DOCUMENT RESUME

ED 385 595 TM 024 043

AUTHOR Enright, Mary K.; And Others
TITLE A Complexity Analysis of Items from a Survey of Academic Achievement in the Life Sciences.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-93-18
PUB DATE Mar 93
NOTE 45p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Academic Achievement; *Biological Sciences; *Difficulty Level; Elementary Secondary Education; Knowledge Level; *Prediction; Rating Scales; *Science Teachers; Science Tests; Surveys; *Test Items
IDENTIFIERS Complex Concepts; National Assessment of Educational Progress; Variance (Statistical)

ABSTRACT

The difficulty of 44 items from the life sciences subscale of the National Assessment of Educational Progress (NAEP) 1985-86 science assessment was analyzed in terms of item attributes and science educators' judgments of difficulty. The attributes included ratings of various characteristics of the items' text and option set, the items' cognitive demand, and the level of knowledge required by items. The mean judgment of three science educators (an instruction supervisor, an experienced teacher, and a young teacher) about item difficulty, which accounted for 52% of the variance, was the best single predictor of item difficulty. Combining item attribute information with educators' judgments of item difficulty improved the prediction of item difficulty on the order of 7% to 15% of the variance. When item difficulty was modeled in terms of discrete item attributes (global judgments of item difficulty not included in the model), the level of knowledge required was an important determinant of difficulty, while cognitive demand was not. The implications of these results for construct validation and for test design are discussed. Two figures and five tables illustrate the discussion. (Contains 23 references.) (Author/SLD)

RESEARCH REPORT

A COMPLEXITY ANALYSIS OF ITEMS FROM A SURVEY OF ACADEMIC ACHIEVEMENT IN THE LIFE SCIENCES.

Mary K. Enright
Nancy Allen
Myung-In Kim

Educational Testing Service
Princeton, New Jersey
March 1993

A complexity analysis of items from a survey of academic achievement in the life sciences.

Mary K. Enright, Nancy Allen, and Myung-In Kim

Educational Testing Service

Running Head: COMPLEXITY ANALYSIS OF SCIENCE ITEMS

Funding for this research was provided by the ETS Office of Corporate Planning and Development.

Abstract

The difficulty of 44 items from the life sciences subscale of the NAEP 1985-86 science assessment was analyzed in terms of item attributes and science educators' judgments of difficulty. The attributes included ratings of various characteristics of the items' text and option set, the items' cognitive demand, and the level of knowledge required by items. Science educators' mean judgment of item difficulty, which accounted for 52% of the variance, was the best single predictor of item difficulty. Combining item attribute information with educators' judgments of item difficulty improved the prediction of item difficulty on the order of 7% to 15% of the variance. When item difficulty was modeled in terms of discrete item attributes (global judgments of item difficulty not included in the model), the level of knowledge required was an important determinant of difficulty, while cognitive demand was not. The implications of these results for construct validation and for test design are discussed.

A complexity analysis of items from a survey of academic achievement in the life sciences.

Standardized achievement tests in science have been criticized as testing primarily lower level skills, such as factual recall, and, consequently, having detrimental effects on science education (Hartwig, 1989). In reality, there is little empirical evidence about the kinds of skills assessed by such tests. Traditionally, validation of achievement tests has been in terms of content coverage with little attention to construct validity (Bejar, 1985). This is not surprising in view of the fact that achievement testing has been carried out in the absence of any well-articulated theory of academic achievement. It is only recently that such theories have emerged and their implications for assessment discussed (Glaser, Lesgold, & Lajoie, 1987; Messick, 1984).

Despite the lack of clearly articulated theories, it has become common to include cognitive or process dimensions in assessment frameworks and item specifications. However, although these assessment frameworks and item specifications guide the test development process, they are not directly subjected to empirical verification. This is unfortunate because examination of the fit between the framework and the items would increase the validity of the assessment, help identify weaknesses in current frameworks and items, provide a basis for comparing different types of items and different tests, and contribute to more systematic test design.

In the present study, we sought to better define what item attributes influenced performance on a national survey of science achievement through an analysis of item difficulty. Understanding item difficulty is a topic that has been neglected until recently (Bejar, 1991). However, there is growing recognition of the usefulness of such knowledge for a variety of purposes: constructing, interpreting, and validating tests (Bejar, 1991; Embretson & Wetzel,1987), comparing different tests (Scheuneman, Gerritz, & Embretson, 1991), equating tests (Mislevy, Sheehan, & Wingersky, 1992), and diagnosing student misconceptions (Tatsuoka, 1990).

NAEP Science Assessment Framework

Because of its design as a survey instrument and because of the approach to developing assessment objectives (based on a consensus of science educators at one point in time), the National Assessment of Educational Progress (NAEP) Science Assessment covers a wide domain of content in a way that reflects educational theory and practice at the time plans were made for the assessment. In past years, the framework for the NAEP Science Assessment has included a cognitive dimension based on Bloom's (1956) taxonomy of educational objectives (NAEP, 1985-86). For example in 1976-77 and 1981-82 this dimension included the levels of knowledge, comprehension, and application plus a fourth level that combined analysis, synthesis and evaluation. In 1985-86 this dimension included three categories: knows, uses, and integrates.

Other dimensions of the framework include descriptions of content in terms of traditional domain categories and topics, and problem context. These dimensions were intended as a guide for constructing test items but they are not particularly helpful in interpreting performance on the test nor in comparing what various versions of the tests have measured over time because their validity is not subjected to empirical verification.

Although Bloom viewed the classes in his taxonomy as hierarchically ordered in terms of complexity and as hypothetically related to problem difficulty, the relationship between the cognitive demand of items and item difficulty is unsystematic for many of the content areas tested on the 1985-86 NAEP Science Assessment. An example of the kinds of relationships that are found between item difficulty and cognitive process categories is presented for the life sciences subscale from the 1985-86 NAEP Science assessment in Figure 1. These results are not really surprising in that performance on test items is likely

---

Insert Figure 1 about here

---

to be a result of a number of factors, not just the "cognitive demand" of an item (Emmerich, 1989; Scheuneman et al., 1991). For example, one of the most striking contrasts between Bloom's taxonomy of educational objectives and emerging theories of achievement or expertise is the role attributed to "knowledge"

(Emmerich, 1989). In Bloom's taxonomy, knowledge is represented at the lowest level of hierarchy and involves recall of facts, methods, principles, and theories. In contrast with this view, the role ascribed to "knowledge" is much more important in descriptions of expertise and achievement in many domains (Glaser, 1981). Messick (1984) noted that research on expertise demonstrates that not only do experts have more knowledge, but it is structured in more complex ways. He summarized the import of such research for our conception of educational achievement as follows:

> "Educational achievement refers to what one knows and can do in a specified subject area. At issue is not merely the amount of knowledge accumulated but its organization or structure as a functional system for productive thinking, problem solving, and creative invention in the subject area as well as for further learning." (Messick, 1984, pp. 155-156).

One implication of these ideas for achievement testing is that we need to think about and analyze the knowledge requirements of items as well as the cognitive or processing demands of the items (Emmerich, 1989).

Related Research. Traditional factor analytical approaches to construct validation rely on the identification of consistencies in the pattern of individuals' responses to group or cluster items and use such information as the basis for inferences

about differences or similarities in the processes or skills assessed. One limitation of this method is an inability to distinguish process or skills that are correlated. However, Embretson (1983) noted how the shift from functionalism to structuralism in psychology has permitted the disentanglement of two aspects of construct validity: nomothetic span and construct representation. Nomothetic span refers to the usefulness of a test in differentiating among individuals while construct representation concerns the identification of theoretical mechanisms such as the processes, skills, and knowledge underlying performance on test items. This latter aspect of construct validity will be the focus of this research. One approach that has been used to clarify the constructs represented by a set of items is the method of complexity factors (Embretson, 1983). In this method individual items are scored or rated on a number of factors representing the items' position on theoretical variables thought to underlie item responses.

For the most part, decomposition of test items in terms of factors that contribute to item difficulty or response accuracy have been conducted for tests of abilities such as reading comprehension (Embretson & Wetzel, 1987), literacy (Kirsch & Mosenthal, 1988), and geometric analogies (Mulholland, Pellegrino, & Glaser, 1980). For example, Embretson & Wetzel (1987) developed a processing model to quantify sources of cognitive complexity in multiple-choice paragraph comprehension items and evaluated the

usefulness of this model for predicting item difficulty. Their cognitive model consisted of two stages, text representation and response decision. Test items were rated in terms of variables thought to affect the difficulty of these stages such as surface structure variables, word frequency, and level of question. They reported that the best model of item difficulty, which accounted for about 37% of the variance, included variables representing both text representation and decision processes. One interesting application of the method of complexity factors in this study was a comparison of cognitive characteristics of item sets from two different tests to illustrate how the constructs represented on the two tests differ.

While items from ability tests can be modeled primarily in terms of stimulus complexity and response selection variables, the nature and accessibility of the knowledge being assessed should be an important, additional factor for achievement test items. The importance of such factors in accounting for the difficulty of achievement test items is illustrated in research by Scheuneman et al. (1991). They used the method of complexity factors to account for the difficulty of items from the GRE Psychology Test (a test of specialized knowledge). In addition to rating items in terms of structural features, Scheuneman et al. also rated items with respect to cognitive processing demands and with respect to the level and aspect of the knowledge being probed. Level of knowledge required to correctly respond to the item was classified

by the researchers into one of five categories that included reading comprehension, popular, basic, intermediate, advanced. Aspect of knowledge categories included theory, criterion, procedure, and relationships. Using multiple regression, Scheuneman et al. accounted for about 65% of the variability associated with item difficulty on the GRE Psychology Test. Four factors were necessary to reach this level and the most important of these were knowledge level (accounting for 21% of the variance in difficulty by itself) and aspect of knowledge assessed by an item.

A Framework for Analyzing NAEP Science Items. In the current project, we sought to identify and quantify factors that contributed to the difficulty of the items that were included on the 1986 NAEP life sciences subscale for 13 year-olds. (The life sciences subscale was selected for study because it had a relatively large number of items when compared to other science domain subscales). A componential model of how test items are solved was used as an organizing framework to identify and group factors that had been shown to be related to item difficulty in previous research or that are hypothetically relevant to the item solution process (cf. Embretson & Wetzel, 1987; Scheuneman et al., 1991). In this model, we assume that in order to answer an item correctly, an examinee needs to understand or interpret the item, to engage in problem-solving activities such as searching for relevant information in long-term memory or reasoning about

information provided or recalled, and, in the case of multiple-choice items, to select an answer from among the set of options available (see Figure 2). Item difficulty is assumed to be a

---

Insert Figure 2 about here

---

weighted sum of the difficulties of the various components, and the difficulties of the components are influenced by different factors. The difficulty of the comprehension component should be affected by the text attributes (e.g., the number of words and sentences, the presence of a figure). The problem-solving component should be influenced by the processing demands implicit in the item (cognitive demand, knowledge level). And response selection difficulty should be affected by factors such as the attractiveness of distractors or similarity between the correct answer and the distractors.

Rating some of these factors required familiarity with the scientific knowledge base of the age group for which the items were designed and knowledge of middle-school science curricula. Therefore, science educators served as consultants and helped analyze the knowledge requirements of the items as well as other item attributes.

Method and Procedure

Items

Forty-four multiple-choice items, which composed the life

sciences subscale for Grade 7/Age 13 on the 1986 NAEP in science, were analyzed in this study. Item parameters for a three parameter IRT model were estimated for the life sciences subscale using samples that typically included at least 1,000 subjects (Beaton, 1988). The IRT parameter estimate b was used as the measure of difficulty for the items in the analyses described below. (The life sciences subscale also included items administered to a younger and an older age group that were used to estimate item parameters but which were not analyzed in the present study.) In accordance with a framework for science objectives that guided the development of the assessment, each item was classified with respect to the cognitive skill it measured (knows, uses, or integrates) and its context (scientific, personal, societal, technological) (NAEP, 1985-86).

The items for Grade 7/age 13 had anywhere from 3 to 6 multiple-choice options although 64% of the items had 4 options. Fifteen of the items included an "I don't know" option.

Interviews with Science Educators

Three local science educators whose specialization was in the area of life sciences were identified and asked to help analyze items from a national science assessment instrument. These consultants included (a) a supervisor of science instruction for grades kindergarten through twelfth in a highly rated, well-to-do suburban district, (b) an experienced middle-school science teacher in an average suburban school district, and (c) a young,

junior high school science teacher in a troubled, urban school district. Thus, these educators had experience with very different student populations that might be expected to influence their judgements of item attributes.

The educators were interviewed individually by the senior researcher. First, the educators were given a self-test with the items to make sure they agreed with the designated correct answer. In order to focus attention on the level of knowledge required to answer a question, the educators were asked to describe what a student needed to know to answer an item correctly and whether the relevant knowledge was usually covered in classes in their school district and at what grade. Then they were asked to sort the items into the following six categories that constituted our scale of knowledge level:

1. Reading Comprehension or Problem Statement. All information required is provided in the item passage though general scientific knowledge might make the material or problem more comprehensible.

2. Popular. Most 13 year-olds would be likely to be exposed to the required knowledge through everyday experience.

3. Elementary (K - 3). Most children would first be exposed to the knowledge necessary to answer the question in the early elementary grades (Kindergarten through 3rd grade).

4. Elementary (4 - 6). Most children would first be exposed to the knowledge necessary to answer the question in grades 4 through 6.

5. Intermediate (7 - 8). Most children would first be exposed to the knowledge necessary to answer the question in grades 7 and 8.

6. Advanced. Items require understanding of more advanced concepts, knowledge of more specific detail or more depth of understanding than those at the previous levels.

Next, the educators were asked to rate how attractive they thought each distractor would be to their students on a scale of 1 (not attractive, not plausible, easily eliminated) to 5 (very attractive, very plausible, hard to distinguish from the correct answer). Finally, the educators were asked to estimate how difficult they thought an item would be overall for their students on a scale of 1 (very easy) to 5 (very difficult).

Each interview took 2 to 3 hours, and the educators were paid a consulting fee of $100.

Other Item Attributes

In addition to gathering information about the items from the interviews with science educators and from the NAEP test framework, we rated the items with respect to other attributes potentially relevant to item difficulty. These included text attributes, which should affect comprehension, such as the total number of words or syllables in the item passage/stem and in the

set of options, and whether the item included a figural material (illustrations, graphs, or tables). A computer program (Micro Power & Light, 1984) was used to obtain counts of the number of syllables, words, 3-syllable words, and sentences in an item's passage/stem and in the item's set of options. For items that included figural material, any labels or numbers included in the figures were entered as words. This computer program also calculated readability indices according to nine formulas. However, these indices were not used in the present study because of the questionable reliability of readability indices for "passages" as short as those found in this item set (Fry, 1990).

The researchers also classified each item according to a cognitive demand classification based on one developed by Emmerich (1989) and used by Scheuneman et al.(1991), in a modified form, in their study of the difficulty of GRE psychology items. Items were classified independently by two researchers into the following main categories and subcategories:

1. Restate given information -- depict, summarize, or translate;

2. Identify a correct piece of information not given - recall, define, exemplify, or clarify;

3. Analyze information -- explain, infer, generalize, simplify, problem-solve, evaluate, resolve, transfer, order, or organize;

4. Support or weaken a claim, procedure, outcome --

substantiate, constrain, or negate;

5. Synthesize components into a new pattern - organize, integrate, or reorganize.

Some attributes potentially relevant to response selection were also coded. These included the number of options, the inclusion of an "I don't know" option, and mean of the ratios of the number of words in the key to the number of words in each of the distractors.

A summary of the item attributes rated or coded in this study, organized in terms of a componential framework, is presented in Figure 2.

Results

Analysis was guided by three concerns that included evaluating the usefulness and appropriateness of ratings of item attributes and difficulty, determining how well item difficulty could be predicted, and establishing how well item difficulty could be decomposed or explained on the basis of item attributes.

Analysis of Item Attributes and Judgments of Difficulty

Ratings by science educators. The science educators rated items with respect to the level of knowledge needed to answer an item, the attractiveness of the distractors, and the overall difficulty of the item.

The level of knowledge invoked by an item was rated by the educators on a scale of 1 to 6. It became evident during the

interviews with the educators and from examination of the data that categories 1 (Reading comprehension) and 2 (Popular knowledge) were either inappropriately placed on this scale or did not belong on the same scale as categories 3 through 6 which related level of knowledge to curriculum and grade level. Agreement among pairs of raters with respect to the use of categories 1 and 2 was very low. Rater agreement, defined as two ratings for an item within +/-1 of each other, was 7% when category 1 was used by at least one rater and 0% when category 2 was used. In contrast, agreement ranged from 50% to 79% when categories 3 through 6 were used by at least one rater. Therefore, only ratings of 3 through 6 were included in the analysis and ratings of 1 and 2 were treated as missing data. Table 1 presents correlations, for the modified scale (3 to 6),

---

Insert Table 1 about here

---

between pairs of raters, of each individual's ratings with item difficulty, and of the mean rating over raters with item difficulty. For the modified scale, a mean rating for each item was calculated only if at least two of ratings were from 3 to 6. (There was only one item assigned a value of 1 or 2 by two of the 3 raters and thus excluded from the analysis.) Correlations among raters were positive but not very high. Nevertheless, the correlations between the individual educator's ratings and item

difficulty were moderate to good. In particular, the correlations between Rater 3's ratings of knowledge level and the mean rating over raters with item difficulty were quite good and account for about 36% - 38% of the variance in item difficulty. (This compares well with Scheuneman et al.'s report that a knowledge level measure accounted for 21% to 31% of the variance in the difficulty of GRE Psychology items.)

The science educators also rated the attractiveness of each distractor on a scale of 1 (not attractive, not plausible, easily eliminated) to 5 (very attractive, very plausible, hard to distinguish from the correct answer). The highest rating among the set of distractors, rather than the mean of the ratings, was used as the measure of distractor attractiveness for each item to distinguish items that had at least one very attractive distractor from those that a had number of equally but only moderately attractive distractors. The correlations among raters for this measure and its relationship to item difficulty are given in Table 1. Agreement was best between raters 1 and 2 and the correlation between ratings of distractor attractiveness and item difficulty were nearly equal for these two raters and accounted for about 20% to 22% of the variance in item difficulty. In contrast, the correlation between distractor attractiveness and item difficulty was very low for rater 3, whose ratings of level of knowledge had correlated the best with item difficulty.

Finally, the raters had also rated the overall difficulty of

each item on a scale of 1 to 5. As can be seen in Table 1, agreement among the raters was moderate. Individually, their judgments of item difficulty accounted for from 17% to 46% of the variance in actual item difficulty and their mean judged difficulty accounted for 52% of the variance in actual item difficulty.

One interesting aspect of the data in Table 1 is that raters appeared to be differentially adept at rating different kinds of information. Rater 3's coding of knowledge level and estimate of item difficulty correlated well with actual item difficulty, but her coding of distractor attractiveness was unrelated to item difficulty. In contrast, Rater 1's coding of knowledge level had only a moderate correlation with item difficulty while her coding of distractor attractiveness and estimate of item difficulty were better predictors of item difficulty.

Other item attributes. Other coded item attributes included textual complexity variables, cognitive demand characteristics, and some characteristics of the option set or response selection attributes. The correlations of these variables with item difficulty are presented in Table 2.

Insert Table 2 about here

The text attributes included counts of the numbers of words and sentences in the items as well as whether or not the item

included figural material. These measures of text complexity were calculated separately for the item passage/stem and for the set of options as a whole. Correlations between these measures and item difficulty were mostly positive but in the low to moderate range. Among this set of variables, the best predictor of item difficulty was the number of syllables in the set of options. Figural material appeared in approximately 30% of the items. Items were slightly more difficult when they included figural material than when they did not.

The type of cognitive demand implicit in an item was categorized by two of the authors into one of five main categories (synthesize, support or weaken a claim, analyze, identify, or restate) and associated subcategories. Initial agreement on the assignment to main categories was 80% and disagreements were resolved through discussion. In effect, only two of the main cognitive demand categories, identify and analyze, were found to be applicable to this set of items and about 57% of the items required some kind of analysis. The mean difficulty of items classified into these two categories and the associated subcategories is presented in Table 3. No systematic

---

Insert Table 3 about here

---

relationships between cognitive demand and item difficulty are evident. There was considerable overlap between the NAEP

cognitive classification of items and the categories used in this study. Almost all (96%) of the items assigned to the "analyze" category in this study were classified as "uses" or "integrates" in the NAEP scheme. Agreement was good, but not as high, for the other categories; 68% of items classified as "identify" in this study were classified as "knows" according to NAEP. The relationship between cognitive demand and item difficulty was trivial whether our classification system ($r = .05$) or the NAEP categories were used ($r = -.04$).

Among the option set attributes coded, the mean of the ratios of the number of words in the key to the number of words in each distractor had the strongest relationship to item difficulty. Items in which the key was shorter than the distractors on the average, were more difficult than those in which the key tended to be longer.

Predicting item difficulty

Clearly, the raters' judgments of item difficulty were the best single predictors of item difficulty, and Rater 3 was better at predicting item difficulty than the other raters. The next set of issues we explored was how to optimize prediction of item difficulty given the information we had gathered. One question we examined was whether to, or how to best combine judgments of difficulty by the three raters. To answer this question, we compared how well item difficulty could be predicted by the "best" rater, by a linear combination of the judgments of all three

raters, and by the mean of the judgments of all three raters. As shown in the first row of Table 4, the percent of variance in item difficulty accounted for by raters' judgments ranged from 46% to 52%. (The estimates in Table 4 are adjusted for the number of variables in the model and the number of items with missing data for any variables in the model.) Combining the judgments of all three raters accounted for 4% to 6% more of the variance than did the judgments of the best rater alone.

Insert Table 4 about here

Our second set of questions concerned whether combining information about item attributes from other sources with raters' judgments would improve the prediction of item difficulty. Separate multiple regressions were conducted combining raters' judgments of item difficulty with information about the items' text attributes, cognitive demand, and option set attributes. From each set of item characteristics, those with the highest correlations with item difficulty in the preliminary analysis were selected for inclusion in the regressions. The text attributes included the number of syllables in the passage/stem and the options, and the presence of figural material. The cognitive demand attribute was included because of its theoretical interest. And the option set attributes included the mean ratio of the number of words in the key and the distractors and whether there

was an "I don't know" option in the item. The results of these analyses are also presented in Table 4. We were able to account for up to 62% of the variance in item difficulty by combining raters' judgements of difficulty with other item attributes. The option set attributes resulted in improvements on the order of 7% to 15% of the variance in the prediction of item difficulty when combined with raters' judgments of item difficulty. In general, smaller improvements were found when text attributes were included, and, as might be expected from the preliminary analysis, no improvement was found when cognitive demand attribute was added. Including both text and option set attributes in the model did not improve prediction more than option set characteristics alone.

Decomposing item difficulty

For purposes such as test development and design, construct validation, comparison of different tests, or equating tests from a psychological perspective, understanding item difficulty is more important then predicting item difficulty. Therefore our next set of questions concerned how well we could decompose item difficulty. Our strategy here was to examine how well we could account for item difficulty in terms of discrete item attributes and without using global judgments of difficulty. Thus, instead of including the raters' judgments of difficulty in the regression models, we included their judgments of knowledge level and distractor attractiveness as well as other item characteristics.

The results of these analyses are presented in Table 5.

Insert Table 5 about here

The best model accounted for 53% of the variance and included information about the items' text and option set characteristics as well as the raters' judgments of knowledge level and distractor attractiveness. Once again, cognitive demand did not contribute very much to the prediction of item difficulty.

Discussion

Developing alternative sources of information about item difficulty has many implications for test development. From a practical point of view, alternative information relevant to item difficulty may reduce the need for pretesting (Mislevy et al., 1992) though it is not likely to replace it (Thorndike, 1982). Understanding what makes items difficult will also contribute to more systematic and principled test design, more meaningful test interpretation, and better construct validation (Bejar, 1991; Embretson & Wetzel, 1987).

In this study we investigated if information about item attributes, obtained from a number of sources including test specifications, expert opinion, and experimenter analysis, was useful in predicting the difficulty of items from a survey of science achievement. We found that global judgments of item difficulty by individual science educators could account for 17%

to 46% of the variance in actual item difficulty. Judgments of item difficulties, pooled over raters, accounted for 52% of the variance, despite the fact that agreement among the raters was not very high. This level of prediction compares well with reports that trained raters could predict 55% to 71% of the variance on aptitude tests (Thorndike, 1982) and that experienced item writers could account for 52% of the variance in item difficulty for analytical reasoning items (Chalifour & Powers, 1988), and 43% of the variance in item difficulty for analogy items (Enright & Bejar, 1989). In the current study, prediction of item difficulty improved to approximately 60% of the variance when pooled judgments of item difficulty were combined with selected information about attributes related to text and option set characteristics.

It should be noted, however, that the level of agreement among raters was not particularly high in the current study. There are a number of ways that rater agreement could be improved in future studies, including increasing the number of raters, training raters, or selecting only raters who demonstrate an ability to predict item difficulty well. However, by focusing on reliability as a standard, diversity among the perspectives and experiences of the raters might be attenuated. Furthermore, in this study, raters appeared to be differentially adept at rating different kinds of information. The issues of how accurate raters are at evaluating different kinds of information, and how

different raters combine this information to estimate item difficulty are probably more important topics for further study than are attempts to improve rater agreement.

Another issue we explored in this study was the extent to which item difficulty could be accounted for or explained by (in a statistical rather than causal sense) discrete item attributes rather than global judgments of item difficulty. These results have a number of implications for the construct interpretation of this test. The best model, which accounted for 53% of the variance in item difficulty, included information about the level of knowledge assessed by the item, the characteristics of the text, and the option set, but not information about the cognitive demand of the item. Of these attributes, level of knowledge, which alone accounted for 38% of the variance, appeared to be most important. This result is not surprising in that achievement tests are supposed to measure the acquisition of knowledge. However, analysis of the knowledge required to answer test items (as we have defined it) is seldom a part of the test development or test validation process. Although the measure of knowledge used in the present study was relatively unsophisticated, these results indicate the importance of this factor and suggest that more rigorous investigations of knowledge structure and acce sibility should be conducted.

The fact that the text characteristics of the items accounts for some, but not a disproportional share, of the variance in item

difficulty is also reassuring and suggests that the test did not simply measure comprehension. The text characteristics used to predict item difficulty accounted for about 8% of the variance beyond that accounted for by knowledge level and distractor attractiveness.

The results concerning option set characteristics are harder to interpret. Two such attributes, distractor attractiveness and the mean of the ratios of the number of words in the key and each distractor, contributed to the prediction of item difficulty, and the latter attribute appeared to be more important. Making fine conceptual distinctions between possible responses would be an appropriate construct-relevant source of item difficulty, but making distinctions among possible responses on the basis of length is a construct-irrelevant source of variance. However, we do not know if the raters' judgments of distractor attractiveness in this study reflected fine conceptual distinctions or other factors. Thus the implications of the results related to response selection-set attributes for construct validity are, at best, ambiguous. At worst, they suggest that the multiple-choice format, in this case, is a source of construct irrelevant variance.

We found no evidence that the items' cognitive demands, as defined in this study, were related to item difficulty. This suggests a number of issues that deserve further exploration including how "cognitive demand" should be defined, and whether we

should expect it to be directly related to item difficulty. The notion of cognitive demand embedded in the 1985-86 NAEP assessment framework in science and used also in this study was influenced by Bloom's taxonomy of educational objectives (1956). A great deal of research on the nature of achievement, expertise, and problem-solving (for summary see Chi, Glaser, & Farr, 1988) has been carried out since Bloom's taxonomy was developed and might serve as the basis for a reevaluation of how the concept of "cognitive demand" can be refined in the context of education and assessment. Furthermore, the fact that "cognitive demand" did not predict item difficulty well cannot be taken as evidence that cognitive demand is unimportant in other respects. A problem here is that we do not have a well-articulated theory of achievement that would allow us to specify how factors such as knowledge level or cognitive demand should relate to item difficulty.

Understanding what makes items difficult is an important component of the construct validation process and has implications for test design and development. Overall, this study produced evidence of the importance of knowledge level and option set characteristics in predicting item difficulty on a national science achievement test. However, the present study was limited in a number of respects, and these limitations suggest further directions for research. First, this study was not an exhaustive exploration of all the factors that contribute to item difficulty, and other characteristics should be explored in further research.

For example, one characteristic that was not included in the present study and that is likely to be very important in science assessment is the level of the vocabulary used in the items. Secondly, this study was exploratory and correlational in nature. One of the greatest potential benefits to be derived from an understanding of item difficulty would be the systematic and principled development of items with known psychometric characteristics (Bejar, 1991; Embretson & Wetzel, 1987). Thus, it remains to be seen if items that are written explicitly to take into account factors identified as important in exploratory studies would achieve an expected level of difficulty. Finally, items can be analyzed cognitively from two perspectives. The perspective taken in this research focused on identifying what problem attributes contributed to problem difficulty. A complementary, alternative perspective is one that describes the attributes of examinees' performance. These perspectives need to be coupled in further research because performance is a result of an interaction between an individual and a problem and needs to be understood in light of both the knowledge and skills the individual brings to the situation and the nature of the demands imposed by the problem. Individuals who get a particular problem wrong may do so for a variety of reasons. Similarly, problems that are equal in difficulty are not necessarily difficult because of identical factors. Describing the varied factors that contribute to problem difficulty and to proficient performance is

an important way of evaluating the construct validity of tests. In addition, a more detailed understanding of the characteristics of problems and performance is critical if tests are to be used to provide more helpful descriptive or diagnostic information to test users.

References

Beaton, A. E. (1988). Expanding the new design: The NAEP 1985-86 technical report. (NAEP Report No. 17-TR-20). Princeton, NJ: Educational Testing Service.

Bejar, I. I. (1985). Speculations on the future of test design. In S. E. Embretson (Ed.), Test design: developments in psychology and psychometrics. Orlando, Fl: Academic Press.

Bejar, I. I. (1991). A generative approach to psychological and educational measurement. (ETS Report No. RR-91-20). Princeton, NJ: Educational Testing Service.

Bloom, B. S., (Ed.). (1956). Taxonomy of educational objectives. New York: Longmans, Green & Co.

Chalifour, C., & Powers, D. E. (1988). Content characteristics of GRE analytical reasoning items. (GRE Professional Report No. 84-14P, ETS Research Report 88-7). Princeton, NJ: Educational Testing Service.

Chi, M. T. H., Glaser, R., & Farr, M. J. (Eds.) (1988). The nature of expertise. Hillsdale, NJ: Erlbaum.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. Psychological Bulletin, 93, 179-197.

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. Applied Psychological Measurement, 11, 175-193.

Emmerich, W. (1989). Appraising the cognitive features of subject tests. (ETS Report RR-89-53). Princeton, N.J.: Educational Testing Service.

Enright, M. K., & Bejar, I. I. (1989). An analysis of test writers' expertise: modeling analogy item difficulty. (ETS Research Report No. RR-86-35). Princeton, N.J.: Educational Testing Service.

Fry, E. (1990). A readability formula for short passages. Journal of Reading, 33, 594-597.

Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. American Psychologist, 36, 923-936.

Glaser, R., Lesgold, A., & Lajoie, S. (1987). Toward a cognitive theory for the measurement of achievement. In R. R. Ronning, J. A. Glover, J. C. Conoley, & J. C. Witt (Eds.), Buros-Nebraska symposium on measurement and testing: Vol. 3. The influence of cognitive psychology on testing. Hillsdale, NJ: Lawrence Erlbaum.

Hartwig, M. D. (1989, February 22). Better testing key to better science learning. The Wall Street Journal, p.16.

Kirsch, I. S., & Mosenthal, P. B. (1988). Understanding document literacy: Variables underlying the performance of young adults. (ETS Research Report RR-88-62). Princeton, N. J: Educational Testing Service.

Messick, S. (1984). Educational achievement testing: The assessment of dynamic cognitive structures. In B. S. Plake (Ed.), Social and technical issues in testing. Hillsdale, N. J: Lawrence Erlbaum.

Micro Power & Light Co. (1984). Readability calculations according to nine formulas. (Computer program). Dallas, TX: Micro Power & Light Co.

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. (1992). How to equate tests with little or no data. (ETS Research Report RR-92-20-ONR). Princeton, N. J: Educational Testing Service.

Mulholland, T., Pellegrino, J. W., & Glaser, R. (1980). Components of geometric analogy solution. Cognitive Psychology, 12, 252-284.

NAEP. (1985-86). Science objectives: 1985-86 assessment. (Objectives Booklet No. 17-S-10). Princeton, N. J: Educational Testing Service.

Scheuneman, J., Gerritz, K., & Embretson, S. (1991). Effects of prose complexity on achievement test item difficulty. (ETS Research Report RR-91-43). Princeton, N. J: Educational Testing Service.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Fredericksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), Diagnostic monitoring of skill and knowledge acquisition. Hillsdale, NJ: Erlbaum

Thorndike, R. L. (1982). Item and score conversion by pooled judgment. In P. W. Holland & D. B. Rubin (Eds.), Test equating (pp.309-326). New York: Academic Press.

Table 1

Interrater Correlations for Ratings of Items' Knowledge Level, Distractor Attractiveness, and Difficulty and Correlations of Raters' Judgments with Actual Item Difficulty

	Rater			
	1	2	3	Mean of Raters
Knowledge Level (Modified Scale 3-6; n=43)				
Rater				
2	.26			
3	.31+	.26		
Actual Item Difficulty	.30+	.42**	.60***	.62***
Distractor Attractiveness (n=44)				
Rater				
2	.48***			
3	.29+	-.06		
Actual Item Difficulty	.47***	.45**	.04	.47**
Judged Item Difficulty (n=44)				
Rater				
2	.23			
3	.44**	.29+		
Actual Item Difficulty	.50***	.41**	.68***	.72***

+p <.10. *p <.05. **p <.01. ***p <.001.

Table 2

Correlations of Other Item Attributes with Item Difficulty

Attributes	r
Text	
Stem	
No. of words	.14
No. of 3 syllable words	.11
No. of sentences	.15
No. of syllables	.15
Options	
No. of words	.33*
No. of 3 syllable words	.19
No. of sentences	.37*
No. of syllables	.20
Presence of figure	.20
Processing	
Cognitive Demand	.05
Option set	
Mean Key/Distractor Ratio	-.27+
"I don't know" Option	.12
Total Number of Options	.05

+$p < .10$. *$p < .05$.

Table 3

Mean Difficulty for Items in Each Cognitive Demand Category

Cognitive Demand Category	n	Mean	Standard Deviation
Identify			
Define	1	1.05	--
Exemplify	3	1.19	1.60
Recall	15	.51	.83
Total	19	.65	.94
Analyze			
Explain	5	1.18	.94
Generalize	3	-.59	.91
Infer	13	.98	.96
Order/Organize	1	1.26	--
Problem Solve	3	.26	.31
Total	25	.76	1.01

Table 4

Adjusted R^2 for Prediction of Item Difficulty
by Raters' Judgments of Item Difficulty and Other Item Attributes

	Best Rater	$R_1+R_2+R_3$	Mean of Raters
Raters Judgment of Item Difficulty	.46	.52	.50
+ Other Item Attributes			
Text Attributes	.44	.55	.57
Cognitive Demand	.44	.51	.50
Option set Attributes	.61	.62	.57
Text & Option set Attributes	.59	.60	.58

Note. R^2 is adjusted for the number of variables and the number of items with missing data for any variables in the model.

Table 5

Decomposing Item Difficulty:
Estimated Regression Parameters and Adjusted R^2 Values

	Alternative Models				
	1	2	3	4	5
Intercept	-4.64	-4.92	-4.67	-4.37	-4.10
Partial regression coefficients for:					
Rater Judgments					
Knowledge Level	.87***	.85***	.87***	.94***	.85***
Distractor Attractiveness	.27+	.28*	.27+	.23	.21
Text Attributes					
Syllables in passage		-.00			-.00
Syllables in option		.01*			.01**
Figural material		.16			-.03
Cognitive Demand			.02		
Response Attributes					
Mean key/ distractor ratio				-.52+	-.59+
"I don't know"				.27	.33
df	(2,40)	(5,38)	(3,39)	(4,36)	(7,31)
Adjusted R^2	.41	.49	.39	47	.53

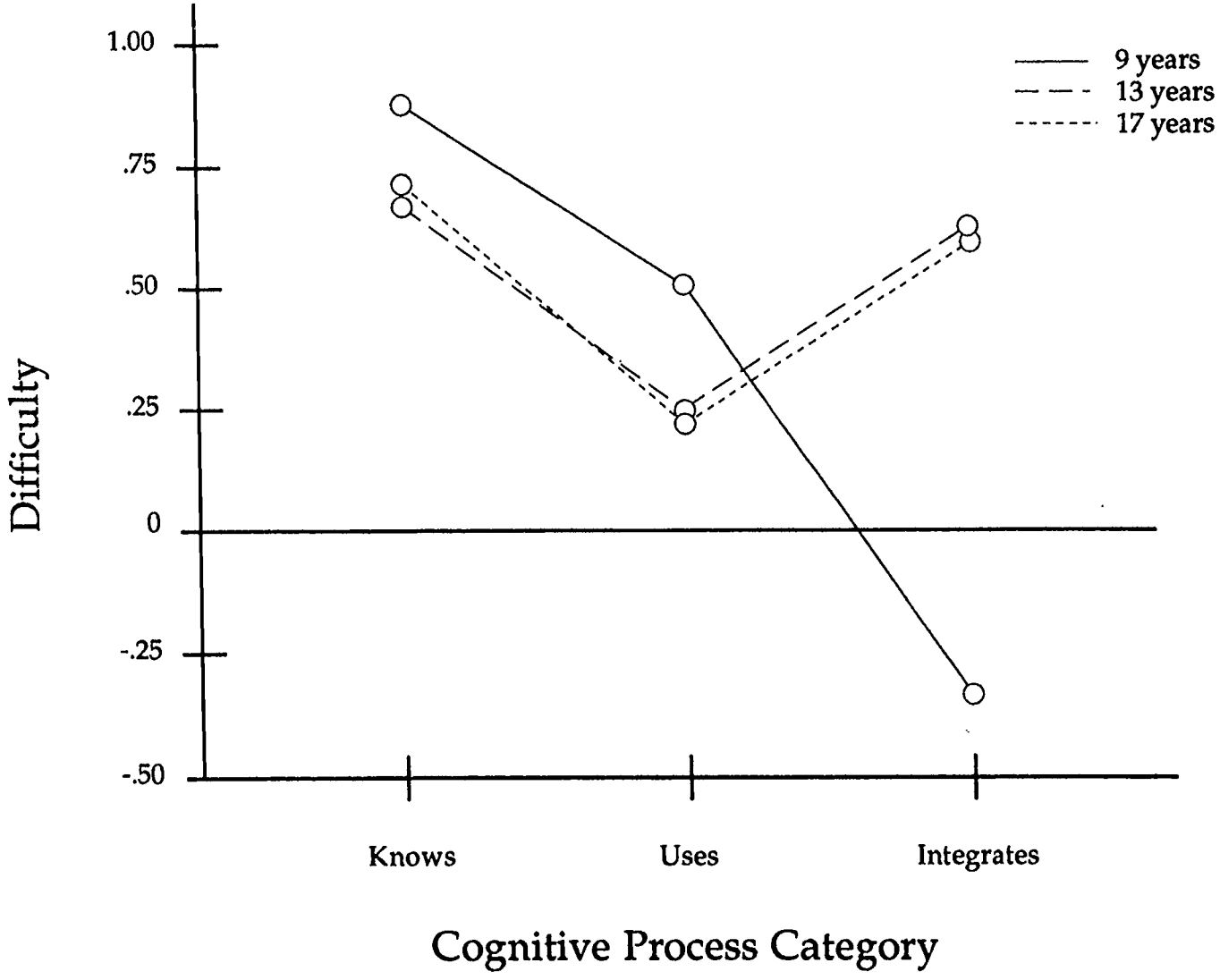
Note. R^2 is adjusted for the number of variables and the number of items with missing data for any variables in the model.

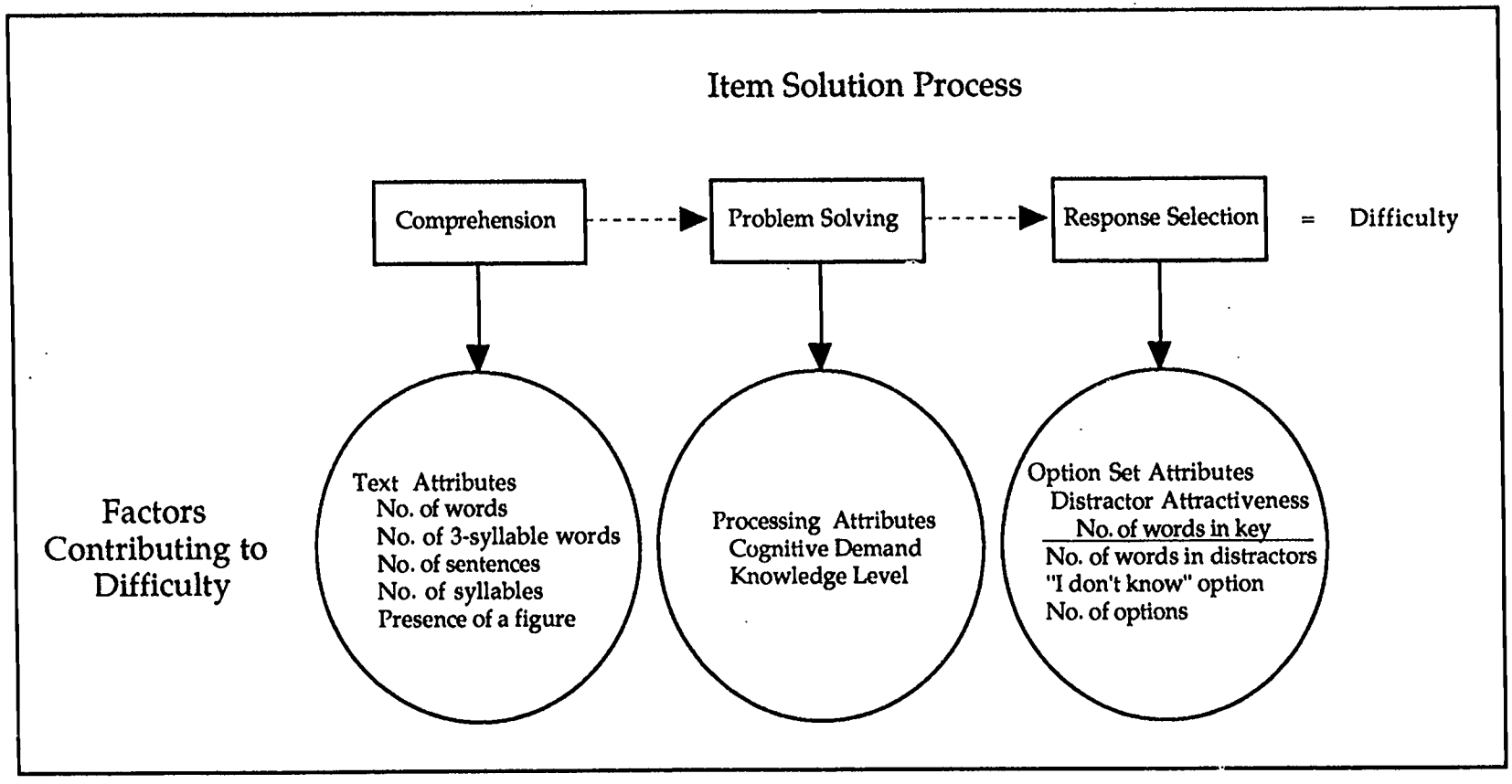
+p <.10. *p <.05. **p <.01. ***p <.001.

Figure Captions

Figure 1. Mean difficulty (IRT b) of items on the 1986 NAEP Life Sciences Subscale for three age groups by three cognitive process category.

Figure 2. Framework for organizing item attributes.


1.00
.75
.50
.25
0
-.25
-.50
Difficulty
9 years
13 years
17 years
Knows
Uses
Integrates
Cognitive Process Category


Item Solution Process
Comprehension
Problem Solving
Response Selection
= Difficulty
Factors Contributing to Difficulty
Text Attributes
No. of words
No. of 3-syllable words
No. of sentences
No. of syllables
Presence of a figure
Processing Attributes
Cognitive Demand
Knowledge Level
Option Set Attributes
Distractor Attractiveness
No. of words in key
No. of words in distractors
"I don't know" option
No. of options