Chapter 14

LITERACY LEVELS AND THE 80 PERCENT RESPONSE PROBABILITY CONVENTION

Andrew Kolstad, National Center for Education Statistics

As described in the preceding chapter of this report, the 1992 National Adult Literacy Survey employed a response probability convention of 80 percent to locate literacy tasks on the three literacy scales, and then used the locations of the tasks to set boundaries between the literacy levels. Many users of this survey do not realize how sensitive the basic survey findings are to small variations in the response probability convention. If an alternative response probability convention were to be used to locate literacy tasks on the scale, the boundaries between the literacy levels would shift.

The initial survey findings reported that large proportions of adults perform in Levels 1 and 2 on the three literacy scales, and that such adults were at risk due to their limited literacy skills. Nevertheless, most adults in Levels 1 and 2 reported that they could read and write English “well” or “very well.” Stitch and Armstrong (1994) were the first to point out that the 80 percent response probability criterion used in the adult literacy survey could be a possible source of the reported gap between performance and perception.

This chapter explores the importance of the response probability convention in reporting prose literacy results from the 1992 National Adult Literacy Survey. The chapter will show that raising or lowering the response probability convention would distribute the population differently across the five levels of literacy. However, such changes would not mean that adults have more literacy skills than previously reported. The true literacy proficiencies of the population remain as reported. What would change is the line that separates those who are classified as “able to do” the literacy tasks and those who are not. The response probability convention influences how the results are interpreted, not how well adults perform on the assessment.

The following sections of this chapter describe the prose literacy tasks and their characteristics; the use of item mapping to anchor the prose literacy scale by locating specific tasks along it (using a response probability convention); the need for a response probability convention; 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. The final part of the paper discusses a few implications of the findings.

14.1 PROSE LITERACY TASKS AND THEIR CHARACTERISTICS

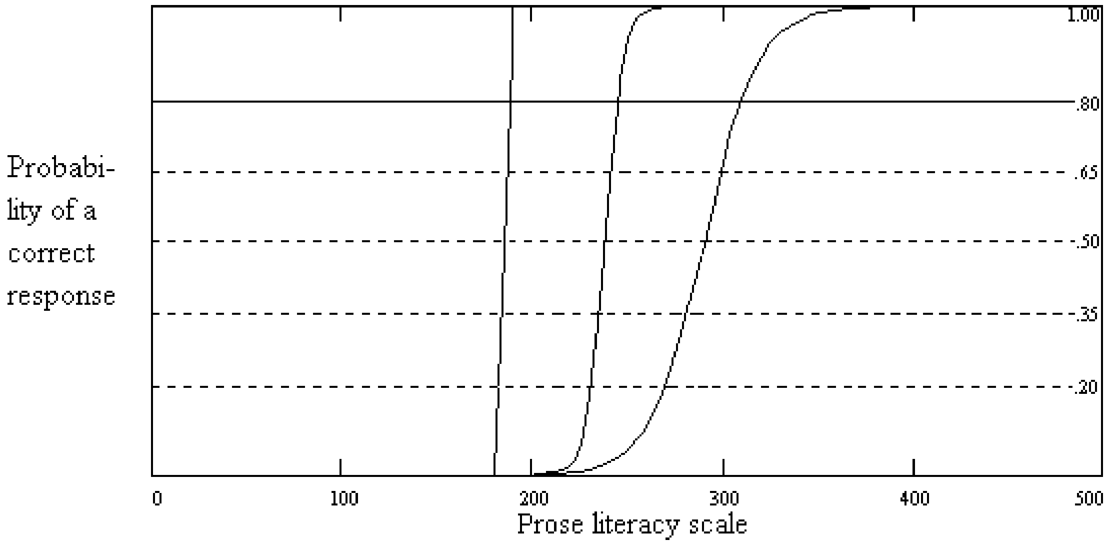
The results of this investigation are based on survey responses and assessment data from the 1992 National Adult Literacy Survey, supplemented with task-specific data developed by Mosenthal and Kirsch that served as the basis of the literacy levels (see Chapter 13). The 1992 National Adult Literacy Survey included 165 literacy tasks, of which 41 were used to measure prose literacy. A typical adult in the survey responded to 11 prose literacy tasks—not enough to measure any particular adult's prose literacy skills with any accuracy, but with a large sample, the 41 litaracy tasks are enough to estimate the distribution of prose literacy skills in the adult population. Only two of the prose literacy tasks were multiple-choice items, while the remainder were answered with a word, phrase, or a few short sentences. All were scored as right or wrong, with no partial credit allowed.

The assessment tasks in the survey were designed to measure prose literacy as a unidimensional scale. Ideally, one ought to be able to predict with a good deal of accuracy the pattern of right and wrong answers to the prose literacy tasks in the assessment from a person's score on the prose literacy scale. Success with prose tasks would also be cumulative. Once any particular skill is mastered, any task needing that skill could be performed correctly. As described in Chapter 9, the 1992 National Adult Literacy Survey used item response theory (IRT) to model the relationship between proficiency—an unobservable variable that is estimated from the responses to many test questions—and correct responses to any particular test question (Lord and Novick, 1968). A feature of such models that provides some realism 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.

If a real literacy task had a high enough discrimination parameter, its item characteristic curve (ICC) would look like a step-function. 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 14-1 displays three item characteristic curves representing hypothetical literacy tasks. Here, the vertical axis represents the probability of a correct response, while the horizontal axis represents the score on the 0-to-500 prose literacy scale. The 1992 National Adult Literacy Survey associated each literacy task with the point on the scale at which an adult would have an 80 percent chance of success with that task. The other horizontal guidelines in the figure (equal to .20, .35, .50, and .65) are presented for comparative purposes, as discussed below. What this figure reveals is that as proficiency increases, so does the likelihood of a correct response to each item.

In Figure 14-1, the ICC of the literacy task on the left was generated by a hypothetical IRT function with its discrimination parameter set high enough to approximate a step function (a = 50), while the one on the right was generated by a hypothetical IRT function with a discrimination parameter set equal to the highest value that actually occurred among any of the 165 literacy tasks (*a* = 2.16). With the task on the left, the boundary between success and failure is very narrow. The probability of success changes from zero to one within a range of about 5 points on the literacy scale. With the task on the right, the boundary between success and failure on this task is more broad. Anyone with a prose literacy score in the range between roughly 250 and 350 has a probability of success that (after rounding) is neither zero nor one. In this range, success or failure with such a task is not a certainty. Yet this task discriminated better than any other.

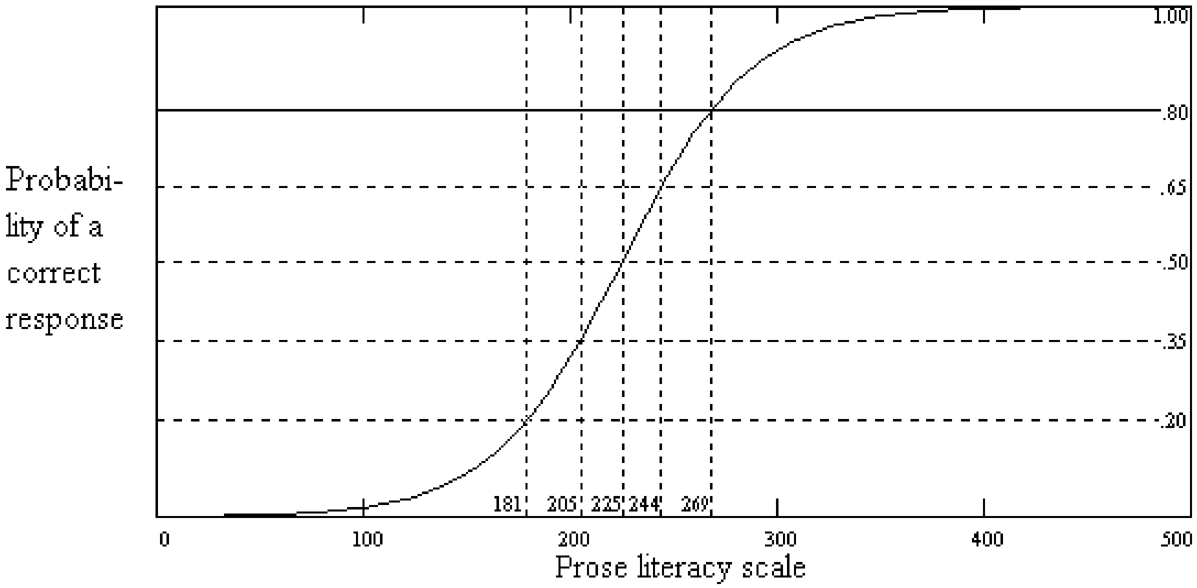
Figure 14-1. Three hypothetical tasks of varying discrimination (and difficulty)



An average prose literacy task from the 1992 National Adult Literacy Survey did not discriminate as well as the best task shown in Figure 14-1. To illustrate performance on an average prose literacy task, Figure 14-2 below presents a hypothetical ICC with average values of both discrimination and difficulty, based on the tasks actually used in the 1992 National Adult Literacy Survey. The intersections between the horizontal guidelines (at probabilities of .20, .35, .50, .65, and .80) and the item characteristic curve locate points on the prose literacy scale. These are the points at which prose proficiency is sufficient to achieve these probabilities of success on this hypothetical task.

What Figure 14-2 reveals is that as proficiency increases, so does the likelihood of a correct response. Typically this increase is gradual. The 1992 National Adult Literacy Survey associated each literacy task with the point on the scale at which an adult would have an 80 percent chance of success with that task. For an average prose task, a proficiency score of at least 269 would be needed to predict success with the 80 percent response probability adopted as a standard by the 1992 National Adult

Figure 14–2. Hypothetical prose literacy task with average difficulty and discrimination

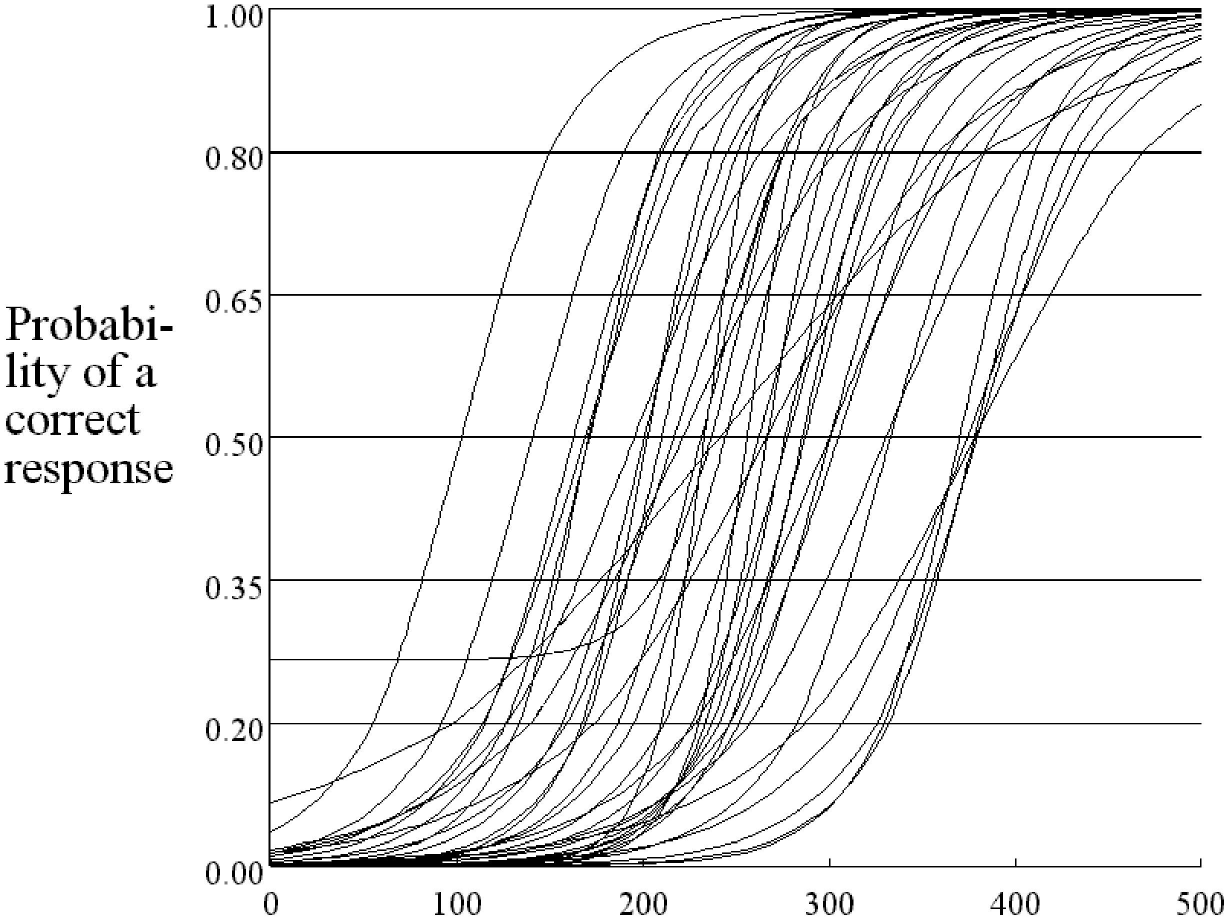


Literacy Survey. Proficiency scores below 269 indicate less than an 80 percent likelihood of success. Still, scores greater than 225 indicate an expectation for performance that is more often than not successful (i.e., greater than 50 percent). If the standard for associating items with the scale were set at 35 percent, an average task would be associated with a score of 205 or more. The score associated with any given literacy task can vary greatly, depending on the response probability adopted as the standard.

The item characteristic curves for 34 of the prose literacy tasks in the National Adult Literacy Survey are displayed in Figure 14-3. (The 7 left out would have been near the middle in difficulty and were omitted to keep the figure legible.) The ICCs are spread along the horizontal axis by empirical differences in their “difficulty.” The more difficult a task, the smaller the likelihood of a correct answer at every point along the proficiency scale.

While the entire curve can be thought of as describing the difficulty of a task, simpler measures of task difficulty are often useful for summary purposes. The “difficulty parameter” in a 3-parameter IRT model indicates the point at which the item characteristic curve changes inflection and discriminates small differences in proficiency best. When it is not possible to guess the correct answer from a short list of options, this “difficulty parameter” corresponds to a 50 percent chance of success on the task. The 1992 National Adult Literacy Survey used a different measure of task difficulty—the scale score associated with reaching a response probability of 80 percent. The “RP80 difficulty” is represented graphically in Figure 14-3 by the point at which the .80 horizontal guideline intersects each ICC curve.

Figure 14–3. Item characteristic curves of 34 prose literacy tasks



In a 1-parameter IRT model, the ICCs are parallel, but the 1992 National Adult Literacy Survey used a 3-parameter IRT model to improve the fit of the model to the data. As a result, the ICCs in Figure 14-3 are not parallel curves. Taking a horizontal section of the ICCs at a different response probability, such as the .50 guideline shown in the figure, would produce a somewhat different ordering of tasks in terms of their “RP50 difficulty.”

14.2 THE NEED FOR A RESPONSE PROBABILITY CONVENTION

If the slope of the item characteristic curves was vertical, it would be easy to interpret points along the prose literacy scale in term of the tasks that people at or above that point could do. The skill difference between success and failure to perform a specific task would be small. In reality, however, assessment tasks often do not provide such an absolute indication of proficiency. The improvement in performance on any given task with prose literacy is gradual, not sudden. The issue of selecting a particular value for a desired response probability arises because the slope of the item characteristic curve is less than vertical.

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 (Beaton, 1987). Anchoring is a way to describe, in general terms, 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. The IRT model's "difficulty parameter" roughly corresponds (except for multiple-choice questions) to a response probability convention of 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. This convention, however, was rejected and a response probability criterion of 80 percent (RP80) was chosen, in order to ensure that students above this criterion would have a sufficiently high probability of success on individual items.

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 scales, the ETS analysts carried over the NAEP RP80 criterion for its reporting (Kirsch, Jungeblut, and others, 1986). The RP80 criterion was subsequently carried over to the 1992 National Adult Literacy Survey, in order that the adult literacy findings remained comparable with the findings from the 1985 Young Adult Literacy Assessment.

The RP80 difficulty criterion was chosen not only because NAEP had been using this criterion at the time of the 1985 survey, but also because the 1992 National Adult Literacy Survey rejected the IRT model's "difficulty parameter" as the point along the scale that should be used to differentiate success and failure. The survey analysts rejected this measure because they believed that examinees with a 50 percent chance of success (the probability associated with the "difficulty parameter") have not demonstrated sufficient mastery of that item. In order to have sufficient confidence that adults of a particular prose proficiency were able to succeed with each task, the 1992 National Adult Literacy Survey adopted NAEP's 80 percent response probability as a criterion. This convention was adopted in order to reduce the proportion of false positives[1] and to assure the accuracy of claims that examinees can meet the demands of the underlying scale at the point at which the literacy task is associated with RP80.

However, during the intervening years between 1985 and 1992, NAEP changed its response probability criterion 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 (1994):

[1] A "false positive" means that an examinee has a scale score above an item's location, implying a positive prediction of success on that item, but actually fails the item. A "false negative" means that an examinee has a scale score below an item's location, implying a negative prediction of failure on that item, but actually succeeds with the item.

> 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 value 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.

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.

During 1994 and 1995, NAEP's Design and Analysis Committee reconsidered the appropriateness of NAEP's response probability convention. In 1995, NAEP 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 (1998) in a paper originally drafted for NAEP's Design and Analysis Committee. Huynh decomposed the item information into that provided by a correct response and that provided by an incorrect response. Huynh showed that the item information provided by a correct 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, rather than the information provided by a *correct response*, 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. The National Adult Literacy Survey staff also made an intuitive judgment, but a judgment that was more stringent. Other testing programs continue to use the IRT model's "difficulty parameter," a less stringent standard, as the criterion for locating items.

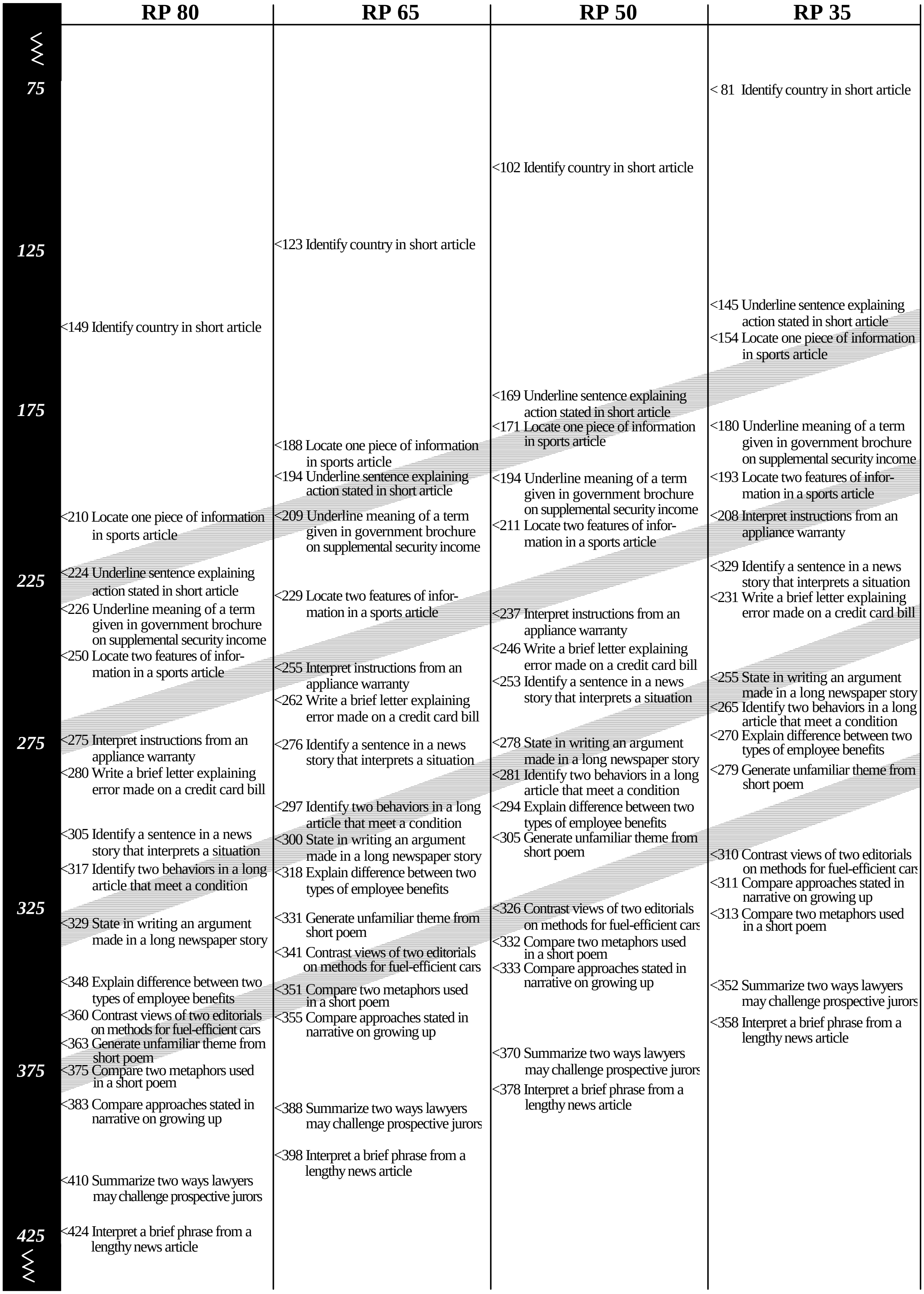
The principal uses of the response probability convention are for item mapping and scale anchoring. The 1994 NAEP reading report, for example, provided an item map that graphically displays the relative difficulty of a select group of reading assessment items. The authors (Campbell, Donahue, Reese, and Phillips, 1996) explained the meaning of their item map 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.)

The first report from the 1992 National Adult Literacy Survey also presented an item map to aid in the interpretation of each literacy scale (Figure 1, page 10). However, little interpretation was provided and the report did not mention the response probability convention used to map the literacy tasks. The report indicated only that this figure "describes some of the literacy tasks and indicates their scale values." An explanation like that above would also be applicable to the prose literacy scale, except that the probability used was 80 percent rather than 65 percent, and the same value was used for all items, including multiple-choice items.

In order to begin an exploration of the impact of the response probability convention on reporting data from the 1992 National Adult Literacy Survey, new item maps were developed for comparison with the item map in the initial survey report. Figure 14-4 displays item maps prepared using response probability conventions of 80 percent, 65 percent, 50 percent, and 35 percent. The left column of Figure 14-4 displays the item map for the 1992 prose literacy scale, using the 80 percent criterion as published in the initial report. Three additional columns in the figure display where the prose literacy tasks would be located had the items been mapped at 65 percent, 50 percent, or 35 percent. The literacy tasks slide up or down the proficiency scale, depending on the chosen value of the response probability criterion.

Figure 14-4. Difficulty values of selected tasks along the prose literacy scale, mapped at four response probability criteria: The 1992 National Adult Literacy Survey



NCES’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. As a result, the 1992 National Adult Literacy Survey is now left with a more stringent standard for describing the skills of adults than that used to describe the skills of elementary and secondary school children. 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.

14.3 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 is derived from three factors: 1) the format of the printed material, 2) the content of the material, and 3) the nature of the request requiring use of the material. The more difficult the literacy task, the higher the literacy skill needed to successfully complete it. Analyzing the sources of the difficulty of literacy tasks helps to understand the nature of literacy skills.

As described in Chapter 13, Kirsch, Mosenthal, and Jungeblut developed a system for measuring different aspects of the cognitive demands that literacy tasks place upon readers: type of match, plausibility of distractors, abstractness of information, and readability of the text. The 41 prose literacy tasks in the 1992 National Adult Literacy Survey were scored. The resulting distribution of scores for each of the skill and process demands were displayed previously in Tables 13-2, and 13-3.

As indicated in Tables 14-1 and 14-2, 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).

Table 14-1. 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 14-2. Distribution of readability 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

If much of the variability in task difficulty can be predicted by these skill and process requirements of the cognitive tasks, then it becomes possible for these skill and process requirements to provide the building blocks for general descriptions of the kinds of skills needed to score in the selected ranges along the literacy scales that constitute the literacy levels. Describing the literacy levels in terms of general skills, rather than specific tasks, is useful because the descriptions can convey the content of the literacy scales in general, rather than how literacy skills are brought to bear on particular literacy tasks.

14.3.1 Predictive Factors and RP80 Task Difficulty

In the preceding chapter, Kirsch, Mosenthal, and Jungeblut used multiple regression to predict the difficulty of the prose literacy tasks in the 1992 National Adult Literacy Survey on the basis of the skill and process requirements of the cognitive tasks 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 ($R^2 = .87$):

$$\begin{array}{lllll} \text{RP80} = 28.9\ \text{TypMatch} & +16.1\ \text{Distract}\ + & 8.8\ \text{Abstract}\ + & .2\ \text{Readability} + \text{Constant} \\ \quad (3.4)\ \text{std.err.} & (3.6)\ \text{std.err.} & (4.2)\ \text{std.err.} & (1.7)\ \text{std.err.} & \end{array}$$

This equation showed that while prose task difficulty was highly predictable by these four factors, ‘type of match’ had a large significant impact (more than 8 times its standard error); ‘plausibility of distractors’ had a 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.

The three authors divided the tasks into levels based on their observation of qualitative shifts in these skill or process requirements of the cognitive tasks, which they had shown were associated with increasing task difficulty. The RP80 criterion provided a method for locating the tasks on the literacy scale, while the observation of qualitative shifts in cognitive requirements provided a method for grouping the tasks into five levels based on similarities in their demands. Both grouping the tasks on the basis of their cognitive demands and locating the tasks on the scale on the basis of the RP80 criterion provided the basis for choosing the boundaries between the five literacy levels. Kirsch, Jungeblut, and Mosenthal described how they grouped the tasks on the basis of qualitative shifts in cognitive demands as follows:

> …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.

To locate the tasks on the literacy scales, Kirsch, Jungeblut, and Mosenthal relied on the 80 percent response probability criterion. The three authors observed several transition points (occurring at roughly 50 point intervals and beginning with 225 on each scale) at which the aggregate task demands shifted upwards. They decided that using equal-interval cutpoints that were the same on each scale (225, 275, 325, and 375) constituted a discretionary degree of rounding that was consistent with the accuracy of predicting the RP80 difficulty of the tasks. They divided each scale into five levels that reflect both the 80 percent response probability criterion and 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).

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 (Table 13-13). The descriptions attempted 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. 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. Nevertheless, the descriptions of literacy levels have served the purpose of providing a general summary of the skills required to succeed (at the 80 percent level) on the literacy tasks included in the 1992 National Adult Literacy Survey.

14.3.2 Predictive Factors, Task Difficulty, and the Response Probability Convention

Kirsch, Mosenthal, and Jungeblut conducted all their analyses using task difficulty as measured at the RP80 response probability convention, and then used their results to justify grouping the literacy tasks into levels for summary desrcriptive purposes. In order to isolate the impact of the response probability convention on reporting results from the 1992 National Adult Literacy Survey, it is necessary to perform similar analyses, using other response probability conventions as a criterion for task difficulty. The new analyses could then indicate whether the prior descriptions of literacy levels would remain appropriate when other response probability conventions are used.

Table 14-3 shows estimates of the corresonding multiple regression coefficients using task difficulty measured at the original 80 percent and fourteen 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 (Kirsch, Jungeblut, and Campbell, 1992).

The coefficients in Table 14-3 display several patterns that could not be seen in a single regression with RP80 as an outcome. The coefficient of explained variance (R^2) 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. The increase in variance explained and the fact that readability becomes a significant factor at lower RP criteria may suggest that there is potentially useful information to be captured in describing literacy levels should a lower criterion be used for reporting.

The boundaries between the prose literacy levels were based on two factors: a clustering of prose tasks with similar cognitive demands and the locations of the literacy tasks on the scale based on the RP80 criterion. In order to isolate the impact of the response probability convention on reporting results from the 1992 National Adult Literacy Survey, it is necessary to find cut points between the levels that will group the literacy tasks together on the same basis, with the same qualitative shifts in the cognitive demands of the tasks, as measured by the four task variables that Kirsch, Mosenthal, and Jungeblut reported. Since the item characteristic curves of the tasks are not parallel, some tasks change their ordering slighlty from one response probability convention to another. It is not possible to divide exactly the same tasks into the exact same five levels, just by changing the response probability criterion. It is possible, however, to use the regression equations to derive alternative cut points between levels that would approximate the same grouping of prose literacy tasks by skill and process demands. The next section describes the process used to derive alternative cutpoints that would ensure, to the maximum extent possible, that the alternative cutpoints reflect only the influence of changing the response probability convention and not other, more substantive considerations.

Table 14-3. Multiple regression coefficients, standard errors, and R-squares for regression equations predicting task difficulty measured at selected response probability criteria for 71 prose literacy task from the 1992 National Adult Literacy Survey and the 1991 study of the literacy of job-seekers

Response probability criterion	Intercept	Type of match	(Std. Error)	Plausibility of distractors	(Std. Error)	Abstractness of information	(Std. error)	Readability of prose text	(Std. error)	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

14.3.3 Alternative Cut Points between Literacy Levels

The goal of deriving alternative cut points between levels first requires developing a general method that reproduces the existing RP80 cut points, then extending the method to other RP values. Since the cut points are derived by grouping similar tasks together, developing sets of hypothetical tasks with difficulty-related characteristics similar to those in the existing literacy levels appeared to be a workable approach. Table 14-4 displays the values of a selected group of hypothetical tasks—about half a dozen for each level. These are “literacy tasks” only in the sense that if real literacy tasks were developed that had the selected values on the variables measuring cognitive demands and that showed the same relationship to task difficulty as the 71 existing tasks used in the regression analyses, the sets of tasks in each level would display the following properties:

- When used to predict the average RP80 task difficulty within each level for the set of hypothetical tasks (using the RP80 equation in Table 14-3 and averaging the predicted values within each level), the predicted averages would come to exactly the midpoints of each of the five existing literacy levels: 200, 250, 300, 350, and 400.
- When the scale distance between pairs of adjacent midpoints are divided in half, the results become the existing RP80 cutpoints between levels: 225, 275, 325, and 375.

- When the hypothetical items are used to predict average task difficulty within each level for alternative response probability criteria, the scale distances between pairs of adjacent midpoints provide alternative cut points between levels, based on the other response probability criteria.

The five sets of hypothetical literacy tasks, along with their scores on the values of the variables affecting task difficulty for each of the hypothetical tasks are shown below in Table 14-5. The values of the task variables were multiplied by the various sets of multiple regression coefficients (shown in Table 14-3) associated with each response probability criterion and averaged within levels to produce new sets of midpoints. The points on the scale half-way between the sets of midpoints provided alternative cut points between levels, and are shown in Table 14-4.

Table 14-4. Alternative cut points between prose literacy levels, by response probability criteria: The 1992 National Adult Literacy Survey

Response probability criterion	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

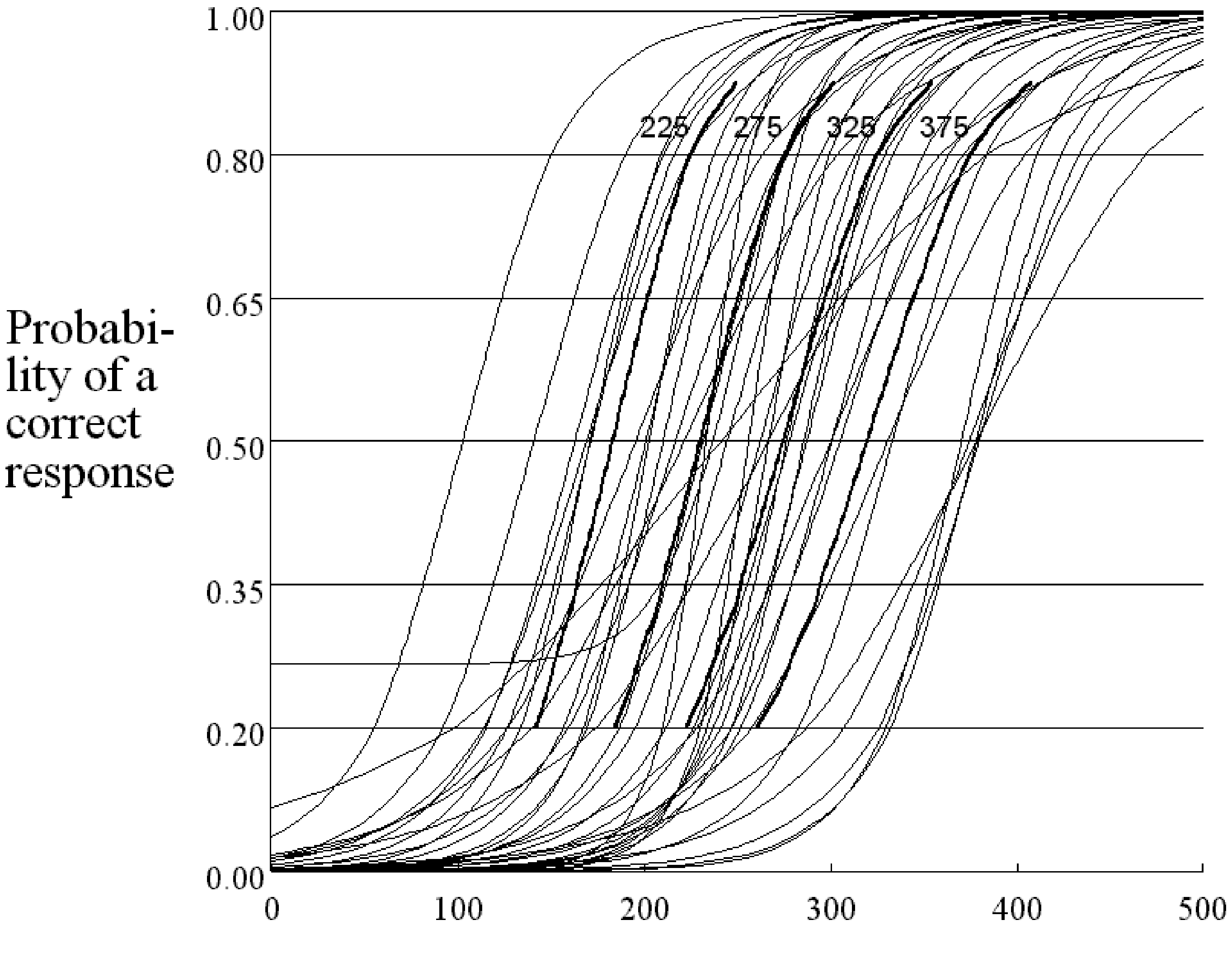
Table 14-5. Scores of hypothetical prose literacy tasks on four factors that predict difficulty

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

Table 14-4 shows how the alternative cut points change as a function of the response probability criterion. Around 50 percent, each increase of five points in the response probability criterion moves the cutpoints between literacy levels about 5 to 8 points higher on the prose literacy scale. Around 80 percent, each increase of five points in the response probability criterion moves the cutpoints between literacy levels about 12 to 14 points higher.

It is instructive to examine the alternative cut points between levels graphically, as well as numerically. Figure 14-4 below plots both the item characteristic curves previously shown in Figure 14-3 and the alternative cut points between the literacy levels as a thick line (with a spline interpolation between the points). Also displayed in this graph are numbers showing the four existing cutpoints along the RP80 horizontal gridline (225, 275, 325, and 375). Visual inspection of this graph leads to the conclusion that

Figure 14-4. Item characteristic curves of 34 prose literacy tasks and literacy level cutpoints between the tasks: The 1992 National Adult Literacy Survey



the majority of prose literacy tasks stay within the same level, regardless of the response probability value used. As a result, the Kirsch-Mosenthal-Jungeblut descriptions of what tasks in Levels 1 through 5 require of adults 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 add elements relating to the readability of the text. If a criterion below RP40 were used, the descriptions would have to be revised to remove elements relating to the plausibility of distractors.

The conclusions that on the one hand, the Kirsch-Mosenthal-Jungeblut descriptions of what tasks in Levels 1 through 5 require of adults do not vary as a function of the response probability criterion, and that on the other hand, the cut points between the levels change drastically, appear to be in conflict. The process of setting alternative cut points has isolated the impact of the response probability convention on setting boundaries between the literacy levels. The current descriptions of the literacy levels, which have carried over into later reports from the International Adult Literacy Surveys, appear to lack a description of the response probability criterion and an understanding of its significance.

The variation in the cut points between levels as a function of the response probablity convention are quite large. To put the sensitivity to this factor in perspective, consider that if the 1992 National Adult Literacy Survey were to adopt the NAEP convention for mapping items onto latent scales (65 percent), the cut points between levels would have to move down by 23 to 28 points, which is about half the 50-point width of a literacy level. The standard error, a measure that incorporates variability due to both statistical sampling and measurement errors, for average prose literacy for the U.S. population as a whole was 0.7 points on the same literacy scale. The variation in the cut points as a function of the response probability convention plays a significant role in estimating the distribution of the adult population over the five literacy levels.

14.4 DISTRIBUTION OF ADULTS AMONG ALTERNATIVE LITERACY LEVELS

The initial report of the 1992 National Adult Literacy Survey found that 21 percent of the 191 million adults in this country demonstrated skills in the lowest, and another 26 percent demonstrated skills in the second lowest of five prose literacy levels, using the RP80 response probability The initial report explained that most adults in Level 1 were able to perform simple, routine tasks involving brief and uncomplicated texts and documents. For example, they were able to identify a piece of specific information in a brief news article. Others in Level 1 attempted to perform the literacy tasks and were unsuccessful. There were individuals who had such limited skills that they were able to complete only part of the survey (Kirsch, Jungeblut, Jenkins, and Kolstad, 1993). Of those who scored in Level 1, 21 percent of adults did not perform a single prose literacy task correctly (Table A.5P in Haigler, Harlow, O’Connor, and Campbell, 1994).

Once the alternative cut points between the prose literacy levels were determined, it was possible to estimate the proportion of U.S. adults who performed in each level under alternative choice of response probability conventions. The results are shown in Table 14-6 below. Each row in Table 14-6 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 time that adults have to be successful with equivalent tasks in order to be counted as “able to do” such tasks.

Table 14-6. Percentages of U.S. adults within each level of prose literacy, defined by alternative response probability values: 1992

Response probability	Level 1		Level 2		Level 3		Level 4		Level 5	
	Prcnt	(st. err.)	Prcnt	(st. err.)	Prcnt	(st. err.)	Prcnt	(st. err.)	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, the National Adult Literacy Survey, 1992

If the 1992 National Adult Literacy Survey had reported the same results using the somewhat lower RP65 response probability convention that is currently used in reporting the educational achievement of our nation’s children in the National Assessment of Educational Progress, the report would indicated that the two lowest levels included only 13 percent and 19 percent of the 191 million adults in this country. If the 1992 National Adult Literacy Survey had reported the same results using 50 percent as a criterion, the combined proportion in Levels 1 and 2 would be reduced to 22 percent. As the criterion response probability is relaxed in Table 14-6, 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). As the response probability convention drops from 80 to 50 percent, the proportion of the population in prose literacy Level 5 increases from 3 to 24 percent, a substantively and statistically significant change.

A factor that has such a large impact on the results of the survey deserves a thorough understanding of the issues and debate over the standard to be adopted. People concerned with measuring literacy accurately need to understand what the response probability convention is and why it matters to reporting the results.

14.5 SELECTING AN APPROPRIATE RESPONSE PROBABILITY CONVENTION

These changes in the distribution across the five literacy levels do not mean that people have more or less literacy skills 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.

The RP80 response probability criterion was adopted for the 1992 survey for several reasons. A sister project at Educational Testing Service, the National Assessment of Educational Progress (NAEP) had been using this criterion at the time of the 1985 Young Adult Literacy Assessment, and the results of 1992 survey were intended to be as comparable to the 1985 assessment as possible. Nevertheless, the traditional argument for a high response probability convention is not satisfactory. The issue is unrelated to the way the data were collected or the way the responses were scored, because the response probability convention is used only during the analysis and reporting stage of the survey. Findings from earlier studies could be reanalyzed using a different convention, yet retain comparability.

The National Adult Literacy Survey analysts rejected the IRT model’s “difficulty parameter” as the point along the scale that should be used to differentiate success and failure because they believed that examinees with a 50 percent chance of success (the marginal probability associated with the “difficulty parameter”) have not demonstrated sufficient mastery of that item. The 80 percent criterion gave the survey analysts sufficient confidence that adults at given points on the prose literacy scale were consistently able to succeed with each task. This convention was adopted in order to reduce the proportion of false positives and to assure the accuracy of claims that examinees can meet the demands of the underlying scale at the point at which the literacy task is associated with RP80. This value of the response probability criterion ensured that when survey reports indicate that adults are able to do a particular task or a generic group of tasks, the adults have mastered that task.

Unfortunately, the substantive argument for a high response probability convention is not very fine-grained. Demonstrating task mastery requires only a response probability convention well above 50 percent, not a specific value. Other choices such as 70, 75, or 85 percent could satisify the logic of the argument, yet would produce substantially different outcomes in terms of both the cut points between and population distribution over the five levels. In addition, the cost of a high response probability convention is an increase in the proportion of false negatives. Those that the scale predicts are unable to succeed with literacy tasks have much more than a negligible probability of actually being able to succeed. Better justification is needed than has heretofore been forthcoming for retaining the 80 percent criterion in future work.

The substantive logic of the mastery argument also contains within it the seeds of an alternative perspective. Many users of adult literacy survey findings want to focus on the *lack* of facility with printed and written material and to report on what adults *can't* do. To be sure that adults are *unable* to perform the literacy tasks, survey analysts might need to use a correspondingly low response probability criterion, such as 20 percent. Such a criterion would assure that adults who do not reach these levels of consistency in their responses have a very low chance of success with the tasks. Data users can be very sure that such a group of adults does not have the skills in question. 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. 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 a low response probability convention. An alternative way to estimate the number of adults who did not have the skills to perform any of the tasks in prose literacy Level 1 would be 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 (See Table A.5P in Haigler, Harlow, O'Connor, and Campbell, 1994). Table 14-6 shows that a similar proportion, 5 percent of the adult population, falls in Level 1 when the response probability convention drops to 20 percent.

A large proportion of the population occupies a middle ground between those who are consistently successful and those who are consistently unsuccessful with the assessment's literacy tasks. 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. Consider the population at the boundary between Level 1 and Level 2. Table 14-4 showed that a score of 225 was the minimum needed to ensure at least a 80 percent chance of success with the tasks in Level 1, while the minimum needed to ensure at least a 20 percent percent chance of success was 142. Table 14-6 shows 20 percent of adults score below 225, and 5 percent score below 142. The difference between the two, 15 percent of adults, represents those for whom it remains uncertain either that they will succeed or that they will fail with the tasks in Level 1.

Using item response theory as a statistical model provides an argument for mapping items at the “difficulty parameter,” which is equivalent to 50 percent for nearly all the tasks in the 1992 National Adult Literacy Survey. The main purpose of creating a scale rather than reporting success on particular items is to focus on the common skills that all the items together require. The more the response probability criterion deviates from 50 percent, the more weight is given to peripheral factors unrelated to the core skill requirements that the items share in common (Kolstad, Cohen, Baldi, Chan, DeFur, and Angeles, 1998).

Because the tasks used in the assessment are not perfect measures of literacy, there may be a substantial proportion of adults in the middle range, who have neither fully mastered the literacy tasks nor are completely unable to succeed with them. The value of a 50 percent response probability convention is that it strikes an even balance between false positives and false negatives. With this criterion, one can have the same degree of confidence in statements about what adults can’t do as in statements about what adults can do. Mapping items in this way both provides a balance between false positive and false negative claims about adult performance and minimizes the total misclassification error.

14.6 CONCLUSIONS

It should be reiterated that while varying the response probability convention would distribute the population differently across the five levels of literacy, adults still have exactly the same literacy skills as previously reported. The true literacy proficiencies of the population are unaffected by variations in the way analysts interpret the data. What changes is the dividing line between those classified as “able to do” the literacy tasks and those not classified as “able to do” the literacy tasks. The substantive argument for the highest possible response probability convention was that maximum practical mastery is needed to describe readers accurately as “able to do” the literacy tasks.

Several alternatives are possible. One would be to use a fixed “mastery” increment (set using professional judgment) to the the IRT model’s “difficulty parameter” to satisfy the mastery criterion, rather than a fixed, high response probability convention. The increment would ensure that examinees have a sufficient likelihood of success, would put more weight on items that discriminate literacy better and less weight on items that discriminate more poorly, and would highlight the judgmental basis of the item mappings. Nevertheless, an imbalance between false positives and false negatives and inconsistencies with the statistical model would remain. A second approach would be to use the IRT model's "difficulty parameter" for mapping, (roughly equivalent to a 50 percent criterion). This would provide a better balance between false positives and false negatives, more consistency with other aspects of the IRT statistical machinery, and equivalent weight on all items, regardless of their discriminating power.

Perhaps what is needed is a fresh approach to interpreting the findings, one that does not rely on item mapping for interpreting the findings. Some have suggested interpreting scale scores in terms of the expected percentage correct for a standard set of released items (a "market basket" approach), as a way of escaping the arbitrariness of choosing a particular value for the mastery increment or the response probability convention. Another alternative, one adopted by the National Assessment of Educational Progress, is to use judgment-based standards for reporting. The National Assessement Governing Board's achievement level setting process, though it has been criticized, does not involve item mapping and would provide a judgmental standard for how much literacy is sufficient, something that the public seems to want.