Text Processing Variables Predict the Readability of Everyday Documents Read by Older Adults

Bonnie J. F. Meyer, University of Washington

Michael Marsiske and Sherry L. Willis, Pennsylvania State University

The research reported in this article was supported by Grant AG08082 from the National Institute on Aging

Abstract

An eight step model, based on research in discourse processing, is presented to predict the difficulty older adults will have in reading documents from an everyday memory test. Test items are part of the Test of Basic Skills (ETS, 1977) and require readers to answer questions about charts (e.g., bus schedules), labels (e.g., plant spray labels and prescriptions), and forms (e.g., tax forms). The components of the model come from theoretical and empirical work on discourse processing and include such factors as discourse structure, emphasis, and position of an answer in a linguistic analysis of the everyday document.

A sample of 483 adults from 52 to 93 years of age took the everyday memory test as well as a psychometric ability battery. The correlation was .54 ($p < .01$) between the readability scores for test items predicted by the model and the percent of older adults correctly answering those items. In addition, the more difficult test items as identified by the model were correlated more highly with fluid intelligence abilities (Figural Relations and Induction), crystallized intelligence abilities (Experiential Evaluation), and with Memory Span.

The findings point to the feasibility of analyzing everyday materials for readability. Also, based on the model, guidelines can be made for improving the readability of everyday materials for older adults. In addition, the model would be useful in designing intervention strategies to assist older adults in attending to salient features of bus schedules, documents, charts, forms, directions and product labels.

Text Processing Variables Predict the Readability of Everyday Documents Read by Older Adults

Bonnie J. F. Meyer, University of Washington

Michael Marsiske and Sherry L. Willis, Pennsylvania State University

An eight step model, based on research in discourse processing, is presented to predict the difficulty older adults will have in reading documents from an everyday memory test. The test items are part of the Test of Basic Skills (ETS, 1977). The components of the model come from theoretical and empirical work on discourse processing and include such factors as discourse structure, emphasis, and position of the answer in a propositional analysis of everyday documents.

Document literacy (Kirsch & Jungeblut, 1986) involves reading skills necessary to understand and use printed materials occurring in a variety of non-prose formats. These formats include charts, schedules, tables, labels, and forms (Kirsch & Mosenthal, in press). Older adults encounter such documents in their everyday lives (e.g., medicine bottle labels, directions for utilizing products, public transportation schedules, financial documents). The opportunity to function independently in our society may rest partly on the ability to comprehend everyday documents.

The Test of Basic Skills (ETS, 1977) requires readers to answer questions about charts (e.g., bus schedules), labels (e.g., plant spray labels and prescriptions), and forms (e.g., tax forms). Many of the items on this test represent instrumental activities required to remain independent in the community (Lawton & Brody, 1969).

The model presented to predict readability of documents does not follow in the tradition of readability research that focused on counting low frequency words and sentence length (Clifford, 1978; Klare, 1963, 1974-75). Readability formulas built on word frequency and sentence length have not satisfactorily characterized the varying difficulty of tests and do not provide productive guidelines for creating readable tests (Davison & Kantor, 1982).

There have been many positive steps to simplify documents in recent years (e.g., Atwood, Baker & Duffy, 1985; Chapanis, 1965; Wright, 1980, 1987; Waller, 1984). Many attempts to improve document readability have followed the "redesign-test-redesign-and-test" approach (Wright, 1979). This approach has been useful in improving individual documents, but lacks generalizability (Kirsch & Mosenthal, in press). The model posited does not follow this approach either.

Instead, the model focuses on theoretical and empirical work on discourse processing and applies it to the reading comprehension of documents. Although the model was initially developed independently, it is similar to the approach of Kirsch and Mosenthal (in press), and ultimately has included some of their components. Mosenthal and Kirsch (in preparation) developed a propositional grammar to describe documents. They explained that this was necessary because existing grammars (e.g., Frederiksen, 1975; Kieras & Dechert, 1985; Kintsch, 1977; Meyer, 1975) tended to apply to a particular type of discourse structure, not documents. However, we found that the Meyer (1975, 1985a) system could be readily applied to documents and this analysis system was utilized for the components in the model requiring propositional analysis.

The grammar utilized by Kirsch and Mosenthal (in press) is similar to the Meyer (1975, 1985a) approach used in the present study.

Kirsch and Mosenthal (in press) identified a set of critical variables that underlie the performance of a national sample of young adult (ages 21 to 25 years) on the National Educational Progress (NAEP) study of Young Adult Literacy. The variables identified accounted for 89 percent of the variance for the total population of 3,618 young adults. Among differing education levels, the variables accounted for 56 percent of variance for young adults with 0 to 8 years of schooling, 81 percent for young adults with 9 to 12 years of schooling, 88 percent for young adults with high school degrees, and 84 percent for young adults with post-high-school degrees.

The variables in the Kirsch and Mosenthal study included six materials variables based on their grammar, three materials-by-task variables, and three response variable.s The only materials variable that predicted performance dealt with the amount of information in the document, a measure of the number of arguments in the document. The model posited in the current study did not have such a variable for the number of ideas in a document. The items on the Basic Skills Assessment Test were approximately the same length, and we did not believe that this would be a critical variable, although it may be for examining everyday documents that vary widely in length.

Kirsch and Mosenthal found two of their materials-by-task variables to predict performance. These variables related to the variable in our model called paraphrases. The Kirsch and Mosenthal work provides a detailed categorization system for this variable, while our approach is simpler.

The Kirsch and Mosenthal work is an important study and provides a foundation for a theoretical model of document processing. Many of the twelve variables identified in their study are complex and required considerable work to ascertain from documents. The current model presented in our study is compatible with the orientation of Kirsch and Mosenthal, but offers a somewhat simpler approach. In addition, we have looked at a different population of document users, older adults rather than young adults.

Figure 1 displays the model developed to analyze the readability of everyday documents over a wide range of formats (e.g., charts, labels) and to derive an item readability score, based on this analysis. Step 1 (Number of Propositions) is a measure of the amount of material to be learned and relates to readability (Kintsch & Vipond, 1979; Kirsch & Mosenthal, in press). Step 2 (Position in the Content Structure) in this scheme deals with the location of the answer to the question in the propositional analysis (content structure) of the document. Position in the content structure attributes greater difficulty and thus gives more points for a question whose answer comes from the lower levels in the content structure and that has no units beneath it in the structure to facilitate retrieval. The rationale for this step comes from the large body of literature showing greater memory for information higher in the content structure than lower in the structure (e.g., Kintsch & van Dijk, 1978; Meyer, 1975; Meyer & McConkie, 1973). Also, Walker and Meyer (1980) found that information in sentences higher in the structure is more easily

---

Insert Figure 1 about here

---

integrated than information in lower level sentences. Step 3 (Paraphrases) involves adding points to the accumulating readability score for each content word in the question that is a paraphrase of a word in the document. The rationale for Step 3 comes from the work on comprehension questions of Anderson (1972). Step 4 (Mismatched Relationships) adds one point for each relationship among the ideas in the question and those in the document that are not identical. The types of relations among the ideas in both the document and question are analyzed. Those relationships among the ideas that are not constant between the analyses of the document and the question are counted and added to the readability score. Facilitation is expected when the relational structure of the question and the document are the same (e.g., Thordyke, 1977; Meyer, 1975). Step 5 (Types of Structures) relates to the types of structures used to organize the answer in the content structure of the document. One point was added for each comparison relationship, and two points were added to the difficulty score for each causation or problem/solution relationship (Meyer, 1985a). No points are given for the easier structures of sequence and enumeration that appear to be more common in documents. The research indicates that sequences are easier to process than causation and problem/solution (Englert & Hiebert, 1984; Hidi & McLaren, 1988; Meyer & Freedle, 1984; Meyer, Young & Bartlett, 1989; Raphael, Englert, & Kirschner, 1986, Richgels, McGee, Lomax, & Sheard, 1987). Step 6 (Emphasis) deals with the highlighting of the answer in the document. If the answer is emphasized through initial placement in the document, signaling, boldface, capitalization, or some other technique, then less difficulty is anticipated for the task. Such emphasis is thought to ease the processing load (Britton, Glynn, Meyer, & Penland, 1982). Facilitation from such emphasis devices has been reported (Brown & Smiley, 1977; Marshall & Glock, 1978-1979; Kieras, 1985; Mayer, 1985; Meyer, 1985b; Meyer, Brandt, & Bluth, 1980). Step 7 (Search Complexity) involves how many places in the content structure a subject would have to examine to find the answer. One point was added for each place in the content structure that it was necessary to examine in order to arrive at an answer. Charts and tables often required a search of two or more places in the content structure. This step is similar to both rational task analysis (e.g., Gagne, 1985) and componential analysis (e.g., Sternberg, 1977). The final step was added for use with multiple choice test items. As the number and relationships among alternative answers increases, the difficulty level of the item increases (Brown, 1986; Drum, Calfee, & Cook, 1981; Kirsch & Mosenthal, in press)

The study examines the efficacy of this proposed model for predicting the reading comprehension of everyday documents by older adults.

Method

Sample

The sample was composed of 483 community-dwelling Caucasian Pennsylvania elders (405 females, 77 males), ranging in age from 52 to 93 years (*M* = 73.39 years, *SD* = 6.60 years). 232 subjects were participants in a 7-year follow-up of the Adult Development and Enrichment Project (ADEPT), conducted in 1986. The rest of the sample (n=251) were subjects from the same geographical locations as the original sample, but tested for the first time in 1986.

Educational level ranged from 3 to 22 years, (*M* = 11.59 years, *SD* = 2.89 years). Mean annual income reported ranged from under $2000 to over $28000 (*Median* = $9876.46, *SD* = $10153.13). Married persons constituted 32.8% of the sample, and the remaining 67.2% of the sample included widowed, single, divorced and separated persons. Only 6.7% of the sample was employed full- or part-time, with the remainder of the subjects reporting themselves as homemakers or retired. Self reported health, hearing and vision ranged from 1 (very good) to 6 (very poor). Mean health self-rating was 2.11 (*SD* = 0.88). Mean hearing self-rating was 2.53 (*SD* = 1.08). Mean vision self-rating was 2.50 (*SD* = 0.92). Subjects reported a mean life satisfaction rating of 2.88 (*SD* = 1.10), with a range of 1 (extremely happy) to 6 (very unhappy).

Subjects were paid for their participation at a rate of $5.00 per hour. Participation involved 5-10 hours of testing (over 2-9 testing sessions), plus a take-home questionnaire packet.

Measures

Ability battery. The psychometric ability battery used in the present study was developed within the fluid (Gf) and crystallized (Gc) model of intelligence (see Cattell, 1971). While multiple marker tests of each of seven primary mental abilities were used, the present study reports results using only one marker for each ability. These seven tests were previously identified as the purest markers (i.e., highest factor loadings) of the abilities they represent (Baltes, et al, 1980). Table 1 presents the ability battery, including the specific measures used, the primary mental abilities they represent, and the broad second-order dimensions on which the primary abilities load.

Insert Table 1 about here.

The primary abilities of Figural Relations (CFR) and Induction (I) were selected to represent fluid intelligence. Subjects must observe patterns of figures or letters in marker tests of these abilities, and educe the pattern of relationships contained within them. Crystallized intelligence (Gc) was represented by marker tests of Experiential Evaluation (EMS), which required subjects to manipulate problems of a social nature, and Verbal Comprehension (V); the measure of Verbal Comprehension was a recognition vocabulary test. Semantic Relations (CMR) measures, which require subjects to complete verbal analogies, were used to represent both fluid and crystallized intelligences. Two other primary abilities were represented in the present study: Memory Span (Ms), which was assessed using a test of digit span; and Perceptual Speed (Ps). Perceptual Speed measures assess the speed with which subjects make simple visual discriminations.

All but one of the measures were adapted versions of previously published psychometric ability tests. Test adaptations consisted mainly of enlargement of test stimuli, modification of response format to simplify test taking for older adults, and reduction of the number of test items. All reduced measures had alpha reliabilities above 0.65. The ADEPT Induction test was not previously published, and was first developed in the initial phase of ADEPT (Blieszner, Willis & Baltes, 1981; Willis, Blieszner & Baltes, 1981).

Everyday task measure. The Educational Testing Service (ETS) Test of Basic Skills (ETS, 1977), was the measure of everyday task performance used in the present study. The measure contains 65 items, and measures subjects' ability to comprehend printed materials, including charts and forms (e.g. bus schedules), labels (e.g., plant spray label), and technical documents (e.g., warranty).

Procedure

Subjects were assessed in small groups (3-12) by a tester and a proctor. All testers and proctors (n=6) ranged from young to middle adulthood. A small number of subjects required individual testing sessions, due to transportation or sensory difficulties. All testing sessions were usually held over a 2-day to 3-week interval. All tests were administered in community settings, usually the senior citizen centers to which participants belonged.

Derivation of scores

Standardization of ability data. Ability data for all subjects in 1985 were standardized to their 1979 base (M = 50, SD = 10). This procedure was done to preserve change in test performance levels from 1979 to 1986.

Score derivations for everyday task measure. To examine the utility of the Readability scoring procedure, a subset of 62 items was selected from the 65-item everyday task measure. These items were representative of a variety of item formats (charts, forms, labels, everyday prose). Three items were eliminated because they involved the interpretation of graphic and pictorial information (e.g., cartoons, maps), and did not fit well into a linguistic parsing scheme.

Individual item responses were recoded into a dichotomous 1=right, 0=wrong format. The maximum total score on the 62 items was therefore 62.

The proportion of subjects (N=483) who correctly answered each of the 62 selected items was computed. This proportion correct ranged from 29.5% to 97.6%, with a mean of 75.88% (SD = 16.63%). The sample proportion answering an item correctly served as an second criterion for determining item difficulty (a higher proportion of subjects would be expected to correct answer an easy item relative to a hard one), against which the Readability score components could be compared.

Derivation of Readability scores.

Figure 1 depicts the model, based on research in discourse processing, that was utilized to derive readability scores. The model includes eight kinds of variables that were expected to influence the difficulty of comprehending the items on the test of everyday memory. As seen in the Figure, the first variable dealt with the number of propositions (as indexed by the number of organizing categories--see Kirsch & Mosenthal (in press)) in the document. The second variable dealt with the position of the answer to the question in the content structure of the analyzed document. The second element of the difficulty model for an item was therefore the number of points equal to level of answer in content structure of the document.

The third variable focused on paraphrases. Points were added for paraphrased words/phrases in the answer, that must be matched with related words/phrases in the document. As seen in Figure 1 from zero to three points could be given for a paraphrase depending on the variance between the words to be matched. No points were given when the match was literal. One point was given for a paraphrase that involved a simple word substitution (e.g., "drug" for "medication; "address" for "where one lives"; "twice a day" for "two times a day"; "a gallon" for "one gallon"). Two points were added to the difficulty score for a paraphrase that involved a string of words or required prior knowledge of common English language usage (e.g., "in combination" for "at same time"; "years of service" for "after working x years"; "other study" for "learned how to repair radios in the navy"). Three points were added for paraphrases that required the reader to have prior specialized knowledge in order to match text material and answer (e.g., "250 mg" for ".250 g"; "gallon" for "4 quarts"; knowing both the generic and brand name of a drug).

The fourth variable was the number of mismatched relationships between the document and the question. First, a propositional analysis (Meyer, 1975, 1985a) was completed to determine the relationship structure of the document and the question. Then, the relationships among contents (verbatim or paraphrased) were examined to see if they were the same in the analysis of the question as they were for the analysis of the document. Each difference noted in this comparison added one point to the difficulty scores. Care was taken to be sure that the document and the question had been analyzed to the same degree of specificity so than any difference did not simply reflect a difference in the level of analysis. The fifth variable related to the types of structures used to organize the answer in the content structure of the document. One point was added for each comparison relationship, and two points were added to the difficulty score for each causation or problem/solution relationship (Meyer, 1985a).

The sixth variable seen in the model shown in Figure 1 relates to emphasis. One point was subtracted from the Readability score each time the answer was emphasized in the document (e.g., an answer could be emphasized by appearing as the first item in the document, by underlining, boldface print, large type, or other types of signaling (Meyer, 1985b)). The seventh variable involved search complexity. One point was added for each place in the content structure that it was necessary to examine in order to arrive at an answer. Charts and tables often required a search of two or more places in the content structure. The eighth and final variable indexed the plausibility of multiple choice foils. When none of the multiple choice distractors appeared in the text, a score of one was assigned. As the number of distractors increased, as the critical features shared by the answer and the distractors increased, and as the proximity of the distractor to the answer in the content structure increased, a higher plausibility score was assigned, to a maximum of five. This model element was adapted from one presented by Kirsch & Mosenthal (in press), although the direction of scoring was reversed.

As will be described below, the individual components of this readability model for each item were simultaneously entered into a multiple regression, with the sample proportion correct for each item as the dependent variable. The obtained beta-weights provided a formula for weighting and summing model elements into a total "readability score". Higher readability scores meant easier items.

Results

Two questions were addressed: 1) What is the relationship between an empirical estimate of item difficulty (i.e., sample proportion correct on each item) and the theory-based estimate of item difficulty (i.e., the model components contained in the Readability score)? 2) What is the relationship between ability performance and item difficulty, as determined by the item Readability scores. That is, is there a stronger relationship with ability performance for items having lower (more difficult) Readability scores than for items having higher Readability scores?

With regard to the first question, the proportion of subjects answering each item correctly was correlated with each of the 8 Readability score components for that item. In addition, the sample proportion correct was regressed on the Readability model components. The average correlations, across the 62 items, and the regression parameter estimates, are shown in Table 2.

Insert Table 2 about here

Subsequently, a Readability score was generated for each item. A reduced regression model was run, containing only those model elements with parameter estimates significant at $p < .20$ (i.e., number of propositions, content level of answer in text, paraphrases, number of mismatched relationships, number of search points, and plausibility of multiple choice foils). The estimated regression weights for this model are displayed in Table 3. These obtained regression estimates were used to create estimated sample proportions correct (i.e., model-based Readability scores) for each item. The correlation between these estimated Readability scores and the sample proportion correct was 0.54.

Insert Table 3 about here

With regard to the second question, the 31 items which received the highest Readability scores were summed together ("Easy" item subset, maximum score = 31), as were the 31 items which received the lowest Readability scores ("Hard" item subset, maximum score = 31). Subjects' scores on each of these two item subsets were then correlated with their scores on the standardized marker tests of each of the seven primary abilities. Table 4 presents the resulting correlation matrix.

Insert Table 4 about here

Easy and hard item subsets differed significantly in their correlations with the psychometric abilities of Figural Relations (CFR), Induction (I), Experiential Evaluation (EMS) and Memory Span (Ms) ($p < .05$). In all cases, the hard item subset was more highly correlated with psychometric ability performance than the easy item subset. These significant differences were obtained for markers of fluid (I, CFR) and crystallized (V, EMS) intelligences, as well as for a marker of general memory (Ms).

Discussion

The results obtained in this study present some interesting findings concerning the utility of applying discourse processing techniques to paper-and-pencil measures of everyday task performance. The high correlation coefficients obtained when Readability scores (obtained through the discourse processing analysis model outlined above) were correlated with sample proportion incorrect on ETS Basic Skills items suggest that a text-processing analysis of test materials may be useful in predicting and accounting for sample response patterns. This has both theoretical and practical significance. Theoretically, it suggests that prior research on the dimensions of text readability _are_ salient for predicting actual comprehension in everyday task materials. Consequently, the present findings are an extension of prior findings, using more ecologically valid materials and text formats than may have been used in prior research. Practically, the text processing model outlined in the present paper presents a useful algorithm for the selection of test items; generation of Readability scores provides an _a priori_ method for varying item difficulty in a text-based test.

Traditionally, items could only be excluded following the results of pilot testing on standardization samples. Unfortunately, such an approach could be very sample dependent; a particular sample may have different response patterns than other samples. Since pilot tests are often conducted on convenience samples, inferences about response patterns in a broader population from a pilot sample may be dangerous. The present analyses suggest that it may be possible to evaluate item properties even before pilot testing, providing an initial estimate of item quality which is less sample dependent. Further, the present results imply that exclusion of items from such a test may now be guided by theoretical considerations; the present item difficulty analysis system provides reasons for excluding particular items. Traditional exclusion heuristics, based solely on sample response patterns, are largely atheoretical.

Two important caveats to the above interpretation should be stated. First, the findings are based on only a limited set of everyday task materials. A larger, broader set of everyday texts and problems is needed to strengthen confidence in the predictive utility of our Readability scores. Second, the findings are based on a sample of 483 older Pennsylvania adults. The sample proportions correct obtained may be salient only for this highly selected population. Thus, the high predictive utility of our Readability scores for predicting sample proportion incorrect could be limited only to the present study. Kirsch & Mosenthal's (in press) findings that the ability to predict item difficulty was reduced with lower-education subsets is germane to the present study. Cohort differences in educational attainment are undoubtedly one reason for the comparatively low variance accounted for in the present study. In addition, a Readability model operates at the item level, and does not take account of individual differences. Consequently it is unlikely that in a highly heterogenous sample of older adults that it would be possible to account for all of the variance in sample performance. Future research needs to examine the usefulness of our text-processing based item-difficulty model for understanding the everyday printed task performance of other groups of older adults, as well as other age groups. Despite these limitations, our findings taken together with those of Kirsch and Mosenthal (in press) with young adults present support for the validity of this technique for predicting the readability of documents and tests querying such documents.

The findings of differential magnitude of correlation with abilities for hard and easy item subsets (as defined by readability scores) provide an additional way of understanding differential item Readability. Hard items correlate more highly with fluid intelligence abilities (Figural Relations and Induction), with crystallized intelligence abilities (Experiential Evaluation), and with Memory Span. These findings suggest that more difficult items may require a larger crystallized knowledge base on which to draw. They may also require a greater application of processing strategies and reasoning to text materials (i.e., greater fluid intelligence). The memory demands that difficult items present may be greater (this last point agrees well with Meyer's previous research, which suggests that the elements which make an item more "difficult" also tend to make text less memorable). The present findings therefore suggest that text elements which make reading materials more difficult may do so because they increase the intellectual demands of text processing.

The present analyses focused specifically on the usefulness of a discourse processing approach for understanding item difficulty on a measure of everyday task performance. The text materials used in the present study were similar to materials encountered in the everyday world: bus schedules, technical documents, charts, forms, directions, product labels. Several intervention implications emerge from our findings. First, it may be possible to analyze the printed materials encountered in everyday life for Readability. Second, it may be possible to suggest a strategy for increasing the readability of everyday text. Third, it may be possible to design intervention strategies which help individuals to attend to salient aspects of everyday text, and to improve their comprehension of these materials.

Since the present study was conducted with a sample of older adults, the intervention implications of these findings become increasingly important. The opportunity to remain independent in the community may rest largely on the ability to comprehend everyday text materials (e.g., medicine bottle labels, public transportation schedules, financial documents, etc.). A heuristic which helps to understand the locus of age-related deficits (if any) in dealing with everyday printed materials can also help in the design of intervention strategies for these older adults. It follows that an effort to improve older adults' abilities to deal with everyday texts may actually help to improve their fitness for an independent life in the non-institutional community. Further research needs to examine these issues.

In summary, the present findings may help to understand three related issues. First, they provide further evidence that text-factors help to determine the readability and comprehensibility of text. Second, the findings from the present study suggest that ecologically valid texts, encountered in the everyday world, may also be more or less difficult due to text-factors intrinsic in them. Third, the results suggest that more difficult texts are more difficult because they impose higher intellectual/processing demands, particularly in the domains of fluid and crystallized intelligences, and in working memory. Fourth, the results suggest that it may be possible to construct highly comprehensible everyday printed materials which reduce the intellectual demands imposed on the reader. A large scale revision of texts found in the everyday world seems like an unlikely short-term goal. Consequently , the results also suggest that it may be possible to instruct older adults in reading strategies geared to comprehending documents. Strategy training has proved helpful for older adults in both the domain of reading expository texts (Meyer, Young, & Bartlett, 1989) and performing on intelligence tests (Willis, Blieszner, & Baltes, 1981). Thus, strategy training focused on reading and using documents successfully may improve older adults abilities to deal with the documents they encounter in everyday life.

Figure 1.

Text based model for assessing difficulty in processing everyday documents.

INPUT: Document and Relevant Question

1\. DOCUMENT SIZE/NUMBER OF PROPOSITIONS
Count the number of organizing categories in the text

2\. POSITION IN CONTENT STRUCTURE
The level at which the answer to the question is embedded in the text

3\. PARAPHRASES

0. Literal match between answer & question
1. Simple word substitution (e.g., drug, medication)
2. Low text-based inference of phrase (e.g., finish all, how long take)
3. High text-based inference (e.g., height with cm)

4\. MISMATCHED RELATIONSHIPS
Add one point for each relationship not identical between analyses of question and document

5\. DISCOURSE STRUCTURE
Discourse structure organizing the answer:
0=description, 1= comparison, 2=causation, problem/solution

6\. EMPHASIS
Subtract one point if answer is emphasized (italics, boldface, capitals, initial item in text)

7\. SEARCH COMPLEXITY
Add 1 point for each place in content structure that must be searched to arrive at answer

8\. PLAUSIBILITY OF MULTIPLE CHOICE FOILS

Table 1.

The ADEPT ability battery: Second-order dimensions, first-order primary mental abilities, and selected marker tests

General Dimension	Primary Ability	Test	Source
Gf	CFR	Culture Fair Test (Scale 2, Form A) and Power Matrices (Scale 3, Form A, 1936 ed., and Form B, 1961 ed.)	Cattell & Cattell (1957, 1961, 1963)
Gf	I	ADEPT Induction Diagnostic Test (Form A)	Blieszner, Willis & Baltes (1981)
Gf/Gc	CMR	Word Matrix	Guilford (1969)
Gc	EMS	Social Translations	Horn (1967)
Gc	V	Verbal Meaning	Thurstone (1962)
Ms	Ms	Auditory Number Span--Delay	After Ekstrom, et al (1976)
Ps	Ps	Number Comparison	Ekstrom, et al (1976)

Table 2.
Relationship between sample proportion correct and Readability model components for 62 ETS Basic Skills items

	Number of propositions	Position in content structure	Paraphrases	# Mismatched Relationships
Sample % Correct	0.194[a]	-0.229[a]	-0.295*	-0.288*
Parameter Estimate	0.748*	-3.163[a]	-2.425*	-2.575[a]

	Discourse structure	Emphasis	Search Complexity	Plausibility of Multiple Choice Foils
Sample % Correct	-0.017	-0.101	-0.094	-0.216[a]
Parameter Estimate	1.720	2.615	0.506[a]	-7.039*

Notes: Intercept = 108.236, $p < .0001$

Model $R^2 = 0.381$, $p < .012$

* $p < .05$

a $p < .20$

Table 3.
Reduced model: Prediction of item difficulty (sample proportion correct) from Readability Model components for 62 ETS Basic Skills items

Dependent variable: Sample proportion correct

Independent variables	Parameter Estimate	T-value for parameter
Intercept	109.947	11.000***
Number of Propositions	0.485	1.897[a]
Position in Content Structure	-2.911	-1.320[b]
Paraphrases	-2.301	-2.006*
# Mismatched Relationships	-2.552	-1.916[a]
Search Complexity	0.551	1.731[a]
Plausibility of Multiple Choice Foils	-7.423	-2.559**

Model $R^2 = 0.29$, $p < .003$

[a] $p < .10$
[b] $p < .20$

* $p < .05$
** $p < .01$
*** $p < .001$

Table 4.
<u>Correlation of Easy and Hard everyday task measure item subsets with primary mental ability marker tests (sample size for each correlation in parentheses)</u>

	Hard item subset	Easy item subset	T value for difference between hard & easy correlations[a]
Figural Relations (Fluid)	0.666 (n=482)	0.628 (n=482)	2.33**
Induction (Fluid)	0.610 (n=482)	0.547 (n=482)	3.00**
Perceptual Speed (Speed)	0.518 (n=445)[b]	0.504 (n=445)	0.52
Verbal Ability (Crystallized)	0.685 (n=353)[b]	0.666 (n=353)	1.15
Experiential Evaluation (Crystallized)	0.589 (n=353)[b]	0.505 (n=353)	-3.13**
Semantic Relations (Fluid/ Crystallized)	0.431 (n=353)[b]	0.395 (n=353)	-1.08
Memory Span (Memory)	0.526 (n=353)[b]	0.466 (n=353)	-2.00*

[a] T-ratio was adjusted for changes in <u>df</u> and correlation between easy & hard item subsets. (Note 1)

[b] Ability data for this measure was available for a subset of individuals.

All correlation coefficients were significant ($p < .0001$).

Difference between the two correlations is significant; "hard" subset score is more highly correlated with ability than "easy" subset score,

* $p < .05$
** $p < .01$

References

Anderson, R. C. (1972). How to construct achievement tests to assess comprehension. Review of Educational Research, 42, 145-170.

Atwood, N., Baker, E., & Duffy, T. M. (1985). Cognitive approaches to the analysis of technical understanding. (CDC Technical Report No. 24). Pittsburgh, P: Communications Design Center, Carnegie Mellon University.

Baltes, P. B., Cornelius, S., Nesselroade, J. R., Spiro, A., & Willis, S. L. (1980). Integration versus differentiation of fluid/crystallized intelligence in old age. Developmental Psychology, 16, 625-635.

Blieszner, R., Willis, S. L., & Baltes, P. B. (1981). Training research in aging on the fluid ability of inductive reasoning. Journal of Applied Developmental Psychology, 2, 247-265.

Britton, B. K., Glynn, S. M., Meyer, B. J. F., Penland, M. J. (1982). Effects of text structure on use of cognitive capacity during reading. Journal of Educational Psychology, 74, 51-61.

Brown, A. L., & Smiley, S. S. (1977). Rating the importance of structural units of prose passages: A problem in metacognitive development Child Development, 48, 1-8.

Brown, J. (1986). Similarity relations in recognition: A model to explain Tulving's data. Journal of Memory and Language, 25, 197-201.

Cattell, R. B. (1971). Abilities: Their structure, growth and action. Boston, MA: Houghton-Mifflin.

Cattell, R. B., & Cattell, A. K. S. (1957). Test of "g": Culture Fair (Scale 2, Form A). Champaign, Ill. Institute for Personality and Ability Testing.

Cattell, R. B., & Cattell, A. K. S. (1961). Measuring intelligence with the Culture-Fair Tests: Manual for Scales 2 and 3. (Scale 2, Form A). Champaign, Ill. Institute for Personality and Ability Testing.

Cattell, R. B., & Cattell, A. K. S. (1963). Test of "g": Culture Fair (Scale 3, Form A and Form B). Champaign, Ill. Institute for Personality and Ability Testing.

Chapanis, A. (1965). Words, words, words. Human Factors, 7, 1-17.

Clifford, G. (1978). Words for schools: the applications in education of the vocabulary researches of Edward Thordike. In P. Suppes (Ed.), Impact of research on education: Some case studies. Washington, DC: National Institute of Education.

Davison, A., & Kantor, R. N. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. Reading Research Quarterly, 18, 187-209.

Drum, P. A., Calfee, R. C., & Cook, L. K. (1981). The effects of surface structure variables on performance in reading comprehension tests. Reading Research Quarterly, 16, 486-514.

Ekstrom, R. B., French, J. W., Harman, H., & Derman, D. (1976). Kit of factor referenced cognitive tests. (Revised ed.). Princeton, NJ: Educational Testing Service.

Englert, C. S., & Hiebert, E. H. (1984). Children's developing awareness of text structures in expository materials. Journal of Educational Psychology, 76, 65-74.

Frederikson, C. (1975). Representing logical and semantic structure of knowledge acquired from discourse. Cognitive Psychology, 7, 371-458.

Gagne, R. M. (1985). The conditions of learning. New York: Hold, Rinehart, & Winston.

Guilford, J. P. (1969). Word Matrix Test. Beverly Hills, CA: Sheridan Psychological Services.

Hidi, S., & McLaren, J. (1988). The development of text structures in children's written expositions. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Horn, J. L. (1967). Social Situations--EP03A. Unpublished test, University of Denver.

Kieras, D. E. (1985). Thematic processes in the comprehension of technical prose. In B. K. Britton, & J. Black (Eds.). Analyzing and understanding expository text (pp. 89-108). Hillsdale, N. J.: Erlbaum.

Kieras, D. E., & Dechert, C. (1985). Rules for comprehensible technical prose: A survey of the psycholinguistic literature. (Technical Report No. 21, TR-85-ONR-21). Ann Arbor, MI: University of Michigan.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. Psychological Review, 85, 363-394.

Kintsch, W., & Vipond, D. (1979). Reading comprehension and readability in educational practice and psychological theory. In L. G. Nilsson (Ed.). Perspectives on memory research. Hillsdale, NJ: Erlbaum.

Kirsch, I. S., & Jungeblut, A. (1986). Literacy: Profiles of America's Young Adults. (Final Report No. 16-PL-02). Princeton, NJ: NAEP and Educational Testing Service.

Kirsch, I. S., & Mosenthal, P. B. (in press). Understanding document literacy: Variables underlying the performance of young adults. Reading Research Quarterly.

Klare, G. R. (1963). The measurement of readability. Ames, IA: Iowa State University Press.

Klare, G. R. (1974-1975). Assessing readability. Reading Research Quarterly, 10, 62-102.

Lawton, M. P., & Brody, E. (1969). Assessment of older people: Self-maintaining and instrumental activities of daily living. Gerontologist, 9, 179-185.

Marshall, N., & Glock, M. D. (1978-1979). Comprehension of connected discourse: A study into the relationship between the structure of text and information recalled. Reading Research Quarterly, 14, 10-57.

Mayer, R. E. (1985). Structural analysis of science prose: Can we increase problem-solving performance? In B. K. Britton, & J. Black (Eds.). Analyzing and understanding expository text (pp. 65-88). Hillsdale, NJ: Erlbaum.

Meyer, B. J. F. (1975). The organization of prose and its effects on memory. Amsterdam: North Holland.

Meyer, B. J. F. (1985a). Prose analysis: Purposes, procedures and problems. In B. K. Britton, & J. Black (Eds.). Analyzing and understanding expository text (pp 11-64; 269-304). Hillsdale, NJ: Erlbaum.

Meyer, B. J. F. (1985b). Signaling the structure of text. In D. H. Jonassen (Ed.), The Technology of Text. (pp 64-89). Englewood Cliffs, NJ: Educational Technology Publications.

Meyer, B. J. F., Brandt, D. M., & Bluth, G. J. (1980). Use of the top-level structure in text: Key for reading comprehension of ninth-grade students. Reading Research Quarterly, 16, 72-103.

Meyer, B. J. F., & Freedle, R. O. (1984). Effects of discourse type on recall. American Educational Research Journal, 21, 121-143.

Meyer, B. J. F., & McConkie, G. W. (1973). What is recalled after hearing a passage? Journal of Educational Psychology, 65, 109-117.

Meyer, B. J. F., Young, C. J., & Bartlett, B. J. (1989). Memory Improved: Enhanced reading comprehension and memory across the life span through strategic text structure. Hillsdale, NJ: Lawrence Erlbaum Associates.

Mosenthal, P. B., & Kirsch, I. S. (In preparation). Defining the structure of literacy materials: Towards a unified grammar of print. Princeton, NJ: Educational Testing Service.

Raphael, T. E., Engler, C. S., & Kirschner, B. M. (1986). The impact of text structure instruction and social context on students' comprehension and production of expository text. (Research Series No. 177). East Lansing: Michigan State University, Institute for Research on Teaching.

Richgels, D. J., Mcgee, L. M., Lomax, R. G., & Sheard, C. (1987). Awareness of four text structures: Effects of recall on expository text. Reading Research Quarterly, 22, 177-196.

Sternberg, R. J. (1977). Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities. Hillsdale, NJ: Erlbaum.

Thorndyke, P. W. (1977). Cognitive structures in comprehension and memory of narrative discourse. Cognitive Psychology, 9, 77-110.

Thurstone, T. G. Primary mental abilities for Grades 9-12 (Revised ed.). Chicago: Science Research Associates, 1962.

Waller, R. (1984). Designing a government form: A case study. Information Design Journal, 4, 36-57.

Walker, C. H., & Meyer, B. J. F. (1980). Integrating different types of information in text. Journal of Verbal Learning and Verbal Behavior, 19, 263-275.

Willis, S. L., Blieszner, R., & Baltes, P. B. (1981). Intellectual training research in aging: Modification of performance on the fluid ability of figural relations. Journal of Educational Psychology, 73, 41-50.

Wright, P. (1980). Strategy and tactics in the design of forms. Visible Language, 14, 151-193.

Wright, P. (1987). Writing technical information. In E. A. Rothkopf (Ed.). Review of Research in Education (pp. 327-385). Washington, DC: American Educational Research Association.

Wright, P. (1979). The quality control of document design. Information Design Journal, 1, 33-42.

Notes

1. The t-test for the difference between correlation coefficients with correlated samples must be adjusted for correlations between the criterion variables being correlated with a common predictor. After Ferguson (1976) the following formula was used:

$$\underline{t} = \frac{(r_{12} - r_{13})\,[(N - 3)(1 + r_{23})]^{1/2}}{\{2[1 - (r_{12})^2 - (r_{13})^2 - (r_{23})^2 + 2r_{12}r_{13}r_{23}]\}^{1/2}}$$

$\underline{df} = N - 3$

Walker & Lev (1953) note that, while the above formula for the t-ratio makes no assumption as to the distribution of X & Y (here, "easy" and "hard" subset scores), generalization is only possible for all samples for which X and Y have exactly the same set of values as those in the observed sample. This stringent assumption makes use of the above formula less than ideal for the present purposes, and caveats to the generalizability of the t-test results obtained must noted.