European Journal of Psychology of Education
2001, Vol. XVI, n° 3, 335-361

The framework used in developing and interpreting the International Adult Literacy Survey (IALS)

Irwin S. Kirsch
Educational Testing Service, Princeton, NJ

This paper offers a framework that has been used for both developing the tasks used to measure literacy in the International Adult Literacy Survey and for understanding the meaning of what has been reported with respect to the comparative literacy proficiencies of adults in participating countries. The framework consists of six parts that represent a logical sequence of steps from needing to define and represent a particular domain of interest, to identifying and operationalizing characteristics used to construct items, to providing an empirical basis for interpreting results. The various parts of the framework are seen as important in that they help to provide a deeper understanding of the construct of literacy and the various processes associated with it. A processing model is proposed and variables associated with performance on the literacy tasks are identified and verified through regression analyses. These variables are shown to account for between 79 and 89 percent of the variance in task difficulty. Collectively, these process variables provide a means for moving away from interpreting performance on large-scale surveys in terms of discrete tasks or a single number towards identifying levels of performance that have generalizability across pools of tasks and towards what Messick has called a higher level of measurement.

Introduction

The International Adult Literacy Survey (IALS) was the first comparative survey of adults designed to profile and explore the literacy distributions among participating countries. It was a collaborative effort involving several international organisations, inter-governmental agencies, and national governments. In 2000, a final report was released (OECD & STATCAN, 2000) which stated that, "by 1998, the survey had covered 10.3 percent of the world population and 51.6 percent of the world GDP" (p. 87).

Who are the constituencies that are likely to use the data from the IALS? It is expected that many individuals including researchers, practitioners, and individual citizens within each of the participating countries will read the survey results and make use of the data for a variety

of purposes. Yet, the primary reason for developing and conducting this large-scale international assessment is to provide empirically grounded interpretations upon which to inform policy decisions. This places the International Adult Literacy Survey in the context of policy research. In their classic volume on this topic, Lerner and Lasswell (1951) argued that the appropriate role for policy research is not to define policy; rather, it is to establish a body of evidence from which informed judgements can be made. Messick (1987) extended this thinking to the area of large-scale assessments and noted that, in order to appropriately fulfil this function, assessments should exhibit three key features: *relevance, comparability, and interpretability.*

Relevance refers to the capability for measuring diverse background and program information to illuminate context effects and treatment or process differences. The IALS developed and administered an extensive questionnaire covering a wide range of issues that can be used to identify characteristics which are correlated with performance and which may differ across a variety of language and cultural backgrounds.

Comparability deals with the capacity to provide data or measures that are commensurable across time periods and across populations of interest. Complex sampling, scaling, and translation procedures are implemented to help ensure that common metrics exist across participating countries so that appropriate comparisons can be made between countries and among major subpopulations of interest within a country. These comparisons are important both in this initial survey and in future assessments where new countries may join the survey and want to be placed onto existing scales or where participating countries may want to measure trends in the distributions of skills among various subpopulations of interest.

Interpretability focuses on collecting evidence that will enhance the understanding and interpretation of what is being measured. In some assessments, the meaning of what is being measured is constructed by examining performance on individual tasks, or by assuming it is inherent in the label that is used to organise one or more sets of tasks – for example, reading comprehension or critical thinking. All too often assessments focus on rank ordering populations or countries by comparing mean scores or distributions. These data tell us that people differ without telling us how they differ. One of the stated goals in the IALS is to try to address the issue of interpretability not only by reporting that countries, groups, or individuals differ in their proficiencies, but also by developing an interpretative scheme for reporting how they differ.

Overview of the framework

While there are many approaches one could take to develop a framework for measuring a particular skill area, Figure 1 represents a process that has been used to construct and interpret the literacy tasks for the National Adult Literacy Survey (Kirsch, Jungeblut, Jenkins, & Kolstad, 1993) and for the International Adult Literacy Surveys (OECD & HRDC, 1997; OECD & STATCAN, 1995, 2000). This process is also being used to develop the reading literacy measure for PISA – the Programme for International Student Assessment (OECD, 1999). The diagram shown here represents a process that consists of six parts. These six parts represent a logical sequence of steps that should be addressed, from needing to define a particular skill area, to having specifications for constructing items, to providing an empirically based interpretation of the scores that are obtained.

Part 1 of the framework focuses on the working definition for literacy along with some of the assumptions that underlie it. In doing so, the definition sets the boundaries for what the survey seeks to measure as well as what it does not measure. Part 2 provides a discussion on how we may choose to organise the set of tasks that are constructed to report to policy makers and researchers on the distribution of a particular skill in the population. Determining how to report the data should incorporate statistical, conceptual, and political considerations. Part 3

deals with the identification of a set of key characteristics that are manipulated by developers when constructing tasks for a particular skill area. Part 4 identifies and begins to define the variables associated with the set of key characteristics that are used in test construction. These definitions are based on the existing literature and on experience with building and conducting other large-scale assessments. Part 5 lays out a procedure for validating the variables and for assessing the contribution each makes toward understanding task difficulty across the various participating countries. The final part, Part 6, discusses how an interpretative scheme was built using the variables that have been shown through the research in Part 5 to account for task difficulty and student performance.

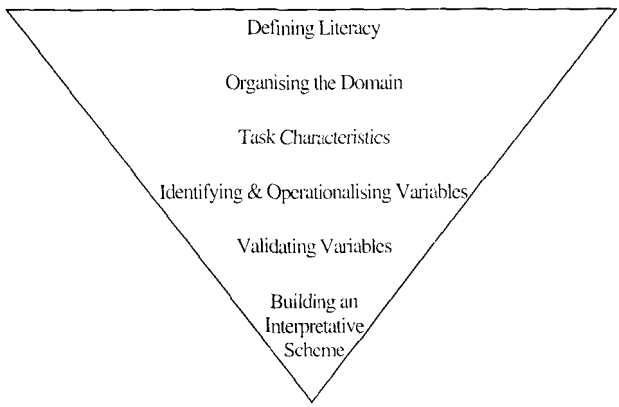


Figure 1. Process used to construct and interpret literacy tasks

The various parts of the framework are important in that they help to provide a deeper understanding of the construct of literacy and the various processes associated with it. Identifying and understanding particular variables that underlie successful performance furthers our ability to evaluate what is being measured and to make changes to the measurement over time. In addition, as we increase our understanding of what is being measured and our ability to interpret scores along a given scale, we have an empirical basis for communicating a richer body of information to various constituencies and a way to link assessments with instruction.

Defining literacy

Definitions of reading and literacy have changed over time in parallel with changes in our society, economy, and culture. The growing acceptance of the importance of lifelong learning has expanded the views and demands of reading and literacy. Literacy is no longer seen as an ability that is developed during the early school years, but is instead viewed as an advancing set of skills, knowledge, and strategies which individuals build on throughout their lives in various contexts and through interaction with their peers and with the larger communities in which they participate.

As Resnick and Resnick (1977) point out, literacy in its earliest form consisted of little more than signing one's name. It was not until much later that fluent oral reading became important and not until the 20th century that reading to gain information was given primary emphasis. Standardised tests became fashionable and reading-grade-level scores became the focus of attention. Through the use of these instruments the term literacy has implied the acquisition of intellectual skills associated with basic academic competencies and with reading

and writing. Standards for literacy increased over the decades from being able to read at a 4th-grade level, to reading at an 8th-grade level and then, by the early seventies, to a 12th-grade level. These measures came under increasing criticism, however, because they did not provide specific information about the kinds of competencies that given levels of literacy imply. Perhaps more important was the recognition that literacy relates not to some arbitrary standard for the purpose of categorising people into literate and illiterate but to what people can do with printed and written materials and how these skills relate to a host of social needs. As Beach and Appleman (1984) noted:

> "The often heard charge, Johnny can't read is a little like saying Johnny can't cook. Johnny may be able to read the directions for constructing a radio kit, but not a Henry James novel, just as Johnny may be able to fry an egg but not cook Peking duck. In discussing reading in the schools, we must recognise that reading involves as wide a range of different types of texts as there are types of food. And, to imply, as does the slogan, 'Johnny can't read', that reading is a single skill suited to all types of texts does not do justice to the range of reading types".

Thus, the multifaceted nature of literacy had often been glossed over through the use of grade-level equivalent scores.

It was from this multifaceted perspective that several large-scale assessments of literacy were conducted in Australia (Wickert, 1989), Canada (Montigny, Kelly, & Jones, 1991), and the United States (Kirsch & Jungeblut, 1986; Kirsch, Junglebut, Jenkins, & Kolstad, 1993).

In 1992, the Organisation for Economic Co-operation and Development (OECD & STATCAN, 1992) concluded that low literacy levels were a serious threat to economic performance and social cohesion on an international level. But a broader understanding of literacy problems across industrialised nations – and consequent lessons for policymakers – was hindered due to a lack of comparable international data. Statistics Canada (STATCAN) and Educational Testing Service (ETS) teamed up to build and deliver an international comparative study. After some discussion and debate, the framework and methodology used in the National Adult Literacy Survey (NALS) was applied to the first large-scale International Adult Literacy Survey (IALS).

NALS, which was funded by the National Center for Education Statistics (NCES) as part of its overall assessment program in adult literacy, was the largest and most comprehensive study of adult literacy ever conducted in the United States. Like all large-scale assessments funded by NCES, NALS was guided by a committee, which was comprised of a group of nationally recognised scholars, practitioners, and administrators who adopted the following definition of literacy:

> *Literacy is using printed and written information to function in society, to achieve one's goals, and to develop one's knowledge and potential.*

This definition captures the initial work of the committee guiding the development of the assessment and provides the basis for creating other aspects of the framework to be discussed. It also carries several assumptions made by panel members and, thus, it is important to consider various parts of this definition in turn.

Literacy is...

The term literacy is used in preference to "reading" because it is likely to convey more precisely to a non-expert audience what the survey is measuring. "Reading" is often understood as simply decoding, or reading aloud, whereas the intention of the adult surveys is to measure something broader and deeper. Researchers studying literacy within particular contexts noted that different cultures and groups may value different kinds of literacy practices (Graff, 1979; Heath, 1980; Sticht, 1975; Szwed, 1981). Heath, for example, distinguished various uses for reading: instrumental, social interactional, news-related, memory supportive, substitutes for oral messages, provision of a permanent record, and personal confirmation. The fact that people read different materials for different purposes implies a range of proficiencies that may not be well

captured by signing one's name, completing a certain number of years of schooling, or scoring at an 8th-grade level on a test of academic reading comprehension.

... using printed and written information...

This phrase draws attention to the fact that panel members view literacy not as a set of isolated skills associated with reading and writing, but more importantly as the application of those skills for specific purposes in specific contexts. When literacy is studied within varying contexts, diversity becomes its hallmark. First, people engage in literacy behaviours for a variety of uses or purposes (Cook-Gumperz & Gumperz, 1981; Heath, 1980; Mikulecky, 1982; Sticht, 1978). These uses vary across contexts (Heath, 1980; Venezky, 1983) and among people within the same context (Kirsch & Guthrie, 1984a). This variation in use leads to an interaction with a broad range of materials that have qualitatively different linguistic forms (Diehl, 1980; Jacob, 1982; Miller, 1982). In some cases, these different types of literacy tasks have been associated with different cognitive strategies or reading behaviours (Crandall, 1981; Kirsch & Guthrie, 1984b; Scribner & Cole, 1981; Sticht, 1978, 1982).

... to function in society, to achieve one's goals, and to develop one's knowledge and potential.

This phrase is meant to capture the full scope of situations in which literacy plays a role in the lives of adults, from private to public, from school to work, to lifelong learning and active citizenship. "To achieve one's goals and to develop one's knowledge and potential" points to the view that literacy enables the fulfilment of individual aspirations – both defined ones, such as graduation or obtaining a job, and less defined and less immediate ones which extend and enrich one's personal life. The phrase "to function in society" is meant to acknowledge that literacy provides individuals with a means of contributing to as well as benefiting from society. Literacy skills are generally recognised as important for nations to maintain or improve their standard of living and to compete in an increasingly global market place. Yet, they are equally important for individual participation in technologically advancing societies with their formal institutions, complex legal systems, and large government programmes.

Organising the domain

Having defined the domain of literacy and having laid out the set of assumptions that were made in developing the definition, it is important to think about how to organise the domain. This organisation needs to focus on how to report the scores that result from administering a pool of literacy tasks. This is an important issue because how the domain is organised can affect test design. Because some believe that reading is not a single, one-dimensional skill, literacy is not necessarily best represented by a single scale or single score along that scale. Yet determining how many and which scales should be used for reporting literacy scores is crucial for ensuring that sufficient numbers of tasks are developed to define and interpret these scales adequately.

Different perspectives can be used to help organise a domain of tasks. Traditionally, literacy skills have been categorised by modality into reading, writing, speaking, and listening. Reading and writing are sometimes combined, as they are thought to require similar processes, and speaking and listening are often grouped in terms of being too costly and difficult to assess. Thus, speaking and learning were not included in the survey. Committee members also wanted to include basic arithmetic calculations as part of the assessment since adults are often required to use printed information that involves these skills. As a result, this aspect of literacy was also included in the surveys.

Work in the area of context of literacy clearly provides one possible organising principle for what may appear to be a disparate set of literacy tasks. There is the familiar academic or

school context (dealing primarily with prose or connected discourse) contrasted with non-school or "everyday life" contexts. And the non-school contexts can be subdivided into the work-related and home-related tasks. However, it is operationally difficult to separate tasks along these latter dimensions since the work and home categories are not mutually exclusive in terms of the relevant literacy tasks.

Another organising principle of some appeal involves categorising literacy tasks in terms of the types of materials or formats in which they occur and to examine the associated purposes or uses both within and across materials. The appeal for this type of organisational scheme stems from a research literature suggesting that different materials or formats are associated with different contexts and that a significant proportion of adult reading tasks in the context of work involve documents (Jacob, 1982; Kirsch & Guthrie, 1984a; Sticht, 1975) – graphs, charts, forms, and the like – rather than prose. Frequently, these documents are embedded in the contexts of home or work and community, as contrasted with prose, which is most frequently associated with school or academia. Moreover, different materials and formats are often associated with different purposes and these purposes are frequently associated with different reading strategies. This line of reasoning led to distinctions such as Sticht's "reading to do" and "reading to learn".

As another instance reflecting similar distinctions, NAEP (1972) came to aggregate reading exercises in terms of "themes" – word meanings, visual aids, written directions, reference materials, significant facts, main ideas, inferences, and critical reading. The areas of reference materials and significant facts were among those in which young adults aged 26 to 35 performed better than did in-school 17-year-olds, while, on the other hand, 17-year-olds performed higher than young adults in inferences and critical reading. These and other NAEP results suggest the utility of a priori classifications that allow for the examination of differential performance for subgroups both within a single assessment and across groups over time.

In the end, a compromise was reached among the various organising concepts that was felt to reflect a number of salient notions from the literature. Three scales were hypothesised – a prose literacy scale, a document literacy scale, and a quantitative literacy scale. In this way, it is possible to acknowledge that the structure of prose passages are qualitatively different from the structures associated with documents such as charts, tables, schedules, etc. and to provide for a separate scale for those tasks involving the processing of printed information in combination with arithmetic operations.

The original data from the NAEP Young Adult Literacy Survey (YALS) was subjected to factor analysis to explore dimensionality (Kirsch & Jungeblut, 1986). Following the logic of Cattell's scree test (1966), the breaks in the pattern of latent roots indicated at least three salient factors with the possibility of as many as five additional factors. Analysis of parallel random data reinforced the judgement that a three-factor solution was appropriate. Thus the empirical data provided by the YALS tended not only to support the a priori judgement for the three literacy scales but also suggested ways in which the assessment could be broadened. It is important to keep in mind that the three literacy scales are not the only salient dimensions of literacy per se. These dimensions are likely to shift as a function of different definitions and different perspectives on literacy.

The advisory committees involved with NALS and IALS agreed that literacy should not be measured along a single continuum and chose to adopt the general definition and three scales defined here. A new adult survey has been developed and will be carried out in 2002 that is titled the Adult Literacy and Lifeskills (ALL) Survey. The oversight committee for this survey decided to replace the Quantitative literacy scale with the broader domain of Numeracy. As a result, this paper focuses on the Prose and Document literacy scales.

Identifying task characteristics

Almond and Mislevy (1998) note that variables can take on one of five roles in an assessment or test. They can be used to: limit the scope of the assessment, characterise

features that should be used for constructing tasks, control the assembly of tasks into booklets or test forms, characterise examinees' performance on or responses to tasks, or help to characterise aspects of competencies or proficiencies. Some of these variables can be used both to help in the construction of tasks and in the understanding of competencies as well as in the characterisation of performance. A finite number of characteristics are likely to influence students' performance on a set of literacy tasks, and these can be taken into account when constructing or scoring the tasks. The following characteristics were manipulated in the development of tasks for IALS:

- *Contexts/Content*: Since adults do not read written or printed materials in a vacuum but within a particular context or for a particular purpose, materials for the literacy assessment are selected that represent a variety of contexts and contents. This is to help ensure that no one group of adults is either advantaged or disadvantaged due to the context or content included in the assessment.
- *Materials/Texts*: While no one would doubt that a literacy assessment should include a range of material, what is critical to the design and interpretation of the scores that are produced are the range and specific features of the text material which are included in constructing the tasks. Thus, a broad range of both prose and document text types are included in this survey.
- *Processes/Strategies*: This refers to the characteristics of the questions and directives that are given to adults for their response. Generally speaking, the questions and directives will refer to a goal or purpose the readers are asked to assume while they are reading and interacting with texts and relate to one or more strategies that the readers are likely to use in producing their response.

Identifying and operationalising the variables

In order to use these three main task characteristics in designing the assessment and, later, in interpreting the results, they need to be operationalised. That is, various values that each of these characteristics can take on must be specified. This will allow item developers to categorise the materials they are working with and the questions and directives they construct so that they can be used in the reporting of the results. These variables can also be used to specify what proportions of the assessment ought to come from each category.

Context/content

Materials that are selected for inclusion in the assessment need to represent a broad range of contexts and contents. Six adult context/content categories have been identified as follows:

- Home and family: may include materials dealing with interpersonal relationships, personal finance, housing, and insurance.
- Health and safety: may include materials dealing with drugs and alcohol, disease prevention and treatment, safety and accident prevention, first aid, emergencies, and staying healthy.
- Community and citizenship: may include materials dealing with staying informed and community resources.
- Consumer economics: may include materials dealing with credit and banking, savings, advertising, making purchases and maintaining personal possessions.
- Work: may include materials that deal in general with various occupations but not job specific texts, finding employment, finance, and being on the job.

- Leisure and recreation: may include materials involving travel, recreational activities, and restaurants.

It is important to note that with respect to this variable, an attempt has to be made to include as broad a range as possible across the six contexts as well as to select universally relevant materials. Following this procedure helps to ensure that the content and materials that are included in the assessment are not so specialised as to be familiar only to certain groups and that any disadvantage for people with limited background knowledge is minimised.

Materials/texts

Reading requires something for the reader to read. In an assessment, that something – a text – must be coherent within itself. That is, the text must be able to stand alone without requiring additional printed material. While it is obvious that there are many different kinds of texts and that any assessment should include a broad range of them, it is not so obvious that there is an ideal categorisation of text types. There are any number of proposals as to the appropriate categories, many of them created for practical rather than theoretical purposes. All of them share the fact that no particular text seems to fit easily into only one category. For example, a chapter in a textbook might include some definitions (often identified as a text type), some instructions on how to solve particular problems (yet another text type), a brief historical narrative of the discovery of the solution (still another text type), and descriptions of some typical objects involved in the solution (a fourth text type).

It might be thought that a definition, for example, could be extracted and treated as a single text for assessment purposes. But this would remove the definition from the context, create an artificial text type (definitions almost never occur alone, except in dictionaries), and not allow item writers to create tasks that deal with reading activities which require integrating information from a definition with information from instructions.

A more important classification of texts, and one at the heart of this assessment, is the distinction between continuous and non-continuous texts. Continuous texts are typically composed of sentences that are, in turn, organised into paragraphs. These may be fit into even larger structures such as sections, chapters, and books. Non-continuous texts are most frequently organised in matrix format, based on combinations of lists.

Continuous texts. Conventionally, continuous texts are formed of sentences organised into paragraphs. In these texts, organisation occurs by paragraph setting, indentation, and the breakdown of text into a hierarchy signalled by headings that help the reader to recognise the organisation of the text. Text types are standard ways of organising the contents of and author's purpose for continuous texts[1].

1) *Description* is the type of text where the information refers to properties of objects *in space*. Descriptive texts typically provide an answer to *what* questions.

2) *Narration* is the type of text where the information refers to properties of objects *in time*. Narration texts typically provide answers to *when* or *in what sequence* questions.

3) *Exposition* is the type of text in which the information is presented as composite concepts or mental constructs, or those elements into which concepts or mental constructs can be analysed. The text provides an explanation of how the component elements interrelate within a meaningful whole and often answers *how* questions.

4) *Argumentation* is the type of text that presents propositions as to the relationship among concepts or other propositions. Argument texts often answer *why* questions. Another important sub-classification of argument texts are persuasive texts.

5) *Instruction* (sometimes called *injunction*) is the type of text that provides directions on what to do.

6) *Document* or *record* is a text that is designed to standardise and conserve information. It can be characterised by highly formalised textual and formatting features.

7) *Hypertext* is a set of text slots linked together in such a way that the units can be read in different sequences, allowing readers to follow various routes to the information.

Non-continuous texts. Non-continuous texts are organised differently than continuous texts and so allow the reader to employ different strategies for entering and extracting information from them. On the surface, these texts appear to have many different organisational patterns or formats, ranging from tables and schedules to charts and graphs, and from maps to forms. However, the organisational pattern for these types of texts which Mosenthal and Kirsch (1998) refer to as documents is said to have one of four basic structures: a simple list, a combined list, an intersected list, and a nested list. Together, these four types of documents make up what they have called matrix documents, or non-continuous texts with clearly defined rows and columns. They are also closely related to other non-continuous texts that these authors refer to as graphic, locative and entry documents[2].

1) Matrix Documents: This set of non-continuous texts consists of four types of increasingly complex documents that have simple lists as their basic unit. A simple list consists of a label and two or more items where the label serves as the organising category and items all share at least one feature with the other items in the list. Next are combined lists that consist of two or more simple lists. One list in a combined list is always primary and, as such, is ordered to facilitate looking up information in it and so that parallel information in the other lists can be located. Intersected lists are the third type of matrix document and comprise exactly three lists. Two of the lists form a row and column defining the cells of the third or intersected list. The fourth and most complex type of matrix document is the nested list. In order to economise on space as well as to display comparative information, designers sometimes combine two or more intersecting lists to form a nested list. In a nested list, one type of information will be repeated in each of the intersecting lists. The intersecting list of unemployment rates, for example, may have separate entries under each month for males and females; in this case, gender would be nested under month.

2) Graphic Documents: A major function of graphic documents is to provide a succinct visual summary of quantitative information. Included in this group of documents or non-continuous texts are: pie charts, bar charts and line graphs. While these appear to be very different types of documents on the surface, they all derive or can be transformed into either a combined, intersecting or nested list.

3) Locative Documents: Like graphic documents, locative documents or maps portray information visually. Unlike graphic documents that display quantitative information, maps portray either the location of persons, places or things in space or depict characteristics of different geographic regions (e.g., types of vegetation or characteristics of a population).

4) Entry Documents: In matrix and graphic documents, the author provides the information that must be read and used. In contrast, entry documents or forms require the reader to provide information that can range from very simple to complex. For example, the reader may be asked to simply check a box, write a single word, number or phrase, or construct a series of phrases or sentences. Generally speaking, forms provide the reader with a label or category for which the reader is asked to provide specifics.

5) Combination Documents: It is important to keep in mind that some displays, especially graphic documents, rely on the use of other documents for their interpretation. Maps and graphs, for instance, often include legends that display important information that must be read and understood. In addition, designers sometimes include more than one document for display or comparative purposes.

Processes/strategies

This task characteristic refers to the way in which examinees process a text to respond correctly to a question or directive. It includes the processes used to relate information in the question (the given information) to the necessary information in the text (the new information) as well as the processes needed to either identify or construct the correct response from the information available. Three variables in the reading/literacy research used to investigate tasks from national and international surveys will be considered here. These are: type of match, type of information requested, and plausibility of distracting information. They are briefly described here and are more fully characterised through a discussion of exemplary tasks in the next section as well as fully operationalised in the appendix at the end of this paper.

Type of match. Four types of matching strategies were identified: locating, cycling, integrating, and generating. Locating tasks require examinees to match one or more features of information stated in the question to either identical or synonymous information provided in the text. Cycling tasks also require examinees to match one or more features of information, but unlike locating tasks, they require respondents to engage in a series of feature matches to satisfy conditions stated in the question. Integrating tasks require examinees to pull together two or more pieces of information from the text according to some type of specified relation. For example, this relation might call for examinees to identify similarities (i.e., make a comparison), differences (i.e., contrast), degree (i.e., smaller or larger), or cause-and-effect relations. This information may be located within a single paragraph or it may appear in different paragraphs or sections of the text. In integrating information, examinees draw upon information categories provided in a question to locate the corresponding information in the text. They then relate the text information associated with these different categories based upon the relation term specified in the question. In some cases, however, examinees must generate these categories and/or relations before integrating the information stated in the text.

In addition to requiring examinees to apply one of these four strategies, the type of match between a question and the text is influenced by several other processing conditions which contribute to a task's overall difficulty. The first of these is the number of phrases that must be used in the search. Task difficulty increases with the amount of information in the question for which the examinee must search in the text. For instance, questions that consist of only one independent clause tend to be easier, on average, than those that contain several independent or dependent clauses. Difficulty also increases with the number of responses that examinees are asked to provide. Questions that request a single answer are easier than those that require three or more answers. Further, questions which specify the number of responses tend to be easier than those that do not. For example, a question that states, "List the 3 reasons..." would be easier than one which said, "List the reasons...". Tasks are also influenced by the degree to which examinees have to make inferences to match the given information in a question to corresponding information in the text, and to identify the requested information.

Type of information requested. This refers to the kinds of information that readers identify to answer a test question successfully. The more concrete the requested information, the easier the task is judged to be. In previous research based on large-scale assessments of adults' and children's literacy (Kirsch & Mosenthal, 1994; Kirsch, Jungeblut, & Mosenthal, 1998), the type of information variable was scored on a 5-point scale. A score of 1 represented information that was the most concrete and therefore the easiest to process, while a score of 5 represented information that was the most abstract and therefore the most difficult to process. For instance, questions which asked examinees to identify a person, animal, or thing (i.e., imaginable nouns) were said to request highly concrete information and were assigned a value of 1. Questions asking respondents to identify goals, conditions, or purposes were said to request more abstract types of information. Such tasks were judged to be more difficult and received a value of 3. Questions that required examinees to identify an equivalent were judged to be the most abstract and were assigned a value of 5. In such cases, the equivalent tended

to be an unfamiliar term or phrase for which respondents had to infer a definition or interpretation from the text.

Plausibility of distractors. This concerns the extent to which information in the text shares one or more features with the information requested in the question but does not fully satisfy what has been requested. Tasks are judged to be easiest when no distracting information is present in the text. They tend to become more difficult as the number of distractors increases, as the distractors share more features with the correct response, and as the distractors appear in closer proximity to the correct response. For instance, tasks tend to be judged somewhat more difficult when one or more distractors meet some but not all of the conditions specified in the question and appear in a paragraph or section of the text other than the one containing the correct answer. Tasks are judged to be most difficult when two or more distractors share most of the features with the correct response and appear in the same paragraph or node of information as the correct response.

In previous surveys, the goal has been to develop pools of prose and document tasks which represent the range of contexts, texts, and processes outlined here with no specific requirement for particular numbers of any type of task. The goal was to draw materials from a wide variety of adult contexts that represented a wide range of linguistic structures such as those outlined in this paper. With respect to continuous or prose texts, the focus has been on expository texts since much of what adults read for work and in their community is associated with this type of discourse. However, some surveys did include narratives and poetry in small numbers. In terms of processes/strategies, the goal was to engage adults in the full range of processes that might reasonably be associated with each type of material. That is, the goal was to use the framework to construct questions/directives that were thought to be authentic with regard to the kinds of information someone might want to understand or use from a particular text.

Validating the variables

In a previous section, three task characteristics labelled context, texts, and process/strategy were introduced. This part of the framework describes a procedure for validating a set of variables developed from these characteristics that have been shown to affect task performance and the placement of tasks along each of the reporting scales. This process borrows heavily from work that has been done in the area of adult literacy where several national and international surveys have reported data that followed this approach:

- the U.S. Department of Labor's Literacy Assessment (Kirsch & Jungeblut, 1992),
- the IEA Reading Literacy Study (Kirsch & Mosenthal, 1994),
- the National Adult Literacy Survey (Kirsch et al., 1993), and
- the International Adult Literacy Survey (OECD & STATCAN, 1995).

Reading tasks for these surveys were developed to represent a broad range of purposes for which students and adults read continuous and non-continuous texts in both school and non-school settings. To identify the variables contributing to adults' reading and task difficulty in the prose and document domains, Kirsch and Mosenthal (Kirsch & Mosenthal, 1990; Kirsch et al., 1998) began by modelling the processes required to complete prose and document tasks in the literacy assessments. This model is shown in Figure 2 and grew out of earlier exploratory work (Fisher, 1981; Guthrie, 1988; Kirsch & Guthrie, 1984b).

In the first step, readers identify a goal or purpose for searching and processing a text or document. In a test or an instructional situation, questions and directives determine the primary purpose for interacting with a text or document, and therefore also determine the information that readers must process in order to complete a cognitive activity.

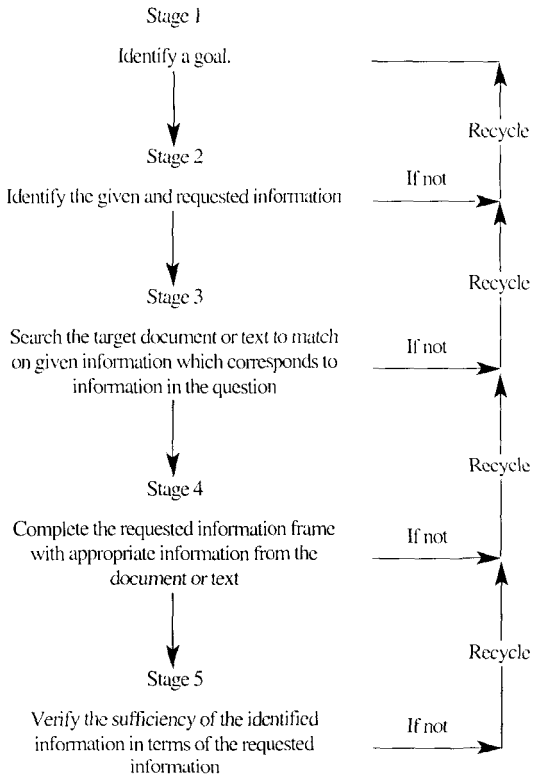


Figure 2. A model of prose and document processing in reading

In the second step, readers must distinguish between "given" and "requested" information in the question (Clark & Haviland, 1977; Mosenthal & Kirsch, 1991). Given information is presumed to be true, and it conditions the requested information. Requested information, on the other hand, is the specific information being sought.

In the third step, readers must search and read (or read and search) a text or document to identify the necessary information that corresponds with information provided in the question and, in the case of multiple-choice items, in the list of choices. In carrying out this search, several matches may be tried before one or more adequate matches are achieved. If a literal or synonymous match is made between requested or given information and corresponding text or document information, readers may proceed to the next step. If such a match is not deemed adequate, readers may choose to make a match based on a low- or high-level text-based inference or on prior knowledge; or readers may recycle to the first step.

In the fourth step, readers complete the requested information frame by identifying the information asked for in the question. In some instances, readers are unable to complete the requested information frame based upon information associated with a current match for given information. In such cases, readers may recycle to an earlier step in the model, searching for information in another part of the text or document. In other instances, readers may once again need to make some sort of inference to relate the requested information to information in the text.

Finally, in the fifth step, readers may recycle to earlier steps to determine that all the conditions specified in a question have been adequately addressed. In some instances, readers may recycle in this step to identify information in different parts of a text or document. In the earlier task, readers may read through the paragraph once again to ensure that the other choices do not represent the main point of the passage.

This test-taking model of reading can be applied to both documents and prose and to multiple-choice as well as open-ended tasks. Based on this model, Kirsch and Mosenthal identified three variables as being among the best predictors of task difficulty for the prose and document scales. These variables (type of requested information, type of match and plausibility of distractors) were described in the previous section and are elaborated in Appendix A which also explainshow each variable is scored.

Several exemplary tasks will be used to further characterise these variables. The next question of interest is how useful these variables are in predicting task difficulty and then in building an interpretative scheme. The examples shown are drawn from the International Adult Literacy Survey (IALS).

Characterising prose literacy tasks

There are 34 tasks ordered along the IALS 500-point prose literacy scale. These tasks range in difficulty value from 188 to 377. The easiest task (receiving a difficulty value of 188) directs the reader to look at a medicine label (see belouw) to determine the "maximum number of days you should take this medicine". In terms of the process variables, type of match was scored a "1" because the reader was required to locate a single piece of information that was literally stated in the medicine label. The label contained only one reference to number of days and this information was located under the label "dosage". Type of information was scored a "2" because it asked for a number of days, and plausibility of distractor received a "1" because there is no other reference to days in the medicine label.

MEDCO ASPIRIN *500*

INDICATIONS: Headaches, muscle pains, rheumatic pains, toothaches, earaches. RELIEVES COMMON COLD SYMPTOMS.

DOSAGE: ORAL. 1 or 2 tablets every 6 hours, preferably accompanied by food, for not longer than 7 days. Store in a cool, dry place.

CAUTION: Do not use for gastritis or peptic ulcer. Do not use if taking anticoagulant drugs. Do not use for serious liver illness or bronchial asthma. If taken in large doses and for an extended period, may cause harm to kidneys. Before using this medication for chicken pox or influenza in children, consult with a doctor about Reyes Syndrome, a rare but serious illness. During lactation and pregnancy, consult with a doctor before using this product, especially in the last trimester of pregnancy. If symptoms persist, or in case of an accidental overdose, consult a doctor. Keep out of reach of children.

INGREDIENTS: Each tablet contains
500 mg acetylsalicicylic acid.
Excipient c.b.p. 1 tablet.
Reg. No. 88246

Made in Canada by STERLING PRODUCTS, INC.
1600 Industrial Blvd., Montreal, Quebec H9J 3P1

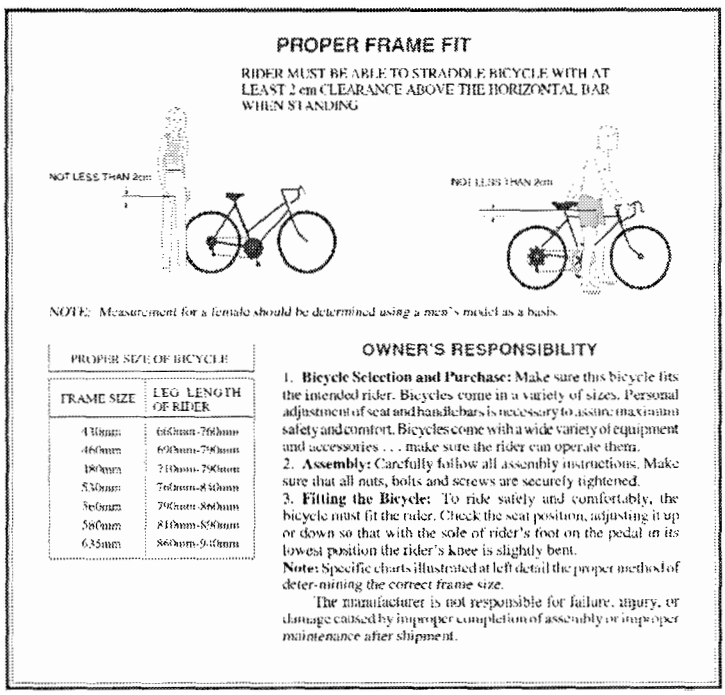
0 67736 11079

Reprinted by permission

A second Prose literacy task directs the reader to look at an article about impatiens. One task receiving a difficulty value of 254 asks the reader to identify "what the smooth leaf and stem suggest about the plant". Again, the task direct the reader to locate information contained

in the text so it was scored "1" for type of information. The last sentence in the second paragraph under the heading *Appearance* states: "The smooth leaf surfaces and the stems indicate a great need of water". Type of information was scored a "3" because it directs the reader to identify a condition. Plausibility of distractor was scored a "3" because the same paragraph contained a sentence which served to distract the readers. This sentence states, "... stems are branched and very juicy, which means, because of the tropical origin, that the plant is sensitive to cold".

Tasks which fall at higher levels along the scale present the reader with more varied demands in terms of the type of match that is required and in terms of the number and nature of distractors that are present in the text. One such task (with a difficulty value of 281) refers the reader to a page from a bicycle owner's manual to determine how to ensure the seat is in the proper position (see below). Type of information was scored a "3" because readers needed to identify and state two conditions that needed to be met in writing. In addition, they were not told how many features they needed to provide from among those stated. Type of information was also scored a "3" because it involved identifying a condition, and plausibility of distractor received a score of "2".



PROPER FRAME FIT

RIDER MUST BE ABLE TO STRADDLE BICYCLE WITH AT LEAST 2 cm CLEARANCE ABOVE THE HORIZONTAL BAR WHEN STANDING

NOT LESS THAN 2cm

NOT LESS THAN 2cm

NOTE: Measurement for a female should be determined using a man's model as a basis.

PROPER SIZE OF BICYCLE

FRAME SIZE	LEG LENGTH OF RIDER
430mm	665mm-760mm
460mm	690mm-790mm
480mm	710mm-790mm
530mm	760mm-840mm
560mm	790mm-860mm
580mm	810mm-880mm
635mm	860mm-940mm

OWNER'S RESPONSIBILITY

1. **Bicycle Selection and Purchase:** Make sure this bicycle fits the intended rider. Bicycles come in a variety of sizes. Personal adjustment of seat and handlebars is necessary to assure maximum safety and comfort. Bicycles come with a wide variety of equipment and accessories . . . make sure the rider can operate them.
2. **Assembly:** Carefully follow all assembly instructions. Make sure that all nuts, bolts and screws are securely tightened.
3. **Fitting the Bicycle:** To ride safely and comfortably, the bicycle must fit the rider. Check the seat position, adjusting it up or down so that with the sole of rider's foot on the pedal in its lowest position the rider's knee is slightly bent.

Note: Specific charts illustrated at left detail the proper method of *deter-mining the correct frame size.*

The manufacturer is not responsible for failure, injury, or damage caused by improper completion of assembly or improper *maintenance after shipment.*

A somewhat more difficult task (318) involves an article about cotton diapers and directs the reader to "list three reasons why the author prefers to use disposable rather than cotton diapers". This task is made more difficult because of several of the process variables. First, type of match was scored a "5" because the reader had to provide multiple responses, each of which required a text-based inference. Nowhere in the text does the author say, "I prefer cotton diapers because...". These inferences are made somewhat more difficult because the type of information being requested is a "reason" rather than something more concrete. This variable received a score of "4". Finally, plausibility of distractor was scored a "3" because the text contains information that may serve to distract the reader.

An additional task falling at an even higher place along the Prose literacy scale (338) directs the reader to use the information from a pamphlet about hiring interviews (see below) to "write in your own words one difference between the panel and the group interview". Here the difficulty does not come from locating information in the text. Rather than merely locating a fact about each type of interview, the readers need to integrate what they have read to infer a characteristic on which the two types of interviews differ. Experience from other surveys of this kind reveal that tasks in which readers are asked to contrast information are more difficult, on average, than tasks in which they are asked to find similarities. Thus, type of match was scored "6". Type of information was scored "5" because it directs the reader to provide a difference. Differences tend to be more abstract in that they ask for the identification of distinctive or contrastive features related in this case to an interview process. Plausibility of distractor was scored "1" because no distracting information was present in the text. Thus this variable was not seen as contributing to the overall difficulty of this task.

The Hiring Interview

Preinterview

Try to learn more about the business. What products does it manufacture or services does it provide? What methods or procedures does it use? This information can be found in trade directories, chamber of commerce or industrial directories, or at your local employment office.

Find out more about the position. Would you replace someone or is the position newly created? In which departments or shops would you work? Collective agreements describing various standardized positions and duties are available at most local employment offices. You can also contact the appropriate trade union.

The Interview

Ask questions about the position and the business. Answer clearly and accurately all questions put to you. Bring along a note pad as well as your work and training documents.

The Most Common Types of Interview

One-on-one: Self explanatory.

Panel: A number of people ask you questions and then compare notes on your application.

Group: After hearing a presentation with other applicants on the position and duties, you take part in a group discussion.

Postinterview

Note the key points discussed. Compare questions that caused you difficulty with those that allowed you to highlight your strong points. Such a review will help you prepare for future interviews. If you wish, you can talk about it with the placement officer or career counsellor at your local employment office.

The most difficult task on the Prose literacy scale (377) required readers to look at an announcement from a personnel department and to "list two ways in which CIEM (an employee

support initiative within a company) helps people who lose their jobs because of departmental reorganization". Type of match was scored "7" because the question contained multiple phrases that the reader needed to keep in mind when reading the text. In addition, readers had to provide multiple responses and make low text-based inferences. Type of information was scored "3" because readers were looking for a purpose or function, and plausibility of distractor was scored a "4". This task is made somewhat more difficult because the announcement is organised around information that is different from what is being requested in the question. While the correct information is listed under a single heading, this information is embedded under a list of headings describing CIEM's activities for employees looking for other work. This list of headings in the text serves as an excellent set of distractors for the reader who does not search for or locate the phrase in the question containing the conditional information – those who lose their jobs because of a departmental reorganisation.

Evaluating the contribution of the variables to difficulty of prose literacy tasks

The Item Response Theory (IRT) scaling procedures that were used in the IALS constitute a statistical solution to the challenge of establishing one or more scales for a set of tasks with an ordering of difficulty that is essentially the same for everyone. Each scale can be characterised in terms of how tasks are ordered along it. The scale point assigned to each task is the point at which individuals with that proficiency score have a given probability of responding correctly. In this survey, an 80 percent probability of correct response was the criterion used. This means that individuals estimated to have a particular scale score are expected to perform tasks at that point on the scale correctly with an 80 percent probability. It also means that they will have a greater than 80 percent chance of performing tasks that are lower on the scale. It does not mean, however, that individuals with given proficiencies can never succeed at tasks with higher difficulty values; they may do so some of the time. Yet, it does suggest that their probability of success is "relatively" low – that is, the more difficult the task relative to their proficiency, the lower the likelihood of a correct response.

An analogy might help to clarify this point. The relationship between task difficulty and individual proficiency is much like the high jump event in track and field, in which an athlete tries to jump over a bar that is placed at increasing heights. Each high jumper has a height at which he or she is proficient – that is, the jumper can clear the bar at that height with a high probability of success, and can clear the bar at lower heights almost every time. When the bar is higher than the athlete's level of proficiency, however, it is expected that the athlete will be unable to clear the bar consistently.

Once the literacy tasks are placed along each of the scales using a response probability criterion of 80 percent (RP80), it is possible to see to what extent the variables associated with task characteristics explain the placement of tasks along the scales. A multiple regression was run using RP80 as the dependent variable[3]. The independent variables were the three process variables used to characterise the prose tasks plus a traditional measure of readability. The results are shown in Table 1.

Table 1

Standardised β and t-ratios representing the regression of readability and process variables against RP80 values on prose tasks along with their zero-order correlation

Variable	β coefficient	t-ratio	Significance	Corr. w/RP80
Type of Match (TOM)	.74	10.0	.00	.89
Type of Information (TOI)	.16	2.3	.03	.55
Plausibility of distractor (POD)	.20	2.8	.01	.54
Readability	.11	1.8	.09	.28

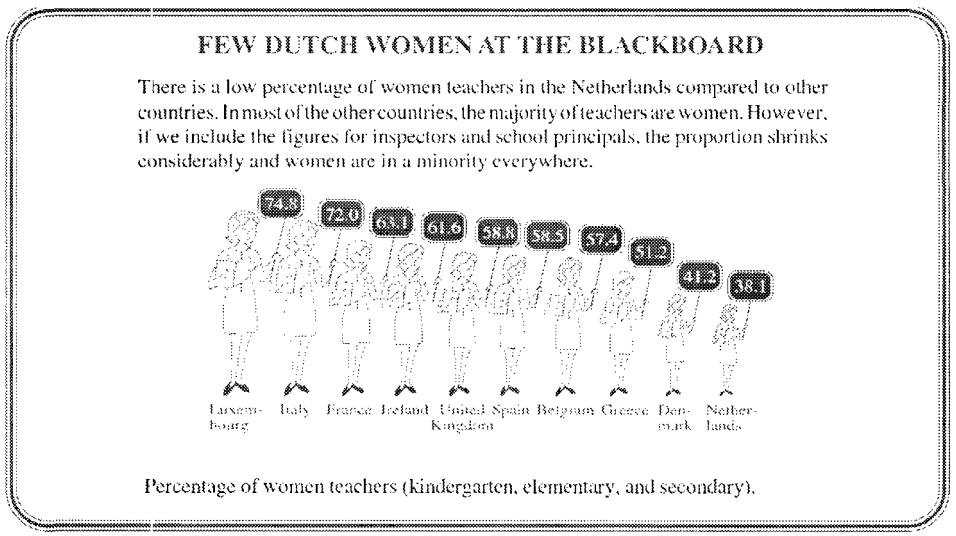
Note. Multiple R: .94; R^2: .89; Adjusted R^2: .87.

Table 1 shows the zero-order correlation of each predictor variable with RP80 along with the results of the regression analysis. These data reveal that type of match had the largest zero-order correlation with RP80 (.89) and received the largest standardised regression weight followed by plausibility of distractor and type of information. Together these variables along with readability accounted for 89 percent of the variance in predicting RP80 values.

Easy tasks on the Prose literacy scale tended to require readers to make a literal match on the basis of a single piece of concrete information where few, if any, distractors were present in the text. Tasks further along the Prose scale become somewhat more varied. While some may still require a single feature match, more distracting information may be present in the text or the match may require a low text-based inference. Some tasks may require the reader to cycle through information to arrive at a correct response. Tasks that are more difficult can take on a variety of characteristics. They may still require the reader to make a match but usually the reader has to match on multiple features, or to take conditional information into account. Tasks may also require the reader to integrate information from within a text or to provide multiple responses. The most difficult tasks typically require the reader to make higher-level inferences, process conditional information, and to deal with highly plausible distracting information.

Characterising document literacy tasks

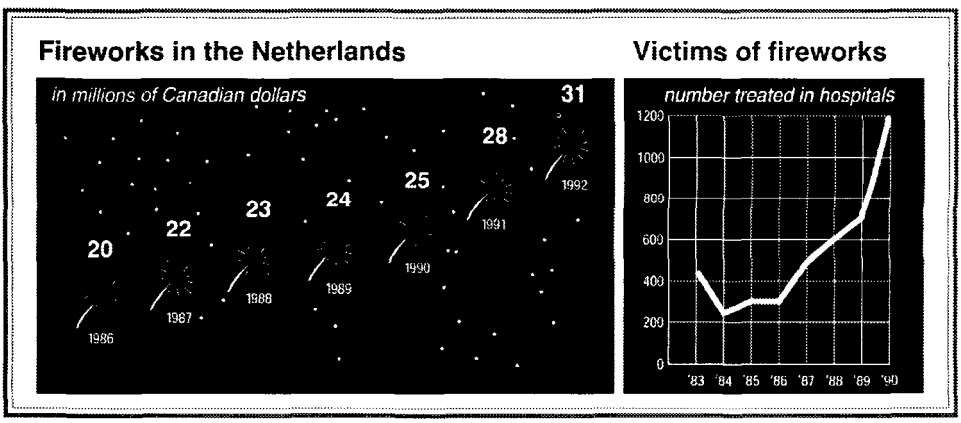
There are 34 tasks ordered along the IALS 500-point Document literacy scale. These tasks range in difficulty value from 182 to 408. One document literacy task with a difficulty value of 188 directs the reader to identify from a chart the percentage of teachers from Greece who are women (see below). The chart displays the percentage of teachers from various countries who are women. In terms of the process variables, type of match was scored a "1" because the reader was required to locate a single piece of information that was literally stated in the chart, type of information received a "2" because it was an amount, and plausibility of distractor was also scored a "2" because there are distractors for the requested information.



A second document task involving this same chart directs the reader to identify the country other than the Netherlands in which women teachers are in the minority. This item received a difficulty value of 234. This task was made a bit more difficult than the first

because rather than searching on a country and locating a percentage, the reader had to know that minority means less than 50 percent. Then, they had to cycle through to identify the countries in which the percentage of women teachers is less than 50 percent. In addition, they had to remember the condition "other than the Netherlands"; otherwise they might have chosen it over the correct response. As a result, type of match was scored a "3"; type of information a "1" because the requested information is a country or place; and plausibility of distractor a "2" because there are distractors associated with the requested information.

A somewhat more difficult task (295) directs the reader to look at charts involving fireworks from the Netherlands (see below) and to write a brief description of the relationship between sales and injuries based on the information shown. Here the reader needs to look at and compare the information contained in the two charts and integrate this information, making an inference regarding the relationship between the two sets of information. As a result, it was scored a "5" for type of match. Type of information received a "4" because the requested information is asking for a pattern or similarity in the data. Plausibility of distractor was scored a "3" primarily because both given and requested information is present in the task. For example, one of the things that may have contributed to the difficulty of this task is the fact that the sales graph goes from 1986 to 1992 while the injuries graph goes from 1983 to 1990. The reader should compare the information from the two charts for the comparable period of time.



Another set of tasks covering a range of difficulty on the document scale involved a rather complicated document taken from a page in a consumer magazine rating clock radios (see below). The easiest of the three tasks, receiving a difficulty value of 287, asks the reader to identify "which two features are not on any basic clock radio". The reader has to cycle through the document, find the listing for basic clock radios, and then determine that a dash represents the absence of a feature. They then have to locate the two features indicated by the set of dashes. As a result, type of match received a score of "4" because it is a cycle requiring multiple responses with a condition or low text-based inference. Type of information was scored a "2" because its features are an attribute of the clock radio and plausibility of distractor is a "2" because there are some characteristics that are not associated with other clock radios.

A somewhat more difficult task associated with this document received a difficulty value of 327 and asks the reader to identify "which full-featured clock radio is rated highest on performance". Here readers must make a three-feature match (full-featured, performance, and highest) where one of the features requires them to process conditional information. It is possible, for example, that some readers were able to find the full-featured radios and the column listed under "performance" but selected the first radio listed assuming it was the one rated highest. In this case, they did not understand the conditional information which is a legend stating what the symbols mean. Others may have gone to the column labelled "overall score", found the highest

number, and chosen the radio associated with it. For this reason, plausibility of distractor was scored a "3". Type of information received a "1" because the requested information is a "thing".

The most difficult task associated with this document (408) asks the reader to identify the average advertised price for the basic clock radio receiving the highest overall score. This task was made more difficult because the readers had to match four rather than three features; they also had to process conditional information and there was a highly plausible distractor in the same node as the correct answer. As a result of these factors, type of match received a score of "5", type of information a score of "2", and plausibility of distractor a score of "5".

RATINGS

Better ◄ ► Worse

Clock radios

Listed by types; within types, listed in order of overall score. Differences in score of 4 points or less were not deemed significant.

1 Brand and model. If you can't find a model, call the company. Phone numbers are listed on page 736.

2 Price. The manufacturer's suggested or approximate retail price, followed by the average advertised price.

3 Dimensions. To the nearest centimetre.

4 Overall score. A composite, encompassing all our tests and judgments. A "perfect" radio would have earned 100 points.

5 Convenience. This composite judgment reflects such things as the legibility of the display, the ease of tuning the radio and setting the alarm, and the presence or absence of useful features.

6 Performance. An overall judgment reflecting performance in our tests of: sensitivity and selectivity; tuning ease; capture ratio, the ability to bring in the stronger of two stations on the same frequency; image rejection, the ability to ignore signals from just above the band, resistance to interference from signals bouncing off aircraft and such.

7 Sensitivity. How well each radio received a station with little interference.

8 Selectivity. How well each radio received clearly a weak station next to a strong one on the dial.

9 Tone quality. Based mainly on computer analysis of the speaker's output and on listening tests, using music from CDs. No model produced high-fidelity sound.

10 Reversible time-setting. This useful feature makes setting clock and alarm times easy. If you overshoot the desired setting, you simply back up.

11 Dual alarm. Lets you set two separate wake-up times.

■ Brand and model	Price ■	Dimensions, HxWxD, cm. 3	Overall Score 4	Convenience 5	Performance 6	Sensitivity 7	Selectivity 8	Tone quality 9	Reversible time setting 10	Dual alarm 11	Warranty, months	Advantages	Disadvantages	Comments
Full-featured clock radios														
RCA RP-3690	$50/$40	6x25x18	86	●	◓	◓	◓	●	✔	✔	12	A,B,D,H,J,L,O,T,U		A
Sony ICF-C303	50/45	5x20x15	84	●	◓	●	○	◒	✔	✔	12	C,E,F,I,N,T		C
Panasonic RC-X220	50/45	10x28x13	82	◓	◓	◓	◓	○	✔	✔	12	A,G,K,M,O,S,T,U	b,c	A
Realistic 272	50/30	5x28x15	79	◓	○	◓	◓	◒	✔	✔	3	A,G,H,K,O,T		D
Magnavox AJ3900	65/—	15x38x13	78	○	●	◓	●	◓	—	✔	3	D,G,K,M,O,R,T	b,g	B
Emerson AK2745	39/20	8x28x15	70	○	◒	◒	◒	○	✔	✔	3	G,O	g	K
Soundesign 3753	20/20	8x23x13	62	○	◒	●	○	○	✔	✔	3	J,Q	d,h	J
Basic clock radios														
Realistic 263	28/18	10x20x10	74	○	◓	◓	○	◒	—	—	3	A,D,H,O,P,U	h	—
Soundesign 3622	12/10	5x20x13	68	◒	◓	●	◒	◒	—	—	3	U	d	L
Panasonic RC-6064	18/15	5x20x13	67	◒	◓	●	○	○	—	—	12	—	b,c	—
General Electric 7-4612	13/10	5x20x13	66	◒	○	●	◒	○	—	—	12	A,D	a,g	—
Lloyds CR001	20/15	5x18x13	64	◒	○	●	○	◒	—	—	3	U	—	—
Sony ICF-C240	15/13	5x18x15	63	◒	○	○	○	○	—	—	12	—	f,g	—
Emerson AK2720	19/10	5x20x13	61	◒	○	◓	●	○	—	—	3	O,T	e	K
Gran Prix D507	15/10	5x18x10	54	◒	●	○	●	●	—	—	3	—	d	—
Clock radios with cassette player														
General Electric 7-4965	60/50	10x30x15	85	◓	◓	◓	◓	◓	✔	✔	12	A,D,G,H,K,O,S,T	—	B,E
Pansonic RC-X250	[1]	10x33x13	76	◓	◓	○	◓	◓	✔	✔	12	A,G,K,O,R,U	b,c	A,H
Sony ICF-CS650	75/65	15x28x15	74	○	◓	●	○	◒	✔	✔	12	G,R,T,U	c,i	A,F,H
Soundesign 3844MGY	40/30	13x30x13	62	○	●	●	●	◒	—	—	3	G,K,J,S,U		F,G,J,M

[1] *Discontinued. Replaced by* **RC-X260**, *$79 list and $60 average advertised sale price.*

Features in Common

All: • Permit snooze time of about 8 min. • Retain time settings during short power failures. *Except as noted, all have:* • Battery backup for clock and alarm memory. • Red display digits 1 cm. high. • Sleep-time radio play for up to 60 min. before automatic shutoff. • Switch to reset alarm.

Keys to Advantages

A–Alarm works despite power failure.
B–Shows actual time plus up to 2 alarm times.
C–Twin alarms settable for 2 different stations.
D–Tone alarm has adjustable volume control.
E–Memory needs no battery.
F–Digital tuner with presettable stations.
G–Tuner can receive in stereo.
H–Battery-strength indicator.
I–Illuminated tuning dial.
J–Illuminated tuning pointer.
K–Earphone jack.
L–Nap timer.
M–Audio input for tape deck or CD player.
N–Display can show date and time.
O–Display has high/low brightness switch.
P–Display has larger digits than most.
Q–Night light—adjusts for room light.
R–Bass-boost tone control.
S–Treble-cut tone control.
T–Better than most in tuning ease.
U–Better than most in image rejection.

Key to Disadvantages

a–Possible to reset time by accident.
b–Controls for time-setting or dimmer inconveniently located on radio's bottom or rear.
c–Display dimmer than most in brightly lit room.
d–Radio volume must be turned completely down for alarm buzzer to sound.
e–Lacks alarm buzzer; radio is sole alarm.
f–Lacks indication alarm is set.
g–Lacks alarm-reset button.
h–Time-setting lacks fast reverse.
i–No slow forward, fast reverse for time setting.

Key to Comments

A–Display shows green digits.
B–Display shows blue digits.
C–Display uses LCD (liquid crystal) digits.
D–Terminals for external antenna.
E–3-position graphic equalizer.
F–Cassette player lacks Record function.
G–Cassette player lacks Rewind function.
H–Model permits wake-up to cassette play.
I–Cassette-deck flutter worse than most.
J–Warranty repairs cost $3 for handling.
K–Warranty repairs cost $3.50 for handling.
L–Warranty repairs cost $6 for handling.
M–Warranty repairs cost $10 for handling.

Evaluating the contribution of the variables to difficulty of document literacy tasks

As with the Prose scale, IRT was used to establish the document literacy scale as well as to characterise tasks along it. Again, a response probability of 80 percent was used as an indicator that someone at a specified point on the Document literacy scale has mastered or is proficient with tasks at that place on the scale. It does not mean that they cannot perform tasks above their estimated proficiency; rather they may do so but with less consistency. Their expected consistency on tasks above their level of proficiency depends on how far the task is from their estimated proficiency.

Once the document literacy tasks are placed along each of the scales using the criterion of 80 percent, it is possible to see to what extent the variables associated with the task characteristics explain the placement of tasks along the scales. A multiple regression was run using RP80 as the dependent variable (see footnote 3). The independent variables were the three process variables used to characterise the prose and document literacy tasks plus a newly developed measure of document readability (Mosenthal & Kirsch, 1998). The results are shown here in Table 2.

Table 2

Standardised β and t-ratios representing the regression of readability and process variables against RP80 values on document tasks along with their zero-order correlation

Variable	*β* coefficient	*t*-ratio	Significance	Corr. w/RP80
Type of Match (TOM)	.43	3.7	.00	.88
Type of Information (TOI)	.13	1.4	.16	.43
Plausibility of distractor (POD)	.40	3.8	.00	.71
Readability	.17	1.7	.09	.55

Note. Multiple *R*: .89; R^2: .79; Adjusted R^2: .76.

Table 2 shows the zero-order correlation between each of the predictor variables and RP80 along with the results from the regression analysis. These data reveal that each of the predictor variables is significantly correlated with RP80, yet only two process variables received significant beta weights. It should be noted that while each of these variables may not be significant in terms of this regression analysis, each was taken into consideration when constructing the literacy tasks and, therefore, each is important for ensuring that the domain is well represented. Together the set of variables accounted for 79 percent of the variance in RP80 values. Type of match received the largest standardised regression weight followed by plausibility of distractors. The regression weights for type of information and readability did not reach significance.

Easy tasks on the Document literacy scale tended to require readers to make a literal match on the basis of a single piece of information. Tasks further along the Document scale become somewhat more varied. While some may still require a single feature match, more distracting information may be present in the document or the match may require a low text-based inference. Some tasks may require the reader to cycle through information to arrive at a correct response. Tasks that are more difficult can take on a variety of characteristics. They may still require the reader to make a match, but usually the reader has to match on multiple features, or take conditional information into account. Tasks may also require the reader to integrate information from one or more documents or to cycle through a document to provide multiple responses. The most difficult tasks typically require the reader to match on multiple features, to cycle through documents, and to integrate information. Frequently, these tasks require the reader to make higher-level inferences, process conditional information, and to

deal with highly plausible distractors. These tasks also tend to be associated with more complex displays of information.

Building an interpretative scheme

Identifying and validating a set of variables that predict performance along each of the literacy scales provides a basis for building an interpretative scheme. This scheme provides a useful means for exploring the progression of information-processing demands across each of the scales and the meaning of scores along a particular scale. Thus, it contributes to the construct validity of a measure (Messick, 1989). This section summarises an interpretative scheme that was adopted by IALS. The procedure builds on Beaton's anchored proficiency procedures (Beaton & Allen, 1992; Messick, Beaton, & Lord, 1983), but is more flexible and inclusive than the one originally developed and used in the 1980s by the National Assessment of Educational Progress (NAEP). It has been used in various large-scale surveys of literacy in North America (Kirsch & Jungeblut, 1992; Kirsch et al., 1993).

As shown in the previous section of this paper, there is empirical evidence that a set of variables can be identified that summarises some of the skills and strategies that are involved in accomplishing various kinds of prose and document literacy tasks. More difficult tasks tend to feature more varied and complex information-processing demands than easier tasks. This suggests that literacy is neither a single skill suited to all types of tasks nor an infinite number of skills each associated with a particular type of task.

In the North American literacy surveys, when researchers coded each literacy task in terms of the process variables described in this paper, they noted that the values for these variables tended to shift at various places along each of the literacy scales. These places seemed to be around 50-point intervals, beginning at approximately 225 on each scale. While most of the tasks at the lower end of the scales had code values of 1 on each of the process variables, tasks with values around 225 were more likely to have code values of 2. Among tasks with scores around 275, many of the codes were 2s and an increasing number were 3s. Among tasks with values of 325, at least one of the three variables had a code value of 4. Code values of 4 or higher predominated tasks at around 375 or higher on the literacy scales.

Although there were some variations across the literacy scales in the points at which the coding shifts occurred, the patterns were remarkably consistent. Further, as was shown in this paper with the IALS tasks, this system of coding tasks accounts for much (although not all) of the variance associated with tasks along the literacy scales. Based on these findings, researchers defined five levels of proficiency having the following score ranges:

Level 1: 0-225
Level 2: 226-275
Level 3: 276-325
Level 4: 326-375
Level 5: 376-500

Once the literacy levels were identified, based on the noted shifts in code values for the three process variables, criteria were identified that describe the placement of tasks within these levels. These criteria are summarised along with the data to which they were applied in a chapter appearing in the IALS technical report (Kirsch et al., 1998). Based on evidence resulting from this work, the five literacy levels were used for reporting results from literacy assessments in both national and international surveys using these literacy scales.

Conclusion

One of the goals of large-scale surveys is to provide information that can help policy-makers during the decision-making process. Presenting such information in a way that will enhance the understanding of what has been measured and the conclusions to be drawn from the data is important for reaching this goal. This paper offers a framework that has been used for both developing the tasks used to measure literacy as well as for understanding the meaning of what is being reported with respect to the comparative literacy proficiencies of adults. The framework identifies a set of variables that have been shown to underlie successful performance on a broad array of literacy tasks. Collectively, they provide a means for moving away from interpreting survey results in terms of discrete tasks or a single number, and towards identifying levels of performance sufficiently generalised to have validity across assessments and groups. As concern ceases to centre on discrete behaviours or isolated observations and focuses more on providing meaningful interpretations of performance, a higher level of measurement is reached (Messick, 1989).

Appendix A: Coding rules for the process variables

Type of information

Type of information requested refers to the nature of information which readers must identify to complete a question or directive. Types of information form a continuum of concreteness, which was operationalised as follows for the purposes of this analysis.

When the requested information pertains to a:

- person, animal, place, or thing, score 1;
- amount, time, attribute, action, or location, score 2;
- manner, goal, purpose, condition, or predicate adjective, score 3;
- cause, result, reason, evidence, similarity, or pattern, score 4;
- equivalent, difference, or theme, score 5.

Plausibility of distracting information

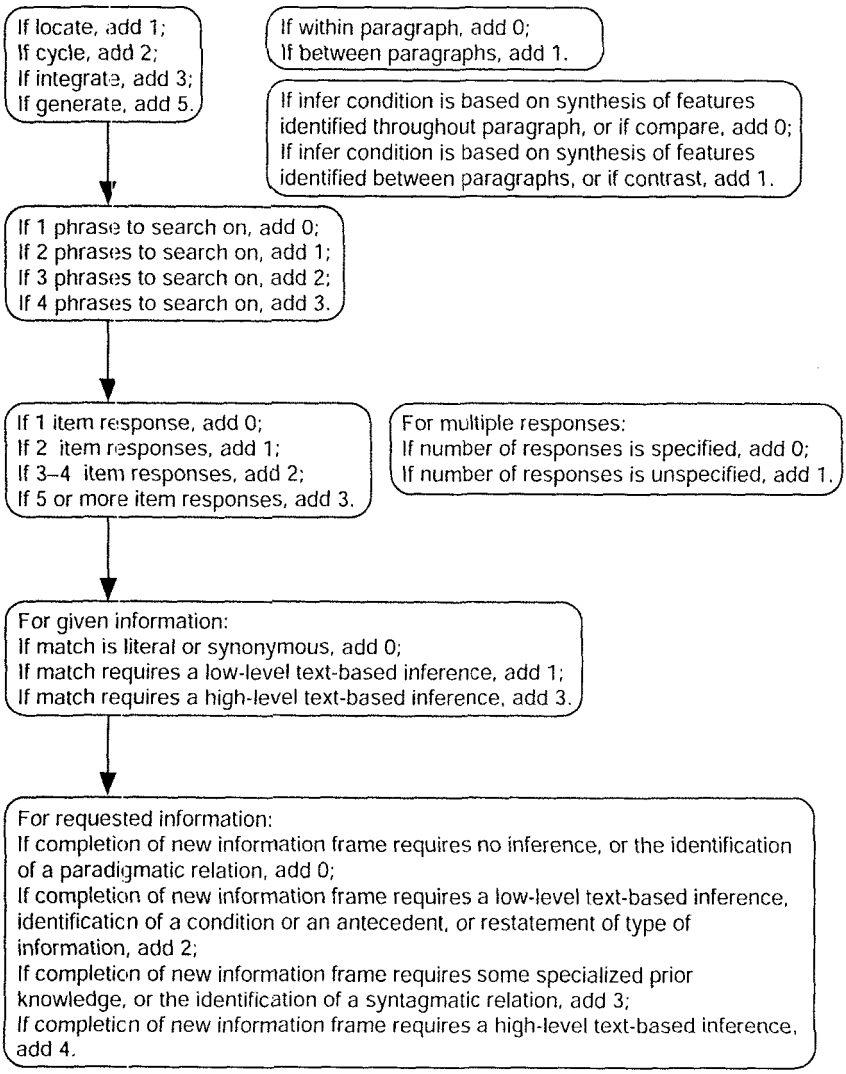
Plausibility of distracting information refers to whether or not an identifiable match exists between information in the question and the text, or between the text and the distractors in a multiple-choice question, which makes it difficult for readers to identify the correct answer. The scoring rules for plausibility of distracting information are as follows:

- when there is no distracting information in the text, score 1;
- when distractors contain information which corresponds literally or is synonymous to information in the text but not in the same paragraph as the answer, score 2;
- when distractors contain information which represent plausible invited inferences not based on information related to the paragraph in which the answer occurs, score 3;
- when one distractor in the choices contain information that is related to the information in the same paragraph as the answer, score 4;

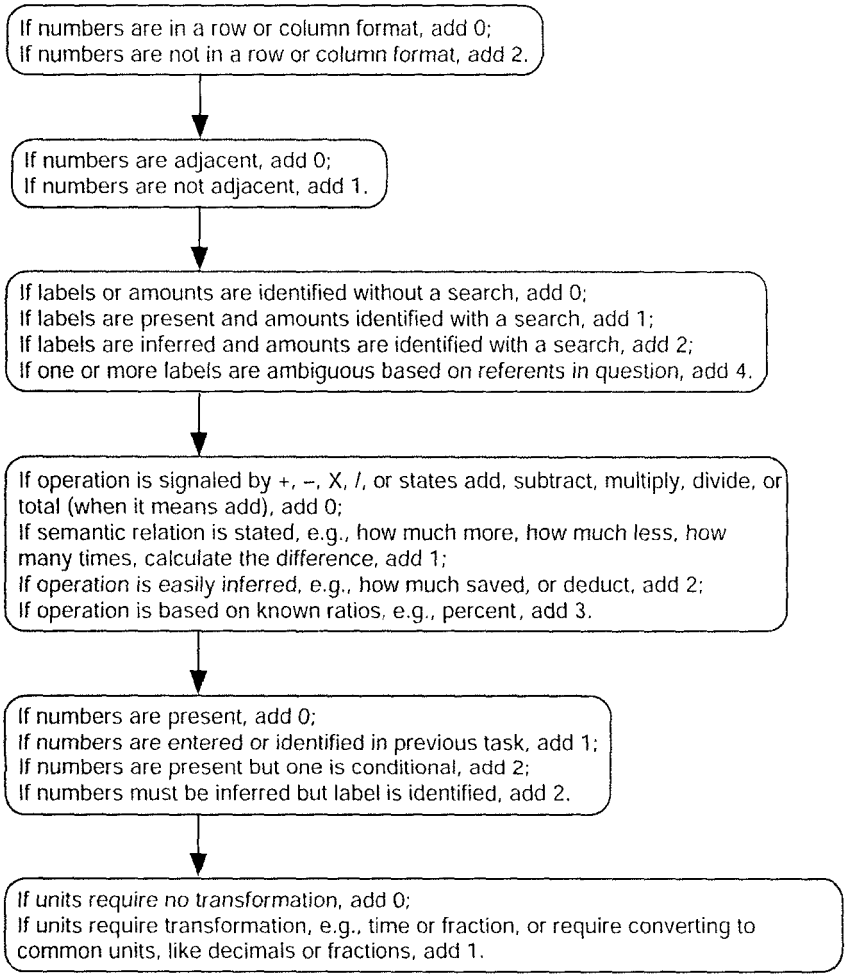
- when two or more distractors in the choices contain information which is related to the information in the same paragraph as the answer, score 5;
- when one or more distractors represent plausible inferences based on information outside the text, score 5.

Type of match

This variable relates to the nature of the task and the level of processing required to respond correctly to a task. The first diagram represents the additive scoring model used to code prose literacy tasks. It is followed by the model used to code document literacy tasks.



An additive scoring model for "type of match" in prose literacy tasks



An additive scoring model for "type of match" in document literacy tasks

Notes

1 This section is based on the work of Werlich (1976).

2 Mosenthal and Kirsch wrote a monthly column on understanding documents which appeared in the Journal of Reading between 1989 and 1991.

3 While most of the tasks in IALS received common RP80 values, a few tasks were assigned values unique to a particular country when warranted by the data. Since the value assigned to each variable used in the regression analyses was based on the evaluation of each task in English, it was decided to use the RP80 values for the U.S. as well.

References

Almond, R.G., & Mislevy, R.J. (1998). *Graphical models and computerized adaptive testing* (TOEFL Tech. Rep. No. 14). Princeton, NJ: Educational Testing Service.

Beach, R., & Appleman, D. (1984). Reading strategies for expository and literacy text types. In A. Purves & O. Niles (Eds.), *Becoming readers in a complex society. Eighty-third yearbook of the National Society for the Study of Education.* Chicago, IL: University of Chicago Press.

Beaton, A.E., & Allen, N.L. (1992). Interpreting scales through scale anchoring. *Journal of Educational Statistics, 17,* 191-204.

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1,* 140-161.

Clark, H., & Haviland, S.E. (1977). Comprehension and the given-new contract. In R.O. Freedle (Ed.), *Discourse production and comprehension* (pp. 1-39). Norwood, NJ: Ablex.

Cook-Gumperz, J., & Gumperz, J. (1981). From oral to written culture: The transition to literacy. In M. Whitman (Ed.), *Writing: The nature, development and teaching of written communication* (vol. 1). Hillsdale, NJ: Erlbaum.

Crandall, J. (1981). *Functional literacy of clerical workers: Strategies for minimizing literacy demands and maximizing available information.* Paper presented at the annual meeting of the American Association for Applied Linguistics (December). New York.

Diehl, W. (1980). *Functional literacy as a variable construct: An examination of the attitudes, behaviors, and strategies related to occupational literacy.* Unpublished doctoral dissertation, Indiana University.

Fisher, D.L. (1981). Functional literacy tests: A model of question-answering and an analysis of errors. *Reading Research Quarterly, 16,* 418-448.

Graff, H.J. (1979). *The literacy myth.* New York: Academic Press.

Guthrie, J.T. (1988). Locating information in documents: A computer simulation and cognitive model. *Reading Research Quarterly, 23,* 178-199.

Heath, S.B. (1980). The functions and uses of literacy. *Journal of Communication, 30,* 123-133.

Jacob, E. (1982). *Literacy on the job: Final report of the ethnographic component of the industrial literacy project.* Washington, DC: Center for Applied Linguistics.

Kirsch, I.S., & Guthrie, J.T. (1984a). Adult reading practices for work and leisure. *Adult Education Quarterly, 34,* 213-232.

Kirsch, I.S., & Guthrie, J.T. (1984b). Prose comprehension and text search as a function of reading volume. *Reading Research Quarterly, 19,* 331-342.

Kirsch, I.S., & Jungeblut, A. (1986). *Literacy: Profiles of America's young adults – Final report* (NAEP Report No. 16-PL-01). Princeton, NJ: National Assessment of Educational Progress.

Kirsch, I.S., & Jungeblut, A. (1992). *Profiling the literacy proficiencies of JTPA and ES/UI populations: Final report to the Department of Labor.* Princeton, NJ: Educational Testing Service.

Kirsch, I.S., & Mosenthal, P.B. (1990). Exploring document literacy: Variables underlying the performance of young adults. *Reading Research Quarterly, 25,* 5-30.

Kirsch, I.S., & Mosenthal, P.B. (1994). Interpreting the IEA Reading Literacy Scales. In M. Binkley, K. Rust, & M. Winglee (Eds.), *Methodological Issues in Comparative Educational Studies: The case of the IEA Reading Literacy Study* (pp. 135-192). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Kirsch, I., Jungeblut, A., & Mosenthal, P.B. (1998). The measurement of adult literacy. In T.S. Murray, I.S. Kirsch, & L. Jenkins (Eds.), *Adult literacy in OECD countries: Technical report on the first international adult literacy survey* (pp. 105-134). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Kirsch, I.S., Jungeblut, A., Jenkins, L., & Kolstad, A. (1993). *Adult literacy in America: A first look at the results of the National Adult Literacy Survey.* Washington, DC: U.S. Department of Education.

Lerner, D., & Lasswell, H.D. (1951). *The policy sciences: Recent developments in scope and method.* Stanford, CA: Stanford University Press.

Messick, S. (1987). Large-scale educational assessment as policy research: Aspirations and limitations. *European Journal of Psychology and Education, 2,* 157-165.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement* (3rd ed.). New York: Macmillan.

Messick, S., Beaton, A., & Lord, F. (1983). *National Assessment of Educational Progress Reconsidered: A new design for a new era* (NAEP Report No. 83-1). Princeton, NJ: National Assessment of Educational Progress.

Mikulecky, L. (1982). Job literacy: The relationship between school preparation and workplace actuality. *Reading Research Quarterly, 17*, 400-419.

Miller, P. (1982). Reading demands in a high-technology industry. *Journal of Reading, 26*, 109–115.

Montigny, G., Kelly, K., & Jones, S. (1991). *Adult literacy in Canadaresults of a national study*. Ottawa: Minister of Industry, Science and Technology (Statistics Canada, Catalogue No. 89-525-XPE).

Mosenthal, P.B., & Kirsch, I.S. (1989-1991). Understanding documents. A monthly column appearing in the *Journal of Reading*. Newark, DE: International Reading Association.

Mosenthal, P.B., & Kirsch, I.S. (1991). Toward an explanatory model of document process. *Discourse Processes, 14*, 147-180.

Mosenthal, P.B., & Kirsch, I.S. (1998). A new measure for assessing document complexity: The PMOSE/IKIRSCH document readability formula. *Journal of Adolescent and Adult Literacy, 41*, 638-657.

National Assessment of Educational Progress (NAEP). (1972). *Reading: Summary* (Report 02-R-00). Denver, CO: Education Commission of the States.

Organisation for Economic Co-operation and Development (OECD). (1999). *Measuring Student Knowledge and Skills: A New Framework for Assessment*. Paris: OECD.

Organisation for Economic Co-operation and Development (OECD) and Human Resources Development Canada (HRDC). (1997). *Literacy Skills for the Knowledge Society: Further Results of the International Adult Literacy Survey*. Paris and Ottawa: OECD and HRDC.

Organisation for Economic Co-operation and Development (OECD) and Statistics Canada (STATCAN). (1992). *Adult illiteracy and economic performance*. Paris: OECD.

Organisation for Economic Co-operation and Development (OECD) and Statistics Canada (STATCAN). (1995). *Literacy, Economy and Society: Results of the First International Adult Literacy Survey*. Paris and Ottawa: OECD and STAT CAN.

Organisation for Economic Co-operation and Development (OECD) and Statistics Canada (STATCAN). (2000). *Literacy in the Information Age: Final report of the International Adult Literacy Survey*. Paris and Ottawa: OECD and STAT CAN.

Resnick, D., & Resnick, L. (1977). The nature of literacy – An historical exploration. *Harvard Educational Review, 43*, 370-385.

Scribner, S., & Cole, M. (1981). *The psychology of literacy*. Cambridge, MA: Harvard University Press.

Sticht, T.G. (Ed.). (1975). *Reading for working: A functional literacy anthology*. Alexandria, VA: Human Resources Research Organization.

Sticht, T. (1978). *Literacy and vocational competency*. Columbus, OH: Ohio State University (Occasional Paper 39, National Center for Research in Vocational Education).

Sticht, T. (1982). *Evaluation of the reading potential concept for marginally literate adults* (Final Report FR-ET50-82-2). Alexandria, VA: Human Resources Research Organization.

Szwed, J. (1981). The ethnography of literacy. In M. Whitman (Ed.), *Writing: The nature development, and teaching of written communication* (vol. 1). Hillsdale, NJ: Lawrence Erlbaum Associates.

Venezky, R.L. (1983). The origins of the present-day chasm between adult literacy needs and school literacy instruction. *Visible Language, 16*, 113-136.

Werlich, E. (1976). *A text grammar of English*. Heidelberg: Quelle & Meyer.

Wickert, R. (1989). *No single measure*. Canberra, Australia: The Commonwealth Department of Employment, Education and Training.

L'article présente un cadre qui permet l'élaboration de tâches pour mesurer la compréhension de l'écrit ainsi qu'un dépistage plus précis du sens des données analogues dont nous disposons relatives à la compréhension de l'écrit parmi les adultes des pays participants. Le cadre consiste de six parties dont la séquence logique va de la nécessité de définir et de représenter un domaine d'intérêt spécifique à l'établissement d'une base empirique pour l'interprétation des résultats, en passant par l'identification et l'opérationnalisation des caractéristiques qui entrent dans la construction des items. L'importance des éléments de ce cadre se manifeste par leur potentiel de contribuer à la compréhension approfondie de la notion de compréhension de l'écrit et des processus divers qui lui sont associés. On propose un modèle de processus mental et procède à l'identification et à la vérification, moyennant des analyses de régression, des variables déterminant la performance dans les tâches de compréhension de l'écrit. On montre que ces variables expliquent de 79% à 89% de la variance relative à la difficulté des tâches. Dans leur totalité, ces variables de processus mental permettent de sortir du mesurage, courant dans les enquêtes à grande échelle, de la performance par tâches discrètes ou par simple échelle numérique pour arriver à une identification des niveaux de performance propres à être généralisés pour des groupes entiers de tâches et, par là, à ce que Messick a appelé un niveaux supérieur de measure.

Key words: Construct validity, Item development, Large-scale assessment, Proficiency levels, Reading literacy.

Received: December 2000

Irwin S. Kirsch. Educational Testing Service, Centre for Global Assessment, Rosedale Road, Mail 02-R, Princeton, NJ 08541, USA. E-mail: ikirsch@ets.org

Current theme of research:

Assessment design. Reading literacy. Test validity. Linking assessment with instruction.

Most relevant publications in the field of Psychology of Education:

Kirsch, I.S., & Guthrie, J.T. (1984). Prose comprehension and text search as a function of reading volume. *Reading Research Quarterly, 19,* 331-342.

Kirsch, I.S., & Guthrie, J.T. (1984). Adult reading practices for work and leisure. *Adult Education Quarterly, 34,* 213-232.

Kirsch, I.S., & Mosenthal, P.B. (1990). Exploring document literacy: Variables underlying the performance of young adults. *Reading Research Quarterly, 25,* 5-30.

Kirsch, I.S., Jungeblut, A., & Mosenthal, P.B. (1998). The measurement of adult literacy. In T.S. Murray, I.S. Kirsch, & L. Jenkins (Eds.), *Adult literacy in OECD countries: Technical report on the first international adult literacy survey.* Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Kirsch, I.S., Jungeblut, A., Jenkins, L., & Kolstad, A. (1993). *Adult literacy in America: A first look at the results of the National Adult Literacy Survey.* Washington, DC: U.S. Department of Education.