DOCUMENT RESUME

ED 397 138 TM 025 324

AUTHOR Kolstad, Andrew
TITLE The Response Probability Convention Embedded in Reporting Prose Literacy Levels from the 1992 National Adult Literacy Survey.
PUB DATE Apr 96
NOTE 30p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adults; Interviews; *Item Response Theory; *Literacy; National Surveys; *Probability; Responses; Scaling; *Test Results; Young Adults
IDENTIFIERS Item Characteristic Function; *National Adult Literacy Survey (NCES); National Assessment of Educational Progress; *Response Probability Convention

ABSTRACT

The role of the response probability convention in reporting results from the 1992 National Adult Literacy Survey is explored, using interviews with more than 26,000 adults and young adults. In order to summarize what respondents of a particular proficiency can do, it is convenient to adopt a convention for a sufficient response probability that is stringent enough to ensure that people at a lower bound can do what the task requires most of the time. Sections of the paper focus on the measurement of literacy in this survey, sources of the data, and item response theory scaling methods and item characteristic curves. The use of item mapping to anchor the prose literacy scale by locating specific tasks along it, using a response probability convention, and the literacy levels created for the survey in order to generalize beyond specific tasks to the more abstract abilities underlying the scales are also discussed. Other sections consider the relationship of the response probability convention to the cut points between the literacy levels and the variation in the proportions of the adult population reported to be in each prose literacy level as a function of the response probability convention. Results indicate that if the adult literacy program were to adopt the same response probability convention as that used in the National Assessment of Educational Progress, the proportion of the population in literacy Levels 1 and 2 would drop by 15% and the proportion in literacy Level 5 would increase by 9 percentage points. (Contains 8 figures and 10 tables.) (SLD)

The Response Probability Convention Embedded in Reporting Prose Literacy Levels from the 1992 National Adult Literacy Survey

by
Andrew Kolstad
National Center for Education Statistics

Paper prepared for presentation to the annual meetings
of the American Educational Research Association
New York, New York, April, 1996

In September, 1993, the National Adult Literacy Survey reported its findings on the distribution of literacy skills in the United States.[1] The survey provided the most detailed portrait that has ever been available on the condition of literacy in this nation. The survey measured the English literacy of adults in the United States based on their performance across a wide array of tasks that reflect the types of printed materials and literacy demands they encounter in their daily lives.[2]

The survey reported that about one fifth of the 191 million adults in this country demonstrated skills in the lowest of five prose literacy levels and that about a quarter of all adults demonstrated prose literacy skills in the next lowest level of proficiency (Level 2). The authors of the initial report did not attempt the task of counting the number of illiterates in this nation, arguing that the conception of literacy as something individuals either have or do not have is misleading and oversimplifies a complex issue. In fact, most of the adults in the lowest literacy level have some literacy skills. More than 70 percent of adults in this level reported reading a newspaper at least once a week, and nearly three-quarters of those in the lowest prose literacy level responded correctly to at least one literacy task in the assessment. Interpreting the adequacy of literacy performance is not easy. The survey collected no data on literacy requirements, so it was impossible to specify what skills are essential for individuals to succeed in society. Still, the authors of the initial survey report interpreted performance in both of the two lowest levels as indicating "limited skills,"using the term at least nine times in the executive summary.

The U.S. Department of Education issued a press release about the survey findings that 47 percent of the U.S. adult population demonstrated "low levels of literacy."[3] The large proportion of the adult population falling in the two lowest levels was widely reported in the media,[4] with some astonishment that this proportion was so large. In the Department's press release, U.S. Secretary of Education Richard Riley commented that "this report is a wake-up call to the sheer magnitude of illiteracy in this country."

Reactions to the findings have sometimes taken the form of challenging the methods used to arrive at the proportions of adults who perform in the five literacy levels. Some have looked to the value of the response probability criterion for a way to change the proportion of examinees who are reported to perform at certain levels. Stitch and Armstrong, for example, have argued that the 80 percent criterion in the adult literacy

---

[1]Irwin S. Kirsch, Ann Jungeblut, Lynn Jenkins, and Andrew Kolstad, 1993, *Adult Literacy in America: A First Look at the Results of the National Adult Literacy Survey*. Washington, DC: National Center for Education Statistics.

[2]Anne Campbell, Irwin S. Kirsch, and Andrew Kolstad, 1992, *Assessing Literacy: the Framework for the National Adult Literacy Survey*. Washington, DC: National Center for Education Statistics.

[3] "Literacy levels deficient for 90 million U.S. adults," U.S. Department of Education press release, September 8, 1993.

[4]David A. Kaplan, 1993, "Dumber than we thought: Literacy: A new study shows why we can't cope with everyday life." *Newsweek*, 122 (September 20): 44-45. Paul Gray, 1993, "Adding up the underskilled: A survey finds nearly half of U.S. adults lack the literacy to copy with modern life." *Time*, 142 (September 20): 75.

survey was too stringent and have recommended a 50 percent criterion.[5] Since the issue of adopting a response probability convention is embedded in technicalities of item response theory and have not received widespread discussion, the issues involved are not widely understood.

The purpose of this paper is to explore the role of the response probability convention in reporting results from the 1992 National Adult Literacy Survey. The following sections of this paper explain 1) the concept and measurement of literacy in this survey; 2) the sources of the prose literacy data; 3) IRT scaling methods and item characteristic curves; 4) the use of item mapping to anchor the prose literacy scale by locating specific tasks along it (using a response probability convention); 5) the literacy levels created for the 1992 National Adult Literacy Survey in order to generalize beyond specific tasks to the more abstract abilities underlying the scales; 6) the relationship of the response probability convention to the cut points between the literacy levels; and 7) the variation in the proportions of the adult population reported to be in each prose literacy level as a function of the response probability convention. The final part of the paper discusses a few implications of the findings.

Defining and Measuring Literacy

Literacy is the set of skills needed to use information contained in printed and written materials. For the past decade, the federal government here in the U.S. and several foreign governments have sponsored a consistent approach to measuring literacy skills in various populations in ways that provide results that are mostly comparable from one study to the next (although this approach is not comparable to earlier surveys).[6] According to this approach, different literacy skills are needed for different materials. Printed and written information, both verbal and quantitative, exists in the form of prose texts and documents Prose literacy skills are needed to use verbal information contained in prose texts; document literacy skills are needed to use information contained in documents, and quantitative literacy skills are needed to use quantitative information contained in prose texts or documents. While these skills share many common features and are highly correlated in the population, they are sufficiently different to require separate measurement scales. For the purpose of brevity, this paper focuses on the prose literacy scale.

---

[5]Thomas G. Sticht and William B. Armstrong, 1994, *Adult Literacy in the United States: A Compendium of Quantitative Data and Interpretive Comments*, Washington, DC: National Institute for Literacy.

[6]Reported in: Irwin S. Kirsch and Ann Jungeblut, 1986, *Literacy: Profiles of America's Young Adults*, Princeton, NJ: Educational Testing Service; Irwin S. Kirsch, Ann Jungeblut, and Anne Campbell, 1992, *Beyond the School Doors: The Literacy Needs of Job Seekers Served by the U.S. Department of Labor*, Princeton, NJ: Educational Testing Service; Irwin S. Kirsch, Ann Jungeblut, Lynn Jenkins, and Andrew Kolstad, 1993, *Adult Literacy in America: A First Look at the Results of the National Adult Literacy Survey*. Washington, DC: National Center for Education Statistics; and Albert Tuijnman, Irwin S. Kirsch, Stan Jones, and T. Scott Murray, 1995, *Literacy, Economy and Society: Results of the first International Adult Literacy Survey*, Ottawa: Statistics Canada and Paris: Organization for Economic Co-operation and Development.

The most literate adults are able to search for and locate information in prose texts while matching on one or more criteria and while simultaneously screening out plausible but incorrect information. The most literate adults can repeatedly search texts as many times as needed to match the information on all necessary criteria. The most literate adults can integrate pieces of information they have located in different part of prose texts and generate new information by writing out a combination of that information, or that information and their own prior knowledge.

The strategies needed to obtain information from printed and written prose texts can be conceptualized as a series of mental steps. First, the reader identifies the information goal in order to focus on what is needed. Second, the reader identifies any information that might already be contained in an information need or request and notices what information is obtainable from the text. Third, the reader searches the text to locate the information. Strategies for searching prose texts and documents are different, because the information is structured differently. While this may be thought of as the essential part of literacy, the preceding steps are critical to organizing the search process efficiently. Fourth, the reader identifies information that might be a candidate to fulfill the information need. Fifth, the reader verifies that the found information is sufficient, and if it is insufficient, goes back to continue the search. Finally, the reader produces a result, either verbally or in written form (by writing something in response to the request).

Data Sources

The results of this investigation are based on survey responses and assessment data from the 1992 National Adult Literacy Survey,[7] supplemented with task-specific data developed by Mosenthal and Kirsch that served as the basis of the literacy levels.[8]

Trained staff interviewed over 26,000 individuals aged 16 and older during the first eight months of 1992. The sample had three components: a national sample of 13,600 participants was randomly selected to represent the adult population in the country as a whole; twelve state samples of about 1,000 adults in states that chose to participate in a special study designed to be comparable to the national data; and a prison sample of 1,100 inmates from 80 federal and state prisons.[9]

Survey participants were asked to spend approximately an hour in their own homes responding to a series of diverse literacy tasks as well as questions about their demographic characteristics, educational background, reading practices, and other areas related to literacy. Based on their responses to the survey tasks, adults received proficiency scores along three scales which reflect varying degrees of skill in prose, document, and

---

[7]Norma Norris, et al., 1994, *National Adult Literacy Survey Public Use Data Tape: User's Guide*. Preliminary version, Washington, DC: National Center for Education Statistics.

[8]Irwin S. Kirsch, Ann Jungblut, and Peter B. Mosenthal, 1994, "Moving toward the measurement of adult literacy." Paper presented at National Center for Education Statistics conference on literacy levels, March 23.

[9]Irwin S. Kirsch, Martha Berlin, Leyla Mohammadjer, Don Rock, Kentaro Yamamoto, and others, forthcoming, *Technical Report of the 1992 National Adult Literacy Survey*. Washington, DC: National Center for Education Statistics.

quantitative literacy. Those adults who did not complete the assessment for literacy-related reasons were later assigned wrong answers to incomplete literacy tasks and scored along with the other respondents.[10]

Prose literacy skills were measured by constructing tasks that simulated everyday demands for the use of information contained in prose texts. The texts were selected so as to represent common prose materials that would not favor any particular group by being more familiar to one group than another. In order to obtain widespread familiarity, six kinds of materials were selected from the following areas: home and family, health and safety , community and citizenship, consumer economics, work, and leisure and recreation.[11] The 41 prose tasks used in this survey required one of three different strategies for successful completion: locating, integrating, and generating information. In order to locate one or more pieces of information, readers had to match the information given in the question with either all the criteria specified in the request. In order to integrate information, readers had to pull together two or more pieces of information located at different points in the text. In order to generate new information, readers had to go beyond locating or integrating by drawing on their knowledge about a subject or by making broad text-based inferences in order to produce new information. Of the total item pool, slightly over half the tasks required locating something, just under a third required generating something new, and the remainder were involved integrating different things.

The prose literacy tasks were included in a design similar to that used by NAEP and other large-scale population assessments. The 1992 National Adult Literacy Survey included 165 literacy tasks, broken into 13 sections, only 3 of which were presented to any particular respondent. Of these tasks, 41 were used to measure prose literacy, 81 were used to measure document literacy, and 43 were used to measure quantitative literacy. As a result, a typical adult in the survey responded to 11 prose literacy tasks—not enough to measure any particular adults prose literacy skills with any accuracy, but with a large sample, enough to estimate the distribution of prose literacy skills in the adult population. In order to represent the large measurement error component inherent in any data analysis based on this design, the data file represents the prose literacy scale with five plausible values rather than a single estimate. Only one of the prose literacy tasks was a multiple-choice item, while the remainder were answered with short constructed responses (scored as right or wrong).

---

[10]See "Missing responses to literacy tasks," pp. 121-130 in Karl O. Haigler, Caroline Harlow, Patricia O'Connor, and Anne Campbell, 1994. *Literacy Behind Prison Walls: Profiles of the Prison Population from the National Adult Literacy Survey.* Washington, DC: National Center for Educations Statistics and Kentaro Yamamoto, "Estimating literacy proficiencies with and without cognitive data," Chapter 8 in I.S. Kirsch, M. Berlin, L. Mohadjer, D. Rock, K.Yamamoto, and others, forthcoming, *Technical Report of the 1992 National Adult Literacy Survey.* Washington, DC: National Center for Education Statistics.

[11]See Chapter 4 in Anne Campbell, Irwin S. Kirsch, and Andrew Kolstad, 1992, *Assessing Literacy: the Framework for the National Adult Literacy Survey.* Washington, DC: National Center for Education Statistics.

The Scale of Prose Literacy

The assessment tasks in the survey were designed to measure prose literacy as a unidimensional scale. If prose literacy consisted of a number of cumulative, particular skills about how to use printed or written information contained in prose, it could be represented by a scale such as that illustrated in Figure 1. In Figure 1, the line represents the prose literacy continuum (designated θ —the Greek letter theta) and the four selected points along the line (labeled a, b, c, and d) represent the differing amounts of literacy possessed by four respondents.

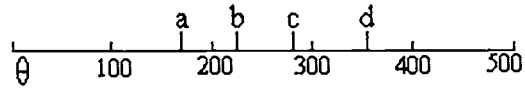


Figure 1: Prose Literacy Scale

From a person's position on the prose literacy scale, one ought to be able to predict with a good deal of accuracy the pattern of right and wrong answers to all the prose literacy tasks in the assessment. The ordering of respondents along a single dimension means that someone like person d, with more prose literacy, can do everything that persons a, b, and c can do, and more. Likewise, person c can do everything that persons a and b can do, and more.

In this and subsequent figures, the prose literacy scale has no inherent unit of measurement. The scale is assumed to have a mean of zero and unit variance, but in order to eliminate decimal points and negative numbers, has been transformed to have a mean of about 250 and a variance near 50. The zero point on the scale has no inherent meaning, and scores below zero could (rarely) occur.

If the prose literacy scale were unidimensional, success with prose tasks would also be cumulative. Once any particular skill is mastered, any task needing that skill could be performed correctly. The skills required by literacy tasks can also be represented along the same prose literacy scale. Suppose a set of three prose literacy tasks were available such that task 1 required one skill, task 2 required that skill and another, and task 3 required the skills of tasks 1 and 2, as well as a third. A large group of responses to these tasks would show a limited set of patterns. There would be a pattern with all three literacy tasks correct, a pattern with tasks 1 and 2 correct, a pattern with task 1 correct, and a pattern with no tasks correct. Given the cumulative nature of the skills involved, so other patterns should appear. Guttman gave the term "scalogram" to ideal scales that exhibit such a pattern.[12]

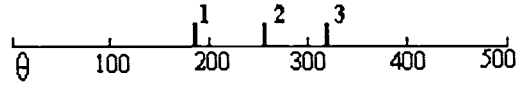


Figure 2: Prose literacy scale

A correct response to a particular prose task requires a certain degree of literacy on the part of respondents. If prose literacy had been measured with perfect tasks, then the likelihood of a correct response to any particular task would show a distinctive pattern.

[12]Louis Guttman, 1950. "The basis for scalogram analysis." Chapter 3 in S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarsfeld, S.A. Star, and J.A. Clausen, *Studies in Social Psychology in World War II: Volume IV, Measurement and Prediction*, Princeton: Princeton University Press.

All adults with the necessary degree of literacy would answer such a task correctly, while none of the adults with less than the necessary amount of literacy would succeed with such a task. Items that compose a scalogram would demonstrate the item characteristic curve of the perfect task shown in Figure 3. In Figure 3, the vertical axis represents the probability of a correct response, and the horizontal axis represents the amount of prose literacy. The step function shown here represents how the likelihood of a correct response changes abruptly as a function of increasing literacy for a perfect test item with a difficulty of 185. In this case, the probability of a correct response is zero until the required degree of literacy is reached, and then the probability of a correct response immediately jumps to one.

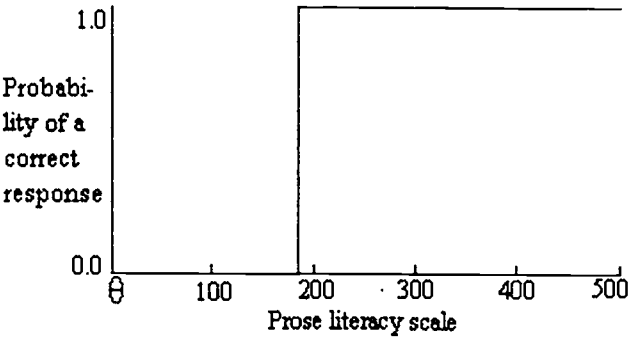


Figure 3: The Item Characteristic Curve of a Perfect Literacy Task Requiring a Score of 185

In an ideal world, the examinees who know the content tested by a question would always answer correctly, and those who do not would never answer correctly. A Guttman scalogram postulates this pattern of responses. However, in the real world, respondents and tasks do not show such ideal patterns. Test developers have been unable to create literacy tasks with item response characteristics that discriminate perfectly among respondents and take the shape of a step function.[13] Errors occur, and psychometricians developed item response theory (IRT) to model the less-than-perfect relationship between proficiency—an unobservable variable that is estimated from the responses to many test questions—and correct responses to any particular test question.[14] The essential feature of such models is that the likelihood of a correct response does not jump immediately from zero to one at some point along the proficiency scale, but rises more gradually as a function of proficiency.

The IRT models in common use today assume an underlying continuum of latent proficiency (designated θ). These models use a logistic function to relate the probability of a correct response to proficiency and to three item parameters. The fundamental equation of the three-parameter logistic model expresses the probability that a person *i* will respond correctly to an item *j* as a function of both that person's *unobservable* proficiency θ_i and of three measurable aspects of item *j* (discrimination, difficulty, and guessing):

[13]Frederic M. Lord, 1953. "The relation of test score to the trait underlying the test." Reprinted P.F. Lazarsfeld and N.W. Henry, (eds.), *Readings in Mathematical Social Science*, Cambridge, MA: Massachusetts Institute of Technology Press, 1966.

[14]Frederic M. Lord and Melvin R. Novick, 1968. *Statistical Theories of Mental Test Scores*, Reading, MA: Addison-Wesley.

$$P(x_{ij}=1 \mid \theta_i, a_j, b_j, c_j) = c_j + \frac{(1-c_j)}{(1+e^{-1.7a_j(\theta_i - b_j)})} \equiv P_j(\theta_i)$$

where

x_{ij} is the response of individual *I* to item *j*, (1 if correct and 0 if not).

θ_i is the *unobservable* literacy proficiency of individual *I*. The individual's pattern of answers to *all* test items is used to estimate this proficiency.

a_j is the discrimination parameter of item *j*, or the ratio of a change in the probability of a correct response to a change in the position along the literacy continuum. The discrimination parameter *a* measures indicates how sharply a literacy task discriminates respondents with a little more literacy from those with a little less. Items that form a perfect scalogram would require an infinite slope to obtain the vertical step function shown in Figure 3.

b_j is the difficulty parameter of item *j*, or its position along the prose literacy continuum. The difficulty parameter *b* measures how much prose literacy is needed to correctly answer item *j*.

c_j is the guessing parameter of item *j*, or its lower asymptote along the prose literacy continuum. The guessing parameter *c* measures the chance of a correct response among those with very low proficiency. With open-ended tasks, guessing is not a factor and this parameter is normally set to zero. The prose literacy scale in the 1992 National Adult Literacy Survey included only one multiple-choice tasks.

The three classes of IRT models differ in that the 1- and 2- parameter models are subsets of the general, 3-parameter model in which the guessing parameter *c* is set to zero or the slope parameter *a* is fixed to a constant value for all tasks.[15]

With a high enough discrimination (*a*) parameter, this IRT model can almost reproduce the step-function item characteristic curve (ICC) of a perfect task. Test developers try to create tasks that vary in their difficulty and have the highest possible discrimination parameters, but they do not achieve this kind of perfection. Figure 4 displays three hypothetical item characteristic curves representing literacy tasks that deviate more and more from the step function ICC of the perfect task shown in Figure 3. The this figure, the ICC of the literacy task on the left was generated by an IRT function with a discrimination parameter set high enough to approximate the item characteristic curve of a perfect task. The ICC in the middle was generated by an IRT function with a

[15]Deborah Harris, 1989. "Comparison of 1-, 2-, and 3-parameter IRT models." *Educational Measurement: Issues and Practice*. 8(Spring): 35-41.

discrimination parameter set to three times the highest actual value that was observed among all the literacy tasks used in the 1992 National Adult Literacy Survey. The ICC of the hypothetical literacy task on the right was generated by an IRT function with a discrimination parameter *a* set equal to the highest value that actually occurred in that survey. With the hypothetical task on the right, anyone with a prose literacy score in the range between 250 and 375 has a probability of success somewhere between zero and one. As literacy increases between 250 and 375, the probability of success increases rapidly, but it remains an intermediate value between zero and one. In this range, success or failure with such a task is not a certainty. Real item characteristic curves are not the perfect step functions hypothesized for a Guttman scalogram.

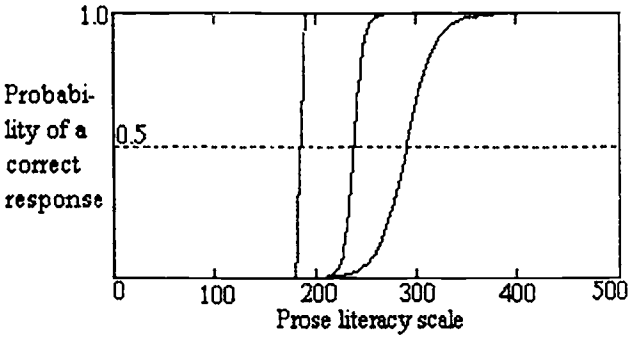


Figure 4: Three hypothetical tasks of varying discrimination (and difficulty)

Figure 4 showed a hypothetical task with a discrimination parameter equal to the best in the survey, but the average task did not discriminate that well. Figure 5 below portrays a hypothetical item characteristic curve for a task with the difficulty and discriminating power of an average task used in the 1992 National Adult Literacy Survey. Figure 5 also

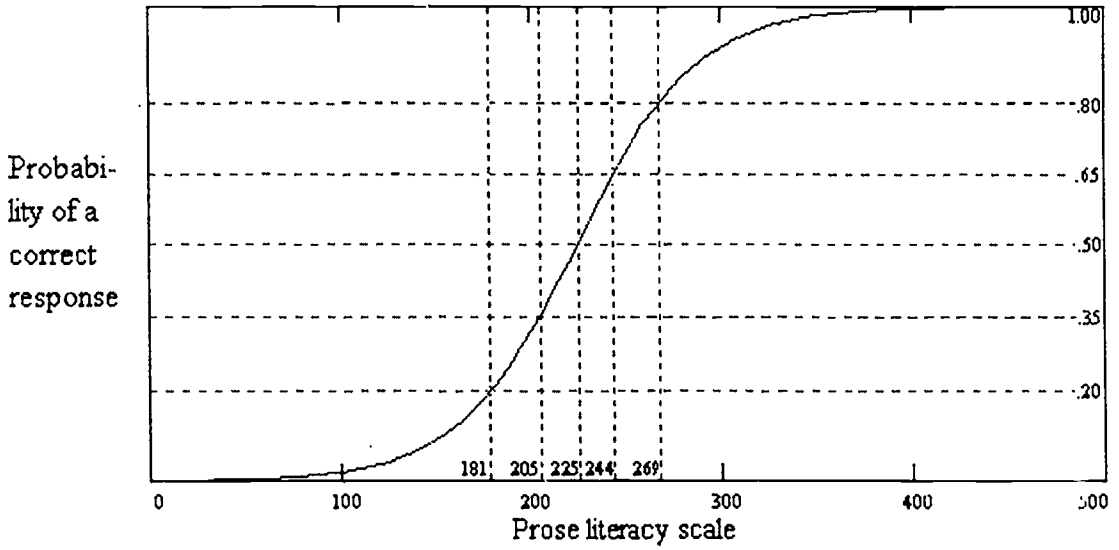


Figure 5: Item Characteristic Curve of Hypothetical Prose Literacy Task with Average Discrimination (*a*) and Difficulty (*b*)

displays horizontal guidelines for response probabilities equal to .20, .35, .50, .65, and .80 and identifies the corresponding points on the prose literacy scale where the proficiency is sufficient to succeed on this hypothetical task with those selected chances of success. These intersections identify the proficiency needed to perform at these response

probability criteria. For a prose literacy task like this hypothetical average task, a proficiency score of at least 205 would be needed to have at least some chance of being successful, or not to be generally unsuccessful (RP35). A proficiency score of at least 225 would be needed to be more likely to be successful than not (RP50). A proficiency score of at least 244 would be needed to be generally successful (RP65). And a proficiency score of at least 269 would be needed to be consistently successful (RP80). Proficiency scores below 269 do not indicate consistently unsuccessful performance—only scores below 181 (the value associated with RP20) would indicate consistent failure with this literacy task. There is a wide range of scores in the middle in which adults are sometimes accurate in their answers and sometimes not.

The item characteristic curves of three-quarters of the prose literacy tasks in the National Adult Literacy Survey are displayed in Figure 6. (Those left out would have been near the middle in difficulty and were omitted to keep the figure legible.) The item characteristic curves are not parallel. They are spread along the horizontal axis by their differences in difficulty. The difficulty of any task can be measured by the *b* parameter in

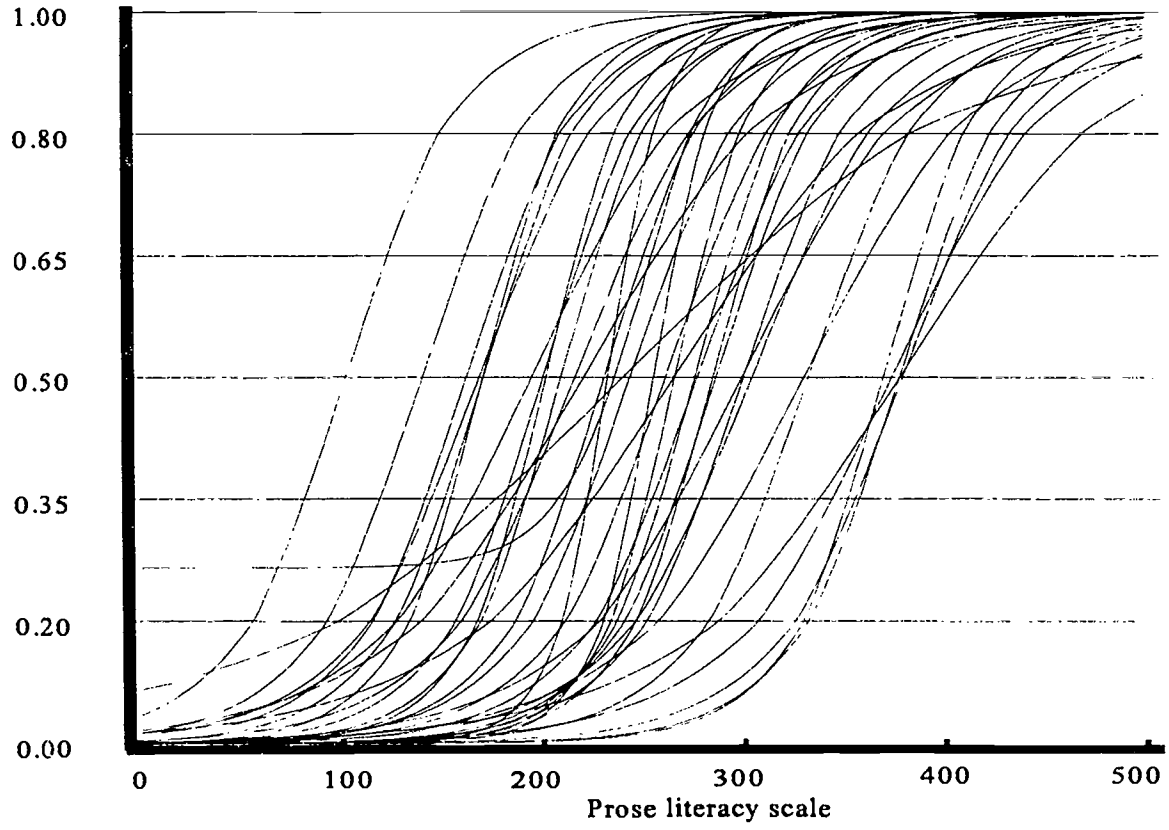


Figure 6: Item characteristic curves of 34 prose literacy tasks

IRT model, the point at which the IRT curve changes inflection. When there is no guessing, this point occurs at a response probability of 0.5. Taking a horizontal section of the curves at any other response probability level would produce a fairly similar ordering

of tasks in terms of their difficulty, but some tasks would become relatively more or less difficult, depending on the response probability criterion chosen.

The Need for a Response Probability Convention

If the slope of the item characteristic curve were vertical, it would be easy to interpret points along the prose literacy scale in terms of the tasks that people at or above that point are able to do. As proficiency improves respondents, respondents can do more and more tasks requiring more and more literacy skills. However, because the assessment tasks do not discriminate perfectly, more proficient respondents are also increasingly successful with any particular literacy task, and with the tasks that currently exist, the improvement is gradual.

In order to summarize what respondents of a particular prose proficiency can do, it is convenient to adopt a convention for a sufficient response probability that is stringent enough to ensure that people at the lower bound can do what the task requires most of the time. While 100 percent success with any particular task might be ideal, some respondents with considerably less proficiency can still be correct with a prose literacy task most of the time. The issue of selecting a particular value for a desired response probability only arises because the slope of the item characteristic curve is less than vertical. If it were vertical, there would be no need to select such a criterion, because the skill differe ce between success and failure would be very small. A conventional criterion for success with literacy tasks needs to be adopted because there are substantial differences in proficiency associated with different response probabilities.

In the early 1980s, the National Assessment of Educational Progress (NAEP), which assesses school-aged children, developed a method of scale anchoring in order to provide descriptions of the kinds of things students know and can do at selected ranges along the NAEP proficiency scales.[16] Anchoring is a way to describe generally those particular assessment items at selected points along the proficiency scales for which students can succeed at least a certain percentage of the time, and for which those at the next lowest point are less successful. This procedure relies on a response probability convention. An obvious choice for the response probability convention was 50 percent. If the convention were set here, those above the boundary would be more likely to get an item right than get it wrong, while those below that boundary would be more likely to get the item wrong than right. However, this convention was rejected on the grounds that having a 50/50 chance of getting the item right showed an insufficient degree of mastery. Instead, a response probability criterion of 80 percent (RP80) was chosen, in order to ensure that students above this criterion would have a high probability of success with any assessment item.

The 1985 young adult literacy assessment was conducted as a part of NAEP, and included one of the 1984 NAEP reading assessment blocks. In order to anchor the literacy

[16]Albert E. Beaton, 1987, "Anchoring scale points." Section 10.5.2, pp. 385-390 in A.E. Beaton and others, *Implementing the New Design: The NAEP 1983-84 Technical Report*, Princeton, NJ: Educational Testing Service.

scales, the ETS analysts carried over the NAEP RP80 criterion for its reporting.[17] The RP80 criterion was continued in the 1992 National Adult Literacy Survey, in order that the adult literacy findings remain comparable with the findings of the 1985 Young Adult Literacy Assessment.

However, during the intervening years between 1985 and 1992, NAEP changed its response probability convention used in anchoring the NAEP scales from 80 percent to 65 percent. Eugene Johnson, the NAEP technical director, described the reasons that NAEP adopted the RP65 convention in an internal ETS memo:

> While the RP percentage of 65 is arbitrary, it was selected after careful consideration of the purpose: describing students' level of performance. A larger RP percentage, such as 80, would result in higher item mapping points for all items. The result would be that smaller percentages of student would exhibit performance consistent with each exercise. For example, in the 1992 writing assessment, using a RP percentage of 65 resulted in most writing tasks having the highest score category being mapped onto the scale well above the proficiency levels exhibited by the vast majority of the assessed population of students. If an RP percentage of 80 had been used, this would likely have been true for both of the two highest score categories. In contrast, a smaller RP percentage, such as 50, would lower the mapping criteria to only a 50/50 chance that students at the scale point could provide the responses of the quality described on the map. The RP vale of 65 was selected as an intermediate value to describe students' level of performance since it corresponded to a reasonably high probability of success on the questions while better matching the observed performance of the assessed population.[18]

Johnson also pointed out in his memo that the public needs to be informed about the criterion level and to understand that the skills ascribed to students are predicated on the degree of success selected.

Over the past two years, NAEP Design and Analysis Committee that advises on technical matters has reconsidered the appropriateness of NAEP's response probability convention. NAEP recently adopted two related response probability conventions: 74 percent for multiple-choice questions (to correct for the possibility of answering correctly by guessing) and 65 percent for constructed response questions (where guessing is not a factor). Some support for the dual conventions was provided by Huynh in a paper written for NAEP's Design and Analysis Committee.[19] Huynh decomposed the item information into that provided by a correct response $[P(\theta)*I(\theta)]$ and that provided by an incorrect response $[(1-P(\theta))*I(\theta)]$. Huynh showed that the item information provided by a correct

---

[17]Irwin S. Kirsch, Ann Jungeblut, and others, 1986, "Describing and anchoring the scales" and "Levels of proficiency." Pp. III-9/III-10 and IV-11/IV-13 in *Final Report: Literacy: Profiles of America's Young Adults*, Princeton, NJ: Educational Testing Service.

[18]Eugene Johnson, 1994, "Description of percentages for anchoring and item mapping." Unpublished internal ETS memo, February 4.

[19]Huynh Huynh, 1995, "On score locations of binary and partial credit items and their applications to scale anchoring or criterion-referenced interpretation," unpublished manuscript.

response to a constructed-response item is maximized at the point along the scale at which two-thirds of the students get the question correct (for multiple-choice questions, information is maximized at the point at which 74 percent get the question correct). Correspondingly, the item information provided by an incorrect response is maximized near the point along the scale at which one-third of the students get the question wrong. It should be noted, however, that maximizing the *item* information $I(\theta)$, rather than the information provided by a *correct response* $[P(\theta)*I(\theta)]$, would imply an item mapping criterion closer to 50 percent.

While Huynh's analyses were influential, NAEP's dual response probability conventions (65 and 74 percent) were based, in part, on an intuitive judgment that they would provide the best picture of reading skills for students at particular points on the reading scales.

An important use of the response probability convention is in item mapping and scale anchoring. The first report from the 1992 National Adult Literacy Survey presented an item map for each scale, reproduced here for prose literacy as Figure 7, but provided little interpretation of the map. The report said only that this figure "describes some of the literacy tasks and indicates their scale values." The response probability convention was not mentioned.

NAEP's most recent reading report card explained the meaning of its item map (one corresponding to Figure 7) in the following terms: "Each reading question was mapped onto the NAEP literacy subscale based on students' performance. The point on the subscale at which a question is positioned on the map represents the subscale score attained by students who had a 65 percent probability of successfully answering the question. Thus it can be said for each question and its corresponding subscale score—student with proficiency scores above that point on the subscale have a greater than 65 percent chance of successfully answering the question, while those below that point have a less than 65 percent chance. (The probability was set at 74 percent for multiple-choice items.)"[20] The same kind of explanation would also be

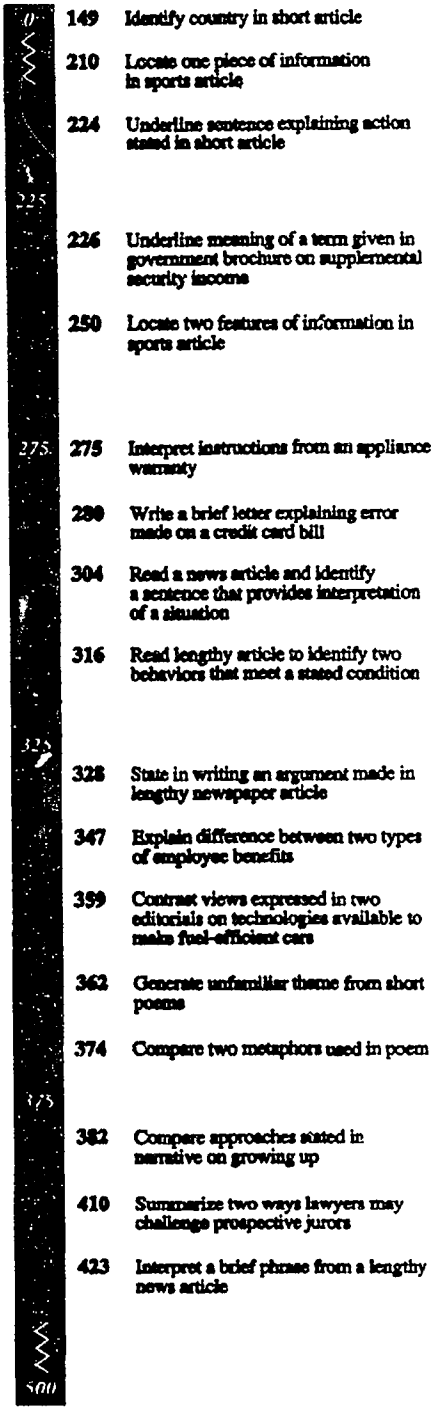


Figure 7: Difficulty values of selected tasks along the prose literacy scale

[20]Figure 6.4 in Jay R. Campbell, Patricia L. Donahue, Clyde M. Reese, and Gary W. Phillips, 1996, *NAEP 1994 Reading Report Card for the Nation and the States: Findings from the National Assessment*

applicable to the prose literacy scale, except that the probability was 80 percent rather than 65 percent.

There is no obvious choice of a point along the probability scale that is clearly superior to any other point. The government's various programs for assessing the skills of children and of adults have set or changed their response probability conventions for reasons unique to the needs of each study with no attempt to maintain a common standard. The result for the 1992 National Adult Literacy Survey is that the reporting standard for what adults can do is now higher than that used to report on what children can do.

The response probability convention plays a significant role in deciding how much ability is needed to qualify as "able to do" some prose literacy task. It is not widely understood how this little-noticed convention fundamentally affects the measurement of the proportions of adults that meet the requirements of the various literacy levels. The next section describes how the response probability convention plays a role in the descriptions of levels of prose literacy used in literacy assessment surveys.

Literacy Tasks and Literacy Levels

The prose literacy tasks in the 1992 National Adult Literacy Survey were developed in order to simulate the everyday literacy activities that people engage in when they use printed materials, and to require of adults the same literacy skills that adults normally encounter in occupational, community, and home settings. Each literacy task consisted of two parts: a selection of printed material, and a request to do something that indicated the adult could use the information contained in that material. The degree of literacy needed to successfully complete the assessment tasks derives from three factors: the format of the printed material, the content of the material, and the information request requiring use of the material. The more difficult the literacy task, the greater the degree of literacy skill needed to successfully complete it. Analyzing the sources of the difficulty of literacy tasks helps to understand the nature of literacy skills.

Beginning with the literacy assessment of job-seekers,[21] Kirsch and Mosenthal developed a system for classifying prose literacy tasks into one of five levels. The literacy levels provide general descriptions of the kinds of skills needed to score in selected ranges along the literacy scales.[22] Mosenthal and Kirsch relied on empirical research that predicted much of the aggregate variability in task difficulty and selected cut points between levels based on their observation of qualitative shifts in the skill or process requirements associated with increasing task difficulty (measured at RP80)..

---

of Educational Progress and Trial State Assessment, Washington, DC: National Center for Education Statistics.

[21]Irwin S. Kirsch, Ann Jungblut, and Anne Campbell, 1992, *Beyond the School Doors: The Literacy Needs of Job Seekers Served by the U.S. Department of Labor*, Princeton, NJ: Educational Testing Service.

[22]Irwin S. Kirsch, Ann Jungeblut, Lynn Jenkins, and Andrew Kolstad, 1994, "Interpreting the literacy scales," Appendix A in K.O. Haigler, C.W. Harlow, P.E. O'Connor, and Anne Campbell, *Literacy Beyond Prison Walls: Profiles of the Prison Population from the National Adult Literacy Survey*. Washington, DC: National Center for Education Statistics.

Factors That Predict Difficulty of Prose Literacy Tasks. According to Kirsch and Mosenthal, the literacy tasks in the 1992 National Adult Literacy Survey varied in the information-processing demands that they placed on adults.[23] Several task-related factors capture these demands: type of match, plausibility of distractors, abstractness of information, and readability of the prose text.

Type of match. An information need often requires readers to relate information in a prose text to a purpose and to select a best-fitting response from a range of response options. Literacy tasks are easiest when matching information to purpose is straightforward.[24] Matching can be more or less complex depending on several factors. The basic rule for scoring type of match is shown in Table 1, but several additional rules, too complex to summarize here, can add points for particular features. Given the basic score

Table 1.--Basic scoring rules for type of match: Nature of the task

Rule	Score
When the task is to *locate* the information in the prose text that corresponds to the features requested.	1
When the task is to *cycle* (that is, perform an iterative series of locate matches) to find the information that corresponds to the features requested.	2
• for prose texts, add 1 point if the answer is located in more than one paragraph.	
When the task is to *integrate* information located in a prose text by comparing, or when the task is to infer a condition based on a synthesis of features found in the same paragraph of text.	3
When the task is to integrate information located in a prose text by contrasting, or when the task is to infer a condition based on a synthesis of features found in more than one paragraph of text.	4
When the task is to *generate* new information (that is, to use prior knowledge to match information requested with that in the prose text).	5

based on the nature of the task, additional points can be added based on the number of phrases or features in the directions, the number of responses requested, the kind of inference needed to answer the question, and how the reader must "complete an information frame."[25] These scoring rules are additive. A prose literacy task, for example, might have a basic score of 2 because it is a cycle task, but have additional points added because the cycling occurs between paragraphs (add 1), involves a two-clause question (add 1), for which the answer should consist of two responses (add 1), but whose actual

[23]Irwin S. Kirsch, Ann Jungblut, and Peter B. Mosenthal, 1994, "Moving toward the measurement of adult literacy." Unpublished paper presented at National Center for Education Statistics conference on literacy levels, March 23.

[24]Peter B. Mosenthal and Irwin S. Kirsch, 1991, "Toward an explanatory model of document literacy." *Discourse Processes*, 14(): 147-180.

[25]Andrew Kolstad, 1996, "Sources of the Difficulty of Literacy Tasks in the 1992 National Adult Literacy Survey," unpublished manuscript.

number is not explicitly specified (add 1). A prose assessment task with these features would have a total type-of-match score of 6.

Plausibility of distractors. An information need often requires readers to screen out irrelevant information that at first may appear to be related to the problem at hand. Such information is referred to as a plausible distractor. Literacy tasks are easiest when there are no plausible distractors in a document.[26] When plausible distractors are closer to the correct answer or share more features with the correct answer, tasks become harder. Defining how close the distractor is to the correct answer involves the decision rules shown in Table 2.

Table 2.–Scoring rules for plausibility of distractors

Rules for prose texts	Score
When no information related to the conditions requested appears, other than the answer (no plausible distractors)	1
When information similar to either given or requested information appear somewhere in the text but not near the answer, or inferences invited by information in the paragraph containing the answer bear a resemblance to the answer	2
When distractors for both given and requested information appear in different paragraphs, though one could occur in the paragraph containing the answer	3
When distractors for both given and requested information, or when plausible distractors represent the opposite condition of what is requested, appear in the same paragraph, but one other than the paragraph containing the answer	4
When distractors for both given and requested information, or when plausible distractors represent the opposite condition of what is requested, appear in the same paragraph as the answer	5

Type of information. An information need often requires readers to locate or identify information contained in a document. Information in documents varies along a continuum of concreteness from very concrete to very abstract; more abstract information is more difficult for readers to use.[27] Scoring the concreteness of the information requested involves the decision rules shown in Table 3.

[26]Irwin S. Kirsch and Peter B. Mosenthal, 1990, "Exploring document literacy: Variables underlying the performance of young adults." *Reading Research Quarterly*, 25 (Winter): 5-30.

[27]Peter B. Mosenthal and I.S. Kirsch, 1991, "Information types in nonmimetic documents: A review of Biddle's wipe-clean slate," *Journal of Reading*, 34 (May): 654-660.

Table 3.--Scoring rules for type of information	
Rule	Score
When the information requested refers to a person, animal, place, or thing (most concrete)	1
When the information requested refers to an amount, time, attribute, type, action, or location	2
When the information requested refers to a manner, goal, purpose, alternative, attempt, condition, pronominal reference, or predicate adjective	3
When the information requested refers to a cause, effect, reason, result, evidence, similarity, or explanation	4
When the information requested refers to an equivalence, difference, theme, or pattern (most abstract)	5

Structural complexity. Prose texts vary in the length of sentences, number of syllables in the words used, and the complexity of the syntax. Literacy tasks may be easier to process when the structure of the text containing the needed information is less complex. The measurement of the complexity of prose texts used came from Fry's research on readability.[28] Readability of prose is based on the average number of syllables per 100 words and the average number of sentences per 100 words. These two continuous variables are then used as coordinates in Fry's readability grade level graph, which portrays a nonlinear the relationship between the two and the resulting readability level. In general, however, the more syllables per word and the more words per sentence, the higher the associated grade level of the text. The grade levels of the texts used in the National Adult Literacy Survey ranged from fourth to fifteenth grade.

The 41 prose literacy tasks in the 1992 National Adult Literacy Survey were scored according to the rules described briefly above, and the resulting distribution of scores for each factor are shown below in Tables 4 and 5. For the prose literacy tasks in the 1992 National Adult Literacy Survey, the most frequent scores on type of match were 3, 4, and 6, with only six tasks in the two easiest categories. The most frequent score on plausibility of distractors was a 2, with eight tasks having no distractors at all (score of 1). The most frequent score on abstractness of information was a 4, with six tasks in the most concrete category (score of 1). The texts used ranged widely in readability, with 4 tasks at grade levels 4 and 5, and 5 tasks at a grade level beyond high school (grade 12).

[28]Edward B. Fry, 1975, "The readability principle." *Language Arts*, 52 (September): 847-851; 1977, "Fry's readability graph: Clarifications, validity, and extension to level 17." *Journal of Reading*, 21(December):242-252; 1977, *Elementary Reading Instruction*, New York: McGraw-Hill, and 1981, "A partial reading model utilizing language unit size by frequency." Pp. 103-107 in M.L. Kamil (ed) *Directions in Reading: Research and Instruction*. Washington, DC: National Reading Conference.

Table 4.–Distribution of predictor scores for 41 prose literacy tasks

Score	Type of match	Plausibility of distractors	Abstractness of information
	Number of tasks		
1	5	8	6
2	1	18	7
3	11	3	9
4	10	8	14
5	5	4	5
6	9	0	0

Table 5.–Distribution of readability grade levels for 41 prose literacy tasks

Grade level	Readability of text
	Number of tasks
4	3
5	1
6	8
7	6
8	8
9	5
10	5
13	2
15	3

Predictive Factors and RP80 Task Difficulty. Kirsch and Mosenthal used multiple regression to predict the difficulty of the prose literacy tasks in the 1992 National Adult Literacy Survey based on the four variables described above: type of match, plausibility of distractors, abstractness of information, and readability of the prose text. Using the prose scale scores evaluated at an 80 percent response probability convention, Kirsch and Mosenthal obtained the following estimates of the regression coefficients ($R^2 = .87$): [29]

RP80 = 28.9 TypMatch + 16.1 Distract + 8.8 Abstract + .2 Readability + Constant
(3.4) std.err. (3.6) std.err. (4.2) std.err. (1.7) std.err.

This equation showed that while prose task difficulty was highly predictable by these four factors. 'Type of match' had a large impact (more than 8 times its standard error); 'plausibility of distractors' had an significant impact (more than 4 times its standard error); and 'abstractness of information' also had a significant impact (more than 2 times its standard error). Readability of the text was not an important factor in explaining task difficulty, after controlling for the other predictors.

Based on the relationships between RP80 task difficulty and their descriptions of the information-processing demands of the tasks as scored by the four variables, Kirsch and Mosenthal decided on where to set cut points between each of five levels on the three scales. These cut points depended on their analyses of the factors that predict RP80 task difficulty. The first report of the 1992 National Adult Literacy Survey described how these cut points were chosen as follows:

[29]Table 11 in Irwin S. Kirsch, Ann Jungblut, and Peter B. Mosenthal, 1994, "Moving toward the measurement of adult literacy." Unpublished paper presented at National Center for Education Statistics conference on literacy levels, March 23.

> Analyses of the interactions between the materials read and the tasks based on these materials reveal that an ordered set of information-processing skills appears to be called into play to perform the range of tasks along each scale.
>
> To capture this ordering, each scale was divided into five levels that reflect the progression of information-processing skills and strategies: Level 1 (0-225), Level 2 (226-275), Level 3 (276-325), Level 4 (326-375), and Level 5 (376 to 500). These levels were determined not as a result of any statistical property of the scales, but rather as a result of shifts in the skills and strategies required to succeed on various tasks along the scales, from simple to complex.[30]

The three factors that predict RP80 task difficulty came into these decisions, as Kirsch, Jungblut, and Mosenthal explained in a subsequent paper:

> ...there appears to be an ordered set of information-processing skills and strategies that may get called into play to accomplish the range of tasks represented by the three literacy domains.
>
> ... As tasks moved up the scales (i.e., became more difficult), the associated [scores on the three factors] also increased. This relationship between [RP80] task difficulty and [scores on the three factors] appeared to be quite systematic. That is, toward the bottom of each literacy scale the [score on the three factors] of 1 was dominant, [scores] of 2 and 3 became more frequent as tasks move up the Prose, Document, and Quantitative Scales, and toward the higher end [scores on the three factors] of 4, 5 and higher became predominant. Although the patterns differed somewhat from scale to scale reflecting differences in the [scores on the three factors], the points on the scale at which major shifts in the processes and skills required for successful task performance were remarkably similar.
>
> Visual inspection of the distribution of [scores on the three factors] along each of the literacy scales revealed several major points occurring at roughly 50 point intervals beginning with 225 on each scale. As with all systems, this one contains some 'noise' and does not account for all of the score variance associated with performance on the three scales. However, using the scales to assign a range from 277 to 319—as would be descriptive of tasks on the Document Scale—or from 331 to 370—reflecting a particular set of processing demands on the Quantitative Scale—implies a precision of measurement that is simply inappropriate.[31]

Once the cut points between the levels were decided, Kirsch and Mosenthal wrote general descriptions of the kinds of demands placed on readers by tasks in each of their five levels. As it came to be used in different surveys, their descriptions evolved over time, as shown in Table 6 on the next page. The description in the middle column was used for

---

[30]Page 73 in Irwin S. Kirsch, Ann Jungeblut, Lynn Jenkins, and Andrew Kolstad, 1993, *Adult Literacy in America: A First Look at the Results of the National Adult Literacy Survey*. Washington, DC: National Center for Education Statistics

[31]Page 33 in Irwin S. Kirsch, Ann Jungblut, and Peter B. Mosenthal, 1994, "Moving toward the measurement of adult literacy." Unpublished paper presented at National Center for Education Statistics conference on literacy levels, March 23.

Table 6.–Evolution of prose literacy levels descriptions

Level	1991 Job-Seekers report	1993 Adult & Intl reports	1995 Profile Approach paper (unpublished)
One	Tasks at this level require a reader to locate a piece of information in which there is a literal match between the question and the stimulus material. If a distractor or plausible answer appears in the stimulus material, it tends to be located away from where the correct information is found.	Most of the tasks in this level require the reader to read relatively short text to locate a single piece of information which is identical to or synonymous with the information given in the question or directive. If plausible but incorrect information is present in the text, it tends not to be located near the correct information.	To complete these tasks, readers must process relatively short text to locate a single piece of information which is identical to, or synonymous with the information given in the question or directive. If distractors appear in the text, they tend to be located in paragraph other than the one in which the correct answer occurs. Most of the tasks in this level require readers to identify information which is quite concrete, including a person, place, or thing, as well as an attribute, amount, type, temporal, action, procedure, or location.
Two	Some of these tasks still require the reader to locate and match on a single literal feature of information; however, these tasks tend to occur in materials in which there are several distractors or where the match is based on synonymous or text-based inferences. These tasks also begin to require readers to integrate information by either pulling together two pieces of information or by comparing and contrasting information.	Some tasks in this level require readers to locate a single piece of information in the text; however, several distractors or plausible but incorrect pieces of information may be present, or low-level inferences may be required. Other tasks require the reader to integrate two or more pieces of information or to compare and contrast easily identifiable information based on a criterion provided in the question or directive.	Tasks at level 2 often require reader to make a low-level inference, or identify a condition or an antecedent in order to identify requested information in a text. Tasks at this level tend to have a distractor for either given or new information present, but not in the same paragraph as the answer. Many tasks in level 2 ask readers to complete information which is fairly concrete. However, in level 2, we find some tasks which also require readers to identify information representing manner, goal, purpose, attempt, alternative, and condition information.
Three	Prose tasks at this level tend to require the reader to search fairly dense text for literal or synonymous matches on the basis of more than one feature of information or to integrate information from relatively long text that does not contain organizational aids such as headings.	Tasks in this level tend to require readers to make literal or synonymous matches between the text and the information given in the task, or to make matches that require low-level inferences. Other tasks ask readers to integrate information from dense or lengthy text that contains no organizational aids such as headings. Readers may be asked to generate a response based on information that can be easily identified in the text. Distracting information is present, but is not located near the correct information.	Level 3 tasks again require readers to make literal, synonymous, and low-level inference matches between the question or directive and the text. Unlike Level 1 and 2 locate tasks, Level 3 tasks usually require readers to identify and list multiple responses (the number of which is specified in the question or directive). Also the questions and directives of Level 3 tasks tend to consist of several phrases. Moreover, these tasks generally require readers to complete requested information by identifying special conditional information stated in a question or directive or by establishing an antecedence between a pronoun and its reference. Distracting information for both given and requested information tends to be present, both of which appear in different paragraphs from one another and neither of which appear in the same paragraph as the answer. Tasks in this Level tend to require readers to identify condition information. In other instances, tasks require readers to identify a reason or explanation.
Four	Not only are multiple-feature matching and integration of information from complex materials maintained, the degree of inferencing required by the reader is also increased. Tasks at this level include conditional information that must be taken into account by the reader in order to integrate or match information appropriately.	These tasks require readers to perform multiple-feature matches and to integrate or synthesize information from complex or lengthy passages. More complex inferences are needed to perform successfully. Conditional information is frequently present in tasks at this level and must be taken into consideration by the reader.	Level 4 tasks generally require readers not only to locate, but also to cycle and integrate. Again, multiple responses may be required but for which the number of responses is not specified. Level 4 tasks often require readers to complete requested information by identifying special conditional information stated in a question or directive, or by establishing an antecedence between a pronoun and its reference. In other cases, high text-based inferences must be made to distinguish the correct requested information from distracting information. Distracting information for both given and requested information tends to be present, both of which may appear in the same paragraph as the answer. Tasks in this Level tend to require readers to identify rather abstract information, including reason, causation, result, comparison, and contrast.
Five	These tasks require the reader to search for information in dense text or complex documents containing multiple plausible distractors, to make high text -based inferences or use specialized background knowledge, as well as to compare and contrast sometimes complex information to determine differences.	Some tasks in this level require the reader to search for information in dense text which contains a number of plausible distractors. Others ask readers to make high-level inferences or use specialized background knowledge. Some tasks ask readers to contrast complex information.	Level 5 tasks often require readers not only to locate, cycle and integrate, but also to generate. Generating may involve the use of specialized background knowledge to interpret a phrase or to synthesize text information. Distracting information for both given and requested information may be present, both of which frequently appear in the same paragraph as the answer. Tasks in this Level tend to require readers to identify quite abstract information, including contrast, equivalence, and theme or summary.

both the 1992 National Adult Literacy Survey in the U.S. and for the 1995 International Adult Literacy Survey, conducted in seven countries. Even though the literacy tasks used in the two surveys were different, they measured the same scale, and their difficulty could be predicted by the same factors. The descriptions attempt to capture the various combinations of the three important predictors of difficulty among typical tasks at each of the five literacy levels.

There are some minor problems with these descriptions. No description of the 'abstractness of information' variable was included in the most well-known version of the level descriptions (the middle column), even though it had a significant impact on task difficulty. The descriptions in the third column (from an unpublished paper by Kirsch and Mosenthal) correct this oversight and include a statement about the type of information typically found in tasks in each level. In addition, the description of Level 1 includes an unwarranted term—"relatively short text"—that describes the readability of the prose stimulus, a factor that their regression analysis showed was not essential to item difficulty when the other factors were included.

Predictive Factors, Task Difficulty, and the Response Probability Convention. Kirsch and Mosenthal conducted all their analyses using task difficulty as measured at the RP80 response probability convention. It is also possible to conduct analyses using alternative response probability conventions as outcomes, especially since there is no universally accepted standard for response probabilities and the choice of a response probability convention is somewhat arbitrary. The regressions will not be identical because the item characteristic curves of the prose literacy tasks are not parallel; they all have different discrimination (*a*) parameters. The differences in the multiple regression coefficients captures the pattern in these variations. Table 7 below shows estimates of the multiple regression coefficients using task difficulty measured at fifteen alternative response probability conventions from 20 to 90 percent. For these regressions, the number of cases was expanded to 71 by including 30 additional prose literacy tasks that had been used in the 1991 study of the literacy of job-seekers.[32]

The coefficients in Table 7 display several patterns that could not be seen in a single regression with RP80 as an outcome. The coefficient of explained variance increases as the response probability falls from RP90 to RP60, then decreases with lower response probabilities. At high response probability levels, the importance of the 'plausibility of distractors' factor is greatest and readability is not a significant factor. However, the importance of these two factors reverses at low response probability levels. At RP35 and below, the coefficient of 'plausibility of distractors' goes below twice its standard error and becomes insignificant. At RP55 and below, the coefficient of readability becomes a significant factor in explaining task difficulty.

[32]Irwin S. Kirsch, Ann Jungblut, and Anne Campbell, *Beyond the School Doors: The Literacy Needs of Job Seekers Served by the U.S. Department of Labor*, (Princeton, NJ: Educational Testing Service, 1992).

Table 7.– Multiple regression coefficients, standard errors, and R-squares for regression equations predicting task difficulty measured at selected response probability criteria for 71 prose literacy tasks from the 1992 National Adult Literacy Survey and the 1991 study of the literacy of job-seekers.

Response probability criterion	Inter-cept	Type of match (std. err.)		Plausibility of distractors (std. err.)		Abstractness of information (std. err.)		Readability of prose text (std. err.)		R^2
0.90	157.2	21.8	(3.4)	21.2	(3.7)	12.4	(4.1)	0.2	(1.9)	0.765
0.85	144.2	22.1	(3.1)	18.6	(3.3)	11.9	(3.7)	0.8	(1.7)	0.794
0.80	134.3	22.3	(2.8)	16.5	(3.0)	11.5	(3.4)	1.2	(1.6)	0.813
0.75	126.2	22.5	(2.6)	14.9	(2.8)	11.1	(3.2)	1.6	(1.5)	0.826
0.70	119.1	22.6	(2.5)	13.4	(2.7)	10.8	(3.0)	1.9	(1.4)	0.835
0.65	112.7	22.8	(2.4)	12.1	(2.6)	10.6	(2.9)	2.2	(1.4)	0.839
0.60	106.7	22.9	(2.4)	10.8	(2.6)	10.3	(2.9)	2.5	(1.4)	0.841
0.55	100.9	23.1	(2.4)	9.6	(2.6)	10.0	(2.8)	2.8	(1.3)	0.840
0.50	95.2	23.2	(2.4)	8.5	(2.6)	9.8	(2.8)	3.0	(1.3)	0.837
0.45	89.6	23.4	(2.4)	7.3	(2.6)	9.5	(2.9)	3.3	(1.4)	0.830
0.40	83.7	23.6	(2.5)	6.1	(2.6)	9.2	(2.9)	3.6	(1.4)	0.821
0.35	77.6	23.8	(2.5)	4.8	(2.7)	8.9	(3.0)	3.8	(1.4)	0.809
0.30	70.5	24.1	(2.7)	3.4	(2.9)	8.5	(3.2)	4.2	(1.5)	0.792
0.25	65.4	24.3	(2.8)	1.8	(3.0)	8.0	(3.4)	4.3	(1.6)	0.767
0.20	64.5	25.3	(3.1)	-0.5	(3.4)	6.8	(3.8)	3.9	(1.8)	0.716

The prose literacy levels were based on a clustering of prose tasks with similar task difficulty and similar factors that predict difficulty. It would be possible to cluster the same tasks together in the five levels, provided the cut points between the levels were adjusted to divide the levels between roughly the same tasks. Since the tasks have different discrimination (*a*) parameters, it is not possible to divide between exactly the same tasks, but it is possible to use the regression equations to derive alternative cut points between levels that would approximate the same grouping of prose literacy tasks in each level as occurred for the RP80 literacy levels.

Alternative Cut Points between Literacy Levels

The method for deriving alternative cut points between levels depended on developing hypothetical tasks with difficulty-related characteristics like those in the existing levels. Table 8 displays the values of a selected group of hypothetical tasks—about half a dozen for each level. These tasks were selected so that when used to predict task difficulty at RP80 using the appropriate equation in Table 7, their average prediction comes to exactly the midpoint RP80 task difficulty of the existing literacy levels: 200, 250, 300, 350, and 400.

Table 8.– Scores of hypothetical prose literacy tasks on four factors that predict difficulty.

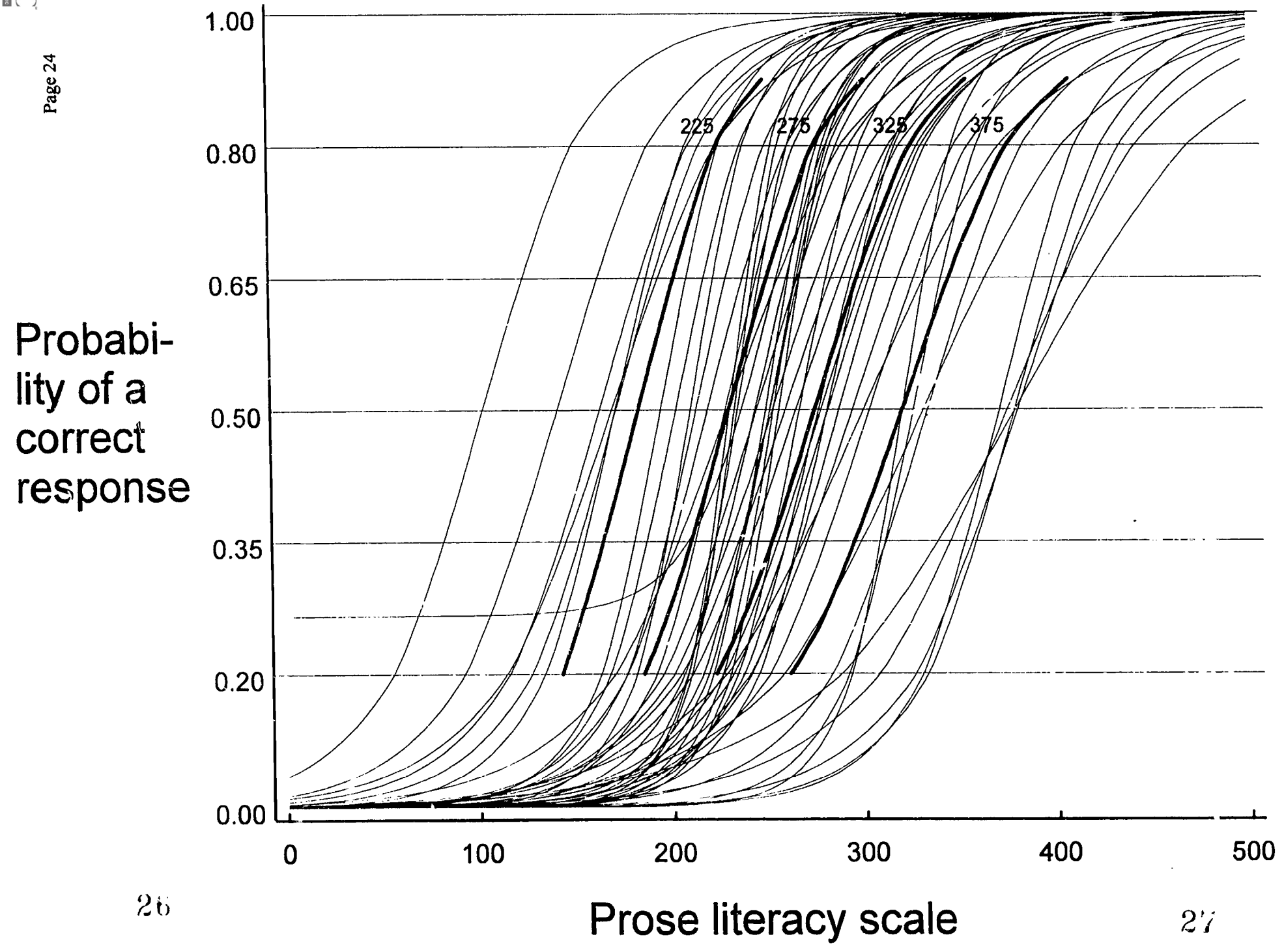
Prose literacy levels	Type of match	Plausibility of distractors	Abstractness of information	Readability in grade levels
Level 1	1	1	1	4
	1	1	1	5
	1	1	1	6
	1	1	2	6
	1	1	2	5
	1	1	3	4
	1	2	1	5
Level 2	2	1	2	6
	2	2	1	7
	2	1	3	6
	2	2	2	7
	3	2	2	6
	3	1	3	7
	3	2	2	8
	2	3	3	7
Level 3	3	3	2	7
	3	3	2	8
	4	2	3	7
	4	2	3	8
	4	3	3	8
	4	3	3	9
Level 4	4	4	3	7
	4	4	4	8
	5	3	4	9
	5	4	2	9
	5	4	3	8
	5	4	3	9
	6	2	4	10
Level 5	4	5	5	13
	6	4	3	10
	5	5	4	8
	6	3	5	10
	6	4	5	9
	7	4	4	10
	6	5	5	13
	7	5	4	10

The scores on the factors of the hypothetical tasks shown in Table 8 were multiplied by the multiple regression coefficients in Table 7 and averaged within levels. Then new cut points between levels were computed, half-way between the midpoints of each level. The alternative cutpoints between levels corresponding to other response probability conventions are shown in Table 9.

Table 9.– Alternative cut points between prose literacy levels, by response probability criteria: 1992 National Adult Literacy Survey

Response probability	Between prose literacy levels ...			
	1 and 2	2 and 3	3 and 4	4 and 5
RP90	250	302	355	408
RP85	236	287	338	389
RP80	225	275	325	375
RP75	216	266	315	364
RP70	209	258	305	354
RP65	202	250	297	345
RP60	195	243	289	336
RP55	189	237	282	328
RP50	183	230	274	320
RP45	177	223	267	312
RP40	171	217	260	303
RP35	164	210	252	295
RP30	157	202	243	285
RP25	150	194	234	275
RP20	142	185	223	261

The full-page graph on the following page plots the above cutpoints between levels (with a spline interpolation between the points) as a thick line, as well as the item characteristic curves previously shown in Figure 6. Also displayed in this graph are the four cutpoints along the RP80 gridline (225, 275, 325, and 375). Visual inspection of this graph leads to the conclusion that the majority of prose literacy tasks stay within the same level, regardless of the response probability value used. As a result, the Kirsch-Mosenthal descriptions of what tasks in levels 1 through 5 require of people do not differ much by the response probability convention adopted, at least for criteria above RP60. If a criterion below RP60 were used, the descriptions would have to be revised to remove elements of plausible distractors and add elements relating to the readability of the text. What differs among levels defined in this way is the consistency of people's success with these prose literacy tasks.


Probability of a correct response
1.00
0.80
0.65
0.50
0.35
0.20
0.00
225
275
325
375
0
100
200
300
400
500
Prose literacy scale

Adult Population Distributions over Alternative Literacy Levels

The 1992 National Adult Literacy Survey has previously reported that 21 percent of the 191 million adults in this country demonstrated skills in the lowest of five prose literacy levels using the RP80 response probability convention.[33] The first report went on to explain that most adults in Level 1 were consistently successful when performing simple, routine tasks involving brief and uncomplicated texts and documents, but were not consistently successful when performing more complicated or difficult tasks. For example, they were able to identify a piece of specific information in a brief news article. Others in Level 1 were not consistently able to perform these types of tasks, and some had such limited skills that they were unable to respond to much of the survey. Of those who scored in Level 1, 21 percent of adults in this level did not perform a single prose literacy task correctly.[34]

Once the alternative cut points between the prose literacy levels were determined, it became possible to estimate the proportion of U.S. adults who perform in each level under alternative choice of response probability conventions. The results are shown in Table 10 below. Each row in Table 10 presents the population distribution of adults across the five prose literacy levels. The rows differ only in the response probability convention used to set the cut points between the levels. For response probability conventions above 60 percent, the same general descriptions of literacy levels can be used. The only difference is the proportion of times that adults have to be successful with equivalent tasks in order to be counted as "able to do" such tasks.

If the 1992 National Adult Literacy Survey had reported the same results using the somewhat lower RP65 response probability convention currently used in reporting the educational achievement of our nation's children by the National Assessment of Educational Progress, the report would indicated that 13 percent percent of the 191 million adults in this country demonstrated skills in the lowest of five prose literacy levels. The remainder of the above description might be modified to read as follows: "Most adults in Level 1 were generally successful when performing simple, routine tasks involving brief and uncomplicated texts and documents, but were not generally successful when performing more complicated or difficult tasks. For example, they were able to identify a piece of specific information in a brief news article. Others in Level 1 were not generally able to perform these types of tasks, and some had such limited skills that they were unable to respond to much of the survey. Of those who scored in Level 1, 32 percent of adults in this level did not perform a single prose literacy task correctly." In these explanations, the term "consistently successful" is an attempt to capture the 80 percent convention and "generally successful" attempts to capture the 65 percent convention.

---

[33]Page 16 in Irwin S. Kirsch, Ann Jungeblut, Lynn Jenkins, and Andrew Kolstad, 1993, *Adult Literacy in America: A First Look at the Results of the National Adult Literacy Survey*. Washington, DC: National Center for Education Statistics

[34]See Table A.5P in Irwin S. Kirsch, Ann Jungeblut, Lynn Jenkins, and Andrew Kolstad, 1994, "Interpreting the literacy scales," Appendix A in K.O. Haigler, C.W. Harlow, P.E. O'Connor, and Anne Campbell, *Literacy Beyond Prison Walls: Profiles of the Prison Population from the National Adult Literacy Survey*. Washington, DC: National Center for Education Statistics.

Table 10.– Percentages of U.S. adults within each level of prose literacy, defined by alternative response probability values: 1992

Response probability	Level 1 Prcnt	(st. err.)	Level 2 Prcnt	(st. err.)	Level 3 Prcnt	(st. err.)	Level 4 Prcnt	(st. err.)	Level 5 Prcnt	(st. err.)
90	32	(0.5)	33	(0.7)	27	(0.4)	8	(0.3)	1	(0.1)
85	25	(0.4)	30	(0.5)	31	(0.5)	13	(0.4)	2	(0.2)
80	20	(0.4)	27	(0.6)	32	(0.7)	18	(0.4)	3	(0.2)
75	17	(0.4)	24	(0.6)	32	(0.8)	21	(0.4)	6	(0.3)
70	15	(0.4)	21	(0.5)	31	(0.6)	24	(0.4)	9	(0.3)
65	13	(0.4)	19	(0.5)	30	(0.6)	26	(0.4)	12	(0.4)
60	12	(0.3)	17	(0.5)	28	(0.7)	28	(0.6)	16	(0.4)
55	10	(0.3)	15	(0.4)	26	(0.5)	29	(0.7)	20	(0.5)
50	9	(0.3)	13	(0.3)	24	(0.6)	30	(0.7)	24	(0.5)
45	9	(0.2)	11	(0.3)	22	(0.6)	29	(0.6)	29	(0.5)
40	8	(0.3)	10	(0.3)	20	(0.5)	29	(0.6)	34	(0.5)
35	7	(0.3)	8	(0.3)	18	(0.5)	27	(0.6)	40	(0.5)
30	6	(0.3)	7	(0.3)	15	(0.5)	25	(0.6)	46	(0.6)
25	6	(0.2)	6	(0.2)	13	(0.4)	23	(0.5)	53	(0.6)
20	5	(0.2)	5	(0.2)	10	(0.3)	19	(0.5)	62	(0.6)

Source: U.S. Department of Education, National Center for Education Statistics, National Adult Literacy Survey, 1992

As the criterion response probability is relaxed in Table 10, larger proportions of adults appear to be able to perform at higher levels of prose literacy. The response probability convention makes the most difference at the upper and lower ends of the scale (Levels 1 and 5). For example, if the adult literacy program were to adopt the same response probability convention as that used by the NAEP, the proportion of the population in prose literacy Levels 1 and 2 would drop from 47 percent (as widely reported in the media) to 32 percent in the same levels, a less distressing figure. The proportion of the population in prose literacy Level 5 would increase from 3 to 12 percent, a substantively and statistically significant increase.

When the purpose of reporting is to discuss what students or adults "can't do," there may be some value in reporting achievement according to low response probability conventions. There is a middle ground between those who are consistently successful and those who are consistently unsuccessful with certain educational achievement questions. Those who are as likely to get a question right as to get it wrong have not mastered certain skills, but they are not unskilled, either. For example, Table 9 showed that a score of 202 was the minimum needed to ensure at least a 65 percent chance of success with all tasks in Level 1, while for those below 164 the chance of success is less than 35 percent. Table 10 shows that six percent of adults score in this middle range of incomplete skills (13 percent in Level 1 at RP65, less 7 percent in Level 1 at RP35). The argument for the

80 percent convention was that a high criterion is needed to ensure mastery. A similar argument could be made that a 20 percent convention is needed to ensure task failure. From this point of view, the size of the population with middling skills in Level 1 is the proportion of adults scoring between 225 (RP80) and 142 (RP20), which Table 10 indicates would be 15 percent.

One way to estimate the number of adults who did not have the skills to perform any of the tasks in prose literacy Level 1 was to compute the proportion of adults who failed to answer correctly a single prose literacy task in the assessment, a number that turned out to be 8.2 million, or 4 percent of the adult population.[35] Table 10 shows that a similar proportion, 5 percent of the adult population, falls in Level 1 when the response probability convention drops to 20 percent.

These changes in the distribution do not mean that people have more skill than previously reported. The underlying skills of the population have not changed. What has changed is the dividing line between those who are said to be "able to do" the prose literacy tasks and those who are not.

Conclusions

The results of this paper indicated that if the adult literacy program were to adopt the same response probability convention as that used by the National Assessment of Educational Progress, the proportion of the population in prose literacy Levels 1 and 2 would drop by 15 percentage points and proportion in Level 5 would increase by 9 percentage points. While these substantial shift are due to reducing the response probability convention from 80 percent to 65 percent, the underlying distribution of prose literacy skills in the population would not change.

The substantive argument for the highest possible response probability convention was that maximum practical mastery is needed to accurately describe readers as "able to do" the literacy tasks. The argument for a lower convention derives from the innovative analytic approach developed by Kirsch and Mosenthal. Prediction of task difficulty at very high probability levels may attribute too much weight to the 'plausible distractors' factor. Prediction of task difficulty at lower probability levels reduces the impact of this factor and allow the 'readability of text' factor to play a role in task difficulty. Further discussion and debate on this issue may clarify the criteria for selecting a standard.

A factor that has such a large impact on the results deserves a thorough understanding of the issues and debate over the standard to be adopted. People concerned with measuring literacy and other educational achievements accurately need to understand what the response probability convention is and why it matters to reporting. There may be an advantage to communicating survey results if all survey programs were to adopt the same convention.

---

[35]See Table A.5P in K.O. Haigler, et al., 1994, *Literacy Beyond Prison Walls: Profiles of the Prison Population from the National Adult Literacy Survey*. Washington, DC: National Center for Education Statistics.