sounding and possibly patronising questions.
15	Is there an element of self-assessment in the instrument? If so, explain briefly.	No.
16	Is extraneous knowledge required (e.g. in terms of literacy levels required to comprehend problems/questions, etc./ numeracy/IT)? Comment briefly as appropriate.	Yes, see text. Literacy, including a sort of document literacy peculiar to numeracy tests. Exam knowledge required.
17	Does the test report results for ■ reading ■ writing ■ spelling ■ punctuation ■ numeracy?	Numeracy.

18	a) At what levels are results reported, if reported in levels?	a) Not reported in levels since each paper tests only one level.
	b) If results are reported in different terms, explain briefly.	b) Achievement in the qualification is reported as a percentage for each level. Candidates receive a statement in terms of an achieved percentage. 'The on-demand nature of these tests makes formal grade awarding meetings impossible as tests are sat and results issued all the year round. However, different versions of the tests are standardised in order to ensure comparability and examiners are standardised and monitored throughout the year in order to maintain a consistent marking standard. It is anticipated that the percentage required to receive an overall pass at each Entry level will be 75 per cent.'
	c) Has the test been explicitly aligned or re-aligned to the new QCA Standards?	c) Yes.
19	Assessment approach and alignment to the Adult Basic Skills Core Curricula and Standards. Discuss, if and as appropriate, as follows: Is alignment	
	a) broad	a) Close.
	b) covering a range of, or all, elements	b) Range of elements.
	c) suitably balanced in terms of emphasis and/or	c) Problematic, see text.
	d) appropriate in terms of required depth of skill, knowledge and understanding?	d) Problematic, see text.
20	Describe the item types used. Comment if appropriate.	All single-answer, right or wrong items based on problems embedded in text, with images intended to contextualise the items
21	Do the materials/items appear to test what they are intended to test?	Yes.
22	What, if any, technical information is available about the test, e.g. pilots, test-retest reliability, reliability between parallel forms, with different assessors, as 'split halves', relationship with results on other instruments, etc. Comment on any likely difficulties in these areas.	None.

23	Is a degree of uncertainty in results, or of limitation in the instrument, recognised, reported, evaluated?	No.
24	Are there any other points which might usefully be made about the instrument, e.g. minor proofing errors, tasks that seem particularly difficult?	See text re the chosen example of Entry 1.
25	Rate the usefulness of the reported results to a) learners; b) tutors; and c) the wider world (high, medium or low).	 a) High - could be a psychological boost; or Low - could mislead and dishearten. b) Low except as above for the learner. c) Medium possibly for employers.
26	Sum up likely positive and negative effects of this instrument in terms of backwash/ teaching to the test.	A teacher would be likely to teach entirely to the test.
27	Form a judgement about the suitability of the instrument a) for its stated purpose b) for NRDC purposes in – cohort studies – research projects (high, medium or low? Comment below if needed)	 a) Own purposes - will provide a result. b) NRDC purposes – cohort studies: unsuitable, doesn't cover levels above Entry. – research projects: unsuitable, doesn't cover Levels 1 and 2.

From September 2002, all previous qualifications at Levels 1 and 2 in adult literacy and numeracy in England were replaced by the national tests. At Entry Level, however, it was left to awarding bodies to develop tests. This item was therefore included here as a partial companion to the practice versions of the national tests (see no.14). Though it was from the outset clearly unsuitable for NRDC purposes precisely because it covers only Entry Level, it was reviewed for any insights it might yield for good assessment practice at that level.

This instrument was designed to be used through centres registered with AQA as the awarding body, with flexibility of use to allow candidates to sit the tests whenever they are ready, with four weeks' notice of the test date required for orders.

A specification booklet gives background information; a scheme of assessment; subject content; a nod towards the wider context; a final section on awarding and reporting. Specimen Papers at Entry 1, Entry 2 and Entry 3 (according to the *National Standards for Adult Numeracy)* are available. The booklet is clear and spare in appearance and content, and the papers give the same impression (however, see below). For an administrator, the procedures would very likely appear non-threatening and easy to apply.

The papers are uncompromisingly 'exam papers' in their appearance. The front-page rubrics match the familiar (in the educational world) design of GCSE and A-Level papers; users are 'candidates' with boxes for their number and signature; there are boxes for Examiner's Use only. For an adult learner newly re-cast as a 'candidate', it might initially appear daunting, but equally, a certain status and an expectation of progression may be given to the Certificate by the implicit link with further stages in the Qualifications framework.

Each Entry Level has two papers. Part A is a Calculator Paper, and Part B is a Non-Calculator Paper.

The Assessment Objectives as listed in the Specification booklet are word-for-word the capabilities from the Adult Literacy Core Curriculum (ANCC), p.12, and the Standards from the QCA Standards. The variant name is not explained; the Assessment Objectives terminology is widespread in other qualifications.

Entry 1 Specimen test paper has been selected for a more detailed review of its suitability for its stated purpose.

Entry 1 Part A has five items with 10 marks available.

Question A1 tests number bonds to 10 (N1/E1.1), though whether it is recall of these facts, the capacity to work them out mentally in some way, or to use a calculator to find the answers would not be clear from a right answer. The route to a wrong answer would be tracked with difficulty, although candidates have been invited to do rough working in this book.

Question A2(a) tests N1/E1.1 again (I think), though that is clear only from the example. Question 2(b) seems to test MSS1/E1.4: use of vocabulary related to length. The 'contextualising' part of the question does not seem to relate to the task of identifying the longest nail. The connexion is implicit, and anyway, you may not need the longest nail to fix the fence. Also one nail never does it. Neither as 'exam knowledge' nor as a 'real-life problem' does this make sense.

Question A3 seems to relate to N1/E1.4: understand the operation of addition with totals to 10.

Question A4 'Sharon has £6. She is buying books that cost £2.00 each. How many books can she buy?' This is difficult to relate to a specific curriculum reference in Entry 1, which contains only the requirement to recognise and select coins in MSS1/E1.1. In N1/E1 the requirement is to count, add and subtract to 10. It is not clear what method the candidate would use to answer this – recall, multiplication/division, counting in twos, calculator? It is more than counting on or back, because a stage of noting the number of £2.00 units would be necessary.

Question A5 relates to HD1/E1.1 and HD1/E1.3 perhaps. Real-life v. exam knowledge issues are raised – there are a lot more colours available than are dreamt of in this philosophy. The learner has to know to ignore that knowledge for this limited purpose; in that respect, an attempt to help by providing a sort of context may obfuscate. The decision to capitalise the colours was taken presumably to clarify, but it goes against the conventional use of capitals in ordinary prose. The phrasing of the question 'There were more White cars than Green' calls to mind Margaret Donaldson's work in the 1970s on embedding (Donaldson, 1978) – I believe in her case it was cows and sleeping cows. The use of language, specifically such use of an adjective, in the question influenced the comprehension and the answers of the children taking part in the experiment. There is, potentially, real difficulty for a reader to make sense of this question, with consequences for the reliability of the results.

Calculators – testing only a range of elements
The ANCC requires at E 1 that a learner should be able to check calculations with a calculator using whole numbers. This could only be tested by an administrator watching the candidate since learners are not directly instructed to do so. The availability of a calculator feels like a token gesture, and the calculator is physically taken away after Part A for the 'real' stuff in Part B.

Entry 1 Part B has 16 items with 20 marks available.

Question B1 involves addition and subtraction in the same sum, with the instruction 'work out'. There is no equals sign, just the house-style dotted line on the right for the answer.

Question B2 uses a single sentence context: 'You empty your pockets and find some money.' Then the question: 'What is the total of which stops in mid-air, leaving the learner to sort out that the simple circle pictures, with '2p' inside, represent money and must all be added together, the answer to be written on the dotted line. A special esoteric kind of reading is necessary here – see other examples below.

Question B3 and B4 both sound babyish in their content and phrasing. There must be few situations, either for children or adults, where someone buys ten bars of chocolate in two goes and gives seven of them away. Similar comment might be made about Question B15 and 16.

Question B7 'What is £7.00 – £2.00?' (N.1/E1.6) requires a sort of reading which can verbalise a comprehensible and workable problem from a mixture of words, numbers and symbols. Is such reading taught anywhere? The ANCC requires adults 'to be taught to interpret +, – and = in practical situations for solving problems'. It does not spell out the possible complexity of the reading exercise, let alone the steps required to link the abstraction of the problem, as formulated on paper, with any real financial transaction as suggested by £7.00 – £2.00.

Question B8 tests the vocabulary 'beneath' and 'on the left of'. However, the task may seem empty, since the learner may normally expect to find ticks and crosses marking *something*, not just floating round a picture of a triangle. Curiously, nothing is made of the presence of the triangle itself.

Question B9 demands reading forwards from 'this' shape to the three-line diagram on the graph. MSS2/E1 requires that the learner can 'recognise and name common 2D and 3D shapes'. This task requires more, in that the learner has to draw in the remainder of the square, having read its name in the question.

Question B10 states with a forward reference: 'These are the prices of vegetables at the supermarket.' Then a list of vegetable names and prices, headed PRICE PER KG, followed by the question 'Which vegetable is the dearest?' To be able to answer this, the learner must understand the implicit qualification: 'according to this chart'. This is an area of exam knowledge. Common sense knowledge has to be held in abeyance. It depends on the time of year, and what the other potatoes in the rack cost; it depends whether you always buy them frozen, depends whether parsnips ever feature in your menu or your vocabulary. The use of the term 'dearest' has overtones, too, which may colour the question differently for different readers.

Another example of imprecise reference comes at question B11. The learner has to know to read, in ' Sam has these shapes', that 'these' refers ahead to the images printed on the page below. It is not clear what Sam has actually *got*, since these are 2D shapes, with no pretence at a concrete form he might get his hands on. The bracketed lower-case (a) convention has to be understood to refer to an impending (b), which will very likely have some connexion. It will, as it happens, require a different process or action. The upper-case and emboldened A and B *inside* the box have to be understood to be altogether different from the lower-case and bracketed a and b. Having imagined and established Sam, the learner has to accept his absence (like Mole, when Otter takes his sudden leave in *Wind in the Willows*), because by 11b, the shapes have become 'your shapes'. There is clearly a difficulty for the item designer to suggest economically that the candidate's own selection of group A and group B shapes is now under discussion, but the half-baked attempt to provide a context leads to a flabby and unusual use of pronouns.

Question B11(b) needs literacy skills to understand a grammatically complex instruction. If the answer were wrong, there would be no way of establishing why. The test would be only of any use for an end-test, little use for a teacher deciding what to do next with this particular learner, supposing the test were failed.

In pursuit of brevity, grammatical and textual consistency has been sacrificed. In question B12, as in question A2(b), there is no connexion between 'You are catching a train at the time shown above' and 'What time does this clock show?' A learner has to understand 'above', and the referential function of 'this' – possibly ambiguous anyway, since the first sentence speaks of 'the time shown', a form of words not repeated directly in the second sentence which has the additional complication of the first use of the word 'clock'. Question B12(b) uses 'this clock' again, but means a different clock, the one below, though this is not stated.

Question B13 has a graphic of 'boys'. In fact, it seems to be several pictures of the same boy, only reducing in size. The instruction says 'picture of the boys'; it would be more accurate to say 'pictures of the boy'. Common sense would say it's the same boy from farther away, not a different boy – potentially misleading.

Forms of questions – terminology, mode and tense of verbs

'*Find* the missing numbers': is this an example of exam knowledge or a mathematical term? This is the first question (Entry 1 Part A) that many adult learners may answer in a formal exam. Specific vocabulary to be taught is not all spelt out in the Standards. In Parts A and B there are several instances of 'how many', some of 'how much'. 'Sort' at question B11(a) is possibly misleading, since it suggests a physical handling of objects, whereas here it demands a mental exercise in recognition. 'Write down' in 11(b) is a specific instruction to act, also in the imperative. At question B13(b), however, the instruction to write an answer is implicit. 'Work out/complete/put/write' also appear as instructions in Part B; they cover a wide range of kinds of meaning and could stretch an inexperienced reader. There is a mixture of present tenses: 'Sharon has £6', 'She is buying', 'You need', and past: 'Jim wrote down', 'You have been shopping'. The use of 'should' on at least two occasions in Entry 2 could be demanding – 'How much change should he get?'

The powerful significance of tense and of modal verbs seems to be disregarded and could put pressure on some adult learners.

It may be, of course, that no learner has any problem with the reading required, but without further research, it cannot be known why any answer is wrong – a fruitful area for teacher, learner and tester alike to explore.

Overall coherence – support for adult learners
The tests *appear* coherent, but do not hang together as texts in their own right. The language demands are not consistent or considered from the point of view of an adult learner.

Contextualisation
This kind of contextualisation is a sad attempt to rehydrate a desiccated learning experience. The richness of the 'real life' envisaged in *A Fresh Start* has become an arid test landscape. The complexity of the learning experience at whatever level is reduced to a small set of right or wrong answers.

In terms of contextualisation, the dramatis personae are various and only fleetingly visible. Amy, Sharon, Jim (the anorak who counts parked cars), Paula, Lee, Lee's friends, the shrinking boy, Jane, Janet, Bob, Iqbal and 'you' (not much of a cultural mix of names, incidentally) provide a thin, potentially bewildering narrative. If competent readers are given scanty information, they will fill in the gaps themselves to arrive at a whole which makes sense to them (if frivolous). It is an observed characteristic of weaker readers that they accept a large measure of incomprehension and do not have the skill or the confidence to *demand* sense from a text. It may further undermine confidence to be presented, in a situation where power-relations are implicitly fixed, with authoritative-looking texts which do not cohere. A good test might pay more attention to the content of each item as a whole narrative, even if that is a sum with numbers and symbols and no words, and to the learner's likely experience of the whole test, and subsequent tests.

Entry 2 and Entry 3 are not studied in detail. Their format is similar.

Suitability for use
In many respects this test does not seem to be useful for its own purposes. Perhaps further information about its trialling and its parallel forms would reassure, but at face value there seem to be problems which might undermine confidence in results.

The tests would not seem to be useful for NRDC studies. The whole range is not covered, which is a requirement for the Cohort study. To measure the effects of teaching would require a much less blunt instrument.

10. Skillscape

1	Title of instrument. Publisher. Date of publication. Edition no.	Smith, P. and Whetton, C. (2000) *Skillscape.* Windsor: NFER-Nelson.
2	Cost of instrument per set/copy.	£109.00 for reference set containing instructions and one copy of each level for each skill; packs of test materials priced between £12.00 and £54.00.
3	Is the instrument 'secure' (unpublished)?	No.

4	Stated purpose of instrument.	Assessment of literacy and numeracy skills in occupational or guidance settings especially if qualifications are not available or possessed, for purposes either of 'remedial help' or as part of a 'low-level selection process'.
5	Are there parallel forms? If so, how many?	No.
6	Are there any obvious manageability problems?	No.
7	What, if any, stipulations about training of administrators are made?	None, but very detailed guidance, on two levels of formality, is given.
8	Time needed/allowed, if known.	Pre-Screen: up to ten minutes. Literacy and numeracy tests, including instructions and practice: 30 minutes. Optional writing test, 30 minutes including instructions.
9	Is the test available in different media? If so, explain briefly.	No.
10	Is there a clear statement of the population for which the instrument is intended? If so, explain briefly.	Major uses expected to be initial screening of trainees or employees, including uses by Careers Guidance Professionals.
11	a) How is the test 'tailored' for the range of achievement, if at all?	a) A pre-screen consisting of ten sentence completion or basic maths tasks is administered, following which one of two levels of the main test is administered.
	b) Is the test pitched at appropriate levels of ease/difficulty for the target learners?	b) The test is not designed for use by educators. Though stated to assess 'basic levels', the combination of the size of the reading material and the task type make this test very challenging for those at Entry Level.
12	If the test is stated to be suitable for ESOL as well as for other learners, what evidence is there of differentiation, including in the reporting of results?	No such evidence available to reviewer.
13	Do the materials seem free from bias? List any difficulties.	Broadly. A detailed investigation into possible bias of particular items is reported in the handbook.
14	Are the materials suitable for adults?	Yes.
15	Is there an element of self-assessment in the instrument? If so, explain briefly.	No.

16	Is extraneous knowledge required (e.g. in terms of literacy levels required to comprehend problems/questions, etc./ numeracy/IT)? Comment briefly as appropriate.	General cultural and background knowledge would be an advantage.
17	Does the test report results for ■ reading ■ writing ■ spelling ■ punctuation ■ numeracy?	Reading, writing and numeracy.
18	a) At what levels are results reported, if reported in levels?	a) An individual profile sheet giving results for the level taken, in terms of *Above Average*, *Average*, and *Below Average*, is provided. These results refer to norms derived from the results obtained with 'comparison groups'.
	b) If results are reported in different terms, explain briefly.	b) N/a.
	c) Has the test been explicitly aligned or re-aligned to the new QCA Standards?	c) No.
19	Assessment approach and alignment to the Adult Basic Skills Core Curricula and Standards. Discuss, if and as appropriate, as follows: Is alignment –	The instrument does not refer to the Adult Basic Skills Curricula and Standards. It aims to assess ' 'functional literacy' – the need to extract required information from a variety of sources, e.g. text, tables or advertisements, and convey this information unambiguously.'
	a) broad;	a) Alignment to levels could, therefore, be described as broad but it would be a matter of judgement which levels of the Adult Curriculum each test level corresponded with.
	b) covering a range of, or all, elements;	b) N/a.
	c) suitably balanced in terms of emphasis; and/or	c) Yes.
	d) appropriate in terms of required depth of skill, knowledge and understanding?	d) Yes.
20	Describe the item types used. Comment if appropriate.	Factual questions are asked, to which a marking guide is provided. Candidates supply their own answers: this is not a multiple-choice objective test. Marks are recorded as correct, incorrect, or not answered.
21	Do the materials/items appear to test what they are intended to test?	See below.

22	What, if any, technical information is available about the test, e.g. pilots, test-retest reliability, reliability between parallel forms, with different assessors, as 'split halves', relationship with results on other instruments, etc. Comment on any likely difficulties in these areas.	Extremely detailed technical information, including score tables, is provided in the accompanying manual. This instrument is based upon an earlier set of *Basic Skills Tests*, and takes into account feedback received from users. Data are provided on comparability with the earlier version. The test was standardised in 1999 using samples from Year 11 in schools and from NVQ and GNVQ courses.
23	Is a degree of uncertainty in results, or of limitation in the instrument, recognised, reported, evaluated?	Yes, both statistically and in words. Suggestions are also made for ways in which local norms for the test may be developed in situations where this is felt to be appropriate.
24	Are there any other points which might usefully be made about the instrument, e.g. minor proofing errors, tasks that seem particularly difficult?	See below.
25	Rate the usefulness of the reported results to a) learners; b) tutors; and c) the wider world (high, medium or low).	a) Low. b) Low. c) Low.
26	Sum up likely positive and negative effects of this instrument in terms of backwash/ teaching to the test.	N/a.
27	Form a judgement about the suitability of the instrument: a) for its stated purpose; b) for NRDC purposes in – cohort studies – research projects (high, medium or low? Comment below if needed).	a) High/medium. b) – cohort studies: not suitable. – research projects: not suitable as it stands (no parallel forms) but could be used as a model for the development of suitable instruments; and in fact the reading and writing tests were so used – see main text.

This is a standardised test, with literacy (reading) and numeracy versions, and an optional writing test. Pre-Screens are available to help decide which level of test individuals should take. The literacy (reading) test uses two versions of a reasonably realistic mock-up newspaper containing adverts, articles, programme times, and other lifelike items. The short version fills six sides of A4, the longer eight. Each version of the test contains 20 questions designed to test the ability to find and report factual information unambiguously. Marks obtained are translated into percentile equivalents using tables supplied in the handbook.

Advice about producing guidelines on adequate levels of performance required is provided for those seeking to use the instrument in particular contexts.

This instrument was not found to be suitable for research projects, as it does not have the parallel forms required. It has some potential for use in cohort studies, though, unlike the materials used in previous studies, it is not aligned with adult literacy and numeracy standards, but norm-referenced in respect of several different populations. Therefore, there would be a lack of continuity with previous studies. The lowest level is above the lowest level of items in previous cohort studies, which means it might not discriminate at those levels as well as previous instruments did. It was judged to be suitable for its own stated purposes as a screening instrument, particularly if those using the test follow the detailed advice provided in the manual. Given the suggestions about the establishment of local norms, lack of reference to the National Standards should not deter an employer or agency thinking of using this instrument. On the contrary, the adequacy criteria suggested for the assessment of writing would, in some situations, be more suitable than the levels set out in the new Standards.

This test provides an interesting comparison with the others reviewed. Though it is standardised, it avoids the artificiality of objective multiple-choice items. The texts bear a resemblance to some IALS prose and document scale items, in that they require factual information to be obtained from texts. However, the statistical techniques used are not those used in the IALS surveys. To some extent, the items resemble those used in previous cohort studies, covering a range of text-types, with a particular difference being that, instead of being displayed one at a time on show cards, they have been assembled into a semi-realistic newspaper. Though it makes no claim to be referenced to the National Standards, the instrument is intended for use with some of those who will be working in adult basic skills provision, being normed with reference both to a sample of Year 11 pupils from 'the lower achieving 60 per cent of schools' and to a sample of NVQ and GNVQ students drawn from 15 FE colleges.

The instructions for this test are clear, and all the materials are well presented for both administrators and those taking the test.

A simple sentence completion Pre-Screen test is used. This is an interesting device, and one that was not encountered in other instruments. The main assessment given depends upon the results on this initial test.

Though perhaps not wholly lifelike, the *Brimfield News*, a mock-up newspaper in which a variety of typical everyday reading items appear, is an imaginative solution to the problem of text for the assessment of reading. Two copies of this newspaper, one more difficult to read than the other, together with a set of 20 questions on each newspaper, are used as the reading material for tests of reading at two levels. Test-takers write their answers in booklets. Answers are scored using a rubric that specifies the information that each answer must contain, and the resultant scores can be converted to results using scaling tables provided.

This is a timed test. It was judged that the speed with which those taking it could work through it was probably a very important factor. The reading tasks themselves were not so obviously graded in difficulty as those in most other materials reviewed.

A minor point that emerged from the analysis of this instrument was that, though candidates' attention is drawn to an index on the front page of the *Brimfield News*, they are instructed to read the paper before answering the questions. Given that most of the questions require discrete pieces of factual information to be located, the use of scanning strategies would be appropriate. In schools, students are often advised to read the whole text in an examination before answering the questions because the information and ideas needed to answer the questions fully may only be available to those who have read the whole text. In real life, when reading to obtain specific information, as opposed to reading to absorb or enjoy a whole text, one would not be likely to read the whole text before searching for the piece of information in which one was interested.

Even the shorter edition of the *Brimfield News* is relatively long, longer certainly than the texts that, according to the Standards, learners at the lower end of Entry Level can deal with. The purpose of the index that appears on the front cover is presumably to help people to find information, and the attention of subjects is drawn to this during the practice session. The use of an index is in itself a task that not all learners will find easy, and it may be that some will not remember to use the index, but will find the answers by skimming and/or scanning, using headings and headlines to help them, or by recalling from their initial reading of the newspaper either the information itself, or the article or section in which it may be found. This means that, in effect, though the instrument is not explicitly aligned with the Standards, many of the Elements within them are effectively covered in a realistic context.

The assessment of writing in this instrument, though aimed at assessing basic adequate levels, was one of the best encountered in the course of the review. There are two important reasons for this judgment. First, actual samples of writing are obtained. Secondly, as was noted explicitly in the draft Adult Basic Skills Curriculum, the importance of accurate spelling, punctuation and grammar is, in many real-life situations, including many workplace situations, less important than whether the writing communicates and serves its purpose. The scoring system used in *Skillscape* takes this into account, providing both an adequacy scale and other, more technical, criteria by which writing may be evaluated. For example, the guideline of functional adequacy provided for evaluating the address on an envelope is 'Has the writer included enough information to get the letter to its destination ...?' This provided a refreshing contrast to the highly artificial, if superficially logical, nature of some of the progression ladders in the Standards. It also contrasted very favourably with the CD-Rom and paper-based assessments of writing, which by their very right/wrong nature, were more likely to use unrealistic criteria based on accuracy and precision.

The test is not explicitly aligned to the levels of the new Standards. This is not necessarily an insuperable problem: statistical information in the Manual enables users to see how scores relate to achievement across the ability range at age 16. Using the correspondence tables in the National Standards for adult literacy and numeracy, it would be possible to estimate very roughly where a cohort member with a particular score on either version of the literacy test might be placed within that framework.

The literacy part of this test provided the principal model for the literacy assessment instrument developed for NRDC by NFER in 2003 – see item 17.

C. Computer-based instruments. Introductory remarks

For additional comment on computer-based instruments, see Heath (2003).

Computerised assessment has a number of potential advantages over more traditional methods.

Given sufficiently large numbers of students, such tests are far cheaper than other assessment methods, as they do not require human markers, assessors, verifiers, moderators, or indeed any of the bodies that recruit, train, organise and employ such personnel. Such tests also seem objective: given the limitations of the technology, answers are either right or wrong. There should be nothing to argue about, though in practice, as those who devise the items are human, there sometimes is.

It is possible in theory to create an adaptive computerised test, thus reducing the time spent on testing. The items presented to a student depend upon their responses to previous answers, and progress through a bank of items is dictated by the operation of a set of algorithms. Such an approach seems to imply a very simple view of the nature of the topic, as if achievement consisted of one dimension along which everybody could easily be ranked.

It is possible to produce excellent graphics on screen, and some good examples were found, such as a thermometer in the *Target Skills* package. However, not all the graphics were directly linked with the task in question; thus students were sometimes faced with extraneous information to sort through in addition to that which was actually necessary. Video clips and audio clips as well as still images were used in one package, though the language of the participants had a somewhat false, scripted feel. In some instances, presentational gimmickry interfered with the clarity of the image. The learndirect Skills Checks provide examples of this.

Compared with other IT products, some of the assessment materials lacked presentational polish. Some items were clumsy, and drag-and-drop techniques were difficult to apply, with potentially off-putting consequences.

Given that not everybody has a home computer, the use of IT for initial assessment and diagnosis poses problems. Though some of the materials included short practice sessions, these rarely seemed likely to make a learner new to the technology feel truly comfortable with it, particularly a learner who found it difficult to learn new things.

11. Target Skills: Initial Assessment

1	Title of instrument. Publisher. Date of publication. Edition no.	*Target Skills Initial Assessment* Cambridge: Cambridge Training and Development. (2001). Version 1.1. Full information not available.
2	Cost of instrument per set/copy.	£199.00.
3	Is the instrument 'secure' (unpublished)?	No.
4	Stated purpose of instrument.	To provide a 'clear picture of each students' current level against the QCA Adult Literacy and Numeracy Standards … a clear basis for developing a learning plan.'

5	Are there parallel forms? If so, how many?	No.
6	Are there any obvious manageability problems?	Yes – the CD version is awkward to use in various ways – see below.
7	What, if any, stipulations about training of administrators are made?	None. It appears to be assumed that learners who have worked through the practice section can work through the test unaided.
8	Time needed/allowed, if known.	Varies with ability of learner. Up to half an hour.
9	Is the test available in different media? If so, explain briefly.	No. However, there is a voice-over in addition to written instructions for some items.
10	Is there a clear statement of the population for which the instrument is intended? If so, explain briefly.	Up to Level 2 of the new Standards.
11	a) How is the test 'tailored' for the range of achievement, if at all?	a) For literacy, a screener is provided consisting of ten items testing 1) ability to pick out letters from a range of symbols and 2) ability to select by clicking with a mouse named letters from an array of 5. Parallel items are provided for numeracy. Those not succeeding on this screener are advised to see their tutor. It is acknowledged that problems using the IT may cause failure on the screener. Learners can quit the program at any time.
	b) Is the test pitched at appropriate levels of ease/difficulty for the target learners?	b) Yes, broadly.
12	If the test is stated to be suitable for ESOL as well as for other learners, what evidence is there of differentiation, including in the reporting of results?	N/a.
13	Do the materials seem free from bias? List any difficulties.	Broadly. There are few images in the instrument. Voices on the sound track reflect to some degree the multi-ethnic make up of society. The on-screen figure who speaks to those who fail the screener is from an ethnic minority background. The materials may discriminate against those unused to the technology and to the assessment methods.
14	Are the materials suitable for adults?	Yes.

15	Is there an element of self-assessment in the instrument? If so, explain briefly.	No.
16	Is extraneous knowledge required (e.g. in terms of literacy levels required to comprehend problems/questions, etc./ numeracy/IT)? Comment briefly as appropriate.	General cultural and background knowledge would be an advantage. For example, one item requires background knowledge connected with bank accounts, such as the ability to recognise the format of sort codes. IT and test-taking skills are also required.
17	Does the test report results for ■ reading ■ writing ■ spelling ■ punctuation ■ numeracy?	Results are reported for reading, writing and numeracy (also for speaking and listening). One set of items seems designed to assess knowledge of punctuation, but no result for punctuation is given. The same applies to spelling. See below.
18	a) At what levels are results reported, if reported in levels?	a) Results are reported in terms of levels for reading (text, sentence and word focuses and overall), for writing (sentence and word focuses and overall) and for numeracy. The rationale for the overall assessment of levels for reading and writing is not explained. In addition, results may show that a learner is 'working towards Entry 1'.
	b) If results are reported in different terms, explain briefly.	b) More detailed information may be supplied for those working with the web-based version of the product, which was not tried out.
	c) Has the test been explicitly aligned or re-aligned to the new QCA Standards?	c) Yes.
19	Assessment approach and alignment to the Adult Basic Skills Core Curricula and Standards. Discuss, if and as appropriate, as follows: Is alignment –	
	a) broad;	a) Broad and loose.
	b) covering a range of, or all, elements;	b) Covers selected aspects of the Curriculum and Standards.
	c) suitably balanced in terms of emphasis; and/or	c) See 21 and below.
	d) appropriate in terms of required depth of skill, knowledge and understanding?	d) No.

20	Describe the item types used. Comment if appropriate.	Mainly drag-and-drop, or multiple-choice. Additionally, some items are timed. See below.
21	Do the materials/items appear to test what they are intended to test?	For numeracy, yes. No actual writing is assessed. A few reading items are based upon lifelike texts. The assessment methodology requires the performance of unrealistic tasks.
22	What, if any, technical information is available about the test, e.g. pilots, test-retest reliability, reliability between parallel forms, with different assessors, as 'split halves', relationship with results on other instruments, etc. Comment on any likely difficulties in these areas	No information available to reviewer. N.B. Items which are not attempted are treated as incorrect. Administration conditions are not wholly standardised: for example, learners may choose to have instructions repeated several times by clicking. However, this is not important given the context of use.
23	Is a degree of uncertainty in results, or of limitation in the instrument, recognised, reported, evaluated?	Yes and no. The instrument claims to be 'powerful', yet the accompanying materials acknowledge that certain aspects of the curriculum are not assessed.
24	Are there any other points which might usefully be made about the instrument, e.g. minor proofing errors, tasks that seem particularly difficult?	No.
25	Rate the usefulness of the reported results to a) learners; b) tutors; and c) the wider world (high, medium or low).	 a) Low. b) Low. c) Low.
26	Sum up likely positive and negative effects of this instrument in terms of backwash/ teaching to the test.	This instrument is part of a commercial product which includes 'pre-prepared packs of content' and uses a range of IT-based teaching methods. It seems clear that, from the point of view of the developers, one purpose of the assessment instrument is to encourage the purchase and use of linked products. This may have influenced the form of the instrument. The instrument should, therefore, be used with full awareness of its strengths and limitations, **especially** when negotiating learning plans. See below.
27	Form a judgement about the suitability of the instrument: a) for its stated purpose b) for NRDC purposes in	 a) stated purpose: low. b)

– cohort studies	– cohort studies: low.
– research projects	– research projects: not suitable.
(high, medium or low? Comment below if needed).	

This assessment comes in CD-Rom and web-based versions. It includes a practice section to make sure that learners can use the necessary IT methods, and a simple screener for each skill aimed at distinguishing learners who cannot, for example, identify letters of the alphabet in upper and lower case by clicking on the correct symbol, or who cannot click on a required colour. It is linked with a range of further products, including on-line and CD-Rom based modules and activities, and can be used at home by learners who have a PC. Learners may not move back to previous items in the test, but can, by clicking, hear the instructions repeated. When a pre-test familiarisation process is taken into account, the whole literacy test takes about an hour, the numeracy test about the same.

This instrument was found to be unsuitable for either of the proposed research purposes. It was also found to be of limited value for its stated purposes, which are 'to provide guidance to students and tutors on the appropriate level to begin using *Target Skills* learning materials.' It would be essential for the assessor to be quite clear as to its limitations and to make sure that the students understood them too.

In the following discussion, general points about the materials are made first, followed by more specific remarks about their suitability in respect of each skill.

No technical or design specifications were available to the reviewers, only the CD containing the assessment materials and a Tutor's Guide that contained more marketing text than useful information, but which makes it clear that the results of the assessment are linked with further electronic products – 'pre-prepared packs of content' – available as part of a package. This raises the possibility that the instrument was designed as much to lead into and encourage further purchase and consumption of related products as to provide learners and/or tutors with useful assessment information.

Though the guide frequently uses the words 'powerful' and 'power' to describe the assessment, and describes the outcome as a 'skills profile' upon which an 'action pack' could be based, writing at text focus is not assessed at all, and no actual writing is required of learners in the course of the assessment. Results are reported for Speaking and Listening, though speaking and discussion are not assessed. Probability at Level 1 is not covered.

It was not always clear which elements and/or descriptors within the standards and curriculum particular items were designed to assess, even when results obtained after completing stages of the assessment were printed out. If the provider has bought the whole package and registered the student, the student can pause the assessment and save the results. Later, it would be possible to chart progress though stored results. Though experienced tutors who are familiar with the product may not be confused by the nature of the reporting, it is possible that students, or others seeing the printouts, may be. The fact that it was not always clear to the reviewers which elements items were linked with suggests that students may gain little understanding of their own strengths and weaknesses from the process of using the materials.

Before taking the test, learners take a practice version, in which a voice-over explains the IT techniques involved. The explanations given are too complicated for some of the learners for whom the test is intended, far more complicated than, for example, those in the screening test itself. There is a great deal of new information for a learner unfamiliar with IT to absorb, especially if the learner has learning difficulties, which is possible, given the levels at which the test is pitched.

Literacy

The test itself begins with an initial screener. The first five of the total of ten literacy items present symbols which, though visually similar, are not letters of the English alphabet. The written instructions and voice-over say: 'Only one of these symbols is a letter of the English alphabet. Click on it.' The second group of five items presents letters of the alphabet, some in upper case and some in lower case. The voice-over instructs learners to click on a named letter. The style of the initial items presented for numeracy is similar. If the learner cannot complete these items satisfactorily, a figure on a video clip advises him or her to seek advice from the tutor. It might be better for the tutor to establish this first, rather than subjecting the student to the test.

At this point, the tutor, if he or she has remained present, may have spent as much as half an hour logging the new learner onto the system, going through the practice test, and waiting while the learner clicks on ten different screens. He or she has now acquired as much information about the learner as could have been gained in two minutes simply by asking and by using a paper-based assessment tool. This raises questions about the utility of the tool. It is difficult to see what has been gained by the use of technology, or how the cost of the instrument up to this point can be justified in terms other than the possibility of the process leading to the purchase and use of the linked products. It would appear to be less time-consuming and potentially less frustrating for both learners and tutors if a short paper-based screening test were to be used instead.

Both reviewers encountered some problems with the IT. Sometimes, when it was necessary to drag and drop an item, it wouldn't 'go' for several tries. This could shake learners' confidence in their IT skills as well as in their literacy or numeracy, and it is possible that some students would conclude that they had the wrong answer.

Another feature of the presentation that was found to be distracting was the use of timed questions: *You will see a picture on screen for a short time only. Which of these is in the picture?* A little light beside the three possible answers (tree/boy/dog) changes from red to green to attract attention before the picture comes up. Then a picture appears, with seconds ticking by. A student not comfortable with computers has a lot of extraneous information to process before being able to answer.

For literacy, the items are presented in batches that are broadly similar in presentation, up until the last batch, in which there is significantly more variety in task type and presentation. This contrasts with the widely differing presentation and nature of tasks on an item-by-item basis in some other electronic tests.

For example, each of the first 15 items in the main literacy assessment is a gap-filling task. For each sentence presented, a list of alternative answers is available. The distractors (wrong answers) vary. Some are words that would not make sense in the context, some are incorrect spellings of the correct answer, and others different forms of verbs. The rationale for the choice of distractors is not always clear, but is often linked with parts of speech. All this must be quite confusing for a learner at initial levels of literacy, and add to the difficulty of the items. The options presented with the sentence *There's no hope ___ him* are *from, with, of, for* and *about.* (Arguably, it is possible to imagine contexts in which more than one of these words might make sense.) The options for the sentence *Have you ___________ to her* are *write, wrote, writing, written* and *speak.* The correct answers to the whole batch are as follows: *feed, drive, birth, o'clock, would, were, for, slowly, written, apply, approve, meeting, equal, depressed* and *recycle*. The reviewer completed the screener and this batch of fifteen items before quitting the test. A printout, while making it clear that the test had not been finished, listed results up to this point as follows: Reading, sentence focus: L1; word focus: E1. Writing, sentence and word focus: working towards E1.

Each of the next set of tasks requires the use of drag-and-drop to re-arrange a given group of words to form a sentence. For example, one group is as follows: *your is number what phone*. It was possible to infer from a print-out of results at this point in the test, which showed an improvement in the score for writing, that these items are intended to assess writing at sentence level. This test lacks validity.

The third batch of items is linked with form-filling. Several of the items assess the understanding of, or ability to read, relevant vocabulary: forename, surname, maiden name, occupation, nationality and next of kin. The IT methods here are multiple-choice and drag-and-drop. For example, learners must click on two words in this list which could be put on a form asking about occupation: *mechanic, cook, English, factory, driver, hospital.* For next of kin, selection of the correct definition from a list is required. A problem with all these items is that it is not possible to be certain whether incorrect answers are caused by a lack of reading skills, or of vocabulary, or by other factors, such as, for example, failure to realise that two selections had to be made for some items. As a whole, the set of items lack face validity. Some of them feature texts, few, if any, of which look realistic. One item requires knowledge of terminology, such as 'sort code', linked with a bank account, and involves the unrealistic task of dragging items into the correct position on an extremely un-lifelike form.

The fourth batch of items seems to be designed to assess punctuation, though, as with other IT-based 'writing' assessments, it could be argued that they assess proof-reading skills and lack validity. A variety of correctly and incorrectly punctuated sentences is presented. The learner is instructed to click on one or more items containing errors.

The scoring system at this point in the test is not clear. An overall reading Level of E3 was gained, though text focus was stated to be at E3 and sentence focus L1. Some kind of averaging seems to be used.

The fifth set of items focuses on spelling. A voice-over provides the word both on its own and in a sentence. Learners must select the correct spelling by clicking on a list that includes a number of incorrect spellings as distractors. The words tested are as follows: *were, come, Saturday, broken, unable, might, writing, would, caught, station, enough, terrible, interested, immediately, necessary*.

No rationale for the inclusion of these words is provided. The list has an arbitrary feel to it. In the last set of items, a variety of realistic texts, including small ads and newspaper articles, is presented, each for a limited length of time, which suggests that scanning and/or speed-reading are being assessed. Before seeing the item, the learner is asked a question, which is good practice: in life we often read with a purpose in mind. These questions require different sorts of information about the text to be gleaned from the time-limited reading, including factual information, the purpose of the text, and the main point of an article. A possible disadvantage of this presentational style is that it might make the student feel pressured, and, therefore, less clear headed than usual.

Numeracy
The numeracy assessment claims to adapt levels in response to performance. Presumably a right answer sends the questions off on a different track from a wrong answer. At such points the test design is difficult to track, without working through all the possible available routes or obtaining an algorithmic diagram from the designer.

Different methods of answering are necessary – typing a figure into a box, clicking on (e.g.) three fractions that are equivalent to one half, a set of multiple-choice answers (usually four options), dragging a number or a symbol. Questions use words, figures and symbols, including one which features a picture of a 50p piece, a 10p piece and a 1p piece with addition signs between them and an equals sign following – not a conventional sum, so assumptions are being made about how people read. Even with audio-assistance and tutor assistance, this test makes demands which may be outside the Core Curriculum and could affect results.

The handbook states that pencil and paper are to be available; it does not specify what is to happen to anything written down during the test. A great deal of evidence for assessing awareness of process, choice of operations, etc., may be wasted, if rough working is not formalised in some way. More subtle interactive steps accumulating a notepad record of workings on screen might provide usable information, especially during a diagnostic assessment. This, though, would require yet more technology-specific skill from the student. Perhaps the handbook could suggest that tutors might ask students to keep any paper they use for inspection later.

The mental aspect of processing mathematical information may be a potential victim of the need to assess in such a way. The questions based on prices displayed in a video hire shop are a case in point. Not many hirers would whip out their pencil and work out the best deal on paper in such a situation.

It would be a long process to sort through the Curriculum Elements in this test to see how well progression could be measured in relation to the chart in the *National Standards* (p.17) or the *Adult Numeracy Core Curriculum* (p.16) *Progression between* (within?) *Elements*. It is difficult to see how a provider could ensure that accurate information was being gathered, and not simply have to rely on the designer's good faith and competence.

Does running through the multiple-choice options count as a checking procedure? The requirement at Entry 3 to 'use given methods to check results' does not have an opportunity to manifest; nor did Level 2 'use approximation to corroborate results' – at least, on the route the reviewer took through the materials.

For example, awareness of the 2-dimensional representation of 3-D objects is necessary. At Entry 1, students need to be able to recognise and name common 2-D and 3-D shapes; at Entry 2, to recognise and name 2-D and 3-D shapes, e.g. triangles, cylinders and pyramids, and to describe the properties of common 2-D and 3-D shapes, e.g. the number of sides, corners, faces; at Entry 3 (where this test steps in), to sort 2-D and 3-D shapes to solve practical problems using properties, e.g. lines of symmetry, side length, angles; at Level 1, to solve problems using the mathematical properties of regular 2-D shapes, e.g. tessellation or symmetry; at Level 2, to recognise and use common 2-D representations of 3-D objects, e.g. in maps and plans.

One set of questions was found to move, without sub-classification or any other intra-test categorisation, from a picture giving information about entry prices to a castle, to a bar chart with sales of crisps, then, two questions later, back to the castle. The nature of reality, as discussed elsewhere in this report, kept presenting itself. How much change from £10.00? Is this a real problem? The student knows nothing of the story behind the problem, and is expected to drop any slight interest in it once the subject has been indicated, only to rouse it again when the subject (or the castle) is revisited without rhyme or reason with a different problem attached. 'Two adults went to the castle at the weekend. Would it have been cheaper on Monday?' So the perfect conditional removes the question even from the temporary fictional reality of a narrative past tense.

A similar sequence emerges around Najeeba who wishes to travel to Fore. Her bus timetable reappears a couple of screens later, making the reviewer think initially that there was a need to correct something that had been done wrong.

The unreal choice of place names which don't look like English – Arding/Fore/Brodge – was found to be unhelpful.

A reviewer commented after working through the numeracy materials:

> "The worst thing about this was not knowing where I was. It was a disorientating and disheartening experience. Tutor supervision would need to be well-prepared and continuous if motivation were to be sustained … Later questions were difficult (for those of us around level 2). For example, a question about seeding a lawn gave a formula and some information, and then the problem; there was a question on area, of a room with a chunk cut out. Several stages had to be gone through. However, it did not seem possible to gain any idea of whether the student had the right process in mind or was applying appropriate operations. If the wrong answer were entered, there would be no knowing why."

General
Overall, the claims made for this assessment instrument are overstated. Neither the items themselves nor the print-outs based upon them provide a suitable basis for formative assessment and the development of learning plans, especially of writing. Assertions that they do, especially when apparently approved by a national agency entrusted with the task of raising standards, are regrettable. Further, not only is it difficult to see what has been gained from supplying a CD-based product in this case, but also the use of technology in the assessment process has the potential to affect the validity of assessment, especially for learners who are not used to using IT.

The need for tutor training to enable tutors to evaluate assessment products is highlighted by the conclusions reached about these assessment materials.

Brooks et al., (2001a) noted that surprisingly little had been found on the value of IT-based literacy teaching. On the basis of the quality of formative assessment encouraged by this instrument, there would appear to be cause for concern, and a need for further investigation. More generally, it is stated in the Tutor Guide that some of the modules or 'pre-prepared packs of content' not only explain and teach skills but also assess them. Given the limitations of IT-based assessment techniques, tutors (and awarding bodies) will need to consider the strengths and weaknesses of these assessments carefully.

The attraction of IT-based activities for many learners is clear, especially for those who do not have computers at home. There is no doubt a place for such activities within a balanced programme of work. In order to ensure that this balance is struck tutors need to familiarise themselves with the software, working through the tasks and evaluating them carefully so that they can provide professional advice to students.

On the basis of the evidence available to the reviewers, it would appear that *Target Skills* has a long way to go before it will be able to supplant person-to-person tuition and support, or do more than provide an occasional change or distraction for learners.

12. Number Skills Check

1	Title of instrument. Publisher. Date of publication. Edition no.	***learndirect*** *Number Skills Check* CD-ROM, with Tutor Guide Version 6.0. **learndirect**/Ufi (Nov 2001) Release 5.
2	Cost of instrument per set/copy.	Not known.
3	Is the instrument 'secure' (unpublished)?	No.
4	Stated purpose of instrument.	'Diagnostic product aimed at basic skills learners... to help them assess their ability in numeracy and plan future learning.'
5	Are there parallel forms? If so, how many?	Only in the sense that the instrument is adaptive.
6	Are there any obvious manageability problems?	The CD-Rom version supplied for review was impenetrable on two different computers. Up-to-date computer equipment required.
7	What, if any, stipulations about training of administrators are made?	Administration via a **learndirect** centre implies trained staff.
8	Time needed/allowed, if known.	Three hours to cover number, measure, shape and space, and data handling, according to the **learndirect** website (accessed 11/9/04), but this depends on the learner; the higher the levels of the learner, the longer the assessment takes.
9	Is the test available in different media? If so, explain briefly.	Yes: now available (11/9/04) on CD-Rom, online at a learndirect centre, in a printed version, and on video and audio cassette.

10	Is there a clear statement of the population for which the instrument is intended? If so, explain briefly.	Basic skills learners of any age or background, likely to be registered with a **learndirect** centre.
11	a) How is the test 'tailored' for the range of achievement, if at all? b) Is the test pitched at appropriate levels of ease/difficulty for the target learners?	a) and b) 'adaptive algorithm'.
12	If the test is stated to be suitable for ESOL as well as for other learners, what evidence is there of differentiation, including in the reporting of results?	Not specified.
13	Do the materials seem free from bias? List any difficulties.	Yes.
14	Are the materials suitable for adults?	Yes.
15	Is there an element of self-assessment in the instrument? If so, explain briefly.	Learning map produced for learners' use. Learning objectives specify: 'Be able to describe personal benefits obtained from using learning tool.'
16	Is extraneous knowledge required (e.g. in terms of literacy levels required to comprehend problems/questions, etc./ numeracy/IT)? Comment briefly as appropriate.	Literacy. Some IT. Learning objectives include learning some IT.
17	Does the test report results for ■ reading ■ writing ■ spelling ■ punctuation ■ numeracy?	Numeracy.
18	a) At what levels are results reported, if reported in levels? b) If results are reported in different terms, explain briefly. c) Has the test been explicitly aligned or re-aligned to the new QCA Standards?	a) from Entry1–Level 2 b) Also reported in subject areas with information available about next step. c) Yes.

19	Assessment approach and alignment to the Adult Basic Skills Core Curricula and Standards. Discuss, if and as appropriate, as follows: Is alignment –	
	a) broad;	a) Close.
	b) covering a range of, or all, elements;	b) Apparently all if enough items accessed.
	c) suitably balanced in terms of emphasis; and/or	c) Yes.
	d) appropriate in terms of required depth of skill, knowledge and understanding?	d) Yes.
20	Describe the item types used. Comment if appropriate.	Item types (from tutor guide): multiple-choice, using mouse to select one option; multiple-choice, using mouse to select two options; gap filling – type in answer; drag-and-drop matching; drag-and-drop reordering; image map, click on image.
21	Do the materials/items appear to test what they are intended to test?	Broadly.
22	What, if any, technical information is available about the test, e.g. pilots, test-retest reliability, reliability between parallel forms, with different assessors, as ‘split halves’, relationship with results on other instruments, etc. Comment on any likely difficulties in these areas.	Extensive tutor guide. Test development and reliability etc not gone into – reference only to learning materials after diagnosis.
23	Is a degree of uncertainty in results, or of limitation in the instrument, recognised, reported, evaluated?	No.
24	Are there any other points which might usefully be made about the instrument, e.g. minor proofing errors, tasks that seem particularly difficult?	No.
25	Rate the usefulness of the reported results to	
	a) learners;	a) High.
	b) tutors; and	b) High.
	c) the wider world (high, medium or low).	c) Low.
26	Sum up likely positive and negative effects of this instrument in terms of backwash/ teaching to the test.	N/a. Tight learning-assessment circle set up by design.

27	Form a judgement about the suitability of the instrument: a) for its stated purpose; b) for NRDC purposes in – cohort studies – research projects (high, medium or low? Comment below if needed).	a) High b) – cohort studies: low – research projects: low

This is an adaptive CD-based test, linked with instructional materials available from the same supplier. It relies strongly on multiple-choice items, though some other types, including drag-and-drop, are also used. Learners are not told whether they have completed items correctly or incorrectly, and results are reported in terms of (idiosyncratically named) levels – see box 18.

These materials were not found to be suitable for cohort studies or research projects. The range was not appropriate.

No detailed technical specifications for these materials were available. Though it sometimes seemed clear that a particular item was designed to assess a particular Element, this was not always the case. A tutor guide contained on the CD-Rom was printed off and read. The Guide suggested that the materials could be used to check on progress. This implies that there are so many items that a learner would not be presented with the same one twice. The reviewer was unable to check this out, and insufficient information was available. The possibility exists that a learner could achieve successively better results simply by learning how to respond to each item in a rote fashion, rather as a certificate can be obtained from some other **learndirect** basic skills software on a trial and error basis.

13. Word Skills Check. Release 2.

1	Title of instrument. Publisher. Date of publication. Edition no.	**Learndirect**/Ufi *Word Skills* Check CD Rom, with Tutor Guide Document Version 6.0 (October 2001) Release 2.
2	Cost of instrument per set/copy.	Information not available.
3	Is the instrument 'secure' (unpublished)?	No.
4	Stated purpose of instrument.	Formative/diagnostic. 'When you have completed the Skills Check, you and your tutor will be able to plan the best course of learning for you to follow using the learning map. As part of an ongoing self-assessment programme, print-outs of learning maps showing progress could be used as evidence of achievement and competency.'
5	Are there parallel forms? If so, how many?	Only in the sense that the instrument is adaptive.

6	Are there any obvious manageability problems?	No.
7	What, if any, stipulations about training of administrators are made?	Intended for use by student with tutor support via tutors, including telephone support.
8	Time needed/allowed, if known.	Three hours to cover speaking and listening, reading and writing, according to the **learndirect** website (accessed 11/9/04), but this depends on the learner; the higher the levels of the learner, the longer the assessment takes.
9	Is the test available in different media? If so, explain briefly.	Yes: now available (11/9/04) on CD-Rom, online at a **learndirect** centre, in a printed version, and on video and audio cassette.
10	Is there a clear statement of the population for which the instrument is intended? If so, explain briefly.	No.
11	a) How is the test 'tailored' for the range of achievement, if at all? b) Is the test pitched at appropriate levels of ease/difficulty for the target learners?	a) The test claims to be adaptive. The items presented to a student depend upon his or her responses to previous items. b) Broadly.
12	If the test is stated to be suitable for ESOL as well as for other learners, what evidence is there of differentiation, including in the reporting of results?	N/a.
13	Do the materials seem free from bias? List any difficulties.	General cultural and background knowledge would be an advantage.
14	Are the materials suitable for adults?	Yes.
15	Is there an element of self-assessment in the instrument? If so, explain briefly.	No.
16	Is extraneous knowledge required (e.g. in terms of literacy levels required to comprehend problems/questions, etc./ numeracy/IT)? Comment briefly as appropriate.	Familiarity with IT skills would be an advantage. Success would also depend upon test-taking skills.
17	Does the test report results for ■ reading ■ writing ■ spelling ■ punctuation ■ numeracy?	Reading. Writing. (Also Speaking and Listening).

18	a) At what levels are results reported, if reported in levels? b) If results are reported in different terms, explain briefly. c) Has the test been explicitly aligned or re-aligned to the new QCA Standards?	a) Starting Out; Stepping Up; On the Road; Getting There; No Limits. b) N/a. c) No.
19	Assessment approach and alignment to the Adult Basic Skills Core Curricula and Standards. Discuss, if and as appropriate, as follows: Is alignment – a) broad; b) covering a range of, or all, elements; c) suitably balanced in terms of emphasis; and/or d) appropriate in terms of required depth of skill, knowledge and understanding?	a) Broad. b) A range of elements. c) Depends largely on the route an individual takes through the test. d) Varying degrees of acceptability. See below.
20	Describe the item types used. Comment if appropriate.	See below.
21	Do the materials/items appear to test what they are intended to test?	See below.
22	What, if any, technical information is available about the test, e.g. pilots, test-retest reliability, reliability between parallel forms, with different assessors, as 'split halves', relationship with results on other instruments, etc. Comment on any likely difficulties in these areas.	No technical information was available to the reviewers.
23	Is a degree of uncertainty in results, or of limitation in the instrument, recognised, reported, evaluated?	No.
24	Are there any other points which might usefully be made about the instrument, e.g. minor proofing errors, tasks that seem particularly difficult?	See below. Some typographical errors were found.
25	Rate the usefulness of the reported results to a) learners; b) tutors; and c) the wider world (high, medium or low).	a) Low. b) Low. c) Low.

26	Sum up likely positive and negative effects of this instrument in terms of backwash/ teaching to the test.	The test is linked with teaching materials produced by the same organisation. These materials were not reviewed.
27	Form a judgement about the suitability of the instrument: a) for its stated purpose; b) for NRDC purposes in – cohort studies – research projects (high, medium or low? Comment below if needed).	a) Low – see below. b) – cohort studies: not suitable, range too limited. – research projects: not suitable, would not. meet criteria for a good test.

This is an adaptive CD-based test, linked with instructional materials available from the same supplier. It relies strongly on multiple-choice items, though some other types, including drag-and-drop, are also used. Learners are not told whether they have completed items correctly or incorrectly, and results are reported in terms of (idiosyncratically named) levels for speaking and listening, reading and writing – see box 18.

These materials were not found to be suitable for cohort studies. The range was not appropriate. Neither were they found to be suitable for research projects: the assessment of writing lacked validity, and the quality of the items used for assessing reading was too variable.

No detailed technical specifications for these materials were available. Though it sometimes seemed clear that a particular item was designed to assess a particular Element, this was not always the case. A tutor guide contained on the CD-Rom was printed off and read. The guide suggested that the materials could be used to check on progress. This implies that there are so many items that a learner would not be presented with the same one twice. The reviewer was unable to check this out, and insufficient information was available. The possibility exists that a learner could achieve successively better results simply by learning how to respond to each item in a rote fashion, rather as a certificate can be obtained from some other **learndirect** basic skills software on a trial and error basis.

A number of questions about reference systems and alphabetical order attempted to contextualise knowledge and skills by using realistic images of, for example, dictionary entries or reference systems, with varying degrees of success. On one such question, success depended upon the student having background knowledge about mother of pearl which was not provided.

The items assessing writing shared the general lack of validity that has been commented upon in other IT-based instruments. One item presented this sentence: 'Mum bought me a long, stripy, red and yellow scarf to wrap around my doll's neck.' The sentence appears under a picture of a child looking at dolls through a shop window. There is no illustration of a scarf. The set task is to choose from a list words that 'best describe' the scarf. The choices are as follows: *long, stripy, red and yellow* and *wrap*. This task is impossible. The word 'wrap' can be eliminated, as it is not a describing word. The plural 'words' suggests that more than one answer must be clicked on, and the word 'best' suggests that some kind of judgement must be made. As there is no picture of the scarf, it is not possible to judge which of the words best describe it. With general knowledge of the Curriculum it is possible to hazard a guess that the intention of the item is to test whether students can pick out adjectives and adjectival phrases, in which case the question is misleadingly worded. The tutor guide lists 'Use adjectives' as an objective at Entry Level 2. This item looks like an attempt to create an IT version of the rather dubious kind of task in which learners are required, for reasons that are not always clear, either to underline descriptive words and phrases or to insert such words and phrases into gaps in pre-prepared texts. If so, the attempt has not been successful, and may not have been worth making. The item shows the problems that arise with validity when attempts are made to assess discrete bits of literacy in this way.

In some but not all items, the task was linked with the image on the screen. The images themselves varied in appropriateness. Sometimes, they were only marginally relevant to the topic of the text. A student taking the test would have to work out for each item whether he or she needed to interpret the image or not. Where the image was supposed to represent a text, line numbers were often placed within it, potentially confusing the reader. For example, an image apparently intended to represent words typed on a screen of some kind read as follows:

Asif

Line 1 jim phoned. He will call thursday
Line 2 to collect the brown file Please leave
Line 3 it is in the office by the window.

Sukhra

Two questions are asked about this text, under which boxes for answers to be typed in are provided. Neither question deals with the superfluous 'is' in line 3.

Items assessing spelling sometimes used multiple-choice lists that included incorrectly spelt words, but poor spellers do not gain from being exposed to incorrectly spelt words.

The items were judged to be of variable quality, and to be, on the whole, more satisfactory at higher levels. One or two multiple-choice items based around the main idea of the text were very good. Such is the variety of ways in which tasks were set and answers could be given that, taken as a whole, the materials tended to resemble a set of puzzles rather than an assessment of literacy. Particularly unsatisfactory were tasks in which blocks of text have to be moved into a new order, using a 'resting place' at the foot of the screen to keep items in while others are moved into their place. Explicit instructions about how to complete each item were not always provided. There was a lack of internal consistency and coherence about the materials.

A number of items had errors of spelling or punctuation. For example, an item about writing clearly and spelling correctly expressed as a question had no question mark. The instructions on another item read: *Fill in the gaps with either a capital letter or with a lower case letter.* An item about lifeguards contained two spelling errors.

14. National Literacy Test, Level 1, Practice Version

1	Title of instrument. Publisher. Date of publication. Edition no.	Practice Versions of the National Test. http://dfes.gov.uk/readwriteplus/learning Accessed 29/05/2002 Also available in hard copy.
2	Cost of instrument per set/copy.	N/a.
3	Is the instrument 'secure' (unpublished)?	No.
4	Stated purpose of instrument.	Summative. To provide valued qualifications based upon the new Standards.
5	Are there parallel forms? If so, how many?	Not of this version, but the live versions will vary every time.
6	Are there any obvious manageability problems?	No.
7	What, if any, stipulations about training of administrators are made?	Guidelines about conditions are given.
8	Time needed/allowed, if known.	One hour.
9	Is the test available in different media? If so, explain briefly.	No.
10	Is there a clear statement of the population for which the instrument is intended? If so, explain briefly.	Adult basic skills learners.
11	a) How is the test 'tailored' for the range of achievement, if at all? b) Is the test pitched at appropriate levels of ease/difficulty for the target learners?	a) Learners will take the tests when they and their tutors feel they are ready – or that was the intention. Anecdotally, it seems that some centres use them as end-of-course tests whether learners are 'ready' for them or not. b) Yes.
12	If the test is stated to be suitable for ESOL as well as for other learners, what evidence is there of differentiation, including in the reporting of results?	N/a.
13	Do the materials seem free from bias? List any difficulties.	Broadly.
14	Are the materials suitable for adults?	Generally.
15	Is there an element of self-assessment in the instrument? If so, explain briefly.	No.

16	Is extraneous knowledge required (e.g. in terms of literacy levels required to comprehend problems/questions, etc./ numeracy/IT)? Comment briefly as appropriate.	General background and cultural knowledge would be an advantage.
17	Does the test report results for ■ reading ■ writing ■ spelling ■ punctuation ■ numeracy?	An overall level is given. The test claims to cover some aspects of writing as well as reading.
18	a) At what levels are results reported, if reported in levels? b) If results are reported in different terms, explain briefly. c) Has the test been explicitly aligned or re-aligned to the new QCA Standards?	a) In levels to correspond both with Key Skills and Basic Skills at Levels 1 and 2. b) N/a. c) Yes.
19	Assessment approach and alignment to the Adult Basic Skills Core Curricula and Standards. Discuss, if and as appropriate, as follows: Is alignment – a) broad; b) covering a range of, or all, elements; c) suitably balanced in terms of emphasis; and/or d) appropriate in terms of required depth of skill, knowledge and understanding?	a) Broad. b) A range of elements. c) See below. The balance on the whole is better for reading than for writing. d) Broadly.
20	Describe the item types used. Comment if appropriate.	Multiple-choice items, involving the selection of an answer by clicking.
21	Do the materials/items appear to test what they are intended to test?	Broadly.
22	What, if any, technical information is available about the test, e.g. pilots, test-retest reliability, reliability between parallel forms, with different assessors, as 'split halves', relationship with results on other instruments, etc. Comment on any likely difficulties in these areas.	None available to reviewer.
23	Is a degree of uncertainty in results, or of limitation in the instrument, recognised, reported, evaluated?	No.

24	Are there any other points which might usefully be made about the instrument, e.g. minor proofing errors, tasks that seem particularly difficult?	No.
25	Rate the usefulness of the reported results to a) learners; b) tutors; and c) the wider world (high, medium or low).	 a) Medium. b) Low. c) Medium.
26	Sum up likely positive and negative effects of this instrument in terms of backwash/ teaching to the test.	Students will need to familiarise themselves with the techniques and conventions of multiple-choice assessments. As their tutors gain experience of the item types, they will be able to provide support in this.
27	Form a judgement about the suitability of the instrument: a) for its stated purpose; b) for NRDC purposes in – cohort studies – research projects (high, medium or low? Comment below if needed).	 a) See below. b) – cohort studies: range not suitable. – research projects: range not suitable.

For additional comment on this instrument, see Heath (2003).

These new tests bear a strong resemblance to the existing QCA Key Skills tests, and serve for both key and basic skills at Levels 1 and 2 (except that gaining the key skills certificate requires submission of a portfolio as well as completion of the test). Designed to be summative, they are perhaps the most successful example of the use of multiple-choice materials reviewed in the course of this project. The website said: 'From September 2002, all adult literacy and numeracy qualifications at Levels 1 and 2 will be based on the new National Tests launched last September. The tests are directly related to the new national standards for adult literacy and numeracy. This means that learners, tutors and employers will have a clear understanding of what has been achieved and will value the resulting qualifications.'

These materials were not judged to be suitable for either proposed research purpose. Each test is aimed at a specific level of achievement. They were judged to be of medium suitability for their stated purposes. No sample of writing is required in these tests.

Paper-based exemplar key and basic skills tests have been available on the internet for some time. These have been made up of 40 multiple-choice questions arranged in groups linked with a particular realistic text or scenario. The pass marks were set at awarding meetings where 'quantitative and qualitative information on each question was used to set the pass boundaries' (QCA website, http://www.qca.org.uk/nq/ks/level_1_2_tests.asp Accessed 01/05/2002). More recently, IT-based practice versions have appeared under the label 'Adult Literacy and Numeracy Tests'. An example literacy test at Level 1 is the main subject of this review.

Perhaps because it is based upon the earlier, paper-based key skills tests, the Level 1 literacy test uses a reasonably predictable set of multiple-choice item types, rather than facing an unprepared learner with a baffling succession of different question formats and IT-based response methods. Tasks are either in the form of a question followed by a list of four possible answers, or in the form of a stem, such as 'The aim of the leaflet is …', followed by a list of four possible endings.

As it is generally understood that learners will be prepared for these tests, as opposed to encountering them as initial or diagnostic assessments, it can be assumed that they will have been thoroughly familiarised with the assessment techniques, will know roughly the kind of item to expect, and will have been given sensible advice about how to tackle multiple choice questions. The fact that whole practice tests are available is in itself a positive sign. This different context for assessment makes the use of the multiple-choice testing technique and the IT technology more acceptable.

The texts used as a basis for the test included a television programme schedule, which had eight questions based upon it, a train company leaflet with four questions based upon it, and a personal profile and job advertisement which, together, had seven questions based upon them. Most questions required the student to find factual information.

Like other multiple-choice items reviewed, some items in this test seemed to imply a rather thin, watered-down interpretation of complex Elements. This applied to two questions about which of a set of images would be most suitable for a particular purpose, and to a question asking whether a particular statement was a fact, an opinion, an idea or an instruction. (Arguably it is an idea and an opinion.) One or two items seemed more like puzzles than tests of literacy; they offered complicated sets of alternative answers that needed to be puzzled out using strategies that a student would not be likely to need to use in a real-life situation. An item giving sets of alternative times for the News is an example here.

Items testing spelling and vocabulary appear through the test: they are small in number, especially when compared with the heavy weighting given to spelling in the new *Initial Assessment* materials.

15. Tests for *Skills for Life* survey of needs

1	Title of instrument. Publisher. Date of publication. Edition no.	National Basic Skills Baseline Survey (2000). Developed by CDELL and unpublished, but items were available to the director of the review project. A few are reproduced in Williams et al., (2003) *The **Skills for Life** Survey*. London: DfES, which also provides much of the information analysed below.
2	Cost of instrument per set/copy.	N/a.
3	Is the instrument 'secure' (unpublished)?	Yes.
4	Stated purpose of instrument.	National baseline survey of adult literacy and numeracy (and ICT).
5	Are there parallel forms? If so, how many?	Only in the sense that the instrument is adaptive.

6	Are there any obvious manageability problems?	No.
7	What, if any, stipulations about training of administrators are made?	The administrators were employees of a market research organisation and were well trained.
8	Time needed/allowed, if known.	About 25 minutes.
9	Is the test available in different media? If so, explain briefly.	No.
10	Is there a clear statement of the population for which the instrument is intended? If so, explain briefly.	A representative (stratified random) sample of the adult population of England. Industry and further education colleges.
11	a) How is the test 'tailored' for the range of achievement, if at all?	a) An initial screening test is used. The respondent is then presented with a series of items depending upon his or her responses to previous items. The test is 'layered', with progress across levels from layer to layer determined by the calculation of a percentage mark. Items on different layers are linked with different assessment criteria from the test specification. For comment on how the items were allocated to levels see below.
	b) Is the test pitched at appropriate levels of ease/difficulty for the target learners?	b) Very broadly. The text used in the symbol-matching task including a skull may be too difficult for learners at that level.
12	If the test is stated to be suitable for ESOL as well as for other learners, what evidence is there of differentiation, including in the reporting of results?	The sample included ESOL adults, and the report gives results for ESOL and native speakers of English. The tests were not adapted for ESOL adults.
13	Do the materials seem free from bias? List any difficulties.	Broadly. There is a possibility of bias in the administration methods: it may be the case that more men than women, or more adults from certain social or age groups than from others, have acquired familiarity with IT through, say, employment outside the home. Most names used are English and the people pictured are all white.
14	Are the materials suitable for adults?	Yes.
15	Is there an element of self-assessment in the instrument? If so, explain briefly.	Not in the literacy and numeracy sections (some in the ICT section).

16	Is extraneous knowledge required (e.g. in terms of literacy levels required to comprehend problems/questions, etc./ numeracy/IT)? Comment briefly as appropriate.	General cultural and background knowledge would be an advantage. Some items, such as that about hotels, seem rather middle class. The electronic format of the test will disadvantage those without previous experience of IT-based presentation of text. Not all adults have experience of multiple-choice assessments, though some will have taken 16+ exams involving them. It was felt that the way in which some instructions were expressed contributed more to the difficulty of items than the actual tasks themselves.
17	Does the test report results for ■ reading ■ writing ■ spelling ■ punctuation ■ numeracy?	Literacy (actually only reading) and numeracy (also ICT).
18	a) At what levels are results reported, if reported in levels? b) If results are reported in different terms, explain briefly. c) Has the test been explicitly aligned or re-aligned to the new QCA Standards?	a) Entry Level 1 or below to Level 2 or above b) N/a. c) Yes, but see below.
19	Assessment approach and alignment to the Adult Basic Skills Core Curricula and Standards. Discuss, if and as appropriate, as follows: Is alignment – a) broad; b) covering a range of, or all, elements; c) suitably balanced in terms of emphasis; and/or d) appropriate in terms of required depth of skill, knowledge and understanding?	a) Yes. b) A selection of assessment criteria, based loosely upon the Standards, is covered. c) Yes except for the omission of any assessment of writing. d) See below and question 22.
20	Describe the item types used. Comment if appropriate.	See below.
21	Do the materials/items appear to test what they are intended to test?	See below.

22	What, if any, technical information is available about the test, e.g. pilots, test-retest reliability, reliability between parallel forms, with different assessors, as 'split halves', relationship with results on other instruments etc. Comment on any likely difficulties in these areas.	The report says very little about piloting (see below) and nothing about reliability.
23	Is a degree of uncertainty in results, or of limitation in the instrument, recognised, reported, evaluated?	No – estimated proportions of the population at the different levels are reported without standard deviations or confidence intervals.
24	Are there any other points which might usefully be made about the instrument, e.g. minor proofing errors, tasks that seem particularly difficult?	Some tasks were found to be unsatisfactory. See below.
25	Rate the usefulness of the reported results to a) learners; b) tutors; and c) the wider world (high, medium or low).	a) Low. b) Low. c) Low.
26	Sum up likely positive and negative effects of this instrument in terms of backwash/ teaching to the test.	Not directly applicable to this version, but the parallel key skills tests may well influence teaching.
27	Form a judgement about the suitability of the instrument: a) for its stated purpose; b) for NRDC purposes in – cohort studies – research projects (high, medium or low? Comment below if needed).	a) Low. b) – cohort studies: not suitable. – research projects: not suitable.

The literacy and numeracy tests were administered using a lap-top computer with sound. Each was made up of an initial screening test, followed by further rounds, or stages, of items – two rounds for literacy, six for numeracy. The level of multiple-choice items presented to any respondent at later stages depended upon their performance at previous stages. Movement across levels from stage to stage, and the final reported level, depended upon percentage scores. Thus the tests were not fully adaptive in the sense that each item presented depended on all those attempted previously, but might be called 'adaptive in blocks' since the score on each batch of items seen determined which further batch each test-taker was routed to next. Each test took no more than 25 minutes to complete.

It cannot be unreservedly recommended that these tests be used for cohort study purposes. The numeracy version was judged to have more potential than the literacy version for this purpose. The literacy test is unsuitable for research projects because of the exclusively multiple-choice response format; for the use of some of the numeracy items in NRDC research projects see the main text. There are questions about the validity and reliability of the tests that make their suitability for their own purposes limited.

Literacy
In addition to a preview version of this test, the materials available to the reviewers included a chart setting out layers and levels, a draft test specification dated 19 December 2001, and the final published report on the survey (Williams et al., 2003).

No information was available about, for example, interviewer scripts, or about the training to be given to interviewers. However, a careful attempt seems to have been made to align the items, subject to the limitations of the assessment techniques used, with the Descriptors in the Standards (the Core Curriculum was not yet available when the test was developed, Williams et al., 2003, p 223).

The draft test specification lists at each level a number of points, referred to as 'Assessment Criteria'. These points do not correspond exactly with the descriptors listed in the standards, though they seem to be condensed and edited versions of (and therefore to some extent different from) them. Often the descriptors have, in effect, been simplified, and it seems possible that this was done to make it easier to use multiple-choice assessment techniques. If so, this is an interesting example of the likely effect of the widespread use of multiple-choice assessment techniques in education and training contexts. One rarely sees these processes actually taking place.

The two versions of reading at Level 1 will now be given and discussed to illustrate this point.

The Standard for reading at Level 1 is as follows:

At this level, adults can:

- read and understand straightforward texts of varying length on a variety of topics accurately and independently; and
- read and obtain information from different sources.

An adult will be expected to:

- trace and understand the main events of continuous descriptive, explanatory and persuasive texts;
- recognise how language and other textual features are used to achieve different purposes *e.g. to instruct, explain, describe, persuade;*
- identify the main points and supporting detail, and infer meaning from images which is not explicit in the text;
- use organisational and structural features to locate information, *e.g. contents, index, menus, subheadings, paragraphs;*
- use different reading strategies to find and obtain information and
- use reference material to find the meaning of unfamiliar words in reports, instructional, explanatory and persuasive texts.

The assessment criteria in the draft test specification are as follows:

- L1.2 Reading

  When reading reports, instructional, explanatory and persuasive texts, the adult can:

  – L.1.2.1 understand and summarise the main points *(e.g. identify main points from a sales advertisement);*

  – L.1.2.2 recognise the different ways text is used for different purposes e.g. to instruct, explain, describe, persuade *(e.g. match text to purpose from range of instructional/explanatory/ descriptive/ persuasive documents);*

  – L1.2.3 identify how images complement text *(e.g. choose from three or four pictures the most appropriate to illustrate a section of a holiday brochure);*

  – L.1.2.4 use a variety of organisational and structural features to locate information *(e.g. use an index in a manual to find an electrical component);*

  – L.1.2.5 use a given source to find the meaning of unfamiliar words *(e.g. look up meaning of specialist words in a glossary).*

The first point to be made here is that the Standard itself could be clearer: in what sense can descriptive texts be said to have 'main events', for example, and why are descriptive texts absent from the list of text types at the foot of the Standards?

It should also be noted that the complex statement that appears before the list of descriptors in the standards about the nature of texts, their length, and familiarity to the learner, and the degree of independence and accuracy of the reading, is absent from the test specification. Such statements appear at each level of the Standards and are intended to 'help identify skill development' (QCA website, http://www.qca.org.uk/nq/ks/level_1_2_tests.asp Accessed 01/05/2002). Neglecting these statements may affect alignment for those achieving both at the higher and at the lower levels of the standards. Within the ***Skills for Life*** survey tests, the length of text is limited by the size of the screen, and by the need to save space for instructions, items to be dragged into place, and so on. This affects alignment at higher levels. At lower levels, where learners are expected to be following relatively individualised programmes of study based around familiar topics, standardised tests may introduce an added element of difficulty in terms of unfamiliarity.

Five assessment criteria in the test specification replace six descriptors in the standards. The first Descriptor has been omitted. So has the fifth, though strategies it refers to underpin success on most, if not all, of the items in the actual instrument.

The second descriptor has been rephrased in such a way as to subtly alter its meaning. Originally, it indicated that learners were expected to be developing the ability to analyse the language and other features of texts to see how these are used to achieve effects and purposes. The assessment criterion has shifted the focus to identifying a single purpose for a whole text, as is emphasised by the kind of example quoted, and something of the richness of the original intention has been lost.

A similar point about the comparative thinness of the assessment criteria could be made in respect of the third descriptor, where the references to 'supporting detail' and to the need to draw inferences rather than merely to read at a literal level have been lost. In both examples, the changes made result in something that appears easier to assess using the very limited assessment techniques available to those using IT-based materials. For L.1.2.1, the assessment task might involve clicking on a selection of alternative 'main points', some of which would be factually incorrect statements about a text, or dragging the correct summary of a section of text into a certain position on the screen. For L.1.2.2 the assessment task might involve presenting a text and requiring the subject to click on one of a list of purposes. Both kinds of task are in fact used in the instrument.

The third assessment criterion seems to be drawn from the requirement in the second descriptor that learners at this level should be able to obtain meaning from images. Again, some sense of depth of understanding and interpretation seems to have been lost, even at Level 1. The suggested task here, while allowing learners to judge which image might be best for a particular text, requires no ability to express ideas about the image or the purpose for which it is used, criteria strongly implied by the descriptor.

To sum up, alignment with the standards could be characterised as broad and fragmentary, rather than being precise at an element-by-element level. There are questions about whether the instrument adequately reflects the depth of knowledge and understanding indicated in the adult literacy standards.

There is also doubt about the rationale for the 'layering' of the items. The Level 1 and 2 literacy (and numeracy) items were taken from the key skills tests, and in the process adapted from paper-based to computer-based format. Also, some Entry Level items were adaptations of Level 1 items from the key skills tests – hence these items were doubly adapted, not only from paper-based to computer-based format, but also from one level to another. In order to validate their use in this survey these items should have been put through extensive trialling, but the report calls the adapted Level 1 and 2 items for both areas 'tried and tested' (p.14) – which they were not, in their new format – and for the literacy items states that 'it was assumed that all Level 1 and 2 items were valid and reliable' (p. 223). And as with many of the instruments discussed in this review, the measure of 'literacy' is in fact a measure only of reading, since no aspect of writing is assessed.

Few of the reading items look realistic. Sometimes this is because graphics have been used to decorate the page. For example, a picture of an injured child is positioned to one side of the absence note sent to school. At other times it is because the area of the screen intended to represent a real life text has not been clearly marked off from the rest of the screen. The buttons that have to be pressed to release drop-down menus in some tasks add to this sense of unreality. Negotiating these difficulties for each successive item requires literacy and/or ICT skills that are not explicitly mentioned in the test specification for the items in question and affect the validity of the tasks as measures of the listed assessment criteria.

In real life, we are rarely faced with reading tasks in which the main aim is the selection of one from a list of four possible items. In the case of writing, a paper-and-pencil test may well be the most valid, since writing is often assessed through tests of technical accuracy which do not represent the real-life needs or experiences of many learners, and often have a somewhat arbitrary feel to them.

The reviewer often used strategies learned through past experience of taking, and of teaching to, multiple-choice tests to help to solve items. It could be argued that some tasks have more than one possible correct answer. In this situation one has to decide which answer the test setter probably had in mind. This test could, therefore, be criticised for involving test-taking and/or problem-solving skills, including awareness of the kind of strategy used by test-setters to achieve good discrimination between items.

It is possible to obtain a correct answer, or at least to eliminate some of the choices proffered, for some items without actually reading the text. Knowledge of the world and common sense suffice. The item about the CD player is one such example. Such items are not valid tests of reading. Though the definition and division of reading into component parts implied in the Standards' reference to 'read and understand' and 'read and obtain information from' is far from perfect, it provides some kind of measure against which to judge the validity of the items used in this test.

The version of the assessment available to the reviewer was long, and was not the adaptive version actually used in the survey.

In the version reviewed, the limitations of the technology meant that there were long pauses between one item and the next, emphasised by the appearance of an irrelevant page on the screen. There was a growing temptation to click any button to get through the items. A learner without well developed skimming and scanning skills would be more liable to become fatigued and impatient than a more practised reader. This could affect the validity of the test.

The quality of the items was variable. One task was based on a postcard from Spain. Punctuation errors remained in the text after all the available punctuation had been dragged into place. Another set of tasks used a mock job advert. One of these asked about the purpose of the advert. Since the previous two questions had asked questions about the name of the firm offering the job and the hours that the successful applicant would be required to work, the order of questions here seems illogical.

The instructions for some tasks were not clear, for example those with the task based upon a note to school. This task involves clicking on a word in the text, which causes a drop-down menu to appear. The respondent had to click again to choose the correct option from this menu. The instructions did not make all the steps in this process clear. The instructions for the text about the invention of television were not well written.

The reviewer was unable to obtain a correct response to her answer to the question about Spotlight torches. This task was carried out by scanning for a key word, though it seems designed to assess ability to use alphabetical order. Trial and error could also play a part in success since clicking on a wrong answer produces no response on screen. The item was not judged to be valid.

The task about 'Kleenquick' could not be answered without reference to the statement in the instructions that the text is an advertising text. On its own, the text could be interpreted as the script for a comedy show. This means that the item is not valid, as it is apparently intended to test understanding of the purposes of texts derived from reading of those texts, and not from reading relatively clear and simple statements of such.

Some of the tasks required quite fine manipulation of the mouse. An item not dragged quite carefully sometimes fell back into place. This could confuse a respondent who might incorrectly assume that they had chosen wrongly.

All the instructions in this instrument should have been checked against the Standards to ensure that they comply with the levels of competence in following written instructions set out therein. There appear to be some mismatches.

Numeracy

Those taking the version of this test seen by the reviewer were given responses to their answers, though this appears not to have been the case in the survey itself. On one item with four parts to the answer, a result of 0 per cent was given even when the reviewer had got one, two or three parts correct. The version of the materials available to the reviewer did not seem to be adaptive – but this was probably due to the 'batching' of items mentioned above.

Sometimes common sense knowledge, rather than numerical knowledge, was required, for example about the likely time of day for a particular activity. Sometimes a hint was signalled, usually indicating that there could be more than one right answer.

Graphic material was often good, using photographs with a three-dimensional feel to them.

The ordering of items in terms of their content was bewildering. One item used a picture of metal brackets, which reappeared several items later.

In one item key words were in upper case letters in a text and in lower case letters in the question.

Many of the items were contextualised using characters, but sometimes the test-taker was asked to imagine him or herself in a certain situation. This lack of consistency is potentially confusing for those taking the test.

Many of the items required understanding of the conventions of assessment. A bar-chart of a doctor's incoming telephone calls, with the task of choosing the true statements from the false ones in a list, required a particular understanding that the question meant 'true in terms of this chart in this test'.

The choice of a fridge-freezer to fit into a space in a kitchen design may well be a real-life task, culminating in the need for just such a drawing and set of information as the test presented, but the context itself was so different, not least in terms of the time-scale, that it was hard to say it felt real. Moreover, it implied a certain economic status, which could be exclusive.

There was obviously increasing complexity as the questions proceeded. Sometimes this was evident in the amount of information which needed sorting before the mathematics began. Considerable reading skill and capacity to select relevant material was required, some of it verging on document literacy, without the distinction from quantitative literacy being made.

The experience of a student with a necessarily limited overview working through this test must be either of a world which doesn't make sense and in which answers don't much matter, or of a fragmented and disjointed set of separate tasks.

D. New instruments

16. Numeracy assessment instrument adapted by NRDC from *Skills for Life* survey

1	Title of instrument. Publisher. Date of publication. Edition no.	NRDC Numeracy assessment instrument. NRDC. 2003 (1st version), 2004 (2nd version). 2nd version is reviewed here.

2	Cost of instrument per set/copy.	N/a.
3	Is the instrument 'secure' (unpublished)?	Yes.
4	Stated purpose of instrument.	Measuring progress of learners in NRDC research projects.
5	Are there parallel forms? If so, how many?	No.
6	Are there any obvious manageability problems?	Because there is just one item at each level for each topic, difficulty increases very sharply.
7	What, if any, stipulations about training of administrators are made?	None.
8	Time needed/allowed, if known.	30 minutes, with 10 extra minutes allowed if learners request more time.
9	Is the test available in different media? If so, explain briefly.	No.
10	Is there a clear statement of the population for which the instrument is intended? If so, explain briefly.	Adult numeracy learners.
11	a) How is the test 'tailored' for the range of achievement, if at all?	a) Learners thought to be at Entry 3 or above are given the complete instrument. Those thought to be at Entry 1 or 2 are given a version containing just the 12 items covering Entry Levels 1–3. If they complete it, and feel like tackling the other 8 items, there is a separate booklet containing these.
	b) Is the test pitched at appropriate levels of ease/difficulty for the target learners?	b) So designed, since facility/difficulty levels were derived from the ***Skills for Life*** survey. But see comment above about cline of difficulty, and text below.
12	If the test is stated to be suitable for ESOL as well as for other learners, what evidence	Not so stated, but in use with both first-language and ESOL learners.
13	Do the materials seem free from bias? List any difficulties.	Yes.
14	Are the materials suitable for adults?	Yes.
15	Is there an element of self-assessment in the instrument? If so, explain briefly.	No.

16	Is extraneous knowledge required (e.g. in terms of literacy levels required to comprehend problems/questions, etc./numeracy/IT)? Comment briefly as appropriate.	Many items (appropriately) assume knowledge of British shopping habits.
17	Does the test report results for ■ reading ■ writing ■ spelling ■ punctuation ■ numeracy?	Numeracy.
18	a) At what levels are results reported, if reported in levels? b) If results are reported in different terms, explain briefly. c) Has the test been explicitly aligned or re-aligned to the new QCA Standards?	a) Entry 1 – Entry 2 – Entry 3 – Level 1 – Level 2. b) Could also be reported as raw scores. Reporting by topics not feasible because of limited number of items. c) Yes.
19	Assessment approach and alignment to the Adult Basic Skills Core Curricula and Standards. Discuss, if and as appropriate, as follows: Is alignment – a) broad; b) covering a range of, or all, elements; c) suitably balanced in terms of emphasis; and/or d) appropriate in terms of required depth of skill, knowledge and understanding?	a) Broad because so few items at each Level. b) Covers numbers, fractions, decimals and percentages, common measures, shape and space, and cata-handling – not quite one item per topic per level since there were not enough items of sufficient reliability to do this, but the items are distributed as equitably as possible across levels. c) Balance limited by number of items. d) Yes.
20	Describe the item types used. Comment if appropriate.	All but three items are four-option multiple-choice (in one item, three of the options are correct and must all be located); two require matching of four items to four response boxes; one requires drawing of two arrows.
21	Do the materials/items appear to test what they are intended to test?	Yes.

22	What, if any, technical information is available about the test, e.g. pilots, test-retest reliability, reliability between parallel forms, with different assessors, as 'split halves', relationship with results on other instruments, etc. Comment on any likely difficulties in these areas.	None directly. However, it is known that the 20 items were the most reliable of the numeracy items in the ***Skills for Life*** survey, but the item values are not publicly reported anywhere. An instrument with only 20 items covering such a wide range of levels is inherently incapable of achieving high internal reliability.
23	Is a degree of uncertainty in results, or of limitation in the instrument, recognised, reported, evaluated?	Not in the available documentation, but the users are well aware of the limitations.
24	Are there any other points which might usefully be made about the instrument, e.g. minor proofing errors, tasks that seem particularly difficult?	No.
25	Rate the usefulness of the reported results to a) learners; b) tutors; and c) the wider world (high, medium or low).	 a) Low. b) Low. c) Medium – the results will have considerable uncertainty because of the lack of reliability.
26	Sum up likely positive and negative effects of this instrument in terms of backwash/ teaching to the test.	Probably none since it is being kept secure.
27	Form a judgement about the suitability of the instrument a) for its stated purpose b) for NRDC purposes in – cohort studies – research projects (high, medium or low? Comment below if needed)	 a) Medium. b) – Not suitable. – Medium, suitable only as stopgap.

In 2002, NRDC recognised that it would need a numeracy instrument for its research projects. The early stages of this review showed that there was no suitable existing instrument. It was recommended that one be developed in parallel with the literacy instrument (see next item), but this was not approved. Diana Coben and colleagues in the NRDC numeracy area therefore explored the possibility of using items from the ***Skills for Life*** survey, decided that some of those would be the best (least bad) available, and applied for and received a licence to use them. As implied above, they selected the 20 most statistically reliable of the ***Skills for Life*** survey items and compiled them into a thematically-organised instrument. In the process they also converted the items from computer-administered to paper form, thus compromising the ***Skills for Life*** survey reliability data. Following experience with the items in their projects in 2003/04, Diana Coben and colleagues revised some of the items slightly for 2004/05. The revised items are now therefore two steps away from the form on which the ***Skills for Life*** survey provided item statistics.

Despite these technical disadvantages, the items are probably still the best available for use in NRDC's research projects. However, the instrument as a whole has significant weaknesses, particularly the absence of a parallel form (there were too few sufficiently reliable items to create two forms) and the paucity of items even in the one form created. These limitations make it highly desirable that this instrument be replaced by something technically better designed.

17. Literacy assessment instrument developed for NRDC by NFER

1	Title of instrument. Publisher. Date of publication. Edition no.	*Go! Literacy Assessment Instrument*. Published by NRDC, 2004. Developed for NRDC by NFER during 2003.
2	Cost of instrument per set/copy.	N/a.
3	Is the instrument 'secure' (unpublished)?	Yes – used by NRDC only in its own research projects.
4	Stated purpose of instrument	Measuring attainment in reading and/or writing of adult literacy learners.
5	Are there parallel forms? If so, how many?	Yes, two forms for each of reading and writing.
6	Are there any obvious manageability problems?	Only if both reading and writing need to be administered to same learners.
7	What, if any, stipulations about training of administrators are made?	None within the documentation, but clear instructions for administration and scoring are given in the manual.
8	Time needed/allowed, if known.	Ten minutes for Locator booklet (see below) if used; ten minutes reading time for main test; 30 minutes working time for reading; 30 minutes working time for writing.
9	Is the test available in different media? If so, explain briefly.	No.
10	Is there a clear statement of the population for which the instrument is intended? If so, explain briefly.	Yes – adult literacy learners at levels between Entry 1 and Level 2.

11	a) How is the test 'tailored' for the range of achievement, if at all?	a) The writing tests are not. The reading tests have two levels (easy and less easy) within each parallel form. Guideline is that the easier level is for learners at Entry level, the less easy level for those at Level 1 or above. If teachers' recommendations for choice of level are not available, there is a very short Locator booklet (total administration time: 10 minutes) which can be marked on the spot, and the manual states the cut-off score for choosing the level of the main reading test for each learner.
	b) Is the test pitched at appropriate levels of ease/difficulty for the target learners?	b) Yes.
12	If the test is stated to be suitable for ESOL as well as for other learners, what evidence is there of differentiation, including in the reporting of results?	Not so stated, but is being so used. No differentiation.
13	Do the materials seem free from bias? List any difficulties.	Yes.
14	Are the materials suitable for adults?	Yes. Each parallel form (both reading and writing) is based on a simulated general-interest magazine called *Go!* with a glossy cover and specially-written content modelled on those in *Skillscape* and literacy level 2 of *Initial Assessment* 2nd edition.
15	Is there an element of self-assessment in the instrument? If so, explain briefly.	No.
16	Is extraneous knowledge required (e.g. in terms of literacy levels required to comprehend problems/questions, etc./numeracy/IT)? Comment briefly as appropriate.	No.
17	Does the test report results for ■ reading ■ writing ■ spelling ■ punctuation ■ numeracy?	Reading and writing, if both are used.

18	a) At what levels are results reported, if reported in levels? b) If results are reported in different terms, explain briefly. c) Has the test been explicitly aligned or re-aligned to the new QCA Standards?	a) Entry 1 or below up to Level 2 or above, for both reading and writing. b) but the reading tests have also been standardised (during piloting by NFER) on a scale with mean of 50 and standard deviation of 10; conversion tables from raw scores to scaled scores for reading are given in the manual. c) Explicitly designed to be so aligned, and conversion tables from raw scores to National Standards levels for both reading and writing are given in the manual. Also, the 'stimulus text plus answer booklet' format was explicitly designed to be similar to the format of the Level 2 national tests in key skills and basic skills.
19	Assessment approach and alignment to the Adult Basic Skills Core Curricula and Standards. Discuss, if and as appropriate, as follows: Is alignment – a) broad; b) covering a range of, or all, elements; c) suitably balanced in terms of emphasis; and/or d) appropriate in terms of required depth of skill, knowledge and understanding?	a) Close – each item is aligned with a particular statement in the Standards/Adult Literacy Core Curriculum b) A range c) Yes d) Yes
20	Describe the item types used. Comment if appropriate.	Balance of multiple-choice and supply (open-ended).
21	Do the materials/items appear to test what they are intended to test?	Yes
22	What, if any, technical information is available about the test, e.g. pilots, test-retest reliability, reliability between parallel forms, with different assessors, as 'split halves', relationship with results on other instruments, etc. Comment on any likely difficulties in these areas.	NFER carried out extensive piloting, and provided detailed technical reports, which are available within NRDC. Reliabilities of the four reading tests (lower and upper levels of the two parallel forms) range between 0.91 and 0.94. Reliabilities are not applicable to the writing tests because they are not scored on discrete items.
23	Is a degree of uncertainty in results, or of limitation in the instrument, recognised,	Yes, but the reading tests have such high reliabilities that they have very little uncertainty.

24	Are there any other points which might usefully be made about the instrument, e.g. minor proofing errors, tasks that seem particularly difficult?	The simulated magazines that contain the stimulus texts have been found interesting by learners.
25	Rate the usefulness of the reported results to a) learners; b) tutors; and c) the wider world (high, medium or low).	a) Low. b) Low because not available to be reported to them. c) High.
26	Sum up likely positive and negative effects of this instrument in terms of backwash/ teaching to the test.	N/a since test is secure.
27	Form a judgement about the suitability of the instrument: a) for its stated purpose; b) for NRDC purposes in – cohort studies – research projects (high, medium or low? Comment below if needed).	a) High because specifically designed to meet the needs of NRDC as determined through the earlier stages of the review which gave rise to this report. b) – cohort studies: low because was not available in time for the 2004 sweep, and the instrument used there needed to be similar to those used in previous sweeps. – research projects: very high.

With the review of this instrument, the project which gave rise to this report comes full circle. The project was intended, among other things, to produce a judgment on the suitability of existing adult literacy assessment instruments for use in NRDC's own research programme. When, in early autumn 2002, we established that none of those available were suitable, we recommended that a new literacy instrument be developed. This was agreed by NRDC and the Adult Basic Skills Strategy Unit within the DfES, and the development of the instrument was commissioned from NFER early in 2003. The finalised instrument was duly delivered to NRDC early in 2004.

Meanwhile, a pilot version was used as a pre-assessment in two NRDC projects during the autumn term 2003. NFER delivered raw and scaled scores based on the pilot version, which enabled the project teams to undertake direct statistical comparisons with post-assessment scores from the definitive versions used later in academic year 2003/04.

The instrument meets all the requirements which NRDC specified to NFER. In particular:

- it is secure;
- it is closely aligned to the QCA Standards; and
- it has parallel forms.

This makes the instrument ideal for use in NRDC's own literacy research projects.

Appendix E. The suitability of instruments for their own purposes

The most important general conclusion reached is the following. The value of each instrument reviewed in supporting teaching and learning will depend upon the degree of skill with which the results are interpreted by tutors and all others using it. What is required is a sensible awareness both of the limits of precision and validity and of the difficulties of chopping literacy and numeracy into discrete fragments, or strands, for individual assessment.

It is important that tutors understand the strengths and weaknesses of assessment materials, especially of those that claim to be diagnostic. Otherwise, there is a danger of a return to a 'death by worksheet' approach focused on isolated aspects or elements of literacy, the level and nature of the worksheet being dictated by the results of diagnostic tests which, after all, can do no more than provide some objective but limited indication of a learner's strengths and weaknesses.

Assessment materials should be appropriate for their purposes; one set of criteria cannot cover all situations. However, some guidelines, suggested on the basis of the experience of carrying out this project, are offered for consideration.

Good materials should:

- have clear instructions;
- have clear print and layout;
- avoid encouraging those using them to set up potentially inhibiting and threatening test-like situations, especially, perhaps, when numeracy is being assessed;
- where appropriate, give clear guidance on practical issues, such as the use of pencil and paper, what to do if a mistake has been made, when to turn a page, and what to do if help is required;
- make sensible use of unambiguous graphics;
- include sufficient practice questions for learners to be clear about the kinds of item to expect and about how these should be dealt with;
- for numeracy, make clear what is to be worked mentally, and when a calculator or pencil and paper may be used;
- be consistent in the relationship implied with their audience by textual and linguistic choices;
- be presentationally consistent in, for example, the relationship of graphic materials on successive pages or screens to the content of the item;
- avoid confusing learners at initial levels with noise arising from irrelevant visual materials;
- as far as possible, have some intrinsic interest;
- be free from bias;
- allow access for all, ensuring that, for example, men and women have equal access for their possibly different numeracy needs;
- allow for possible economic and cultural differences, between, for example, men and women, in respect of ownership of computers and familiarity with IT;
- signal differences in item type and, where this is appropriate for the literacy levels of learners, provide clear textual cues to answering conventions, with proper differentiation between straightforward instructions and items in the form of (for example) narrative statements;
- present questions, or items, in an obvious, internally logical order and, where appropriate, explicitly link such questions or items with the curriculum area they are supposed to test, or at least support the learner by providing categorising information. This would match a requirement for assessment to 'make sense of … learning' for the learners;

- avoid the use of items in which difficulty arises more from the manner in which the task is presented to the learner or in which the answer must be provided, than from the skill or knowledge supposedly being assessed;
- where appropriate, report results in terms that provide useful formative information for learner and tutors
- be based upon a thought-through and explicitly acknowledged approach to alignment with the National Standards and Adult Literacy and Numeracy Core Curricula;
- acknowledge their own limitations; and
- where there are parallel forms, look parallel and include similar task and text types.

Many of these recommendations chime with those of Clausen-May (2001).

Possible staff development activities suggest themselves:

Take some examples of learner writing and decide as individuals which level the learner is working towards. Compare and discuss judgments.

Take example items from various instruments and try to decide which 'Descriptors', 'Elements' or even 'Additional Skills and Knowledge' the items seem to be diagnosing. List all possibilities. Compare and discuss judgements.

Discuss how far the mental processes involved in completing each item correctly reflect those involved in a real-life reading or numeracy activity.

Take materials which have not yet been labelled with Level and Element and a) decide at which level the materials might be used and then b) think of an activity using the same materials that could provide evidence of achievement both for the level below and the level above that at which the materials had been placed.

In small groups, list as many kinds of information as possible about a learner which it would be useful to have, and to discuss with a learner, that a particular assessment instrument cannot provide.

In small groups, list and discuss as many methods of formative assessment as possible. Discuss the advantages and disadvantages of each method. Don't forget that you are allowed to talk to learners, and to read and to think about what they have written.

Such training might provide useful insights for staff using pre-packaged assessment materials in the development of learning plans or as a basis for target-setting.

Experienced tutors have the skills to interpret and explain the strengths and limitations of assessment materials to learners who may be uncritically accepting of the results, especially if they are produced by a computer. This is particularly important in the case where materials offer not only the diagnosis but also, supposedly, the cure, and where the tests include marketing text aimed at learners. Learners could use records of such discussion as evidence towards several of the Elements of the Adult Literacy Curriculum. For example, tests are themselves texts, which may have several purposes.

With the possible exception of tests reporting results in reading and spelling ages, there seems to be little reason why any of the earlier, more broadly aligned initial literacy assessment materials should not continue to be used by appropriately experienced staff, though it appears that such practice might not be acceptable in terms of the new Inspection Framework.

This report is funded by the Department for Education and Skills as part of ***Skills for Life***: the national strategy for improving adult literacy and numeracy skills. The views expressed are those of the author(s) and do not necessarily reflect those of the Department.

www.nrdc.org.uk

NRDC
Institute of Education
University of London
20 Bedford Way
London WC1H 0AL
Telephone: +44 (0)20 7612 6476
Fax: +44 (0)20 7612 6671
email: info@nrdc.org.uk
website: www.nrdc.org.uk

NRDC is a consortium of partners led by the Institute of Education, University of London with:

- Lancaster University
- The University of Nottingham
- The University of Sheffield
- East London Pathfinder
- Liverpool Lifelong Learning Partnership
- Basic Skills Agency
- Learning and Skills Development Agency
- LLU+, London South Bank University
- National Institute of Adult Continuing Education
- King's College London
- University of Leeds

Funded by the Department for Education and Skills as part of ***Skills for Life***: the national strategy for improving adult literacy and numeracy skills.