DOCUMENT RESUME

ED 378 570 CS 011 992

AUTHOR	Binkley, Marilyn, Ed.; And Others
TITLE	Methodological Issues in Comparative Educational Studies: The Case of the IEA Reading Literacy Study.
INSTITUTION	Westat, Inc., Rockville, MD.
SPONS AGENCY	National Center for Education Statistics (ED), Washington, DC.
REPORT NO	NCES-94-469
PUB DATE	Jan 95
NOTE	293p.
PUB TYPE	Collected Works - General (020)
EDRS PRICE	MF01/PC12 Plus Postage.
DESCRIPTORS	*Comparative Analysis; *Cross Cultural Studies; *Data Interpretation; Global Approach; Grade 4; Grade 9; Intermediate Grades; Junior High Schools; Methods Research; *Reading Achievement; *Reading Research; Research Design; *Research Methodology
IDENTIFIERS	*Cross National Studies; International Assn Evaluation Educ Achievement

ABSTRACT

This report discusses various methodological issues confronted in the Reading Literacy Study conducted under the auspices of the International Association for the Evaluation of Educational Achievement (IEA) and issues relating to analysis of the data. The study analyzed in the report involved fourth- and ninth-grade students (9-year-olds and 14-year-olds) in 32 countries. Chapters in the report are: (1) "Issues in Sampling for International Comparative Studies in Education: The Case of the IEA Reading Literacy Study" (Keith Rust); (2) "Estimation, Sampling Errors, and Design Effects" (Edward Bryant); (3) "Handling Item Nonresponse in the U.S. Component of the IEA Reading Literacy Study" (Marianne Winglee and others); (4) "Assessing the Dimensionality of the IEA Reading Literacy Data" (Nadir Atash); (5) "Exploring the Possibilities of Constructed-Response Items" (Barbara Kapinus and Nadir Atash); (6) "Interpreting the IEA Reading Literacy Scales" (Irwin S. Kirsch and Peter B. Mosenthal); (7) Creating a Measure of Reading Instruction" (Marilyn R. Binkley and others); (8) "Hierarchical Models: The Case of School Effects on Literacy" (Steve W. Raudenbush); and (9) "Synthesizing Cross-National Classroom Effects Data: Alternative Models and Methods" (Steven W. Raudenbush and others). Contains 30 references. An appendix presents empirical Bayes and Bayes estimation theory for two-level models with normal errors. (RS)

National Center for Education Statistics

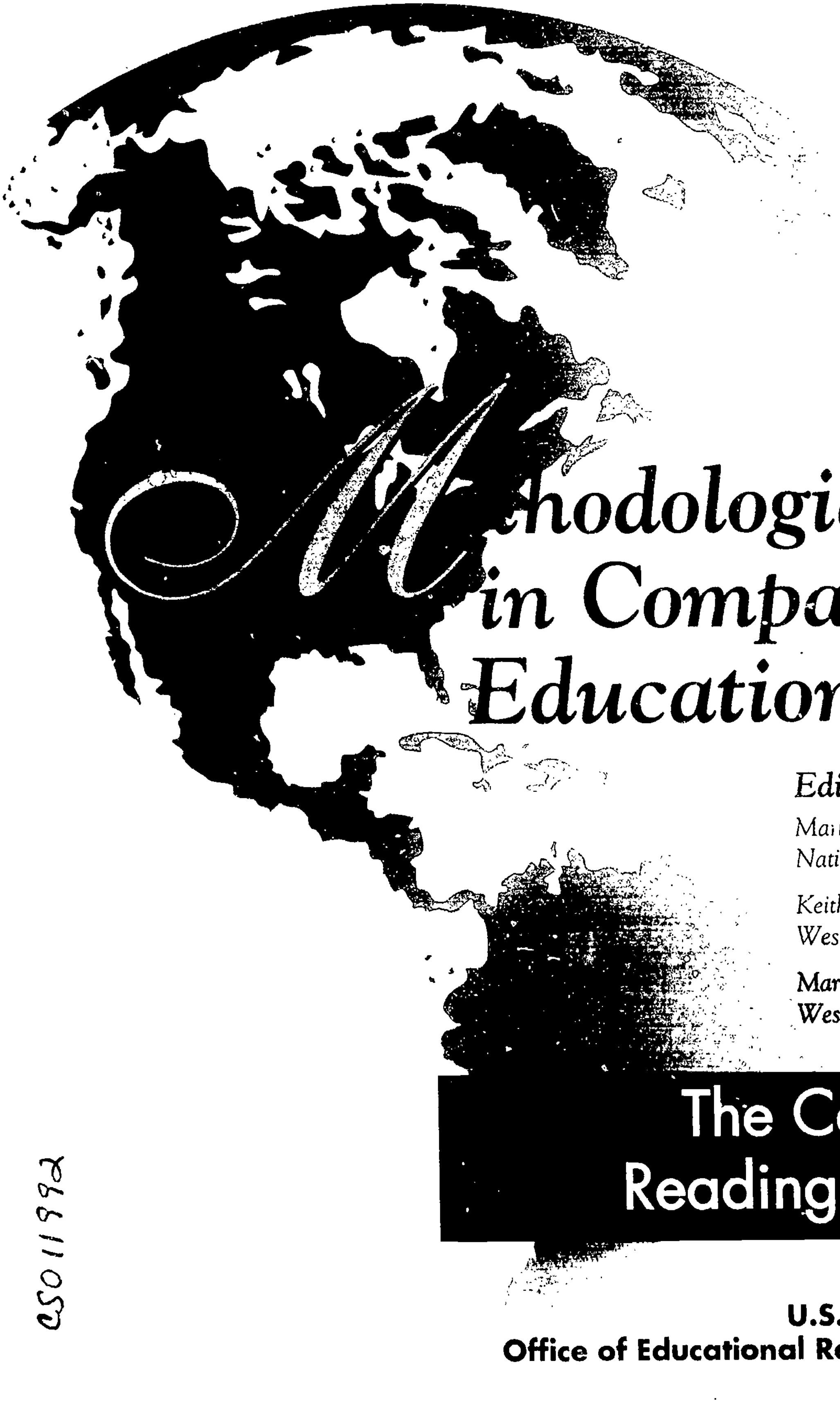
Methodological Issues in Comparative Educational Studies

The Case of the IEA Reading Literacy Study

U.S. Department of Education
Office of Educational Research and Improvement

NCES 94-469

National Center for Education Statistics



Methodological Issues in Comparative Educational Studies

Editors:

Marilyn Binkley
National Center for Education Statistics

Keith Rust
Westat, Inc.

Marianne Winglee
Westat, Inc.

The Case of the IEA Reading Literacy Study

U.S. Department of Education
Office of Educational Research and Improvement

U.S. Department of Education
Richard W. Riley
Secretary

Office of Educational Research and Improvement
Sharon P. Robinson
Assistant Secretary

National Center for Education Statistics
Emerson J. Elliott
Commissioner

Data Development Division
Jeanne E. Griffith, *Associate Commissioner*

International Activities Group
Eugene Owen, *Chief*

National Center for Education Statistics

"The purpose of the Center shall be to collect, and analyze, and disseminate statistics and other data related to education in the United States and in other nations." - Section 406(b) of the General Education Provisions Act, as amended (20 U.S.C. 1221e-1).

Suggested Citation

U.S. Department of Education, National Center for Education Statistics, *Methodological Issues in Comparative Educational Studies: The Case of the IEA Reading Literacy Study*. Washington, D.C.: 1994.

January 1995

Technical Editors: Barbara Brickman and Carol Litman, Westat, Inc.
Cover Design: Ana Horton, Luann Moore, Westat, Inc.

PREFACE

In 1991, 32 countries participated in a study to evaluate the reading literacy skills of their school students and to assess factors thought to be related to Reading Literacy. The study was conducted under the auspices of the International Association for the Evaluation of Educational Achievement (IEA). Two populations of students were assessed in the study: those in the grade with the most 9-year-old students, (Population A, grade 4 in the United States), and those in the grade with the most 14-year-old students (Population B, grade 9 in the United States). Most of the countries involved, including the United States, participated at both populations. The IEA published an initial set of results for all countries in 1992 (*How in the World do Students Read*, by W.B. Elley), and has subsequently published two other volumes directed at specific topics of interest (*Effective Schools in Reading*, by N. Postlethwaite and K. Ross, and *Teaching Reading Around the World*, by E. Lundberg and P. Linnakyla).

While analyses and reports were being carried out by the IEA, within the United States the National Center for Education Statistics (NCES) sponsored an intensive and extensive analysis of the U.S. national data. NCES also initiated a number of comparative studies in partnership with the study representatives from a number of European countries and instigated some cross-national analyses of the data from many countries aimed at comparing the relationships of certain factors to educational achievement. The findings of the NCES studies are being released in a set of four reports. A comprehensive technical report (*Reading Literacy in the United States: Technical Report)* covers the conduct of the study within the United States and the methods of analysis employed with the U.S. data on students, teachers, and schools. A more general report (*Reading Literacy in the United States: Findings from the IEA Reading Literacy Study)* describes the findings from these analyses. A third report *(Reading Literacy in an International Perspective)* describes the results of the cross-national investigations discussed above. This methodological report completes the set of four.

This report contains nine distinct chapters that discuss both the various methodological issues confronted in the conduct of the study and the analysis of the data. The chapters were written by individuals with extensive experience in the general methodological area they discuss, and who dealt with the specific issues first hand for the U.S. portion of the Reading Literacy Study. A perusal of the topics covered in this volume will reveal that many of them are relevant not only to the Reading Literacy Study and to other comparative educational studies, but, in fact, to educational assessment and sociological surveys in general. For example, the question of how best to define a target population for a survey, and the importance of doing this carefully and precisely, are covered in Chapter 1. The definitions of target populations are always relevant in survey research, but the failure to address them adequately can have particularly severe consequences in an international comparative study. The process of how best to develop a parsimonious and useful multilevel model with many potential predictor variables is covered in Chapter 8. This question was central to the philosophy and technical approach that the U.S. team, led by Trevor Williams of Westat, used in conducting a multivariable analysis of the U.S. data from the study. It is equally important in any investigation of data with a hierarchical structure.

The nine chapters of this report can be broadly classified into three categories. The first three chapters address various aspects of the survey design and preparations of the data for analysis. In Chapter 1, Rust discusses issues of the definition of the target population for a multinational study of educational achievement and how to operationalize the definition. The chapter also discusses the choice of sampling unit in a multistage design of school students: specifically, the choice of whole classrooms of students as the ultimate sampling units. In Chapter 2, Bryant discusses the impact of the sample design on the precision of estimation (design effects), how to estimate this precision via the jackknife method of variance estimation, and the reliability of such sampling error estimates. In the third chapter, Winglee and colleagues describe the procedures used to impute for missing responses to survey items in the Student, Teacher, and School Questionnaires. This imputation was intended to maintain statistical power in multivariable statistical analyses and to reduce the influence of item nonresponse as a source of bias in estimation. An evaluation of the imputations conducted for the U.S. data is included.

The next three chapters deal with different aspects of the literacy assessment instruments and the abilities that the instruments purport to measure in the U.S. context. In Chapter 4, Atash addresses the question of whether the three literacy scales used to report the results (narrative, expository, and document) do in fact behave as three distinct dimensions of literacy (the related question as to whether each of the scales is itself composed of just a single dimension is addressed partially in Chapter 10 of the U.S. Technical Report). In Chapter 5, Kapinus and Atash examine several properties of items in the assessment that required an extended written response from the student. In Chapter 6, Kirsch and Mosenthal provide an analysis of the assessment items themselves, with the aim of revealing the sources of differences in the level of difficulty of multiple-choice response items, which constituted the majority of the assessment material. Readers who wish to look at the assessment instruments are referred to the Technical Report for the study. All the instruments have been reproduced in the Attachments to that publication, which is available from the U.S. Government Printing Office.

The final three chapters address issues related to the analysis of data from the Reading Literacy Study. In Chapter 7, Binkley addresses an important aspect of the data collected from teachers, via a Teacher Questionnaire, as part of the study. This is the issue of distilling information about teachers' instructional practices and beliefs from a large set of questions that ask teachers about their opinions and behaviors with regard to a variety of specific circumstances. Two chapters, Chapter 8 by Raudenbush, and Chapter 9 by Raudenbush and colleagues, discuss issues involved in developing multilevel, or hierarchical, linear models to describe the relationships between students' reading literacy capabilities and the characteristics of students, classrooms, teachers, and schools. Chapter 8 discusses the theory behind the application of these models to the Reading Literacy Study for the United States. In particular, the chapter describes the approach to reducing the set of available student, teacher, classroom, and school variables to produce a parsimonious, useful, and defensible model of reading literacy. Chapter 9 develops a methodology for applying three-level models to analyze data across countries. This approach goes beyond the "standard" approach to three-level models. It implicitly recognizes the uniqueness of each contributing country and provides the means to investigate the similarities and differences of influences on literacy within the different participating countries.

The topics covered in this methodological volume go "behind the scenes" of the methods presented in the technical report to describe and evaluate approaches to various technical issues and problems encountered in the design and analysis of the study. The purpose of documenting these efforts in this way was at least to open up the discussion of the options and concerns considered during the conduct of the study, and perhaps to provide some insight and guidance to those involved in future comparative studies of educational achievement.

ACKNOWLEDGMENTS

The volume *Reading Literacy in the United States: Technical Report of the U.S. Component of the IEA Reading Literacy Study*, released in August 1994, acknowledged the contributions of the many people whose special efforts helped to make the U.S. component of the IEA Reading Literacy Study a success. Obviously that same assistance from these individuals contributed to our ability to produce this methodological volume from the same project, and we owe each of them our thanks.

At this point, however, it seems appropriate to recognize particularly those individuals who have made a special contribution to the manuscripts included in this present volume, over and above the many important general contributions that have been made to the project.

Support and encouragement for the concept of a special methodological volume arising from the study was provided by Jeanne Griffith, Associate Commissioner for Data Development, and Eugene Owen, Chief of the International Activities Group, at the National Center for Education Statistics. Their foresight enabled this unique volume to become a reality.

Several experts from across the country provided invaluable comments on drafts of various manuscripts in this volume. Our thanks go to James Guthrie, Edward Haertel, Sylvia Johnson, David Judkins, Scott Paris, and Ramsay Selden for their advice and suggestions.

There were many at the National Center for Education Statistics who made special efforts to provide suggestions for improvements to earlier versions of these manuscripts. Susan Ahmed, Robert Burton, and Mary Frase in particular gave considerable advice and assistance, and we thank Nabeel Alsolom, Sharon Bobbitt, Peggy Carr, Paula Knepper, and Andrew Kolstad for their contributions.

Assembling a volume of manuscripts from authors at widely scattered sites, each using various pieces of the full Reading Literacy Study dataset, demanded considerable skill and persistence. Our thanks go to those at Westat who undertook this task carefully and efficiently. Stephen Roey successfully undertook the complicated process of meeting the various specialized data requirements, and he conducted a number of the analyses. Barbara Brickman and Carol Litman provided expert editorial assistance, unifying an initially fragmented collection of manuscripts into one volume. Document preparation was carefully carried out by Susan Locke, Lisa Moore, Ann Price, Debra Reames, and Sylvie Warren.

Finally, our thanks go to all of the authors for their patience during the complex and time-consuming process of turning their manuscripts into a collected volume. We think that the effort was worthwhile; we hope that they do.

TABLE OF CONTENTS

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

TABLE OF CONTENTS (continued)

1 Issues in Sampling for International Comparative Studies in Education: The Case of the IEA Reading Literacy Study

Keith Rust

1.1 Introduction

This chapter deals with sampling issues for international comparative studies in education, with a focus on the IEA Reading Literacy Study as implemented in the United States. The IEA Reading Literacy Study was conducted in 1991 by the International Association for the Evaluation of Educational Achievement (IEA). A total of 32 countries (or parts of countries) participated. In most countries, sampled students from two grade levels (one grade only in some countries) took an assessment in reading literacy. Apart from translation into the appropriate national language, the assessments were essentially identical across countries for each of the two student populations sampled. A report of the results of the study is presented in Elley (1992).

In designing a study that can provide data for a cross-national comparison of student literacy, a number of key design issues had to be addressed. These same issues have been faced by those involved in all international comparative education studies conducted during the past 25 years. First, there are issues of instrument, assessment administration, scoring, and scaling comparability across countries, and especially across languages. Then there are questions of the definition of the populations to be surveyed and the procedures used to ensure that appropriate valid samples of those populations were drawn. Finally, there are the technical issues related to the sample design and quality control of the sampling process to ensure that the samples drawn, when subjected to analysis, will yield valid and reliable estimates of the characteristics and parameters of interest.

We focus on the second and third issues: the implications of the population definitions and sample design requirements when implemented in the United States. For a discussion of issues relating to multinational comparability of the assessment instruments and conditions, scoring, and scaling, the reader should consult the forthcoming international technical report on the study to be published by IEA.

For the IEA Reading Literacy Study, the samples of students in each country were drawn under the supervision of an International Sampling Coordinator (Dr. Kenneth Ross of Deakin University, Australia), following the directions provided in a sampling manual that he authored (Ross 1991). Control of the sampling process was a vital component of the effort to obtain results from the study that were comparable across countries and useful to those in each country responsible for educational policy and reading instruction. In each country, a target population (or populations) was established. This target was defined in terms of the school grade of the students; the appropriate grade for each country was based on the age distribution of the student population across grades. Each country then prepared a stratified list of schools with the relevant grade. A sample of schools was drawn (between 100 and 200 per population), typically with probability proportional to size within strata. Then a sample of intact classrooms was drawn from within each selected school, generally one classroom per school. All students within selected classrooms were to be assessed in the study.

This chapter addresses a number of issues related to the definition of the target population for each country and the population actually assessed, as well as the choice of sampling procedure and its impact on the analysis of the data. Section 1.2 discusses the method of defining a suitable target population for an international comparative study, its practical effects on the population actually surveyed in each country, and the comparability across countries. Section 1.3 discusses issues of exclusion of students from the target population within each country, intentional or otherwise. The importance of carefully reporting the sources and rates of exclusion is emphasized. Sections 1.4 and 1.5 introduce the topic of the choice of an appropriate multistage sampling procedure and its relationship to the types of analyses that are to be performed.

1.2 The Definition of a Target Population

In attempting to define a suitable target population within each participating country that will lead to the most useful comparisons across countries, designers of international comparative studies attempt, explicitly or implicitly, to strike a balance among three educational components that are sometimes in conflict. They must also be constrained by practical concerns in the implementation of designs to sample the target population. The three components can be summarized as age, grade, and curriculum.

In comparing students across countries, it is desirable to assess students of similar mental, physical, and emotional maturity; that is, students should be of the same age. At the same time, it is desirable to compare students with similar amounts of formal education; that is, students should be in the same grade. Finally, it is desirable that students should have received exposure to a comparable breadth and depth of material in the subject being assessed; that is, the curriculum to which students are supposed to have been exposed should be similar from country to country. In the ideal case these three requirements converge to give rise to a natural definition of the target population in each country, but in reality they often conflict to some extent and a choice must be made.

Target Populations for the IEA Reading Literacy Study

For the IEA Reading Literacy Study, the primary component used to define the target population in each country was age. Two populations were defined for the study: one of students at about 9 years of age, and one of students at about 14 years of age. Having two populations may have helped to overcome some of the problems of comparability, since if the 9-year-olds in two countries are not really comparable with respect to grade and curriculum, perhaps the 14-year-olds might be, and vice versa.

For two reasons, however, the definitions of the target populations adopted were not just in terms of age. The first was that defining students by a given age would mean that in most countries two or more grades would have to be sampled. This leads to practical difficulties in sampling and especially administration, and exacerbates the problem of grade comparability across countries. The second, related issue was the decision to sample intact classrooms of students. This choice of sampling unit, and its consequences for analysis, are discussed below, but clearly the use of intact classrooms meant that a target population could not be defined in terms of age alone.

It should be noted that it is feasible to define target populations for international studies in terms of age alone. The two International Assessments of Educational Progress (IAEP) studies have done just this (Lapointe, Mead, and Phillips 1989; Lapointe, Mead, and Askew 1992). These studies have used a simple age definition to define the target population, and both have successfully handled the sampling and administrative issues associated with this approach. With these studies, however, it is more difficult to analyze teacher and classroom effects than is the case for studies where whole classrooms are sampled.

The approach used for the IEA Reading Literacy Study was to define the target population for each country in terms of grade, but with the choice of grade being determined by the age distribution of the students. Thus the two target populations were defined as follows: Population A consisted of the students in the grade level containing the most 9-year-olds; Population B consisted of students in the grade level containing the most 14-year-olds. The formal definitions adopted are presented in appendix D of Elley (1992):

Population A: All students attending school on a full-time basis at the grade level in which most students were aged 9:00-9:11 years during the first week of the eighth month of the school year.

Population B: All students attending school on a full-time basis at the grade level in which most students are aged 14:00-14:11 years during the first week of the eighth month of the school year.

In fact, these definitions proved difficult to implement in the United States, and possibly in other countries as well. Three factors led to difficulty in establishing the appropriate grades for the United States for this study. First, with the decentralized school system in the United States, it was not at all clear exactly when the first week of the eighth month of the school year would be. Second, there was a lack of good data available about the distribution of students by age and grade at that time of year. Census data were available for April 1, 1980, and annual data were available through the late 1980s for whole years of age as of early October. Data from the National Assessment of Educational Progress (NAEP) gave information on the age distribution (in months) of students in grades 4 and 8 in 1988. Third, after dealing with these two issues it was apparent that no grade contained "most" 9-year-olds (or 14-year-olds); there appeared to be a grade in each case with just under 50 percent of the students of the appropriate age, and a second grade with over 40 percent of the students of that age.

In the United States, these target definitions were operationalized as, "The best available *estimate* as tc the modal grade foı 9-year-olds and the modal grade for 14-year-olds at the time of assessment (February-March 1991)." This operational definition resulted in the choice of fourth grade students for Population A and ninth grade students for Population B. (For more discussion on the choice of target grade for the United States, see Rust and Bryant 1991). Table 1-1 shows the age distribution of the U.S. students in the IEA Reading Literacy Study. For the Population A (grade 4) sample, 45 percent of the students were 9 years old, and the mean age of students was 10.1 years. For the Population B (grade 9) sample, almost half of the students were 14 years old and the mean age of students was 15.1 years. For the United States, the "estimate" was evidently the correct one.

There are, inevitably, substantial variations in the age distributions across countries, which raise questions about the comparability of assessment results among countries (Elley 1992 discusses this issue in his appendix E). It is interesting to note that this combined grade/age definition of the target population was intentionally varied in at least one instance because of curricular issues. In Indonesia, grade 4 was used for Population A, even though the students in that grade are predominately 10 years old (mean age 10.8 years). That was because grade 4 was the lowest grade at which most students are educated in the national language, the language of assessment in that country.

Table 1-1. Age distribution of U.S. students in the IEA Reading Literacy Study, by population, region, and school control: Percentage of students at age level and mean age

Age	U.S. total		Region				Control	
	Estimate	Standard error	Northeast	Southeast	Central	West	Public	Private
Population A (Grade 4)								
8 years	7.7%	0.9%	7.1%	7.7%	5.8%	9.9%	8.3%	3.8%
9 years	45.0	1.2	51.1	42.1	44.9	43.1	43.9	53.1
10 years	41.7	1.2	36.0	41.9	45.0	42.4	41.9	39.8
11+ years	5.6	0.5	5.8	8.3	4.2	5.6	5.9	3.3
Mean age	10.1	0.02	10.1	10.2	10.1	10.2	10.2	10.1
Population B (Grade 9)								
13 years	4.7%	0.6%	4.2%	5.5%	6.1%	3.5%	4.9%	3.6%
14 years	48.6	1.3	58.0	47.3	45.8	45.3	47.7	54.2
15 years	39.6	1.3	33.9	36.7	43.5	42.8	40.8	32.6
16+ years	7.0	0.8	3.8	10.4	4.6	8.4	6.6	9.6
Mean age	15.1	0.03	15.0	15.2	15.1	15.2	15.1	15.1

NOTE: Percentages may not add to 100 due to rounding.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Among other participating countries, the mean age of Population A students ranged from 8.9 years to 10.4 years (data from Elley 1992, table 3.1). The two extremes in this range are for countries that appear to have assessed the "wrong" grade. When the two extremes are excluded, the mean age ranged from 9.2 years to 10.1 years, with a median of 9.8 years. In terms of grade, 12 countries assessed grade 3 students, 11 countries assessed grade 4 students, and 1 country assessed grade 5 students. For Population B, the mean age of students ranged from 13.9 to 15.6 years (Elley 1992, table 4.1); in terms of grade, 17 countries assessed grade 8 students, 13 countries assessed grade 9 students, and 1 country assessed grade 10 students.

A difficulty with the procedure of defining a target grade as being modal for a particular 12-month age span is that small differences between countries in the age by grade distribution of students can lead to large differences in the target population across countries defined with respect to age and grade simultaneously. Thus, although most countries apparently correctly ascertained the target population for Population A, this results in the inclusion of grade 3 students with a mean age of 9.2 from the Netherlands, grade 3 students with a mean age of 9.3 from Singapore, and yet grade 4 students with a mean age of 10.1 from both the United States and France. This variation is fine if one wishes to compare the United States with France, or Singapore with the Netherlands, but the choice of grades is not very suitable for comparing the United States with the Netherlands. Note that the overlap in age distribution between the United States and the Netherlands (say) is so small that the approach of using age as a covariate in analysis, often a useful technique to control variations in age distribution, will be of doubtful validity, resting heavily on untestable assumptions.

An Alternative Approach to Obtaining Target Populations

There is no simple solution to the problem of obtaining samples from across countries that are grade-based and yet age-comparable. One approach has been proposed for the Third International Mathematics and Science Study (TIMSS), to be conducted by IEA in 1995. This approach is to select the pair of adjacent grades that best covers the age population of interest (Wolfe and Wiley 1992). This would ensure for most countries that a very high proportion of students defined by a 12-month age span would be included, and thus comparison across countries of students of the same age could be carried out.

For example, had this approach been used for the Reading Literacy Study, students from grades 3 and 4 would have composed Population A, and over 85 percent of 9-year-olds in the United States would have been included. Such an approach increases the chance that pairwise comparisons between countries could be made using comparable grades. For example, the United States could compare grade 4 students with France and grade 3 students with the Netherlands. In fact, it is likely that many countries would select the same pair of grades (say 3 and 4) rather than splitting into two distinct groups based on their choice of a single grade.

The paired grade approach also permits cross-country comparisons of the nature and extent of differences between adjacent grades. It is postulated that change in student achievement from one grade to the next is more strongly associated with teacher background and practices than is cumulative achievement at a single grade. Thus, such a design may enhance the ability to discern important teacher characteristics. The use of a paired grade design does present administrative and cost implications that must be considered in the design of a study.

1.3 The Definition and Reporting of Excluded Students

In a study such as the IEA Reading Literacy Study, students can be excluded from eligibility for the assessment at a variety of points in the sampling and administration process. Inevitably, exclusions will not be comparable across countries, because of the different natures of the populations and school systems that are involved. Attempts to make countries comparable with respect to exclusions will generally prove fruitless. Much more meaningful in practice, and ultimately no doubt more useful to those making comparative analyses across countries, is an accounting for each country of the size and nature of the excluded student population. This approach was recommended by the Board on International Comparative Studies in Education of the National Research Council (Bradburn and Guilford 1990, 24).

It seems useful to classify excluded students according to the stage in the process of sampling and administration that they are either intentionally or inadvertently excluded from having an appropriate chance of inclusion in the sample of assessed students. An attempt at such a classification follows.

1. *Students excluded from the Desired International Target Population.* Although the Desired International Target Population defines the populations of interest for the study across countries, there will be students in some countries who are not in this population, whereas equivalent students in another country are in this population and could be excluded in one of the stages discussed below. The Desired International Target Population for the IEA Reading Literacy Study for Population A (for example) was "All students attending schools on a full-time basis at the grade level in which most students were aged 9:00-9:11 years during the first week of the eighth month of the school year" (Elley 1992). To interpret the

results of the study, it is important to know at least approximately in aggregate for the country the proportion of persons in the following categories:

a. Persons aged 9:00-9:11 who are noi enrolled in a mainstream school.

b. Persons aged 9:00-9:11 who are enrolled in school, but are not graded.

c. Students in the relevant grade who are not in school full time.

In many countries these will be small proportions of the population. But, for example, countries vary as to how they educate their disabled students; in some countries they are taught in regular graded classes for the most part, and so might be considered as being included in the Desired Population. Other countries educate many or most such students in settings that do not correspond to any grading system, and so would **exclude** such students from the Desired Population.

Another aspect that must be addressed under this heading or the next is the question of what constitutes a country. Specifically, what territories, protectorates, possessions, etc. are included. For example, in the IEA Reading Literacy Study, a proportion of the identified excluded population in the United States consisted of students who are residents of Puerto Rico and various U.S. territories (although the rates of exclusion reported in Elley (1992) do not treat this group of students as exclusions). These outlying populations could perhaps have been considered as not being included in the Desired International Target Population.

2. *Students excluded from the Defined National Target Population before implementation of school, classroom, and student sample selection.* This is the component referred to as the Excluded Population in the Sampling Manual (Ross 1991). It includes that portion of the Desired International Target Population that a country explicitly recognizes as being excluded from the assessment. Usually this is a school-level exclusion in that certain schools are excluded as a whole from the assessment, even though they have students in the relevant grade. Typical cases are geographic regions, language minority schools, and nongovernment schools. For example, in the IEA Reading Literacy Study, France and Italy excluded nongovernment schools from the assessment. Portugal assessed only mainland dwellers, and the United States excluded students from Puerto Rico, the U.S. Virgin Islands, and other territories, as discussed above.

   In presenting the results of exclusion in the IEA Reading Literacy Study, Elley (1992) recognizes each of these sources of exclusion. In reporting the proportion excluded in each country, however, exclusions based on geography were ignored. Thus, these are effectively treated as exclusions from the Desired International Target Population. (Examples include Indonesia, Philippines, Portugal, and, as mentioned above, the United States.) Exclusions of students from nongovernment schools (France and Italy) or language minority students within schools (Flemish speakers in Belgium, non-English speakers in the United States) are counted in the rates of exclusion reported.

3. *Students explicitly excluded in the process of drawing samples of classrooms and students.* These would typically be students from special classes for language minority students or individual language minority students (in schools having predominantly students whose native language was the language of the assessment), and students with learning or other disabilities, who may or may not be in special classrooms within regular schools. It is very easy to fail to capture an indication of the size and nature of this group unless careful

procedures are adopted for identifying them at the classroom and student sampling phase. Thus the sampling procedures must list every classroom that has students who qualify under the definition of the Defined National Target Population, and not just those that the school deems appropriate or convenient to assess. The school can then explicitly exclude such classrooms, giving the reason for exclusion, so that the magnitude and characteristics of this portion of the excluded population can be identified. Similar procedures are needed to identify students to be excluded from the selected classrooms. Even with whole classroom sampling, a full list of students within each selected classroom should be compiled and used to account for each student.

4. *Students excluded because they did not attend any of the assessment sessions.* Some students will inevitably be unable to attend the assessment sessions because of a temporary health problem or other reason. Such students would not normally be regarded as excluded. If, however, no effort has been made to identify excluded students during the classroom and student sampling process, but rather all students in all selected schools are treated as being included, it seems inevitable that some disabled or language minority students will fail to attend the assessment (or will not be invited to attend) even though their circumstances at the time of the assessment are no different than they are at other times. For other absent students, provided that they are recorded correctly, nonresponse weighting adjustments can be made to the data to remove substantially the potential for bias in the study estimates. If no such adjustments are made, such absent students in effect are also a part of the excluded student population. In either case it is important to report the magnitude of such absences, as the extent and nature of student absenteeism may have moie impact on the results than does explicit student exclusion.

In summary, a useful description of the magnitude and effect of student exclusion for comparative assessments needs to address several aspects. First, the meaning of the Desired International Target Population must in fact be spelled out for each country, together with a consideration of the size and nature of the population that could conceivably have been included but was not (e.g., persons of the appropriate age range not enrolled in school, students in outlying territories). Second, the Defined National Target Population must be delineated at the student level as well as the school level, and the type and magnitude of the difference between the national population and the Desired International Target Population must be identified (e.g., students in schools not using the language of assessment, language minority students in other schools, physically disabled students, learning disabled students). Third, strict procedures are needed at the stage of classroom and student sampling to ensure that only students explicitly excluded are in fact not included in the assessment process, and to give an account of the size and nature of the population that is excluded at this stage. Fourth, procedures are needed to obtain the rate of absenteeism from the assessment of nonexcluded students.

1.4 Consequences of the Choice of Sampling Unit on Descriptive Statistics

The IEA Reading Literacy Study adopted as the standard sampling approach the use of a stratified two-stage sample design. The first stage was to select a stratified sample of schools, and the second was to select a sample of intact classrooms from within the selected schools. Other large educational surveys, such as the International Assessment of Educational Progress and the National Assessment of Educational Progress (NAEP), also used a two-stage sampling approach, although the approach to sampling of students within schools is different from that of the IEA Reading Literacy Study. Both the IAEP and the NAEP select a stratified sample of schools, but within selected schools, samples of individual students are drawn from a list of all eligible students enrolled in the school. The classroom is not used as a unit of sampling.

This section discusses some issues concerning the analysis of the assessment data, related to the use of intact classrooms as sampling units. In particular, we examine issues of effective sample size and design effects in the estimation of descriptive statistics, such as means, standard deviations, and proportions, for the whole population and for subgroups.

Effective Sample Size and Design Effect for Whole Population

It is well known among survey researchers and educational analysts that sample data on students nested within samples of schools, and perhaps classrooms within schools, will generally give less reliable estimates than data from a sample of the same size that is not clustered in this way. This is because students within a given school or classroom tend to be more similar than students across classrooms and schools. This tendency can be quantified by the intraclass correlation coefficient, ρ. To calculate ρ, the first step is to compute the variance (of a proficiency score, say) among students within classrooms (say) as a proportion of the total between-student variance. Then ρ is given as 1 minus this proportion. Thus, ρ is low (near 0) for characteristics for which students tend to be heterogeneous within classrooms, and is high (near 1) for characteristics that tend to be relatively homogeneous within classrooms.

For a single stage of clustering, the variance of a mean estimate of a characteristic y, $Var(\bar{y})$, can be expressed approximately in terms of the population variance (σ_y^2), ρ, and the sample sizes of clusters (e.g., classrooms) and students. Denoting m as the sample size of clusters, and $\bar{n}$ as the average sample size of students within cluster, then

$$Var(\bar{y})=\frac{\sigma_y^2}{\bar{n}m}\{1+(\bar{n}-1)\}\rho.$$

When two stages of clustering are involved (schools and classrooms), this expression generalizes, but the variance remains a function of intraclass correlation and sample size.

For studies that use samples of intact classrooms within schools, the intraclass correlation tends to be quite high. This means that unless care is taken to achieve an adequate sample size of schools and classrooms, the sampling precision of estimates will be inadequate. Participants in the IEA Reading Literacy Study were instructed to account for this phenomenon in designing their samples. The target was to achieve an "effective sample size" of 400 students (Ross 1991); that is,

$$Var(\bar{y})=\frac{\sigma_y^2}{400}\ .$$

The effective sample size of students is given by the expression $\bar{n}m/\{1+(\bar{n}-1)\rho\}$. Assuming an average classroom size of 25, this means that the number of classrooms needed in the sample is given by $m=400\{1+(\bar{n}-1)\rho\}/\bar{n}=16(1+24\rho)$.

Most countries assumed large values of ρ in their design phase and typically selected 100 to 200 schools, resulting in several thousand students in total. In the United States, for example, ρ was assumed to be 0.4. This yielded $m = 170$ in the above equation, and ultimately the numbers of schools participating were 167 for Population A and 165 for Population B.

This conservative approach was well founded in most countries. In the United States, for example, the *true* sample size for Population B (grade 9) was 3,209 students. For the means for the narrative, expository, and document literacy scales (the three proficiency scales developed for the study), the *effective* sample sizes were 383.4, 338.9, and 450.5, respectively. Similarly, for Population A (grade 4), the total sample size was 6,248 students, as two classrooms per school were selected where possible. The effective sample sizes for the mean scores on the literacy scales were 1,061.3, 618.8, and 993.4, respectively.

The ratio of the true sample size to the effective sample size is known as the design effect. For this study, the design effects on the mean literacy scores ranged from 5.9 to 10.1. In studies that do not draw whole classroom samples of students, the design effects can be much lower. For NAEP, for example, a typical design effect of 3 would be found for this type of mean (Johnson and Rust 1992).

Rationale for Using Whole Classroom Sampling and Impact on Subgroup Estimation

One reason for adopting a design with whole classroom sampling is cost of administration. For many countries the major costs are determined by the numbers of schools and classrooms involved in the study, not the number of students. In fact, in many cases, once a few students from a given classroom have been included in the sample (as will occur if direct student sampling within schools is used), it is much more convenient to include all students from that classroom in the assessment than to make arrangements for the nonselected students during the time that the assessment is being conducted. Thus, to obtain an effective sample size of 400 students (say), it may well be that the cheapest and easiest method is to select one intact classroom from each of 100 to 200 schools, even though the actual sample size of students is several thousand.

Another reason is that not all analyses of the data involve the calculation of simple means and proportions for the whole population. One important feature to note is that although large design effects are encountered for estimates for the whole population, design effects are invariably lower for subgroup estimates. This is because the value of $\bar{n}$ is lower for a subgroup than for the whole population. The design effect is given by the expression $\{1+(\bar{n}-1)\rho\}$. Therefore, when ρ is relatively large, a smaller value of $\bar{n}$ will give rise to a reduced design effect. For example, for the narrative scale at grade 4, the design effects for both males and females are about 3.5, compared with 5.9 for the total population.

Put another way, this means that even though the effective sample size for the whole population is only 400 or so in many cases, for subgroups the effective sample size is much greater than would result from a true unclustered sample of 400 students. Suppose an estimate for grade 9 has a design effect of 8 (and an effective sample size of 400). With a sample of 3,209 students in the United States from 165 schools, $\bar{n}=20$ approximately, so that we have $8 = 1 + 19\rho$, or $\rho = 0.37$. For the same estimate of a mean or proportion restricted to males, say $\bar{n}=10$, it is likely that ρ for males will be very similar to that for the whole population at 0.37. The design effect for males will be about $1 + 9\rho = 4.3$, and the effective sample size will be about 370. That is, the effective sample size for this subgroup is almost as large as the 400 for the whole sample.

Thus, a clustered sample such as that used for the IEA Reading Literacy Study will provide much more precision across subgroups than is apparent from a simple consideration of the overall effective sample size. Since it is the precision of such subgroup estimates that will generally be the poorest, this means that such a clustered sample is relatively more efficient than an unclustered sample for estimates for which such efficiency is most important.

1.5 Consequences of the Choice of Sampling Unit on Estimates from Linear Models

The question of whether or not to use sampling of intact classrooms, or, more generally, what is the desirable extent of clustering for the sample, may have quite a different answer when one is estimating parameters for models. We saw above that for the IEA Reading Literacy Study, when estimatıng mean proficiencies for the whole population, sampling intact classrooms led to very high design effects and would have been an appropriate procedure only because of cost efficiencıes. The situation for estimating models is quite different.

Kish and Frankel (1974) have suggested that design effects for parameters in linear regression equations often will prove to be much closer to 1.0 than for, say, subclass means of the dependent variable, with the subclasses formed from the independent variables. This phenomenon also has been observed for NAEP data (Johnson and Rust 1992). At the same time, if one is interested in analyzing the modifying effect that classroom, teacher, and school variables have on the influence of student characteristics on literacy, it is necessary to assess many students in each classroom. The large within-classroom samples are needed to estimate the within-classroom effects of student characteristics with sufficient precision that it will be feasible to model the variation across classrooms (see discussion in Bryk and Raudenbush 1992).

In this section, we will consider the consequences of sampling intact classrooms for developing linear models for predicting scores on the reading scales (narrative, expository, and document). We will briefly discuss the design effects for some simple linear and hierarchical linear models and examine the impact of the use of clustered sampling, especially whole classroom sampling, for the U.S. data from the IEA Reading Literacy Study.

Simple Linear Model Involving Student Characteristics

Consider the following model for the U.S. grade 4 student population that was first presented by Atash and Rust (1992):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \qquad (1)$$

where

Y denotes score on the narrative scale

X_1 denotes minority status:

$X_1 = 1$ for black and Hispanic students

$X_1 = 0$ for all other students

X_2 denotes hours per day of .levision viewing (integer values from 0 to 7)

β_0 denotes the intercept

β_1, β_2 denote the slope coefficients associated with X_1 and X_2

ϵ denotes residual term with mean 0, independent of X_1 and X_2.

Table 1-2 shows the results of fitting this model, using a method that takes account of the intraclass correlation of the error term, ϵ, within classrooms, schools, and geographic primary sampling units (PSUs). The jackknife method (Wolter 1985; Rust 1985; Ross 1991) is used throughout the IEA Reading Literacy Study, both for the U.S. data and internationally, to estimate sampling errors, accounting for the sample design. The statistical significance of the coefficient estimates is assessed using a t-distribution with 33 degrees of freedom. This choice of the number of degrees of freedom is dictated by the nature of the sample design and the method of implementing the jackknife procedure (see Chapter 2 of this volume).

Table 1-2. Coefficient estimates, standard errors, and design effects for a linear model of narrative scale score of grade 4 students, with minority status (X_1) and hours of TV viewing (X_2)

Parameter	Coefficient estimate[1]	Standard error	Design effect	t (33)	Significance
β_0	266.571	1.8547	2.23	143.73	$p<0.001$
β_1	-27.500	1.8743	1.71	-14.67	$p<0.001$
β_2	-2.247	0.4018	1.79	-5.59	$p<0.001$
$R^2 = 0.0740$					

[1] For this analysis, the narrative scale scores are scaled differently from the final international scale reported in Elley (1992), but the difference is essentially a linear transformation and so has no effect on the statistical significance or design effects of terms in a linear model.

SOURCE: Data from N. Atash and K. Rust, "A Comparison of Hierarchical Linear Models and Jackknife Methods for Estimating Standard Errors," presented at the Annual Meetings of the American Educational Research Association, San Francisco, California, 1992.

The analysis summarized in Table 1-2 was conducted on a subset of 6,220 students. The reduction was obtained by eliminating classrooms with fewer than 10 assessed students, as these cases would have been problematic for fitting the hierarchical linear models discussed below. The table shows that the design effect for each of the three parameters is about 2 in each case. Thus, the effective sample was approximately 3,000 for each parameter, in comparison to the effective sample size for the mean of the narrative scale of about 1,000.

Wright and Williams (1992) also presented design effects for a number of other linear models for the U.S. data from the IEA Reading Literacy Study. Their models generally involved more independent variables than the simple one presented above, and the results from one of their models are summarized in Table 1-3. Again the design effects for slope parameters are consistently of size 2 or lower, but the design effect is higher for the intercept term (3.83).

Thus, for linear models of student characteristics, the use of whole classroom sampling appears to have much less consequence for precision of parameter estimates than is the case for estimates of descriptive statistics. However, for simple linear models, there are no analytic benefits to the use of whole classroom sampling—the only benefits are those of cost and administrative convenience, as is the case for descriptive statistics.

Table 1-3. Coefficient estimates and design effects for linear model of narrative scale scores of grade 4 students

Variable	Coefficient estimate[1]	Design effect
Intercept	251.70	3.83
Gender 1	-4.52	1.72
Gender 2	4.58	1.72
Race 1	9.88	0.75
Race 2	-0.06	2.32
Race 3	-5.51	1.16
Race 4	-20.74	1.41
Race 5	5.05	1.08
Father's Education 1	-6.64	1.41
Father's Education 2	-3.99	0.73
Father's Education 3	-0.52	1.23
Father's Education 4	3.94	1.03
Mother's Education 1	-7.32	1.38
Mother's Education 2	0.28	1.81
Mother's Education 3	0.54	1.24
Mother's Education 4	1.14	1.45
Language 1	1.29	1.16
Language 2	-13.51	0.38
Language 3	-2.52	0.59
Language 4	-5.08	1.36

[1] For this analysis, the narrative scale scores are scaled differently from the final international scale reported in Elley (1992), but the difference is essentially a linear transformation and so has no effect on the statistical significance or design effects of terms in a linear model.

SOURCE: Data from D. Wright and T. Williams, "Effects of Sampling Design on Regressions of IEA Reading Literacy Data," presented at the Annual Meetings of the American Educational Research Association, San Francisco, California, 1992.

Simple Linear Model Involving Student and Teacher Characteristics

The IEA Reading Literacy Study collected a wealth of data about teachers and their teaching practices, principals, and schools. In developing models for student performance on the literacy assessments, we may wish to include variables relating to whole classrooms of students, derived from the Teacher and School Questionnaires. One method of doing this is to use linear regression models at the student level, attaching to each student the values of the classroom, teacher, and school variables that are associated with the classroom to which the student belongs. Such models are likely to suffer potentially serious misspecification effects, discussed below. More fundamentally, however, the impact of clustering on parameter estimate design effects is likely to be much greater. Consider the following simple model that includes a student characteristic, a teacher characteristic, and an interaction between teacher and student characteristic:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \epsilon \qquad (2)$$

where

Y denotes score on the narrative scale

X_1 denotes minority status:

$X_1 = 1$ for black and Hispanic students

$X_1 = 0$ for all other students

X_2 denotes the qualifications of student's classroom teacher:

$X_2 = 1$ if teacher has advanced qualifications

$X_2 = 0$ if otherwise

β_0 denotes the intercept

β_1 denotes the slope coefficient associated with X_1

β_2 denotes the slope coefficient associated with X_2

β_{12} denotes the slope coefficient associated with interaction between X_1 and X_2

ϵ denotes a residual term with mean 0, not dependent on X_1 or X_2.

Table 1-4 shows the result of fitting this model to the same 6,220 student records as before. Notice that the design effect for the coefficient estimate for the classroom level variable (X_2), β_2, is 4.81, a value much greater than the design effects for slope coefficients seen in the previous model, which were less than 2. The intercept term has a similar change in design effect, being 5.08 compared to 2.23 in the previous model.

Table 1-4. Coefficient estimates, standard errors, and design effects for linear model of narrative scale score for grade 4 students with minority status (X_1), teacher qualifications (X_2), and their interaction

Parameter	Coefficient estimate[1]	Standard error	Design effect	t(33)	Significance
β_0	257.125	2.0907	5.08	122.99	$p<0.001$
β_1	-30.184	2.5090	1.91	-12.06	$p<0.001$
β_2	4.174	3.1231	4.81	1.34	$p=0.20$
β_{12}	2.122	4.3020	2.21	0.49	$p=0.60$
$R^2 = 0.0679$					

[1] For this analysis, the narrative scale scores are scaled differently from the final international scale reported in Elley (1992), but the difference is essentially a linear transformation and so has no effect on the statistical significance or design effects of terms in a linear model.

SOURCE: Data from N. Atash and K. Rust, "A Comparison of Hierarchical Linear Models and Jackknife Methods for Estimating Standard Errors," presented at the Annual Meetings of the American Educational Research Association, San Francisco, California, 1992.

This example demonstrates that when estimating for linear models that involve independent variables for aggregate units (classrooms and schools), heavy clustering of the sample results in substantial loss of precision in measuring the effect of such aggregate variables. Simply put, if one wishes to learn about classroom-level effects on mean student proficiency, it is advantageous to include many classrooms, rather than many students per classroom, in the sample.

It is also important to note that an analysis that fails to account for the effect of sampling intact classrooms is very likely to falsely conclude that aggregate level effects are significant. For example, in this analysis the significance of the coefficient for teacher qualifications, β_2, was $p = 0.20$. An analysis conducted using a standard regression package, failing to adjust for the design effect, would have given a test statistic of 6.43, rather than 1.34, and the analyst would erroneously have concluded that teacher qualifications have a highly statistically significant effect. The use of a relatively unclustered design, with only a few students sampled per classroom and school, would provide protection against the possibility of such misleading analyses.

For populations of students clustered within schools and classrooms, the use of the linear modeling approach to examine the effect of aggregate level variables has been criticized as being inappropriately specified. Two primary sources of misspecification are likely to be encountered in practice. The first is that the distribution of the residual term is dependent upon the values of the independent variables. The second is that the model fails to account for the fact that the classroom variable may impact the intercept, as well as the slopes of the student-level variables. In the above example, teacher qualifications may affect not only the average student achievement within the groups of minority and nonminority students, but also the difference in achievement between minority and nonminority students. This second issue can be addressed by introducing interaction terms in the linear model, between student-level and classroom-level variables, as was used in equation (2). The likelihood then is that not only will the residual term be dependent on the classroom-level variables, but also upon the interactions. In the example above, Atash and Rust (1992) evaluated the interaction between minority status and teacher qualifications and found it to be nonsignificant (see Table 1-4). There was evidence, however, that the residual terms were substantially correlated with the minority-status variable.

Hierarchical Linear Models

To deal with these deficiencies in linear models, the technique of hierarchical linear modeling (HLM) has been developed. For a discussion of its application to the IEA Reading Literacy Study data, see Chapter 4 of this volume. This approach proposes models at two or more levels. In this case the first-level model is a simple linear model of the effect of student level variables. The second-level model includes the impact of classrooms and classroom-level variables on the intercept and slope coefficients of the first model. Consider the following example, using the same variables as above:

$$\begin{aligned} Y &= \beta_0^* + \beta_1^* X_1 + \epsilon * \\ \beta_0^* &= \gamma_{00} + \gamma_{01} X_2 + \delta \\ \beta_1^* &= \gamma_{10} + \gamma_{11} X_2 + \theta. \end{aligned} \tag{3}$$

The terms ϵ^*, δ, and θ are residual terms, assumed to be independent of X_1 and X_2. Note that this model can be re-expressed as

$$Y = \gamma_{00} + \gamma_{10} X_1 + \gamma_{01} X_2 + \gamma_{11} X_1 X_2 + \theta X_1 + \epsilon * + \delta.$$

This reduces to a standard linear model with interaction, if and only if θ is always 0 (*i.e.*,$\sigma_\theta^2=0$).

In estimating terms for this model, a crucial element is to estimate the value of σ_θ^2 . That is, the variation in β_1^* from classroom to classroom must be measured reliably. To do this we need to obtain reliable within-classroom estimates of the differences in achievement between minority and nonminority students for each of a sample of classrooms. Thus, to realize the advantages of the hierarchical linear modeling approach over a linear model, it is essential to have large within-classroom sample sizes—that is, whole-classroom sampling is needed.

As an aside, it is of interest to consider the effect of the sample design on parameter estimates for a hierarchical linear model. If the sample were a simple random sample of intact classrooms, the design effects for the parameter estimates would be 1.0, since this is the design assumed in the development of HLM theory. The software currently available assumes such a design. The IEA Reading Literacy Study data are clustered by school and geographic PSU (and also stratified), so that it is sensible to examine the design effects for HLM parameter estimates. Atash and Rust (1992) did this for a hierarchical model that was equivalent to the linear model in equation (1). That is, the HLM took the following form:

$$\begin{gathered} Y=\beta_0^*+\beta_1^*X_1+\beta_2^*X_2+\epsilon* \\ \beta_0^*=\gamma_{00}+\delta \\ \beta_1^*=\gamma_{10} \\ \beta_2^*=\gamma_{20}. \end{gathered} \tag{4}$$

The results are shown in Table 1-5. It can be seen that the design effects are quite close to 1 for all parameter estimates. If this result generalizes to other models of student variables (level 1 models), there are two implications. The first is that the use of standard HLM software to analyze IEA Reading Literacy Study data will seldom lead to erroneous conclusions of significance. The second, related point is that using such an HLM approach is a viable alternative to fitting ordinary linear models and accounting for the complex sample design. It should be noted that HLM offers no real advantages for a model such as (4). We have not yet investigated the more crucial issue of the design effects for coefficients of true hierarchical models (such as (3)) with the IEA Reading Literacy Study data.

Table 1-5. Coefficient estimates, standard errors, and design effects for hierarchical linear model of narrative scale score for grade 4 students with minority status (X_1) and hours of TV viewing (X_2)

Parameter	Coefficient estimate[1]	Standard error	Design effect	t (33)	Significance
γ_{00}	252.313	1.6378	1.39	154.06	$p<0.001$
γ_{10}	-14.095	2.2191	1.16	-6.35	$p<0.001$
γ_{20}	-1.103	0.3025	0.80	-3.65	$p<0.001$

[1] For this analysis, the narrative scale scores are scaled differently from the final international scale reported in Elley (1992), but the difference is essentially a linear transformation and so has no effect on the statistical significance or design effects [illegible] terms in a linear model

SOURCE: N. Atash and K. Rust, "A Comparison of Hierarchical Linear Models and Jackknife Methods for Estimating Standard Errors," presented at the Annual Meetings of the American Educational Research Association, San Francisco, California, 1992.

The conclusion from this discussion is that in designing comparative education studies effectively, it is essential to have an understanding of the types of analyses that are to be carried out, where the data are to come from, and the relative importance of each type of analysis before finalizing the design. For data such as those obtained in the IEA Reading Literacy Study, for estimates of mean proficiency for the whole population and large subgroups, the use of whole-classroom sampling is highly inefficient unless justified on the grounds of administrative convenience. For standard linear models, such as (1) and (2), such a design is likely to be inefficient also, especially when mixed-level variables are involved. Hierarchical linear models are designed to handle data having a true hierarchical structure, so that the use of whole-classroom sampling is not a problem and is often desirable for obtaining good estimates of the model terms.

1.6 Concluding Remarks

In this chapter, we have shown that for international comparative studies of educational achievement, the definition of the target population has implications for the definition and identification of excluded students and on the choice of sampling units in a multistage design. The choice of sampling unit, in turn, has consequences for efficiency in the analysis of the data, and may in fact preclude certain kinds of analysis. Thus, it is not desirable to reach decisions about the choice of target population, the definition of the excluded population, the choice of sampling units, and the methods of analysis, in isolation or even in sequence. To realize the full potential of such a study, these decisions must be linked. A single, overall strategy must be developed for the study that best addresses the research interests that motivate the study, and the resource limitations that constrain it.

References

Atash, N., and Rust, K. (1992). A comparison of hierarchical linear models and jackknife methods for estimating standard errors. Presented at the Annual Meetings of the American Educational Research Association, San Francisco, CA.

Bradburn, N.M., and Guilford, D.M. (eds.) (1990). *A framework and principles for international comparative studies in education*. Washington, DC: National Academy Press.

Bryk, A.S., and Raudenbush, S.W. (1992). *Hierarchical linear models*. Newberry Park, CA: Sage.

Elley, W.B. (1992). *How in the world do students read? IEA Study of Reading Literacy*. The Hague: International Association for the Evaluation of Educational Achievement.

Johnson, E.G., and Rust, K.F. (1992). Population inferences and variance estimation for NAEP data. *Journal of Educational Statistics*, 17(2), 175-190.

Kish, L., and Frankel, M.R. (1974). Inference from complex samples. *Journal of the Royal Statistical Society*. B36, 1-37.

Lapointe, A.E, Mead, N.A., and Askew, J.A. (1992). *Learning mathematics*. Princeton, NJ: Educational Testing Service.

Lapointe, A.E., Mead, N.A., and Phillips, G.W. (1989). *A world of differences: An international assessment of mathematics and science*. Princeton, NJ: Educational Testing Service.

Ross, K.N. (1991). *Sampling manual for the IEA International Study of Reading Literacy*. University of Hamburg, Hamburg, Germany: International Coordinating Center, IEA Reading Literacy Study.

Rust, K. (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics*, 1(4), 381-397.

Rust, K., and Bryant, E. (1991). The IEA Reading Literacy Study design and implementation: National and international perspectives - population definitions and sample design. Presented at the Annual Meetings of the American Educational Research Association, Chicago, IL.

Wolfe, R., and Wiley, D. (1992). *Third International Mathematics and Science Study: Sampling plan*. International Association for the Evaluation of Educational Achievement.

Wolter, K.M. (1985). *Introduction to variance estimation*. New York: Springer-Verlag.

Wright, D., and Williams, T. (1992). Effects of sampling design on regressions of IEA Reading Literacy data. Presented at the Annual Meetings of the American Educational Research Association, San Francisco, CA.

2 Estimation, Sampling Errors, and Design Effects

Edward Bryant

This chapter addresses the issues involved in making estimates of population means, totals, proportions, and other relatively simple statistics from a complex design such as the one used for the IEA Reading Literacy Study. The design for the IEA Reading Literacy Study called for the selection of a sample of classrooms at grades 4 and 9 to represent the U.S. population at each of these grade levels. To accommodate the plans for assessing the students in their classrooms using centrally trained administrators, a complex multistage sample design was required. The use of such a design has implications for the analysis of the data. First, appropriate estimation formulas require the use of survey weights. Second, estimates of sampling error require special methods that account for the design features. More generally, inferences about population parameters can only be made appropriately .f proper account is taken of the impact of the sample design and estimation procedures on both thι parameter estimates themselves and on their accompanying estimates of sample error.

Sections 2.1 through 2.5 provide more detailed information on the various steps in the sample design and development of sampling weights. Sections 2.6 through 2.8 examine the issues of estimation. Section 2.6 discusses the use of the IEA Reading Literacy Study data to estimate simple averages, ratios, and proportions. Section 2.7 examines methods of estimating standard error. Section 2.8 explores methods for assessing the appropriate degrees of freedom in confidence tests.

2.1 Basic Structure of the Design

The U.S. component of the IEA Reading Literacy Study was designed to collect test scores and data on student, teacher, school, and family characteristics; family, school, and classroom environments; instructional strategies; and student reading acι.vities and behaviors on a sample of grade 4 students (Population A) and grade 9 students (Population B). The first stage sample was drawn from the primary sampling units (PSUs) constructed for the National Assessment of Education Progress (NAEP) surveys, after some changes in stratification described below. The sample was allocated to the strata in proportion to 1980 population, which was the basis for construction of the NAEP PSUs. The schools in the sample of PSUs were further stratified by enrollment in grade 4 or grade 9 (the two populations were handled independently), and by school sector (public and private).

The structure of the sample design differed somewhat from the models suggested by the international referee (Ross 1991). The United States adopted the unusual approach, approved by the referee, of arranging for centrally trained personnel from outside the school system to administer the assessments. This approach was adopted to maximize school participation by minimizing the burden on schools and to assist in maintaining uniformly high standards of assessment administration throughout the sample. In most other countries, school personnel administered the assessments in the interest of minimizing costs. As a consequence, the sample of U.S. schools was concentrated in selected areas to reduce travel costs.

The sample was designed to select 200 schools from each of grade 4 and grade 9. The basic sample plan called for sampling intact classrooms or classes as follows:

- **Grade 4.** If there are fewer than an estimated 50 grade 4 students in the school, take all. If there are 50 or more, sample two classrooms at random.

- **Grade 9.** If there are fewer than an estimated 25 grade 9 students, take all. If there are 25 or more, take one classroom (typically, one home room).

The numbers of grade 4 and grade 9 students in the school were estimated by dividing the total enrollment, as reported on the 1989 Quality Education Data (QED) file, by the grade span of the school. The QED file is a commercial file of schools often used as a sampling frame for national or regional samples of schools. Enrollment by individual grades is not reported, so dividing total enrollment by the number of grades in the grade span provides an estimate of the enrollment in each grade. Even though enrollment sometimes is either overestimated or underestimated by this procedure, no error is introduced if the under- or overestimation is reasonably consistent across schools. The purpose of the estimates is simply to provide a measure of size of the schools for the purpose of allocating the sample. For purposes of targeting the sample size, it was estimated that the average enrollment per classroom would be in the neighborhood of 25. Thus, with 200 schools representing each grade, and taking one classroom per school in grade 9 and (typically) two classrooms per school in grade 4, one would expect to sample about 4,000 students in grade 9 and somewhat fewer than 8,000 students in grade 4. After nonresponse and allowing for the variability in class size, the number of sampled students from grade 9 was 3,209 and from grade 4 was 6,248.

2.2 Stage 1 Stratification

The Stage 1 stratification involves a regrouping of PSUs defined for the NAEP. The NAEP PSUs are counties (or independent cities) and groups of counties with a minimum size of 60,000 population as of the census of 1980. The counties constituting metropolitan areas are kept together. Other aggregations of counties avoid mixing urban and rural counties.

The NAEP PSUs were restratified for use in the IEA Reading Literacy Study because estimates that are required by various subgroups (such as minorities) in the NAEP surveys are not required in the IEA survey. In the IEA survey, the first-level stratification was by NAEP region (four geographic strata) and two degree-of-urbanization strata (Metropolitan Statistical Area--MSA--and non-MSA). In addition, the PSUs in the Southeast and West regions were stratified by percent minority, divided into those with less than 20 percent minorities and those with 20 percent or more. Minorities are relatively less significant in the Northeast and the Central regions, so the minority stratification was not used in those regions. In the West the high minority, non-MSA stratum contained so few schools that it was combined with the low minority, non-MSA stratum. The 50 PSUs to be selected for the sample were allocated across the strata in proportion to the 1980 population; the 1990 census population was not available at the time of the allocation. The numbers of PSUs allocated into each sampling stratum are shown in Table 2-1.

A sample of PSUs was drawn according to the above allocation using probability proportional to size (PPS) sampling, where the size measure was the 1980 population. Any PSU that contained more than 50 percent of the total measure of size of the stratum divided by the number of PSUs allocated to the stratum was included with certainty, and the measure of size to be allocated PPS was adjusted accordingly. Sampling weights of PSUs are the inverse of the probabilities of selection.

2.3 Stage 2 Stratification

It was believed that control (public and private) and size of school might be related to reading literacy, so the schools in the sampled PSUs were extracted from the 1989 QED file and substratified by control and size which, in some cases, cross-cut the first level of stratification. It was presumed that the distinction between private and public schools was so important that the design should adequately represent the relatively thin population of private schools. It was also thought important to have an appropriate representation of the large number of small schools with small enrollments.

The substrata that included control and size are shown in Table 2-2 for grade 4 (Population A) and in Table 2-3 for grade 9 (Population B) Enrollment was attributed to schools for the fourth and ninth grades by dividing total enrollment for the school by the grade span covered by the school. The enrollments thus constructed were put into three classes, less than 15, at least 15 but less than 50, and 50 and more. The amount of collapsing of first-stage stratifying factors necessary to effect the second stage of stratification is evident from the tables. Note that the last stratum in each table consists of the large number of schools with small enrollments. These schools were sampled at a lower rate in order to increase the efficiency of the design. To compensate for the lower sampling fraction of the schools with small enrollments, the weights of the sampled small schools were increased so that their effect on national projections would be proportionate to the total enrollment of the stratum.

The sample of 200 schools from each grade was allocated to the deeply stratified universe in proportion to the number of students in the given grade projected from the sampled PSUs, since, at the time the sample was drawn, total counts for the universe were not available in time to meet the deadline for the design work. This required a later adjustment in the sampling weights, as is discussed in Section 2.5.

2.4 Selection of Schools and Classes of Students

The schools, as identified above, were coded by substratum number, as shown in Tables 2-2 and 2-3. Within each substratum, each school was given a measure of size that reflected the way in which the within-school sample was to be drawn. The measures of size were determined by multiplying the PSU weight by a measure of the per-school enrollment for the schools in the grade level corresponding to the study population. The measures of estimated per-school enrollment for schools in particular size strata are shown in Table 2-4.

Table 2-1. Number of PSUs allocated to sampling strata

Region[1]	Urbanicity	Certainty	Minority level[2]	Number of PSUs[3]
Northeast	MSA	Certainty	All	7
		Noncertainty	All	4
	Non-MSA	Noncertainty	All	2
Southeast	MSA	Certainty	High	2
		Noncertainty	High	2
		Noncertainty	Low	2
	Non-MSA	Noncertainty	High	2
		Noncertainty	Low	2
Central	MSA	Certainty	All	3
		Noncertainty	All	6
	Non-MSA	Noncertainty	All	4
West	MSA	Certainty	High	2
		Noncertainty	High	4
		Noncertainty	Low	4
	Non-MSA	Noncertainty	All	4
Total PSUs				50

[1] Region definitions (note that these region definitions are those used by NAEP, and hence were used for forming strata for the Reading Literacy Study).

Northeast
Connecticut
Delaware
District of Columbia
Maine
Maryland
Massachusetts
New Hampshire
New Jersey
New York
Pennsylvania
Rhode Island
Vermont
Virginia (the part in Washington, DC MSA)

Southeast
Alabama
Arkansas
Florida
Georgia
Kentucky
Louisiana
Mississippi
North Carolina
South Carolina
Tennessee
Virginia (outside Washington, DC MSA)
West Virginia

Central
Illinois
Indiana
Iowa
Kansas
Michigan
Minnesota
Missouri
Nebraska
North Dakota
Ohio
South Dakota
Wisconsin
Utah
Washington
Wyoming

West
Alaska
Arizona
California
Colorado
Hawaii
Idaho
Montana
Nevada
New Mexico
Oklahoma
Oregon
Texas

[2] Minority level of primary sampling unit (PSU) only used in Southeast and West regions: Low = less than 20%, High = 20% or more.

[3] The PSUs constructed for NAEP were restratified for this IEA survey.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 2-2. Substrata for grade 4 (Population A)

Substratum number	NAEP stratum characteristics				Substratum		Number of schools	
	Region[1]	Urbanicity	Minority level[2]	Certainty status	Control	Grade size[3]	Sample	Population[4]
1	Northeast	MSA	NA	Certainty	Public	15-49	2	1,029
2	Northeast	MSA	NA	Certainty	Public	50+	12	2,627
3	Northeast	MSA	NA	Certainty	Private	15+	4	1,685
4	Northeast	All	NA	Noncertainty	Public	15-49	6	2,268
5	Northeast	All	NA	Noncertainty	Public	50+	13	3,221
6	Northeast	All	NA	Noncertainty	Private	15+	3	1,408
7	Southeast	MSA	High	Certainty	All	15+	4	915
8	Southeast	MSA	High	Noncertainty	Public	15+	9	2,282
9	Southeast	MSA	Low	Noncertainty	Public	15+	11	2,323
10	Southeast	All	All	Noncertainty	Private	15+	4	1,579
11	Southeast	Non-MSA	High	Noncertainty	Public	15+	8	1,920
12	Southeast	Non-MSA	Low	Noncertainty	Public	15+	9	2,393
13	Central	MSA	NA	Certainty	Public	15+	7	2,067
14	Central	MSA	NA	Certainty	Private	15+	2	782
15	Central	All	NA	Noncertainty	Private	15+	5	2,304
16	Central	MSA	NA	Noncertainty	Public	15-49	4	1,880
17	Central	MSA	NA	Noncertainty	Public	50+	14	3,718
18	Central	Non-MSA	NA	Noncertainty	Public	15-49	8	3,106
19	Central	Non-MSA	NA	Noncertainty	Public	50+	6	1,728
20	West	MSA	High	Certainty	All	15+	9	2,081
21	West	All	All	Noncertainty	Private	15+	4	1,696
22	West	MSA	High	Noncertainty	Public	15+	17	3,543
23	West	MSA	Low	Noncertainty	Public	15+	19	4,383
24	West	Non-MSA	All	Noncertainty	Public	15-49	5	1,538
25	West	Non-MSA	All	Noncertainty	Public	50+	7	1,630
26	All	All	All	All	All	<15	8	10,408

[1] Region definitions (note that these region definitions are those used by NAEP, and hence were used for forming strata for the Reading Literacy Study).

Northeast
Connecticut
Delaware
District of Columbia
Maine
Maryland
Massachusetts
New Hampshire
New Jersey
New York
Pennsylvania
Rhode Island
Vermont
Virginia (the part in
Washington, DC MSA)

Southeast
Alabama
Arkansas
Florida
Georgia
Kentucky
Louisiana
Mississippi
North Carolina
South Carolina
Tennessee
Virginia (outside
Washington, DC MSA)
West Virginia

Central
Illinois
Indiana
Iowa
Kansas
Michigan
Minnesota
Missouri
Nebraska
North Dakota
Ohio
South Dakota
Wisconsin
Utah
Washington
Wyoming

West
Alaska
Arizona
California
Colorado
Hawaii
Idaho
Montana
Nevada
New Mexico
Oklahoma
Oregon
Texas

[2] Minority level of primary sampling unit (PSU) only used in Southeast and West regions: Low = less than 20%, High 20% or more.

[3] Enrollment in the given grade was estimated by dividing the school enrollment for the school as listed in the 1989 Quality of Education Data (QED) file by the number of grades in the grade span of the school. This was used as the grade size.

[4] Tabulated from QED file.

NOTF NA - Not applicable.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 2-3. Substrata for grade 9 (Population B)

Substratum number	NAEP stratum characteristics				Substratum		Number of schools	
	Region[1]	Urbanicity	Minority level[2]	Certainty status	Control	Grade size[3]	Sample	Population[4]
1	Northeast	MSA	NA	Certainty	Public	15+	18	961
2	Northeast	MSA	NA	Certainty	Private	15+	3	599
3	Northeast	MSA	NA	Noncertainty	Public	15+	14	1,265
4	Northeast	MSA	NA	Noncertainty	Private	15+	3	453
5	Northeast	Non-MSA	NA	Noncertainty	All	15+	5	726
6	Southeast	MSA	High	Certainty	All	15+	4	278
7	Southeast	All	All	Noncertainty	Private	15+	3	882
8	Southeast	MSA	High	Noncertainty	Public	15+	9	750
9	Southeast	MSA	Low	Noncertainty	Public	15+	12	680
10	Southeast	Non-MSA	High	Noncertainty	Public	15+	10	1,003
11	Southeast	Non-MSA	Low	Noncertainty	Public	15+	9	1,078
12	Central	MSA	NA	Certainty	All	15+	10	619
13	Central	All	NA	Noncertainty	Private	15+	3	602
14	Central	MSS	NA	Noncertainty	Public	15+	22	1,695
15	Central	Non-MSA	NA	Noncertainty	Public	15+	14	2,826
16	West	MSA	High	Certainty	All	15+	9	471
17	West	All	All	Noncertainty	Private	15+	2	588
18	West	MSA	High	Noncertainty	Public	15+	18	857
19	West	MSA	Low	Noncertainty	Public	15+	19	1,103
20	West	Non-MSA	All	Noncertainty	Public	15+	11	1,863
21	All	All	All	All	All	<15	2	4,088

[1]Region definitions (note that these region definitions are those used by NAEP, and hence were used for forming strata for the Reading Literacy Study).

Northeast
Connecticut
Delaware
District of Columbia
Maine
Maryland
Massachusetts
New Hampshire
New Jersey
New York
Pennsylvania
Rhode Island
Vermont
Virginia (the part in
  Washington, DC MSA)

Southeast
Alabama
Arkansas
Florida
Georgia
Kentucky
Louisiana
Mississippi
North Carolina
South Carolina
Tennessee
Virginia (outside
  Washington, DC MSA)
West Virginia

Central
Illinois
Indiana
Iowa
Kansas
Michigan
Minnesota
Missouri
Nebraska
North Dakota
Ohio
South Dakota
Wisconsin
Utah
Washington
Wyoming

West
Alaska
Arizona
California
Colorado
Hawaii
Idaho
Montana
Nevada
New Mexico
Oklahoma
Oregon
Texas

[2]Minority level of primary sampling unit (PSU) only used in Southeast and West regions: Low = less than 20%, High = 20% or more.

[3]Enrollment in the given grade was estimated by dividing the school enrollment for the school as listed in the 1989 Quality of Education Data (QED) file by the number of grades in the grade span of the school. This was used as the grade size.

[4]Tabulated from the QED file.

NOTE: NA - Not applicable.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 2-4. Measures of estimated per-school enrollment by grade and size stratum

Grade and size stratum	Measure of estimated per-school enrollment[1] at grade level
Grade 4 (Population A)	
Fewer than 15 students[1]	7.6
15-49 students	
In MSA and private	26.0
In MSA and public	38.0
Not in MSA and private	21.0
Not in MSA and public	29.0
50 or more students	Actual grade size
Grade 9 (Population B)	
Fewer than 15 students	7.9
15 or more students	Actual grade size

[1] Enrollment in the given grade was estimated by dividing the school enrollment for the school as listed in the 1989 Quality of Education Data (QED) file by the number of grades in the grade span of the school.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

After assignment of the measures of size, the schools were drawn with probability proportional to size within the substrata. The first stage of the process was to identify the schools selected with certainty, which were defined as any schools with measures of size equal to or greater than three-fourths of the sampling interval used in the sample selection. These schools were given probability 1.00 of inclusion at the substratum level. All other schools selected in the substratum were given the probability of inclusion of 1 divided by the revised sampling interval, after exclusion of the certainty selections. The overall probability of selection of the schools is the product of the within-substratum probability of selection and the PSU probability of selection.

The sampling was done using WESSAMP, Westat's proprietary package for sample selection. This software also provides the overall probability of selection of the schools. The base weight of the selected school is the inverse of the probability of selection. These base weights were adjusted for school nonresponse, as described below.

Within the selected schools, classes (one to two) were drawn with equal probability without replacement. For grade 4, two classes were selected per school. For grade 9, one class was selected per school.

2.5 Weight Adjustments

School Weight Adjustments

The allocation of the sample of schools to the secondary stratification that employed control and size of school was based on estimates of the measures of size in those secondary strata. These estimates were made from the sample of NAEP PSUs and, hence, were not the true measures of size of the strata. Since the time that the design was determined, it has been possible to tabulate the entire QED file by the characteristics that define the substrata. This made it possible to adjust the sample weights so that the number of schools in the responding sample would weight up to the number of schools in the QED file within each substratum—a straightforward, poststratification procedure.

The enrollments in the sampled schools were multiplied by the school weights and compared with estimated enrollments for grades 4 and 9 produced by the Current Population Survey (CPS). The differences were judged to be large enough that a second adjustment to the sampling weights was made so that the estimated enrollments in the two grades would equal the CPS estimates within each NAEP region.

The two weight adjustments automatically correct for school nonresponse to the survey. In making the first adjustment, the weighted number of sampled schools was adjusted to equal the number of schools listed in the QED file, with no account taken of the number of closed schools. This handling of closed schools in the file counts was considered appropriate since there was no opportunity to include newly opened schools after the time of collection of the data for the QED file. A 1989 QED file was used in the sample allocation and drawing of schools. Some schools are closed every year, and some new schools are opened. Experience has shown, however, that turnover rates are low and, for the purposes of obtaining a cross-section of reading literacy, there is relatively little impact of failing to include every new school that has been opened since the frame for the design was determined. The added cost of updating the QED file was judged not to be worthwhile.

Student Weight Adjustments

The student weights within each school reflect both the subsampling of classrooms in the school and the individual student nonresponse within the school. That is, the school weight was multiplied by the number of classrooms in the school at the target grade and divided by thc number of classrooms sampled. This weight was multiplied by the number of students in the selected classrooms and divided by the number of responding students, to compensate for student nonresponse.

The distribution of student weights after adjustment is shown by substratum in Tables 2-5 and 2-6. Note that the range in weights within the substratum is never more than twice the average weight. The last substratum for each class represents the schools with an estimated enrollment of less than 15 students in the class. These two substrata were sampled thinly to conserve costs, so one expects their average weights to be high. However, the weighted sum of students in substratum 26 for grade 4 is only 3 percent of the total, so that the contribution to the average is small from the large weights in that substratum. For grade 9, substratum 21 weights up to only about 1.5 percent of the total grade 9 students.

2.6 Estimation of Averages, Ratios, and Proportions

With data from surveys such as the IEA Reading Literacy Study, the estimation of averages, ratios, and proportions can be derived by the same method. The estimated average over all students in the given grade is the weighted sum of the sample scores divided by the weighted number of all sample students taking the test. The same rule applies if the sum is taken over any subset of the sample, say, males, or Hispanics, or students who watch television more than 2 hours per day. For example, let

w_{ijk} = adjusted student weight for the *k*th student in the *j*th school in the *i*th PSU (or variance replicate as defined in the next section), and

y_{ijk} = narrative reading score of the *k*th student in the *j*th school in the *i*th PSU (or variance replicate).

Table 2-5. **The maximum, minimum, average, and sum of student weights, by substratum for grade 4 (Population A)**

Substratum	Maximum weight	Minimum weight	Average weight	Sum of weights
1	644.2	596.5	620.4	60,247
2	1,031.3	633.0	889.9	231,655
3	998.9	451.9	666.0	64,723
4	944.6	674.3	850.3	124,004
5	907.2	113.4	572.3	245,555
6	494.7	494.7	494.7	16,821
7	877.4	565.3	714.4	141,318
8	1,203.4	242.8	573.3	207,809
9	628.4	327.0	461.4	224,020
10	744.1	269.5	519.5	52,061
11	753.8	203.3	427.4	117,354
12	554.7	214.1	333.2	91,740
13	728.7	208.2	415.5	160,344
14	769.1	270.7	436.9	40,215
15	892.5	640.9	707.3	98,517
16	1,029.6	489.8	644.7	101,438
17	679.8	177.7	506.1	269,221
18	621.2	512.5	570.3	83,026
19	805.8	555.7	691.1	121,170
20	1,035.9	228.9	496.0	142,455
21	569.6	470.8	496.8	46,841
22	1,013.3	187.6	514.2	305,071
23	975.3	279.2	458.8	388,262
24	395.7	353.6	369.1	66,485
25	844.7	671.2	744.8	146,182
26	4,048.1	2,576.0	3,100.2	110,397
Total				3,656,929

NOTE: Details may not add to totals due to rounding.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Then the average score for the subset *i*, *j*, and *k* in the subset *s* is estimated as follows.

$$\bar{y}=\sum_s w_{ijk}y_{ijk}/\sum_s w_{ijk}. \qquad (1)$$

The same formulation is appropriate for any ratio of one variable to another. Th difference is that one replaces an estimate of the weighted total number of students in the denominator by the weighted sum of a variable.

For example, the ratio of hours of TV watching per hour spent on homework can be estimated as follows:

$$\bar{y}=\sum_s w_{ijk}y_{ijk}/\sum_s w_{ijk}x_{ijk} \qquad (2)$$

Table 2-6. **The maximum, minimum, average, and sum of student weights, by substratum for grade 9 (Population B)**

Substratum	Maximum weight	Minimum weight	Average weight	Sum of weights
1	2,483.9	671.1	1,378.7	213,508
2	2,628.8	1,376.7	1,868.7	134,824
3	2,071.8	1,180.6	1,663.7	289,340
4	1,019.3	329.6	674.4	16,516
5	3,562.4	263.6	1,147.8	63,120
6	1,145.6	376.6	844.8	60,409
7	1,251.0	518.5	834.5	47,657
8	1,723.3	316.2	797.2	117,513
9	1,436.4	671.0	958.6	239,497
10	2,208.1	228.3	1,318.3	219,710
11	978.6	588.9	800.1	133,714
12	1,576.8	534.3	1,064.5	187,838
13	1,592.2	678.0	1,154.4	217,428
14	1,786.1	671.9	983.5	327,937
15	1,445.4	418.6	935.5	229,549
16	1,837.8	256.5	1,219.8	224,310
17	610.7	610.7	610.7	10,383
18	3,038.2	343.4	710.4	185,294
19	2,742.3	326.2	1,375.9	380,447
20	1,563.5	679.5	1,044.0	199,549
21	4,599.9	4,599.9	4,599.9	55,199
Total				3,553,741

NOTE: Details may not add to totals due to rounding.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

where

y_{ijk} = hours of TV watching for the *k*th student in the *j*th school in the *i*th PSU (or variance replicate), and

x_{ijk} = hours spent on homework for the *k*th student in the *j*th school in the *i*th PSU (or variance replicate).

The formulation in (1) also works for estimating the proportion of students having a given characteristic, such as having a single parent. In that case, the variable in the numerator is 1 if the student has a single parent and 0 otherwise (y_{ijk} = 1 or 0).

2.7 Estimation of Sampling Errors

The sampling for the IEA Reading Literacy Study in the United States was designed so that standard errors could be estimated using the "ultimate cluster" method (Hansen, Hurwitz, and Madow 1953). The ultimate cluster is a grouping of sampled students for variance estimation purposes. For students in schools that are in PSUs that were not selected with certainty, the appropriate ultimate cluster is the PSU, since the aggregate for the PSU takes into account the stratification and allows for variation between PSUs within strata and between schools within the PSUs. For PSUs selected with certainty, the appropriate ultimate cluster is the school or aggregates of schools in the same certainty PSU. In this case there is no contribution from the variation among PSUs since the PSU was selected with certainty. In general, the use of ultimate clusters for sampling error estimation reflects the gains in precision from stratification and the loss in precision from clustering of the students within classrooms or within schools.

2.7.1 Jackknife Method of Variance Estimation

Sampling errors for the descriptive statistics were computed by the jackknife method (Rust 1985) using Westat's WESVAR software. To use this method, the noncertainty PSUs were grouped into pairs within the substrata, and within the certainty PSUs the schools were grouped into pairs. These pairings are termed variance estimation strata.

To compute the jackknife estimate, the mean or ratio or proportion, as defined in Section 2.6, is computed for the whole file (or for a given subset). Then, the first variance replicate is constructed by deleting one ultimate cluster at random in the first variance estimate stratum. That member of the pair is given zero weight and the weight of the other member is doubled. The average or ratio or proportion is recomputed for this set of weights. This constitutes the first replicate estimate. Call it E_i and denote the overall estimate computed above by E. Repeat this process for each of the variance replicates. Then, the standard error of the overall estimate, E, is given by

$$SE(E)=\sqrt{\sum_i (E_i-E)^2} \qquad (3)$$

where the summation is over all of the variance replicates and E_i denotes the estimate for replicate i.

2.7.2 Construction of the Variance Estimation Strata

Table 2-7 shows the 15 original sampling strata and the number of variance strata for grade 4 and grade 9. The contribution to variance of the noncertainty PSUs was estimated by pairing noncertainty PSUs. Each pair consisted of PSUs that were sampled from strata having similar characteristics. In total, 18 variance estimation strata were constructed from the noncertainty PSUs. One of the NAEP strata contained only one sampled PSU, so within-stratum pairing was impossible. This stratum was collapsed with the succeeding PSU in the next stratum in the ordered sequence (which contained three NAEP PSUs) to form a replicate pair, and the replicate weights were adjusted accordingly.

The contribution to variance of the certainty strata was estimated by pairing groups of schools in the certainty strata. Variance estimation strata were formed by grouping responding schools in the same region, with the same type of control (public or private), and the same enrollment class. Pairs, and in a few cases triplets, were formed within these classes. The sample of grade 4 schools in certainty strata consists of 32 responding schools in 12 PSUs. Fifteen variance strata were formed from these schools. Thirteen of them had two schools each, formed into two replicate pairs of one school each. Two had three schools each, grouped into three members of one school each. The grade 9 sample of certainty schools contained 33 responding schools in 13 certainty PSUs. Schools were grouped into 12 variance strata with one school in each member of the variance pair and three strata with three schools in each variance stratum.

Table 2-7. Summary of variance strata for jackknife estimation

Region	Urbanicity	Certainty	Minority	Number of variance strata	
				Grade 4	Grade 9
Northeast	MSA	Certainty	Al'	5	5
		Noncertainty	All	2	2
	Non-MSA	Noncertainty	All	1	1
Southeast	MSA	Certainty	High	2	1
		Noncertainty	High	1	1
		Noncertainty	Low	1	1
	Non-MSA	Noncertainty	High	1	1
		Noncertainty	Low	1	1
Central	MSA	Certainty	All	4	4
		Noncertainty	All	3	3
	Non-MSA	Noncertainty	All	2	2
West	MSA	Certainty	High	4	5
		Noncertainty	High	2	2
		Noncertainty	Low	2	2
	Non-MSA	Noncertainty	All	2	2
Total				33	33

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

When there were three PSUs (or groups of schools) per variance stratum, the method specified in Function 3 of WESWGT, a Westat proprietary program, was used to compute the weights to be attached to the two comparisons within the variance stratum. When there are three members within the variance stratum, say *a*, *b*, and *c*, one can form three estimates of the stratum from taking 1.5 times *a* + *b*, or 1.5 times *a* + *c*, or 1.5 times *b* + *c*. Function 3 of WESWGT chooses two of these estimates at random and, in effect, forms two replicates representing the *i*th variance stratum, say *i*(1) and *i*(2) where *i* represents the stratum with three members. The variance is computed as in (3), above, adding over all of the replicates, including the additional ones created by splitting the three members.

2.7.3 Jackknife Estimates of Standard Errors for Selected Subclasses

The jackknife estimation method of variance estimation serves quite well for most estimates, and the estimate, *E*, can take many forms. It is known not to be efficient in some circumstances for the estimates of quantiles such as medians or percentiles. However, empirical research (Hansen 1989) suggests that for multistage samples of PSUs, schools, and students, it has sufficient reliability for such measures for large samples such as estimates for the total student population, but may not be satisfactory for small samples, such as for estimates for small subgroups of the population.

Jackknife estimates of standard errors of means of reading scales were computed for the subclasses of family composition, language spoken, ethnicity, father's education, mother's educati‹n, gender, whether the student lives within a nuclear family or an extended family, region of residence, and degree of urbanization. Other subclasses could have been chosen, but these are likely to be used in many analyses of the data.

The jackknife estimates of standard errors are shown in Table 2-8 for grade 4 and in Table 2-9 for grade 9. The estimated standard errors differ somewhat by the reading scale used and by the variable defining the subgroup. There is a general tendency for standard errors to decrease as the sample size, n, increases, but the relationship is not linear. The fact that the sample is clustered (all of the students in a classroom were taken into the sample) causes the subclasses with large numbers of students in them to have relatively larger standard errors than can be accounted for by the sample size alone. With large subgroups there are, on the average, many students in a classroom who are members of the subgroup and, hence, the effect of the intraclass correlation (see Hansen, Hurwitz, and Madow 1953) is magnified in comparison with small subgroups with an average of one or two students in the classroom.

2.7.4 Estimated Design Effects

The design effect (DEFF) is the ratio of the variance of a statistic (square of the standard error), taking into account the stratification and clustering in the design, to the variance of the statistic that would have been achieved if the sample had been drawn as a simple random sample of the same size, i.e., without stratification or clustering. Except for binomial variables (and then with some limiting assumptions), the variance that could have been achieved under simple random sampling can only be estimated with substantial error, particularly when the number of PSUs in each stratum is small. To avoid this problem, the achieved variance is often compared with the variance computed by ignoring the design, that is, using the data drawn from the design but considering those data as a simple random sample.

This method of estimating DEFF is good if the design is self-weighting. The Reading Literacy Study design is not self-weighting, but the differences in weights across the strata are not great, as shown in Tables 2-5 and 2-6, except for the quite small stratum of very small schools. This approximation of the variance of a randomized design contains the positive effect of stratification, but ignores the effect of clustering. Since the effect of clustering tends to dominate the difference between the design variance and the simple random sample variance, the approximation yields estimates of the design effects that are useful in evaluating the design. The design effects for the three scales, and the same population subclasses as for the standard error analyses, are given in Tables 2-10 and 2-11.

The estimated design effects are increased when the cluster sizes (number of students in the classroom) increase and when the sampling weights vary from self-weighting. The design effects decrease as the effectiveness of the stratification increases.

A few variables have unusually large design effects. They include the Northeast region and cities with over 500,000 population for both grades 4 and 9, and cities with from 50,000 to 100,000 population for grade 9. These large values indicate a homogeneity within schools and a lack of homogeneity between schools in the strata from which these students were drawn. The sample sizes, in terms of schools, are so small, however, that one cannot generalize broadly from these data.

One way to use the design effects is to divide the actual sample size, n, by the design effect to achieve an "effective" sample size, that is, the size of a simple random sample that would have produced the same precision as the design sample size. For example, 1,047 grade 4 students were black. The design effect for the narrative reading scale for this subgroup was estimated to be 2.45, so the effective sample size was about 427. When making such interpretations, it should be remembered that the DEFF estimates are subject to a substantial amount of sampling error since the number of schools producing members of the subclass is small. Design effects of less than 1.0 typically are associated with small

Table 2-8. Jackknife estimates of standard errors for grade 4

Variable	Category	Sample size (n)	Standard errors		
			Narrative	Document	Expository
Family composition	No parents[1]	159	9.43	6.95	8.34
	One or both stepparents	209	5.85	6.59	5.01
	Mother only	671	5.29	4.37	4.03
	Mother & stepfather	428	5.04	3.45	5.53
	Father only	224	5.86	6.08	4.97
	Father & stepmother	165	7.38	6.64	7.97
	Mother & father	3,590	3.13	2.55	2.75
	Other groupings	802	4.89	4.26	4.54
Student's language at home and first language	English/English	4,657	3.13	2.63	3.06
	Other/English	86	7.80	6.40	7.60
	English/other	1,004	3.20	2.90	3.37
	Other/other	501	5.75	4.82	5.13
Race/ethnicity	White	4,219	2.26	2.11	2.63
	Black	1,047	4.28	3.06	4.82
	Hispanic	541	3.94	5.29	5.14
	Asian	246	8.39	7.14	7.21
	American Indian	195	11.74	8.57	7.43
Father's education[2]	Less than high school	607	6.11	4.77	4.21
	High school	1,454	3.47	2.69	3.42
	Some college	1,058	4.10	3.27	2.96
	College/university	2,926	3.36	3.07	3.44
Mother's education[2]	Less than high school	547	4.72	4.13	4.50
	High school	1,631	4.15	3.10	3.20
	Some college	1,274	4.41	3.22	3.66
	College/university	2,739	3.06	3.02	3.09
Gender	Male	3,153	3.63	3.02	3.18
	Female	3,095	3.10	2.81	3.03
Kind of family	Nuclear family	4,016	2.61	2.47	2.74
	Extended family	2,232	3.64	2.57	3.39
Region	Northeast	1,008	9.77	9.67	9.57
	Southeast	1,622	6.47	4.25	5.65
	Central	1,568	5.74	4.36	5.40
	West	2,050	3.54	2.74	3.39
Community[3]	Rural or farm	1,099	7.62	6.38	5.92
	Small town or city (<50 k)	1,290	6.18	6.12	6.34
	Medium size city (50k-100k)	774	7.56	6.75	7.16
	Suburb of medium size city	512	7.31	6.98	6.29
	Large city (100k-500k)	808	8.22	8.10	6.85
	Suburb of large city	641	10.14	6.78	9.24
	Very large city (Over 500k)	432	15.04	11.04	13.91
	Suburb of very large city	644	10.23	9.20	9.57
	All	6,248	2.94	2.85	2.57

[1] This category includes students who lived with siblings, grandparents, relatives, or nonrelatives.

[2] Sample sizes add to less than totals due to missing data.

[3] Sample sizes add to less than total because 48 students from a school on a military base were not included.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 2-9. Jackknife estimates of standard errors for grade 9

Variable	Category	Sample size (n)	Standard errors		
			Narrative	Document	Expository
Family composition	No parents[1]	58	16.00	9.18	17.16
	One or both stepparents	100	10.45	6.43	7.13
	Mother only	422	6.69	5.47	7.48
	Mother & stepfather	318	7.05	7.09	7.83
	Father only	83	10.80	11.02	12.11
	Father & stepmother	114	13.66	11.27	12.57
	Mother & father	1,945	5.54	3.99	6.29
	Other groupings	169	12.35	9.95	12.20
Student's language at home and first language	English/English	2,480	4.96	4.04	5.86
	Other/English	56	12.11	9.52	14.11
	English/other	388	6.47	6.90	8.17
	Other/other	285	9.01	7.81	12.30
Race/ethnicity	White	2,338	4.50	3.73	5.36
	Black	399	11.60	9.49	13.06
	Hispanic	269	11.29	7.97	10.58
	Asian	114	12.12	9.19	12.55
	American Indian	89	17.63	12.27	21.25
Father's education[2]	Less than high school	359	7.76	8.38	8.72
	High school	1,044	5.49	4.24	6.52
	Some college	622	7.08	5.90	7.12
	College/university	1,138	5.03	4.17	5.65
Mother's education[2]	Less than high school	346	8.24	6.81	7.13
	High school	1,104	6.09	4.59	7.54
	Some college	781	5.48	4.95	6.81
	College/university	970	4.81	3.79	5.31
Gender	Male	1,583	6.23	4.89	7.50
	Female	1,626	4.99	3.97	5.74
Kind of family	Nuclear family	2,691	4.90	3.79	5.68
	Extended family	518	7.54	5.54	8.15
Region	Northeast	524	15.68	10.52	17.70
	Southeast	878	7.70	7.37	10.02
	Central	819	8.22	7.40	10.00
	West	988	8.58	6.09	9.33
Community	Rural or farm	635	7.53	6.99	10.11
	Small town or city (<50 k)	831	7.41	6.21	7.86
	Medium size city (50k-100k)	320	25.44	15.99	26.46
	Suburb of medium size city	166	28.18	20.49	30.90
	Large city (100k-500k)	268	12.36	10.76	15.82
	Suburb of large city	259	18.36	11.09	19.97
	Very large city (Over 500k)	257	26.34	23.19	36.34
	Suburb of very large city	473	14.01	10.28	16.15
	All	3,209	4.98	5.71	3.87

[1] This category includes students who lived with siblings, grandparents, relatives, or nonrelatives.

[2] Sample sizes add to less than total due to missing data.

NOTE: Details may not add to totals due to rounding.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 2-10. Estimated design effects for grade 4

Variable	Category	Sample size (n)	Design effect		
			Narrative	Document	Expository
Family composition	No parents[1]	159	1.56	1.17	1.66
	One or both stepparents	209	0.87	1.93	1.12
	Mother only	671	1.92	2.05	1.63
	Mother & stepfather	428	1.30	0.89	2.39
	Father only	224	0.86	1.27	0.98
	Father & stepmother	165	1.00	1.35	1.63
	Mother & father	3,590	4.00	3.61	4.39
	Other groupings	802	2.13	2.70	2.75
Student's language at home and first language	English/English	4,657	4.90	4.90	6.77
	Other/English	86	0.74	0.71	0.75
	English/other	1,004	1.10	1.35	1.94
	Other/other	501	2.00	1.94	2.26
Race/ethnicity	White	4,219	2.45	3.02	4.66
	Black	1,047	2.45	2.15	5.02
	Hispanic	541	1.09	2.64	2.75
	Asian	246	1.66	1.53	1.74
	American Indian	195	2.68	2.16	2.10
Father's education[2]	Less than high school	607	2.52	2.28	1.98
	High school	1,454	2.04	1.87	2.93
	Some college	1,058	1.96	1.75	1.55
	College/university	2,926	3.59	4.02	5.31
Mother's education[2]	Less than high school	547	1.52	1.69	1.98
	High school	1,631	3.14	2.58	2.84
	Some college	1,274	2.79	2.14	2.80
	College/university	2,739	2.71	3.65	3.97
Gender	Male	3,153	4.33	3.99	4.89
	Female	3,095	3.38	4.12	4.63
Kind of family	Nuclear family	4,016	2.99	3.68	4.67
	Extended family	2,232	3.46	2.61	4.62
Region	Northeast	1,008	11.38	14.20	14.77
	Southeast	1,622	7.40	4.83	8.78
	Central	1,568	5.59	4.65	7.18
	West	2,050	2.73	2.24	3.60
Community[3]	Rural or farm	1,099	7.01	6.73	5.88
	Small town or city (<50 k)	1,290	5.11	7.59	8.75
	Medium size city (50k-100k)	774	4.79	5.47	6.33
	Suburb of medium size city	512	3.51	4.19	3.24
	Large city (100k-500k)	808	5.96	8.25	6.49
	Suburb of large city	641	7.01	4.60	8.46
	Very large city (Over 500k)	432	10.70	7.98	12.61
	Suburb of very large city	644	7.47	8.38	9.28
	All	6,248	5.90	7.95	6.29

[1] This category includes students who lived with siblings, grandparents, relatives, or nonrelatives.

[2] Sample sizes add to less than totals due to missing data.

[3] Sample sizes add to less than total because 48 students from a school on a military base were not included.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 2-11. Estimated design effects for grade 9

Variable	Category	Sample size (n)	Design effect		
			Narrative	Document	Expository
Family composition	No parents[1]	58	1.51	0.85	1.11
	One or both stepparents	100	1.33	0.91	0.62
	Mother only	422	2.08	1.94	2.20
	Mother & stepfather	318	1.98	2.65	1.94
	Father only	83	1.32	1.12	1.11
	Father & stepmother	114	2.01	2.93	1.70
	Mother & father	1,945	6.32	4.54	6.81
	Other groupings	169	2.46	2.33	2.43
Student's language at home and first language	English/English	2,480	6.37	5.94	7.55
	Other/English	56	0.73	0.75	1.04
	English/other	388	1.88	2.73	2.39
	Other/other	285	3.05	3.40	4.31
Race/ethnicity	White	2,338	5.39	5.26	6.31
	Black	399	6.09	6.45	7.01
	Hispanic	269	4.62	2.93	3.30
	Asian	114	1.81	1.42	1.66
	American Indian	89	2.96	1.92	5.08
Father's education[2]	Less than high school	359	2.47	4.10	2.98
	High school	1,044	3.52	3.07	4.25
	Some college	622	3.55	3.22	2.81
	College/university	1,138	3.14	3.01	3.39
Mother's education[2]	Less than high school	346	2.67	2.76	2.05
	High school	1,104	4.63	3.81	6.22
	Some college	781	2.66	2.96	3.33
	College/university	970	2.28	1.91	2.29
Gender	Male	1,583	6.37	5.17	7.59
	Female	1,626	4.53	4.13	5.01
Kind of family	Nuclear family	2,691	7.00	5.77	7.81
	Extended family	518	3.01	2.47	3.24
Region	Northeast	524	13.99	9.23	15.67
	Southeast	878	5.69	7.83	8.52
	Central	819	5.94	6.53	6.46
	West	988	7.95	5.36	8.48
Community	Rural or farm	635	4.48	4.57	6.45
	Small town or city (<50 k)	831	5.13	5.02	4.88
	Medium size city (50k-100k)	320	21.31	12.04	18.59
	Suburb of medium size city	166	13.09	8.94	13.85
	Large city (100k-500k)	268	3.91	4.24	5.27
	Suburb of large city	259	9.19	6.02	10.51
	Very large city (Over 500k)	257	20.42	20.96	26.97
	Suburb of very large city	473	9.10	7.50	11.34
	All	3,209	8.37	9.35	7.15

[1] This category includes students who lived with siblings, grandparents, relatives, or nonrelatives.

[2] Sample sizes add to less than totals due to missing data.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

subgroup sizes and with characteristics that are thinly distributed over the entire sample, that is, that are not clustered. In general, because of the sampling error, these estimates should be considered as being near 1.0.

2.7.5 Effect of Weight Variations on Design Effects

In order to estimate the effect of variation in the sampling weights on the design effect, the inflation factor due to variation in the weights was estimated (Kish 1965). To apply the method, one squares the unit weights for each PSU and multiplies by the proportion of the units in the stratum in that PSU, summing over the stratum. This sum is divided by the square of the sum of the PSU weights times their proportion in the stratum. In a self-weighting sample the ratio (i.e., the inflation factor) is equal to 1.0. In a non-self-weighting sample it is the factor by which the variance of the mean is increased due to unequal sample weights. Sample proportions rather than population values were used, so the estimates are not exact. However, they are indicative of the amount of variance due to variation in the weights. For the grade 4 sample, the stratum ratios varied from 1.00 to 1.30, with an average of 1.10. That is, for grade 4 the variance of the mean was increased by about 10 percent due to variation in the weights. For the grade 9 sample, the ratios by strata varied from 1.00 to 3.04, with an average of 1.25. That is, for grade 9, the variance of the mean was increased by 25 percent. As an approximation, one can divide the design effects in Table 2-10 by 1.10 and in Table 2-11 by 1.25 to obtain estimated design effects that could have resulted had all sampling weight been equal.

The square roots of the above factors represent the factors by which the standard error increases. Thus, on the average, the standard error of estimate from the IEA Reading Literacy Study is increased because of variation in the weights by 5 percent in grade 4 and by 12 percent in grade 9.

2.7.6 Effect of Clustering on Estimates of Variance and Design Effect

The clustering effect arises because students in the same class tend to be more homogeneous with respect to their scores than students in other classrooms in the same school and certainly than students in other PSUs. The general rule for measuring the effect of clustering is explained by the formula

$$Var(\bar{y}_c) = \frac{Var(y_r)[1+\rho(\bar{n}-1)]}{m\bar{n}} \tag{4}$$

where $Var(y_c)$ is the variance of the mean from the clustered sample, $Var(y_r)$ is the unit variance of a random sample, ρ is the intraclass correlation coefficient resulting from the clustering, m is the number of PSUs, and $\bar{n}$ is the average sample units (students) per cluster (classroom). As one can see, if the average classroom size is (say) 25, the variance of the mean will be increased by 24 times the intraclass correlation coefficient. Even a small intraclass correlation of (say) 0.01 will cause an inflation of about 24 percent in the variance of the mean, or about 11 percent in the standard error.

The intraclass correlation coefficients tend to be small (in the neighborhood of 0) for characteristics that do not differ greatly from cluster to cluster. An example is the gender of students in public schools. However, school policies, neighborhood environments, and instructional methods may combine to cause cluster-to-cluster variation in test scores, thus causing a substantial design effect. Also, clustering effects tend to be small for small subsets of the population and large for large subsets. The reason is that for large subsets the clusters tend to contain larger numbers of the members of the subset, that is, $\bar{n}-1$ becomes large and, when multiplied by ρ, the effect on the variance is large.

It should also be noted that for the grade 4 sample, two classrooms were usually drawn from the sampled schools. Only one was drawn when the enrollment was small. Thus, the clustering effects contain variation due to differences between classrooms in the same school and classrooms in different schools within the PSU. The net effect of this merging of difference in classrooms within and between schools is not known. One could argue that classrooms in the same school should be more alike than classrooms in different schools, but if there have been attempts to stratify classrooms within the school, either intentionally or unintentionally, the effect could be quite the opposite. This problem does not occur with the grade 9 sample since only one classroom was selected per school.

2.7.7 Generalized Standard Error

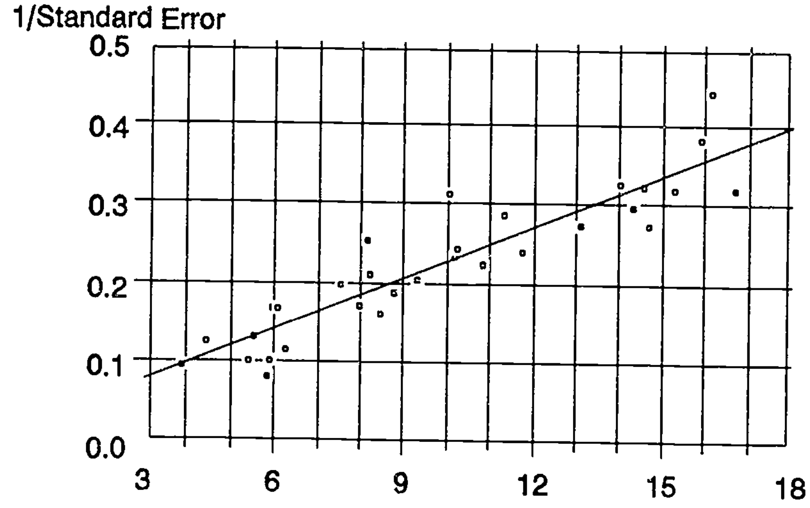
Since a substantial part of the amount of the design effect is related to sample size, it seems possible that one might be able to make estimates of standard errors, as functions of sample size, that would be sufficiently accurate for most analytic purposes. Various transformations of both subgroup sample size and the standard errors were tried in order to find a linear relationship between the transformed standard errors and the transformed subgroup sample sizes.

The result shows that the inverse of the standard errors was approximately a linear function of the cube root of subgroup sample size for variables that are well distributed over the population, that is, for variables that are not identified with one or more specific geographic areas, such as region or urbanicity. Variable categories (or subclasses) may be of three different kinds. Cross classes are those subclasses that are approximately evenly spread across the PSUs. Examples are gender and family composition. Mixed classes are those that appear in most PSUs, but that have an uneven distribution across the PSUs. Segregated classes are those that appear in a subset of the PSUs. An exampie is region. Generalized standard errors can be applied to estimates by cross classes, to a lesser extent for estimates for mixed classes, and not at all for estimates for segregated classes.

Figures 2-1 and 2-2 show the relationship between the inverses of the subgroup standard errors and the cube roots of the subgroup sample sizes for the variables that are not geographic in nature (that is, excluding region and urbanicity) for the narrative scales of both grades 4 and 9. The charts for the other two reading assessment scales were similar and are not shown here.

For grade 4, the fitted line for narrative scale is $100\,y = 1.5914 + 2.1354x$ where y = inverse standard error and x is the cube root of n, the subclass size. This line is a good fit; the squares of the correlation coefficients being 0.84 indicate that about 84 percent of the variation in the transformed standard errors is accounted for by variation in the subgroup sample size, n. The R^2 estimates for the expository and document scales were 0.85 and 0.86, respectively. For grade 9, the fitted line for narrative scale is $100\,y = 1.3916 + 1.5175x$. The R^2 estimates for the three assessment scales are 0.86, 0.88, and 0.77, respectively.

Figure 2-1. A plot of the inverse of standard errors as a function of sample size (n) for narrative scales[1] of grade 4



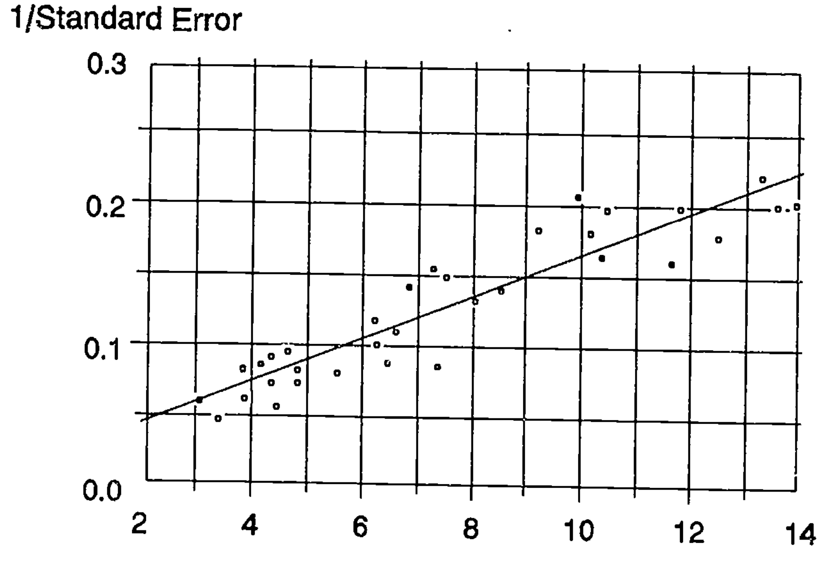
Cube Root of n

Fitted line: $100y = 1.5914 + 2.1354x$ $R^2 = 0.843$

[1] The plots for document and expository scales were very similar.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Figure 2-2. A plot of the inverse of standard errors as a function of sample size (n) for narrative scales[1] of grade 9



Cube Root of n

Fitted line: $100y = 1.3916 + 1.5175x$ $R^2 = 0.864$

[1] The plots for document and expository scales were very similar.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

The generalized estimates as converted to the untransformed scales are shown in Tables 2-12 and 2-13. These estimates may be used, with linear interpolation between adjacent subclasses, in lieu of computing the standard errors from the data. They are appropriate for means estimated for cross classes and, to a lesser extent, for mixed classes. However, generalized estimates should not be used for segregated classes. Estimates for sampling errors of means for such classes require estimation from the sample data.

The standard errors estimated from this relationship are, of course, subject to errors in the estimation of the true relationship. But individual estimates of the standard errors are also subject to a substantial amount of sampling error. Thus, in some instances, more credibility can be attached to the generalized standard errors than to the individually estimated standard errors. To provide some measure of the reliability of the estimates, the average absolute deviation around the fitted curves (in untransformed units) was computed for various classes of sample size.

Table 2-12. Generalized standard errors[1] of Reading Assessment scales, by subclass size for grade 4

Subclass size (n)	Narrative scale	Document scale	Expository scale
100	8.7	7.2	7.4
200	7.1	5.9	6.3
300	6.3	5.3	5.7
400	5.8	4.9	5.3
500	5.4	4.6	5.0
600	5.1	4.3	4.7
700	4.9	4.1	4.5
800	4.7	4.0	4.4
900	4.5	3.8	4.2
1000	4.4	3.7	4.1
1200	4.1	3.5	3.9
1400	3.9	3.4	3.7
1600	3.8	3.2	3.6
1800	3.6	3.1	3.5
2000	3.5	3.0	3.4
2500	3.3	2.8	3.2
3000	3.1	2.7	3.0
3500	2.9	2.5	2.9
4000	2.8	2.4	2.8
4500	2.7	2.3	2.7
5000	2.6	2.3	2.6

[1] These estimates are appropriate for cross classes that are evenly spread across the PSUs (such as gender and family composition) and, to a lesser extent, for mixed classes that appear in most PSUs but have an uneven distribution across the PSUs. However, they are inappropriate for segregated classes that appear in a subset of the PSUs only (such as region and urbanicity).

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 2-13. Generalized standard errors[1] of Reading Assessment scales by subclass size for grade 9

Subclass size (n)	Narrative scale	Document scale	Expository scale
100	11.9	9.8	12.3
200	9.8	8.0	10.5
300	8.7	7.1	9.5
400	8.0	6.5	8.8
500	7.5	6.0	8.3
600	7.1	5.7	7.9
700	6.7	5.4	7.6
800	6.5	5.2	7.4
900	6.2	5.0	7.1
1000	6.0	4.9	6.9
1200	5.7	4.6	6.6
1400	5.5	4.4	6.3
1600	5.2	4.2	6.1
1800	5.0	4.0	5.9
2000	4.9	3.9	5.7
2500	4.6	3.6	5.4
3000	4.3	3.4	5.1
3500	4.1	3.3	4.9
4000	3.9	3.1	4.7
4500	3.8	3.0	4.5
5000	3.7	2.9	4.4

[1] These estimates are appropriate for cross classes that are evenly spread across the PSUs (such as gender and family composition) and, to a lesser extent, for mixed classes that appear in most PSUs but have an uneven distribution across the PSUs. However, they are inappropriate for segregated classes that appear in a subset of the PSUs only (such as region and urbanicity).

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Tables 2-14 and 2-15 show the average absolute values of the deviations for grades 4 and 9. In general, the deviation decreases as sample size increases, and this trend is consistent across reading assessment scales. For both grades, the difference between the individually estimated standard error was typically less than 0.5 for subclasses with a sample size in excess of 1,000.

Table 2-14. The average absolute deviation of individual estimates of standard error from generalized estimate, by subclass size for grade 4

Subclass size (n)	Number of cases	Average absolute deviation		
		Narrative	Document	Expository
Less than 250	9	1.60	1.00	1.30
250-999	7	0.63	0.53	0.34
1,000-1,999	6	0.42	0.46	0.57
2,000 or more	9	0.29	0.26	0.16

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 2-15. The average absolute deviation of individual estimates of standard error from generalized estimate, by subclass size for grade 9

Subclass size (n)	Number of cases	Average absolute deviation		
		Narrative	Document	Expository
Less than 250	14	1.92	1.69	2.71
250-999	10	1.30	0.79	1.43
1,000 or more	9	0.48	0.31	0.61

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

2.8 Estimation of Degrees of Freedom

Survey data are often used to test hypotheses about the significance of differences among various subsets of the universe, although the fact that the data are based on a sample from the universe raises questions about the validity of such tests. More generally, confidence intervals are placed around estimates of the means of subsets of the data to provide the reader with a sense of the reliability of the means. These confidence intervals take the form of the estimated mean plus and minus a constant times the estimated standard error. For subsets of the universe having a large sample size, the normal distribution often provides the constants to be applied to the standard errors, for example, 1.96 for a two-sided 95 percent confidence limit, and so on. The assumption is that the means of samples drawn by the same method and same design from the same population will be normally distributed.

Complications arise for complex surveys when the standard errors are computed from the jackknife method. In this method the standard error is computed from a set of *r* variance replicates. If *r* is large, the assumption of normality may be a reasonable one, but if *r* is small, particularly when the subclass of interest is a segregated subclass, the assumption of normality of the distribution of standardized means may be unreasonable. That is, the sampling distribution of the ratio of sample mean to the estimated standard error is not well approximated by the normal distribution. In such cases, the *t* distribution is often used to supply the constant multiplier for the standard error in the computation of the confidence interval. But the appropriate value of *t* to be used as the multiplier for the estimated standard error depends on the degrees of freedom in the data from which it was estimated. For a sharply segregated subclass, for example, region of the country, it may be relatively easy to estimate degrees of freedom. If, say, eight variance strata are in a given region, one might assume that there should be eight degrees of freedom for estimates of the standard error of the mean for that region. But severe problems arise in estimating degrees of freedom for mixed classes and even for cross classes when the distribution is only approximately uniform across the PSUs.

For surveys with complex sample designs, special methods are required to estimate sampling errors and the degrees of freedom associated with the estimates. The degrees of freedom associated with estimates of sampling errors from complex surveys are not the degrees of freedom from a simple random sample. If the total sample size is large enough that the variance replicates for the estimation of sampling errors for cross classes by the jackknife or balanced randomized replication methods permit a large number of such replicates, say, 30 or more, the estimation of degrees of freedom is generally not critical. The numbers of replicates will, in general, be large enough that the assumption of normality of an estimate divided by its estimated standard error will be a reasonable one. However, when the sample design will not support a large number of variance replicates, or when the sample is wildly, unequally distributed across the strata, the effective degrees of freedom are often substantially smaller than the number of variance replicates. The effective number of degrees of freedom becomes quite small when estimating the sampling error of a statistic for a subgroup that does not appear in every stratum, or at least appears infrequently in some strata but frequently in others.

The problem was recognized by Satterthwaite (1941) and more recently by a number of researchers concerned with survey data: Cochran (1977), Notz and Bart (1990), Kott (undated), Johnson (1989), Johnson and Rust (1992), and others. Johnson (1989) suggests two approaches to the estimation of degrees of freedom, one that is appropriate for the variance of estimates of percent of test items answered correctly and one that is appropriate for the variance of estimates of mean test scores or scales, possibly including estimates of model parameters. The second is the one applicable to the analysis of the IEA Reading Literacy Study data and is described here.

2.8.1 An Approach to Estimating Degrees of Freedom

The approach is to create k independent random subsets of the replicates used for the estimation of variances, thereby creating k independent estimates of the variance ($k = 10$ was used in the analysis). Each of the k variances, when divided by d and multiplied by the true variance, is assumed to be distributed as chi-square, with d degrees of freedom. The true variance is unknown but can be estimated by the average of the k independent estimates. The number of degrees of freedom, d, is also unknown, but the chi-square distribution that matches most closely the distribution of the sample of the k sample values provides the estimate of d.

The estimation proceeded by comparing the average squares of the standard error for the 10 replicates with the theoretical values of chi-square for distributions having from 2 to 60 degrees of freedom. The match was truncated at 2 degrees of freedom if the value of the criterion for matching was still decreasing as the matching reached 2 from above. Similarly, if the value of the criterion was still decreasing as the matching reached 60 from below, the value of 60 was used as the estimated degrees of freedom. It should be noted that chi-square divided by degrees of freedom approaches normality as the number of degrees of freedom increases. There is little difference in the distribution of chi-square divided by degrees of freedom between 30 degrees of freedom ($t = 2.04$) and 60 degrees of freedom ($t = 2.00$). The normality assumption will serve adequately in this range. The primary reason for estimating degrees of freedom is to warn against making normality comparisons when the number of degrees of freedom is small, perhaps less than 10 or 15.

The above methods were applied to subsets of students having the characteristics used in the analysis of standard errors, as described in the previous section. The results were highly variable. It was observed, in looking at graphs comparing the theoretical chi-square distribution with the distribution derived from the data, that the poor fit was often due to a poor fitting of the end points. Dropping the end points, that is, the 1st and 10th deciles, stabilized the estimates somewhat.

It should be emphasized that the proper use of the estimated degrees of freedom is to serve as a caution when normality assumptions are used in tests of significance, rather than reliable estimates of the correct degrees of freedom. However, it was felt that the above method for estimating degrees of freedom produced estimates that were too variable, even for this limited purpose.

It was reasoned that the differences among the estimated degrees of freedom for the three reading scales, narrative, expository, and document, should be dominated by the differences within those scales, so the estimated degrees of freedom for the three scales were averaged to stabilize the estimated degrees of freedom still further. A method was needed to transform degrees of freedom for the purpose of averaging them across the three scales, so that a difference of, say, 5 degrees of freedom in the range of 5 to 10 degrees of freedom would be more important than a difference of 5 degrees of freedom in the range of 30 to 60 degrees of freedom. The two-sided t value necessary for significance at the 0.05 level was used as the transformation for this purpose. Both the average t values and the average degrees of freedom, as well as their ranges across the three scales, are shown in Tables 2-16 and 2-17. The average degrees of freedom were found by averaging the *t* values and transforming back to degrees of freedom. No categories with fewer than 100 students were included in the tables.

Table 2-16. Estimated average and range in *t* values and degrees of freedom (DF) for grade 4[1]

Variable	Category[1]	Sample size (n)	Average t(.05)	Range in t(.05)	Average DF	Range in DF
Family composition	No parents[2]	159	2.48	1.07	6	15
	One or both stepparents	209	2.25	0.36	10	43
	Mother only	671	2.07	0.13	24	45
	Mother & stepfather	428	2.40	0.31	7	3
	Father only	224	2.23	0.25	11	10
	Father & stepmother	165	2.98	2.04	4	6
	Mother & father	3,590	2.09	0.09	20	18
	Other groupings	802	2.47	0.58	6	8
Language at home and first language	English/English	4,657	2.10	0.16	19	21
	English/other	1,004	2.14	0.29	15	42
	Other/other	501	2.77	0.82	5	5
Race/ethnicity	White	4,219	2.18	0.29	12	14
	Black	1,047	2.16	0.20	13	27
	Hispanic	541	2.38	0.68	7	15
	Asian	246	2.10	0.16	19	33
	American Indian	195	3.29	1.53	3	1
Father's education	Less than high school	607	2.14	0.13	15	12
	High school	1,454	2.43	0.58	7	8
	Some college	1,058	2.14	0.36	15	47
	College/university	2,926	2.04	0.07	31	27
Mother's education	Less than high school	547	2.29	0.33	9	10
	High school	1,631	2.21	0.10	11	4
	Some college	1,274	2.38	0.62	7	9
	College/university	2,739	2.12	0.13	17	16
Gender	Male	3,153	2.10	0.15	18	34
	Female	3,095	2.12	0.05	17	6
Kind of family	Nuclear family	4,016	2.06	0.11	26	28
	Extended family	2,232	2.06	0.10	25	41
Region	Northeast	1,008	2.24	0.35	10	12
	Southeast	1,622	2.29	0.43	9	9
	Central	1,568	2.58	0.60	5	9
	West	2,050	2.42	0.67	7	14
Community	Rural or farm	1,099	2.57	0.41		4
	Small town or city (<50 k)	1,290	2.11	0.26	17	51
	Medium size city (50k-100k)	774	2.30	0.25	9	5
	Suburb of medium size city	512	3.31	1.86	3	4
	Large city (100k-500k)	808	2.19	0.33	12	24
	Suburb of large city	641	2.71	0.82	5	5
	Very large city (Over 500k)	432	2.24	0.15	10	5
	Suburb of very large city	644	2.49	0.66	6	13

[1] Subclasses with fewer than 100 students were eliminated.

[2] This category includes students who lived with siblings, grandparents, relatives, or nonrelatives.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 2-17. Estimated average and range in *t* values and degrees of freedom (DF) for grade 9

Variable	Category[1]	Sample size (n)	Average *t*(.05)	Range in *t*(.05)	Average DF	Range in DF
Family composition	One or both stepparents	100	3.42	1.53	3	56
	Mother only	422	2.12	0.23	16	49
	Mother & stepfather	318	2.27	0.14	9	47
	Father & stepmother	114	3.18	0.00	3	52
	Mother & father	1,945	2.00	0.01	57	7
	Other groupings	169	2.14	0.22	15	43
Language at home and first language	English/English	2,480	2.01	0.02	52	21
	English/other	388	2.18	0.28	12	42
	Other/other	285	2.57	0.92	6	46
Race/ethnicity	White	2,338	2.09	0.16	19	27
	Black	399	2.09	0.19	20	20
	Asian	114	2.29	0.22	9	28
	Hispanic	269	2.60	0.92	5	31
Father's education	Less than high school	238	2.32	0.14	8	21
	High school	1,044	2.09	0.04	20	9
	Some college	622	2.09	0.06	20	10
	College/university	1,138	2.12	0.31	17	52
Mother's education	Less than high school	246	2.21	0.36	11	16
	High school	1,104	2.05	0.08	28	9
	Some college	781	2.31	0.52	8	19
	College/university	970	2.19	0.19	12	12
Gender	Male	1,583	2.00	0.00	60	41
	Female	1,626	2.09	0.15	20	28
Kind of family	Nuclear family	2,691	2.01	0.04	49	43
	Extended family	518	2.15	0.26	14	21
Region	Northeast	524	2.06	0.13	26	33
	Southeast	878	2.23	0.25	11	39
	Central	819	2.31	0.32	8	39
	West	988	2.23	0.10	10	36
Community	Rural or farm	635	2.47	0.51	6	36
	Small town or city (<50 k)	831	2.04	0.07	34	27
	Medium size city (50k-100k)	320	2.37	0.62	7	38
	Suburb of medium size city	166	2.64	0.21	5	37
	Large city (100k-500k)	268	2.64	0.21	5	34
	Suburb of large city	259	2.46	0.21	6	33
	Very large city (Over 500k)	257	2.25	0.53	10	23
	Suburb of very large city	473	2.21	0.37	11	30

[1] Subclasses with fewer than 100 students were eliminated.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Although the degrees of freedom shown in Tables 2-16 and 2-17 are still quite variable, they do serve as warnings concerning the subclasses for which one should use the *t* distribution rather than the normal distribution in setting confidence limits on the mean. The results from the estimates of degrees of freedom appear to call into question the validity of the assumption that the variances of the 10 random subsets are distributed as chi-square. More experience in using the method on the results from other complex designs is needed before one can generalize about the usefulness of the method. A more complete discussion of the technical matters involved in the estimation of degrees of freedom appears in the appendix to this chapter.

2.9 Concluding Remarks

This chapter summarizes the sampling procedure used by the United States to select samples of schools and classes of students to participate in the IEA Reading Literacy Study. Since a complex sample design was used, special methods are required to estimate standard errors and the degrees of freedom in hypothesis testing. The jackknife method of variance estimation was used, and variance estimation strata were defined for this purpose. Generalized standard errors were calculated for subclasses of sizes ranging between 100 and 5,000. The generalized standard errors can be used for cross classes (subclasses that are evenly spread across PSUs, such as gender and family characteristics), and for mixed classes that appear in most PSUs. However, they are not recommended for segregated classes that appear in selected PSUs (such as region and urbanicity) and when the subclasses are very small (fewer than 100 in sample).

When survey data are used to test hypotheses about significant differences between subclasses, the degrees of freedom associated with estimates of standard error are smaller than the degrees of freedom from a simple random sample. Current methods to estimate degrees of freedom are not very reliable; however, they serve as a warning that normality assumptions may be inappropriate for certain subclasses.

References

Cochran, W.G. (1977). *Sampling techniques*, 3rd ed., New York: John Wiley & Sons.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). *Sample survey methods and theory*, Vol. 1, New York: John Wiley & Sons.

Hansen, M.H. (November 1989). Comparison of jackknife estimates of standard errors of the mean and median. Memorandum to Eugene Johnson.

Johnson, E.G., and Rust, K.F. (1992). Population inferences and variance estimation for NAEP data, *Journal of Educational Statistics*, 17(2), 175-190.

Johnson, E.G. (1989). Considerations and techniques for the analysis of NAEP data. *Journal of Educational Statistics*, 14(4), 303-334.

Kish, L. (1965). *Survey sampling*. New Yori.: John Wiley & Sons.

Kott, P.S. (undated). Hypothesis testing of linear regression coefficients with survey data. Bureau of the Census, unpublished manuscript.

Notz, W., and Bart, J. (1990). Degrees of freedom for estimates of the variance in multistage sampling plans, *Biometrics*, 46(3), 873-4.

Ross, K.N. (1991). *Sampling manual for the IEA International Study of Reading Literacy*. University of Hamburg, Hamburg, Germany: International Coordinating Center, IEA International Study of Reading Literacy.

Rust, K., (1985). Variance estimation for complex estimators in sample surveys. *Journal of Official Statistics*, 1(4), 381-397.

Satterthwaite, F.E. (1941). Synthesis of variance, *Psychometrika*, 6, 309-316.

Appendix
Estimation of Degrees of Freedom

Various methods are used in estimating degrees of freedom in experimental data. The typically complex design of surveys makes the traditional methods less relevant in survey applications because of the greater uncertainty about the distribution of variances in such data. The method chosen for investigation in this report is the one suggested by Johnson[1] in which degrees of freedom are estimated from the standard errors of random subsets of the sample data. Two questions arise in employing the method. First, how accurately can one determine the degrees of freedom from *k* values independently drawn from a chi-square distribution with *d* degrees of freedom where *d* is unknown? If the answer to this question is satisfactory, the second question is, For a selection of sample estimates from the survey data, how reasonable is it to assume that the *k* sample variances are distributed as chi-square?

An answer to the first question was provided by a small simulation experiment. From chi-square distributions with each of 5, 10, and 15 degrees of freedom, 100 samples of 10 each were drawn and the number of degrees of freedom was estimated by three methods. The first method selected the chi-square distribution (determined by the parameter *d*) such that the sum of the squares of differences between the 10 ordered sample values and the corresponding values of the test distribution was a minimum. The second method used the sum of the absolute deviations as a criterion. The third method used the maximum difference between the cumulative sample and the test distributions, thereby using the criterion employed in the Kolmogorov-Smirnov test. The results are shown in Appendix Tables 1, 2, and 3.

There is little difference between methods 1 and 2. Method 3 provides similar results, but tends to be biased downward by about half a degree of freedom. Although the number of simulations run was small (100), running more than 100 simulations appeared not to be worthwhile since no major differences were shown among the methods. The sum of squares of deviations was used as the basic criterion in this methodology report, although some modifications were introduced later. From Appendix Tables 1 through 3 one can see that nearly all of the estimated degrees of freedom will be within plus or minus 3 degrees of freedom when the "true" degrees of freedom is 10. This appears to be sufficient accuracy for the determination of degrees of freedom since the principal objective is to provide a wider confidence interval when the assumption of normality is inappropriate. For example, if the correct degrees of freedom is 10, the appropriate *t* value to use in the construction of a 95 percent confidence interval is 2.23. Allowing for estimation error, estimating 7 degrees of freedom would produce a *t* value of 2.36, while estimating 13 degrees of freedom would produce a *t* value of 2.16. Both estimates are substantially higher than 1.96, which would be the appropriate *t* value for a large number of degrees of freedom, i.e., using the normality assumption.

Another method investigated was to average the 10 sample values in each data set and, since the mean of a chi-square distribution is equal to the degrees of freedom and the variance is twice the degrees of freedom, to estimate the degrees of freedom via the method of moments. This method produced highly variable estimates. It was clear that any of the methods discussed above would produce better estimates, so this method was not considered for the analysis.

The estimation proceeded by comparing the average squares of the standard error for the 10 replicates with the theoretical values of chi-square for distributions having from 2 to 60 degrees of freedom. The match was truncated at 2 degrees of freedom if the value of the criterion for matching was

[1] Eugene G. Johnson. Considerations and Techniques for the Analysis of NAEP Data. *Journal of Educational Statistics*, Vol. 14, No. 4, pp. 303-334, 1989.

still decreasing as the matching reached 2 from above. Similarly, if the value of the criterion was still decreasing as the matching reached 60 from below, the value of 60 was used as the estimated degrees of freedom. It should be noted that chi-square divided by degrees of freedom approaches normality as the number of degrees of freedom increases. There is little difference in the distribution of chi-square divided by degrees of freedom between 30 degrees of freedom (t=2.04 for 95 percent confidence) and 60 degrees of freedom (t=2.00). The normality assumption will serve adequately in this range. The primary reason for estimating degrees of freedom is to warn against making normality comparisons when the number of degrees of freedom is small, perhaps less than 15 or 20. If the estimated degrees of freedom is 30 or more, there is little harm in using the normality assumption.

Table 1. Frequency distribution of difference between the actual and estimated degree of freedom: Results from 100 simulation trials per method, using a chi-square distribution with d = 15

Difference	Method 1	Method 2	Method 3
-3	7	5	16
-2	17	16	16
-1	15	20	22
0	19	17	18
1	22	20	19
2	13	13	6
3	5	5	3
4	2	4	0
	100	100	100

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 2. Frequency distribution of difference between the actual and estimated degree of freedom: Results from 100 simulation trials per method, using a chi-square distribution with d = 10

Difference	Method 1	Method 2	Method 3
-3	1	1	6
-2	15	15	20
-1	24	25	28
0	28	26	24
1	16	18	13
2	10	7	7
3	5	7	2
4	1	1	0
	100	100	100

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 3. Frequency distribution of difference between the actual and estimated degree of freedom: Results from 100 simulation trials per method, using a chi-square distribution with d = 5

Difference	Method 1	Method 2	Method 3
-2	9	5	9
-1	19	25	33
0	39	40	38
1	23	20	18
2	8	9	1
3	2	0	1
4	0	1	0
	100	100	100

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

The above methods were applied to subsets of students having the characteristics used in the analysis of standard errors, as described in Section 2.7. The results were highly variable. Although it is possible for a subgroup to have a larger number of degrees of freedom for a given scale than for another scale, such variation is likely to be small when compared to the sampling and structural error in the computation. Several of the estimates of degrees of freedom seemed inconsistent from scale to scale, much more variable than one would have expected. Further, some of the degrees of freedom defied logic when compared with the number of variance strata used in the estimation of standard errors by the jackknife method. What one would expect is that degrees of freedom for categories of students that are reasonably well distributed across the United States, (i.e., cross-classes, as defined in Section 2.7.7) such as gender categories, would range between about 18 and 33. One might assume that a maximum of 33 degrees of freedom is possible since there were 33 variance replicate pairs. However, only 18 of these replicate pairs (for grade 4) were in the noncertainty PSUs. Schools, rather than PSUs, were paired in the certainty PSUs. The pairing is somewhat arbitrary, and since the certainty PSUs had probability one of being included, one could argue that a single pair, consisting of random halves of all schools in the certainty PSUs, could have represented the variation among all of the schools in such strata. Thus, one would expect the degrees of freedom to be somewhere between 18 (the number of noncertainty pairs) and 33 (the total number of pairs). Hence it is not surprising that the degrees of freedom is, on the average, less than the number of pairs.

It should be emphasized again that the proper use of the estimated degrees of freedom is to serve as a caution against using normality assumptions in the construction of confidence intervals. It is unrealistic to expect the estimates to be precise. However, it was felt that the above method for estimating degrees of freedom produced estimates that were too variable, even for this limited purpose. Consequently, some other methods were investigated. Appendix Tables 1 through 3, above, showed that a variation of plus or minus 3 degrees of freedom is to be expected when the correct degrees of freedom is *known* to be around 10. When this variation is coupled with uncertainty about the distribution of the random subsample variances and with the amount of sampling error inherent in the estimates, the variation in degrees of freedom among the three scales is not surprising.

Two methods were tried in an effort to stabilize the estimates of degrees of freedom. One method involved fitting only the middle eight or the middle six deciles of the chi-square-over-degrees-of-freedom distribution to the middle eight or the middle six sample values of the variances divided by their average. It had been observed that the end points tended to contribute the greatest amount to the sum of squares of differences between the theoretical and actual values, and it was believed that truncating the fit in this way would tend to stabilize the estimates. The second method involved fitting a simple regression to both the middle eight cumulative chi-square-over-degrees-of-freedom values and the middle eight sample values, and choosing as the estimated number of degrees of freedom the chi-square distribution with the slope most nearly equal to the slope of the cumulated sample values. It can be demonstrated easily that the slope of the line of regression of the deciles of chi-square reduces monotonically as the number of degrees of freedom increases, and dropping off the lowest and highest deciles causes the chi-square-over-degrees-of-freedom cumulated distribution to be reasonably well fitted by a straight line.

The criterion used to distinguish among the methods was the range in degrees of freedom among the three test scales—narrative, document, and expository—under the assumption that the true differences among the test scales would be dominated by the sampling and model errors in the methods used in estimating degrees of freedom. Appendix Table 4 summarizes the results.

Using the sum of squares of differences after dropping the end points gave the most consistent estimates for both populations (line 2, Table 4) and, hence, was the method used in presenting the estimates that appear in Tables 2-16 and 2-17. As discussed above, a method was needed to transform degrees of freedom for the purpose of averaging them across the three scales, so that a difference of, say, 5 degrees of freedom in the range of 5 to 10 degrees of freedom would be more important than a difference of 5 degrees of freedom in the range of 30 to 60 degrees of freedom. The two-sided *t* value necessary for a 95 percent confidence level was used as the transformation for this purpose. Both the average *t* values and the average degrees of freedom are shown in Tables 2-16 and 2-17, as well as their ranges across the three scales. The average degrees of freedom were found by averaging the *t* values and transforming back to degrees of freedom.

Table 4. Comparison of methods of estimating degrees of freedom

Method	Percent with maximum range	
	Grade 4	Grade 9
Minimum sum of squares, 10 deciles	45	49
Minimum sum of squares, 8 deciles	10	8
Minimum sum of squares, 6 deciles	15	15
Sum of squares from regression	30	›27

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

It is evident from Tables 2-16 and 2-17 that the estimates of degrees of freedom are still highly variable. A measure of variability is available for a part of the grade 9 sample in which two different randomizations of the values within the variance strata were used in order to choose the 10 random subgroups in the estimation of degrees of freedom. These two independent randomizations provide an estimate of the standard error in the estimation of degrees of freedom that is due to sampling error. The results are shown in Appendix Table 5 in terms of the standard deviation of the estimated *t* value for the 95 percent confidence interval. Since it is clear that there can be greater variation when *t* is large (i.e., when degrees of freedom is small) than when *t* is small, the results of the analysis are shown for five size classes of average degrees of freedom, where the average was derived from the two randomizations.

Table 5. Standard errors of estimated degrees of freedom, grade 9

Average estimated degrees of freedom	Average *t* values			Number of pairs			Standard deviations			Average standard deviations
	Narrative	Document	Expository	Narrative	Document	Expository	Narrative	Document	Expository	
Under 5	3.51	3.35	3.56	10	11	9	0.46	0.60	0.47	0.51
5-9	2.50	2.55	2.41	10	11	13	0.50	0.43	0.21	0.38
10-14	2.31	2.26	2.21	5	8	8	0.35	0.25	0.12	0.24
15-24	2.13	2.12	2.11	9	8	5	0.08	0.11	0.08	0.09
25 and higher . .	2.06	2.05	2.02	11	7	10	0.14	0.07	0.03	0.08

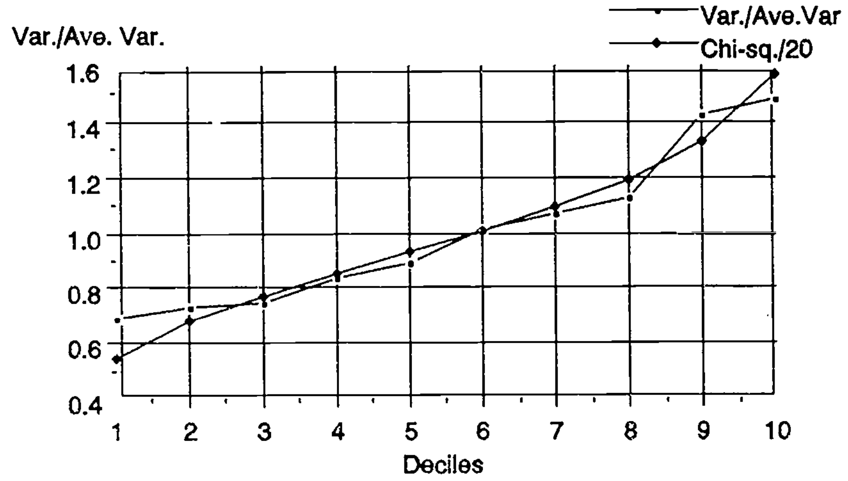
SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

To be conservative, it appears that one needs to add about 20 percent to the *t* value for estimates of degrees of freedom under 5, about 15 percent for degrees of freedom greater than 5 but less than 15, and about 10 percent for estimated degrees of freedom equal to or greater than 15. These are rough figures, but they may be useful for some purposes.

A remaining issue is whether it is reasonable to assume that the variances of the *k* replicates divided by their average are distributed as chi-square. The assumption appears to be reasonable for variables that have similar estimates of degrees of freedom for the three scales. In these cases the Kolmogorov-Smirnov test fails to discriminate the scaled distribution of replicate variances from a chi-square distribution. Appendix Figure 1 shows th fit between the replicate variances of narrative reading scores for the variable "Other" family compositi n and a chi-square distribution with 20 degrees of freedom. All 10 deciles were used in the fitting. Note the departure from a good fit for deciles 1 and 10. This problem is avoided by truncating the smallest and largest deciles.

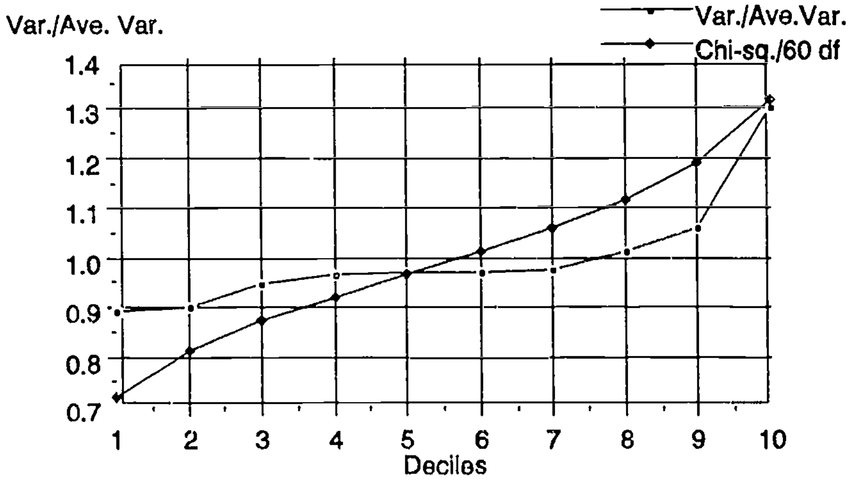
When the estimates for the three scales differ widely, generally one or more of the scales shows a poor fit. An example is shown in Appendix Figure 2 where the variable is the existence of an extended family for students in grade 4. It is clear that the assumption that the sample variances divided by their average is distributed as chi-square is unjustified. The source of the problem is unknown, but the analysis above with the two randomizations of a single sample indicate that sampling error may play a large role. There is, of course, the distinct likelihood that estimates made from a stratified and highly clustered sample will often depart from normality. When that occurs, the validity of this method of estimating degrees of freedom is called into question.

Figure 1. Actual versus theoretical chi-square distributions for "other" family composition, narrative scale, grade 4



SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Figure 2. Actual versus theoretical chi-square distributions for extended family variable, narrative scale, grade 4



SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

3 Handling Item Nonresponse in the U.S. Component of the IEA Reading Literacy Study

Marianne Winglee, Graham Kalton, Keith Rust, and Dan Kasprzyk

3.1 Introduction

This chapter discusses the handling of item nonresponse in the U.S. component of the IEA Reading Literacy Study. As in any survey, there are some nonresponses to the questionnaire items. There are many reasons for nonresponses. Sometimes a respondent will not remember or will not know an answer, will refuse to give it, or will inadvertently skip answering an item. Sometimes responses will need to be deleted because they fail to satisfy edit constraints. For some of the nonresponses, answers can be deduced through logical deductions and data edits. However, not all item nonresponses can be completed this way. The presence of nonresponses complicates analyses of the survey data. Therefore, imputation procedures (hot-deck and modal imputations) have been used to produce a completed data file for analysis.

With the completed data file, the most straightforward way to analyze the data is to proceed as if the completed data were actually reported. This study compares regression models estimated from the dataset completed by imputation with the corresponding models estimated using three other methods of handling the missing data. These other methods are the complete case analysis (CC), where cases with missing values for any of the variables involved in the analysis are discarded (also known as the listwise deletion method); the available case (AC) analysis, where all the reported data are used to derive the sample means and variance-covariance matrix employed in the regression analysis (also called the pairwise deletion method); and a method that assumes that the data come from a multivariate normal distribution and estimates the parameters of this distribution by a maximum likelihood method using the EM algorithm (estimation-maximization method). In addition, hierarchical linear model (HLM) analyses are estimated for the dataset completed by imputation and using the complete case analysis.

Section 3.2 of this paper describes the analytic objectives for the U.S. component of the IEA Reading Literacy Study and nonresponse in this survey. The methods of imputation are covered in Section 3.3. Section 3.4 discusses the methodology used to compare the outcomes of multiple regression analyses and hierarchical linear modeling analysis using the imputed data and three other methods of handling missing data. Section 3.5 is a discussion of the efficiencies of the various methods.

3.2 The U.S. Component of the IEA Reading Literacy Study

The U.S. component of the IEA Reading Literacy Study (Elley 1992) was conducted in the 1990-91 school year. The study involved national samples of over 6,000 grade 4 and 3,000 grade 9 students. The grade 4 students were sampled from over 160 schools nationwide, and two complete classes of students were selected per school. The grade 9 students were sampled from about the same number of schools over the country, and one class of students was included per school. Students sampled for the study were given performance tests to evaluate their reading levels and comprehension. In addition, the students, their teachers, and school principals completed questionnaires about background factors related to the students' reading achievement. Student performance on the cognitive tests was scored using the Rasch scaling method, and nonresponses to the cognitive test items were handled within the context of the Rasch model (see Elley 1992). The item nonresponses discussed in this paper refer to nonresponses to the questionnaire items only.

The aims of the IEA Reading Literacy Study were to assess school children's reading proficiency in the language of their own country and to collect information from students, teachers, and schools about the factors that lead some students to become better readers than others. While the prime focus of the study is on international comparisons, an additional objective for the United States is to develop conceptual models of which factors are effective and which are ineffective in improving reading skills in the U.S. school systems. In order to develop these models, questionnaire items are often used as independent variables for predicting student's reading performance. Therefore, it is important to have complete data on the questionnaire items to facilitate these analyses.

Survey Nonresponse

Like most surveys, the IEA Reading Literacy Study experienced two types of missing data--unit nonresponse and item nonresponse. Unit nonresponse occurred when a sampled school refused to participate in the study or a student from a participating school failed to complete the reading performance tests for reasons of absenteeism, health, or language problems. At the school level, response rates in the United States were about 87 percent for grade 4 schools and 86 percent for grade 9 schools. (These rates exceeded the international requirement of at least 85 percent for each grade.) At the student level, about 7 percent of the grade 4 students and 14 percent of the grade 9 students were unit nonrespondents. Weighting class adjustments were used to compensate for unit nonresponse at both the school and student levels (see Chapter 2 of this volume).

Item nonresponse to the questionnaire items occurred when a student who completed the reading performance test failed to complete an item on the student background questionnaire, or when a teacher or principal failed to complete an item on the questionnaires that they completed. Possible reasons for item nonresponse include lack of knowledge, inadvertent omissions, refusals, and edit failures. As discussed below, the level of item nonresponse was generally low, but there were some items that were not answered by 10 percent or more of the respondents.

Questionnaire items that were unanswered by respondents were reviewed, and efforts were made to locate the missing responses or to deduce the responses by means of logical edits. For example, for schools that failed to report the type of school or communities they served, hard copies of the questionnaire form were retrieved to check the address of the school and to deduce the missing response. For items for which data are available from other sources, those data were used to replace the missing values. For example, for schools that failed to report enrollment, the enrollment was taken from the 1989 Quality Education Data (QED) file, a comprehensive database of schools in the United States that was used as the sampling frame of schools for this study.

A small amount of deductive editing was used to complete the responses for items for which unique responses could be deduced from responses to other items on the questionnaires. Examples include the following: (1) for children who reported their dates of birth but not their ages, the missing ages were deduced from birth dates; (2) if a grade 9 student responded that he or she did not have a regular job, a nonresponse to the question about the time spent on jobs was assigned a response of "not applicable"; (3) for certain list checking items requiring a "yes" or "no" response to multiple parts of a question, partial nonresponses to some parts were deduced to have a "no" response. For example, when school principals were asked whether they use students' standardized tests to evaluate student progress, curriculum, teachers, textbooks/ materials, and special programs, they tended to circle "yes" choices but left other items unanswered. The unanswered items were inferred to have "no" responses.

After the process of data review and edits, the amount of missing data in the U.S. component of the IEA Reading Literacy Study is relatively small. Table 3-1 summarizes the extent of item nonresponses in each of the six datasets corresponding to the three questionnaires for each of the grade 4 and grade 9 samples. Items in each questionnaire are separated into three categories according to the amount of missing data: 5 percent or less of missing data, between 6 and 10 percent of missing data, and 11 percent or more of missing data. Generally, the percentage of items with 11 percent or more of missing data is small. Close to 90 percent of the items on the School Questionnaire have no more than 5 percent of missing responses. Similarly, 92 percent of items on the grade 4 Teacher Questionnaire and 84 percent of the items on the grade 9 Teacher Questionnaire have 5 percent or less of missing data. The grade 9 students too provided reasonably complete responses, with 87 percent of the items having less than 5 percent missing data. The grade 4 students had a higher level of item nonresponses, with 20 percent of the items on their questionnaires having 11 percent or more of missing data. Even though the item nonresponse rates are higher for grade 4 students, about 35 percent of these students completed all items, and about 97 percent of them completed at least 80 percent of the items on the questionnaire.

Table 3-1. Percentage of items with different levels of missing data

Percentage of questionnaire items with given level of missing data	Grade 4	Grade 9
Student Questionnaire		
5 percent or less	54%	87%
6-10 percent	26%	5%
11 percent or more	20%	8%
Total number of items	134	241
Teacher Questionnaire		
5 percent or less	92%	84%
6-10 percent	5%	14%
11 percent or more	3%	2%
Total number of items	250	153
School Questionnaire		
5 percent or less	89%	87%
6-10 percent	4%	12%
11 percent or more	7%	1%
Total number of items	113	117

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

There are three types of questionnaire items with high nonresponse rates. The first type comprises factual items that require a certain degree of effort in information retrieval. For example, teachers frequently omitted the number of education courses they had taken. This information should have been available to all teachers provided that they were willing to make the effort to review their training records. The second type comprises items that may have been unclear to some respondents. For example, teachers were asked about the amount of time they spent teaching "ESOL"—English as second language. Since the term ESOL was not defined in the questionnaire, some teachers may have been confused and therefore did not provide a response. The third type comprises items in which the response categories may be inappropriate. When none of the choices was suitable, respondents may have decided to skip the question. For example, school principals were asked to rate their levels of satisfaction with various forms of student assessment. The principal may have omitted the item because the form of assessment was not used in the school.

3.3 Methods of Handling Missing Data in Surveys

The missing responses in the questionnaire items can be handled in one of two ways: they can be filled in by some imputation procedure, or they can be left as missing with missing data codes assigned in the data files. The use of imputation to assign values for item nonresponses in large surveys has a number of advantages (see Kalton and Kasprzyk 1982; Kalton 1983). One is that carefully conducted imputations can reduce the biases in survey estimates arising from missing data. Second, with data assigned by imputation, analyses can be conducted as if the dataset were complete. Thus, analyses are made easier to conduct and the results easier to present. Third, the results obtained from different analyses with missing data are likely to be somewhat inconsistent, a feature that need not apply with analyses of a dataset with imputed data.

The alternative to imputation is to leave to the secondary data analyst the task of compensating for the missing data. This alternative is preferred by those who maintain that methods of handling missing data should be developed to satisfy specific analytic models or objectives. Such an approach may be preferable in some cases, but it is often impractical and unsatisfactory. Most analysts who are confronted with this task have to rely on the options available in software packages for handling missing data. For most surveys, it is unrealistic to believe that efficient compensation procedures could be developed for each separate analysis.

Imputations for the IEA Reading Literacy Study

Imputation has been chosen as the method to deal with missing data items in questionnaires administered to students, teachers, and school principals in the U.S. component of the IEA Reading Literacy Study. The primary method of imputation selected for this study was a form of hot-deck imputation. This method was selected because it is most suitable for the type of data collected in the study (see Kalton and Kasprzyk 1986 for a review of imputation methods). For a few items with very low rates of missing data, modal imputation was used to expedite the process.

Hot-Deck Imputation. The hot-deck imputation procedure WESDECK (a SAS macro developed by Westat, Inc.) was used in this study. This procedure starts with the definition of imputation classes according to a cross-classification of auxiliary variables chosen for use in imputing for missing responses to a particular item. Then, for each missing response, a value is assigned from a respondent in the same imputation class. The auxiliary variables used in constructing the imputation classes were those that were related to the item to be imputed. One advantage of this approach is that it preserves the relationships between the item being imputed and the auxiliary variables used to form the imputation classes. An assumption of the hot-deck approach is that after controlling for the auxiliary variables, the distribution of responses for the nonrespondents is the same as that for the respondents.

The following examples illustrate how hot-deck imputation was carried out. To impute for the item "the amount of time a student spent watching television on a school day," imputation classes were formed by cross-classifying the following variables: home possesses a TV and VCR (three classes were defined—those with no TV, those with TV but no VCR, and those with both); the highest educational level of the parents (college, high school, or less than high school); school control (private or public); and race/ethnicity (white/Asian, or other minorities). Within each of the 36 classes, respondents were chosen at random to donate their values to nonrespondents in that class. The process was subjected to a constraint that no respondent was allowed to donate his or her response to more than three nonrespondents.

Items that were strongly related to each other were imputed together, assigning values from the same respondent (donor). For example, because the educational levels of parents are correlated, father's and mother's education were imputed together. Thus, for students with both parents present in the household, the following procedure was employed. For those with father's education reported but mother's education missing, father's education was used as an auxiliary variable along with race of student, school control, and community size in forming the imputation classes. Hence, donors for the students with missing mother's education were selected from within imputation classes where all students had the same level of father's education. Likewise, when mother's education was reported but father's education missing, mother's education was used in forming the imputation classes. When the educational levels of both parents were missing, the imputation classes were formed using only race, school control, and community size. The imputation of both variables was performed jointly, taking both values from a given donor for each recipient and restricting the choice of donor to those students with both father's and mother's education reported.

Modal Imputation. Modal imputation assigns the modal value of a certain imputation class to replace any missing value within that class. This method was used for items with very low item nonresponse rates (less than 2-3 percent) and where respondents showed a clear preference for one of the response categories. Modal imputation, which is easier to implement but somewhat less effective than hot-deck imputation, was used to expedite the process of imputation. Under the conditions in which it was used, modal imputation is likely to have produced similar results to those that would have been obtained from a hot-deck procedure. Modal imputation was employed with about one-fifth of the items on the School Questionnaire and a half of the items on the Teacher Questionnaire.

As an illustration of modal imputation, consider an item on the grade 4 Teacher Questionnaire that reads, "Do you regularly (i.e., at least once a week) do the following activities to encourage your students to read outside school?" followed by a list of statements such as "suggest books to students to read," and "read stories to students." The teachers were requested to select a "yes" or a "no" response for each statement on the list. There were only one or two nonrespondents for each part of the question, and these nonresponses were replaced by the modal values. For example, since 82 percent of the responding teachers replied "yes" to the statement "suggest books to read," the single nonrespondent to this statement was assigned a "yes" response. Had a hot-deck procedure been applied, there is a very high probability that a "yes" response would have been imputed.

The Process of Imputation. The items in the questionnaires were imputed sequentially, following roughly the logical sequence of the questionnaires. The imputed values of some variables were used for subsequent imputation of other variables. For example, for the Student Questionnaires, race and parents' education were imputed first. The imputed values of these variables were then used to classify students into imputation classes for subsequent imputations of other items. Similarly, for nested items, the filter items that led into skip patterns were always imputed first, and the responses to the items that followed were imputed so as to be consistent with the responses to the filter items.

The frequency distributions of all items subject to imputation were monitored. For each item imputed, the distributions of the reported cases, the imputed cases, and the two combined were compared. Table 3-2 shows the distributions for several items from the grade 4 Student Questionnaire that are often used to explain student's reading performance: father's education, mother's education, the use of a language other than English at home, hours spent watching television on a school day, and frequency of reading story books. For all of these variables, the frequency distributions for the imputed cases are slightly different from those for the reported. These differences are not surprising given that the imputations were conducted within imputation classes of similar characteristics. When hot-deck imputation is used, such differences are to be expected when the item nonresponse rates and the distributions of the responses differ across imputation classes. However, when the imputed and the reported data are combined, the overall distributions are similar to the distributions of the reported data. This occurs because the missing data rates for these items are relatively small.

Table 3-2. Frequency distributions of selected variables from the grade 4 Student Questionnaire based on reported data, imputed data, and both sources combined

Characteristic	Reported	Imputed	Combined
Father's education			
Less than high school	10%	11%	10%
High school	24	21	24
Some college	18	17	17
College	48	51	49
Number of students[1]	5,441	577	6,018
Mother's education			
Less than high school	7%	10%	9%
High school	27	25	26
Some college	21	19	21
College	44	46	44
Number of students[2]	5,607	556	6,163
Use of language other than English at home			
No	75%	72%	75%
Yes	25	28	25
Number of students	6,106	114	6,220
Hours spent watching TV on a school day			
Low (0-1)	17%	14%	16%
Medium (2-4)	50	44	50
High (5+)	33	42	34
Number of students	5,991	229	6,220
Frequency of reading story books			
Almost never	17%	16%	17%
Once a month	14	17	14
Once a week	26	22	26
Once a day	43	45	43
Number of students	5,777	443	6,220

[1] Excluding 202 students with no father in the household.

[2] Excluding 57 students with no mother in household.

NOTE: Percentages may not add to 100 due to rounding.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Imputation flag variables were added to the data file to identify the data elements that were imputed and the method of imputation. Flags are important to inform analysts that some of the data items are not directly reported by the survey respondents. Analysts with specific analytic goals can choose whether to include the imputed values in their analysis or to use alternative methods to handle the missing data.

3.4 Evaluations of the Effects of Imputation on Data Analysis of the IEA Reading Literacy Study

A convenient way to analyze the fully imputed datasets for the IEA Reading Literacy Study is to proceed as if the completed data contained only reported responses. The impact of treating imputed data as real reported data in these levels of analysis has been studied by a number of researchers (Santos 1981; Jinn and Sedransk 1987, 1989a, and 1989b; Jinn, Sedransk, and Wang 1991; and Wang, Sedransk, and Jinn 1992). This section examines the effects of imputation on regression analysis and hierarchical linear models, the main forms of analysis being applied to the survey data.

Regression equations were estimated using four different estimation approaches: the hot-deck imputed data analysis, the complete case analysis, the available case analysis, and the EM algorithm analysis. These methods are outlined below and the mathematical details are provided in Appendix 1 to this chapter:

- **Hot-deck imputation (HD).** The analysis of the dataset with hot-deck imputation included the data imputed through the hot-deck procedure as if they were reported. The analysis is conducted on the full sample using standard procedures for a dataset with no missing values.

- **Complete case analysis (CC).** The CC analysis includes only students who have provided complete data for all the variables involved in the regression model. Students with missing responses for one or more of these variables are excluded. The analysis is then conducted using standard procedures for a dataset with no missing values.

- **Available case analysis (AC).** The basic quantities involved in conducting a regression analysis are the means of each of the variables, the variances of each of the variables, and the covariances between all pairs of variables. The AC analysis estimates each of these quantities from the incomplete data, using as much data as possible. Thus the mean and variance of a particular variable are estimated from all students who provided a response for that variable. Covariances between two variables are derived using data from students who responded to both variables.

- **The EM algorithm (EM).** The EM algorithm (Dempster, Laird, and Rubin 1977) uses an iterative maximum likelihood procedure to provide estimates of the mean and variance-covariance matrix based on all the available data for each respondent. The algorithm assumes that the data come from a multivariate normal distribution and that, conditional on the reported data, the missing data are missing at random. Although the assumption of multivariate normality may appear restrictive, Little and Rubin (1987) have shown that the EM algorithm can provide consistent estimates under weaker assumptions about the underlying distribution. The predictor variables can, for instance, be categorical variables treated as dummy variables in a regression analysis.

Effects of Imputation on Regression Analysis

The regression model used to predict a student's performance on a particular reading literacy test is the following:

$$y_i = b_o + b_1x_{i1} + \dots + b_px_{ip} + e_i \qquad i = 1, \dots, n,$$

where y_i is the performance score for student i on the reading test, $x_{i1}, \dots, x_{ip}$ are the predictor variables, $b_o, \dots, b_p$ are the regression parameters, and e_i is the error term.

The three reading performance scores used as the dependent variable in the regression models were the narrative, expository, and document performance scores. These performance scores, which are described in Appendix 2 to this chapter, were derived using Item Response Theory (IRT) models scaled for international comparisons (see Elley 1992). The predictor variables used in all models were gender, age, race, father's and mother's education, family structure (presence or absence of both parents), family composition (nuclear or extended family), family wealth and possessions, and use of a language other than English at home. Other than family wealth and age, all predictor variables were coded as dummy variables.

Table 3-3 shows the rate of missing data on the predictor variables. In most cases, the rates are less than 10 percent. The rate for family wealth index is somewhat higher because it is a factor score based on many data elements. Although the missing data rates for individual items are low, over 30 percent of students had data missing for one or more of the variables. As noted earlier, students without reading performance scores were excluded from the dataset. There are, therefore, no missing data for the performance scores.

Table 3-3. Percentage of missing data on variables used to predict the reading performance of grade 4 students

Predictor variable	Percent
Gender	0
Age	2
Race	0
Father's education	9
Mother's education	9
Family wealth index (derived from factor score)[1]	18
Family composition (living with both parents)[2]	3
Nuclear or extended family[2]	3
Use of a language other than English at home	2
One or more variables	31

[1] Factor scores derived from factor analysis of items related to family possessions.

[2] Composite variable based on the responses to several items.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

The four estimation approaches were applied first to unweighted data, using standard software packages for the HD, CC, and AC analyses and using a FORTRAN routine ROBMLE (Little 1988b) for the EM analysis. For the EM analysis, cases with complete data were used to provide the initial estimates of the means and the variance-covariance matrix employed to start the chain of the estimation-maximization process. Convergence was specified to occur when the successive log likelihood values differed by less than 0.001 and estimates of all parameters differed (proportionally) by less than this amount. The means and variance-covariance matrix produced by the EM method then served as input for a standard regression analysis.

The unweighted regression analyses ignore the weights resulting from the unequal selection probabilities associated with the complex sample design and the weighting adjustments made to compensate for unit nonresponse. To examine the effects of the weights, weighted regression analyses were conducted using the HD, AC, and CC approaches. The variances of the weighted regression coefficients were estimated using a jackknife method of variance estimation that took account of the complex sample design (see Chapter 2 of this volume). The EM method was not included in the weighted analyses because of a lack of software to handle weighted data, and the fact that the method gave comparable results to those obtained from the HD and AC approaches in the unweighted analyses.

Since the findings obtained from the regression analyses for the three reading performance scores produced similar conclusions about the comparability of the regression results from the four estimation methods, only the regression analyses for the narrative performance scores are presented here. The results for the unweighted and weighted analyses are presented below.

Results from the Unweighted Regression Analysis. Tables 3-4 to 3-6 present the results of the unweighted ordinary least squares (OLS) regression analyses from the four estimation approaches for the three independent variables (the narrative, expository, and document performance scores). For each table, the regression coefficients estimated using the HD, EM, and AC methods are very similar. The estimates from the CC analysis, however, are somewhat different from those produced by the other methods.

One use of the regression model is for computing adjusted mean performance scores for subgroups of students with various characteristics, after controlling for other characteristics. These adjusted mean scores may be computed as $\bar{y} = xb$, with certain x's in the regression model set to values that specify the subgroup of interest, the remaining x's set to their mean values, and b as the estimated regression coefficients. The adjusted mean scores can also be compared with unadjusted scores for the subgroups to see the effect of controlling on other variables.

Table 3-7 shows the unadjusted and the adjusted mean narrative performance scores for the following subgroups: males, females, minority students, nonminority students, students with no father living in the household, students with fathers with less than high school education, students with fathers with college education, male minority students not living with parents, and female nonminority students living with both parents. The unadjusted means were computed in two ways: using all available cases as in the AC analysis, and using only complete cases as in the CC analysis.

Overall, the unadjusted mean for the complete cases is about 10 points (or a 10th of a standard deviation) higher than that for the available cases. The difference is roughly of this magnitude for all the subgroups in the table except for the category father absent, where the difference is markedly higher at 24 points (almost a quarter of a standard deviation). The regression adjusted means estimated by the HD, EM, and AC methods are very comparable to one another, but they differ from the adjusted means based on the CC analysis by approximately the same magnitude as observed with the unadjusted means. The larger difference in the father absent category may have resulted from a high nonresponse rate among these students. Only about 50 of the 202 sampled students in this category had responses for all the analysis variables.

Table 3-4. Unweighted regression estimates predicting the narrative performance scores for grade 4 students: Results from four estimation algorithms

Predictor variable	Hot-deck imputation		EM algorithm		Available cases		Complete cases	
	b	s.e.	b	s.e.	b	s.e.	b	s.e.
Intercept	744.7	22.4	731.0	22.6	726.5	25.0	729.3	27.9
Gender (1=female)	16.9	2.3	17.2	2.3	17.1	2.5	16.4	2.7
Age	-1.8	0.2	-1.8	0.2	-1.7	0.2	-1.6	0.2
Minority (1=black or Hispanic)	-36.0	2.7	-36.0	2.7	-36.3	2.9	-36.3	3.3
Father education - high school	9.8	4.7	9.3	4.7	9.2	5.2	9.9	5.7
Father education - some college	15.3	5.0	17.4	5.0	18.0	5.5	16.9	6.0
Father education - college	23.3	4.6	25.4	4.6	25.7	5.1	28.2	5.6
No father in household	18.1	7.5	20.2	7.2	18.7	8.0	31.6	10.3
Mother education - high school	16.3	4.7	19.0	4.7	19.7	5.2	17.6	6.0
Mother education - some college	17.5	4.9	19.2	5.0	19.4	5.4	16.5	6.2
Mother education - college	18.7	4.7	21.0	4.8	21.0	5.2	17.1	6.0
Family wealth index	9.3	1.2	8.6	1.3	8.5	1.4	7.2	1.5
Family composition	20.2	2.4	21.2	2.4	21.3	2.7	19.4	2.9
Extended family	-23.0	2.4	-24.1	2.4	-23.8	2.7	-27.6	2.9
Use of foreign language at home	-8.0	2.7	-7.5	2.7	-7.7	3.0	-9.3	3.2
Sample size		6,220		6,220		5,105		4,280
Model R²		0.15		0.16		0.16		0.15

b = coefficient; s.e. = standard error.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 3-5. Unweighted regression estimates predicting the expository performance scores for grade 4 students: Results from four estimation algorithms

Predictor variable	Hot-deck imputation		EM algorithm		Available cases		Complete cases	
	b	s.e.	b	s.e.	b	s.e.	b	s.e.
Intercept	688.2	18.6	679.4	18.8	676.9	20.8	685.2	23.2
Gender (1=female)	7.1	1.9	7.2	1.9	7.2	2.1	5.7	2.2
Age	-1.4	0.1	-1.4	0.1	-1.4	0.2	-1.3	0.2
Minority (1=black or Hispanic)	-29.8	2.2	-30.2	2.2	-30.4	2.4	-28.0	2.7
Father education - high school	11.9	3.9	8.9	3.9	9.1	4.3	8.8	4.7
Father education - some college	15.7	4.2	15.3	4.2	15.7	4.6	15.4	5.0
Father education - college	22.1	3.8	21.2	3.8	21.5	4.2	24.9	4.6
No father in household	11.4	6.2	11.1	6.0	9.7	6.6	16.1	8.6
Mother education - high school	7.3	3.9	10.9	3.9	11.4	4.3	12.0	4.9
Mother education - some college	9.0	4.1	12.1	4.2	12.3	4.5	9.7	5.2
Mother education - college	13.6	3.9	17.1	4.0	17.3	4.3	13.9	5.0
Family wealth index	8.4	1.0	7.3	1.1	7.1	1.2	5.9	1.2
Family composition	15.6	2.0	15.8	2.0	15.8	2.2	15.2	2.3
Extended family	-17.0	2.0	-18.4	2.0	-18.2	2.2	-21.9	2.4
Use of foreign language at home	-5.6	2.2	-5.3	2.2	-5.4	2.5	-5.0	2.7
Sample size		6,220		6,220		5,105		4,280
Model R²		0.14		0.14		0.14		0.14

b = coefficient; s.e. = standard error.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 3-6. **Unweighted regression estimates predicting the document performance scores for grade 4 students: Results from four estimation algorithms**

Predictor variable	Hot-deck imputation		EM algorithm		Available cases		Complete cases	
	b	s.e.	b	s.e.	b	s.e.	b	s.e.
Intercept	663.3	18.7	651.6	18.8	649.1	20.8	655.0	23.2
Gender (1=female)	-1.7	1.9	-1.6	1.9	-1.6	2.1	-3.2	2.2
Age	-1.1	0.1	-1.0	0.1	-1.0	0.2	-1.0	0.2
Minority (1=black or Hispanic)	-34.0	2.2	-34.3	2.2	-34.4	2.4	-34.4	2.7
Father education - high school	12.4	3.9	9.8	3.9	10.0	4.3	10.7	4.7
Father education - some college	17.1	4.2	16.7	4.2	17.0	4.6	16.3	5.0
Father education - college	18.7	3.8	18.3	3.9	18.6	4.2	22.5	4.6
No father in household	15.3	6.2	16.4	6.0	15.2	6.6	32.0	8.6
Mother education - high school	10.0	3.9	14.0	4.0	14.4	4.3	14.2	5.0
Mother education - some college	13.1	4.1	17.5	4.2	17.5	4.5	14.3	5.2
Mother education - college	19.3	3.9	22.7	4.0	22.3	4.3	21.0	5.0
Family wealth index	8.6	1.0	7.1	1.1	7.0	1.2	5.1	1.2
Family composition	22.4	2.0	23.2	2.0	23.3	2.2	22.1	2.4
Extended family	-18.9	2.0	-20.1	2.0	-19.9	2.2	-21.7	2.4
Use of foreign language at home	-10.4	2.2	-10.7	2.2	-10.8	2.5	-9.4	2.7
Sample size		6,220		6,220		5,105		4,280
Model R^2		0.17		0.17		0.17		0.16

b = coefficient; s.e. = standard error.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 3-7. **Unweighted adjusted and unadjusted mean narrative performance scores for grade 4 students**

Student group	Unadjusted mean		Adjusted mean			
	Available cases	Complete cases	Hot-deck imputation	EM algorithm	Available cases	Complete cases
All students	555	565	552	552	552	563
Males	543	555	544	544	544	557
Females	562	575	561	561	561	573
White or Asian	567	578	562	563	563	574
Other minority students	514	524	526	527	526	537
Father not in household	532	556	554	555	553	577
Father had less than high school education	515	525	536	535	534	545
Father had college education	568	577	559	560	560	573
Males, minority, not living with parents	495	504	506	506	506	518
Females, white or Asian, living with both parents	588	596	580	580	584	590

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Results from the Weighted Regression Analysis. This section presents results for weighted regression analyses for the narrative performance scores using the survey weights developed to compensate for unequal selection probabilities and unit nonresponse. The regression analyses are based on the HD and AC methods and two versions of the CC method.

The unweighted analyses reported above showed that the CC analysis produced appreciably different results from the three other methods. These differences are presumably explained by the fact that the CC analysis excludes the more than 30 percent of students who failed to answer one or more of the items involved in the regression model. With a weighted analysis, as well as compensating for unit nonresponse, there is the possibility of attempting to compensate for these partial respondents through an additional weighting adjustment. To examine the effectiveness of a weighting adjustment for the partial nonrespondents, an additional weighting adjustment was developed as described below. The CC regression analysis was then conducted twice, once using the original survey weights and once using the adjusted weights that compensate for the partial respondents.

The first step in developing the adjusted weights for the CC analysis was to perform a CHAID analysis (Magidson 1989) to identify subgroups of students with different rates of complete response for the set of items in the regression model. Using a sequential procedure, the analysis partitioned the sample into 12 subgroups involving the following variables: the narrative reading performance scores, race/ethnicity of student, community served by school (urban, suburban, nonurban), the region of the country (Northeast, Southeast, West, Central), and control of school (public, private). Table 3-8 shows that the levels of complete response for the 12 subgroups vary substantially from a high of 88 percent to a low of 43 percent. In particular, the level of complete response varies markedly by performance level, from about 88 percent for students in the highest quintile of performance to about 58 percent in the lowest quintile. To compensate for the differential loss of students in the complete case analysis across the 12 subgroups, the adjusted weights for the CC analysis were constructed by multiplying the survey weights for students in a subgroup by the inverse of the weighted percentage of complete cases in that subgroup.

Table 3-8. The percentage of complete cases for different subgroups of students

Subgroup	Percent
Highest quintile on narrative performance score:	
South, public schools	70
South, private schools	88
Other regions, black and Hispanic	70
Other regions, white and Asian	83
4th quintile on narrative performance score:	
Black and Hispanic	63
White and Asian	77
2nd and 3rd quintile on narrative performance score:	
Black and Hispanic	57
White and Asian	70
Lowest quintile on narrative performance score:	
Urban, black	43
Urban, other races	58
Suburban, West	51
Suburban, other regions	64

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 3-9 shows the regression coefficients estimated by the HD and AC analyses and the CC analyses using both the original and the adjusted sampling weights; and Table 3-10 shows the corresponding unadjusted and adjusted subgroup means. The results of the weighted analyses confirmed the finding from the unweighted analyses that the HD and AC analyses produce similar results, but that the CC analysis appears to overestimate student performance. The use of the adjusted weights with the CC analysis reduces the overestimation to some extent, but the results still deviate from those produced by the HD and AC analyses.

Table 3-9. Weighted regression estimates predicting the narrative performance scores for grade 4 students: Results from four estimation algorithms

Predictor variable	Hot-deck imputation		Available cases		Complete cases (adjusted weights)		Complete cases (unadjusted weights)	
	b	s.e.	b	s.e	b	s.e.	b	s.e.
Intercept	747.3	23.5	728.6	24.4	730.7	24.8	737.7	27.4
Gender (1 =female)	16.0	3.1	16.3	3.3	16.9	3.9	15.8	3.9
Age	-1.8	0.2	-1.7	0.2	-1.7	0.2	-1.7	0.2
Minority (1 =black or Hispanic)	-35.4	3.4	-35.5	3.4	-35.2	4.4	-36.2	4.8
Father education - high school	8.5	5.8	7.9	6.9	9.1	6.0	8.4	6.5
Father education - some college	12.9	5.9	15.0	6.6	14.5	6.6	13.7	6.9
Father education - college	21.9	6.1	24.0	7.1	25.5	7.0	25.9	7.6
No father in household	25.8	11.6	25.9	12.4	36.2	13.6	39.1	13.3
Mother education - high school	14.8	5.6	18.3	6.4	16.1	5.2	18.8	5.3
Mother education - some college	16.0	6.8	18.7	7.3	14.8	7.7	18.3	7.8
Mother education - college	16.2	6.4	18.7	7.1	14.3	6.8	17.5	7.2
Family wealth index	8.7	1.6	8.2	1.9	7.0	2.0	6.8	1.9
Family composition	20.7	3.2	21.7	3.2	20.7	3.7	20.6	3.7
Extended family	-22.8	2.9	-23.6	3.0	-25.8	4.2	-26.3	4.0
Use of foreign language at home	-6.4	3.4	-6.2	3.3	-8.7	4.6	-9.3	4.7
Sample size		6,220		5,105		4,280		4,280
Model R^2		0.14		0.15		0.15		0.15

b = coefficient; s.e. = standard error.

NOTE: The regression estimates were computed using weighted least squares methods, and standard errors were estimated using a jackknife replication procedure.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 3-10. Weighted adjusted and unadjusted mean narrative performance scores for grade 4 students

Student group	Unadjusted mean		Adjusted mean			
	Available cases	Complete cases	Hot-deck imputation	Available cases	Complete cases (adjusted weights)	Complete cases (unadjusted weights)
All students	557 (3.5)	566 (3.8)	555 (2.8)	555 (12.3)	558 (12.7)	566 (3.0)
Boys	546 (3.7)	556 (4.4)	547 (3.1)	547 (2.8)	550 (3.4)	559 (3.8)
Girls	564 (3.2)	576 (3.2)	563 (2.4)	563 (1.9)	567 (2.1)	574 (2.2)
White or Asian	570 (2.4)	579 (2.8)	565 (2.2)	565 (1.8)	567 (2.3)	575 (2.4)
Other minority students	517 (3.7)	526 (4.4)	529 (3.7)	529 (3.4)	532 (3.9)	539 (4.8)
Father not in household	543 (9.7)	567 (11.6)	565 (10.3)	564 (10.3)	577 (12.1)	588 (12.1)
Father had less than high school education	519 (6.8)	528 (7.2)	540 (5.7)	538 (6.3)	541 (6.4)	549 (7.0)
Father had college education	570 (3.4)	580 (3.6)	561 (2.5)	562 (2.3)	567 (2.6)	575 (2.7)
Males, minority, not living with parents	499 (5.5)	503 (6.3)	509 (4.5)	509 (4.3)	511 (4.7)	519 (5.7)
Females, white or Asian, living with both parents	588 (3.3)	596 (3.6)	581 (2.9)	582 (2.6)	584 (3.0)	591 (3.1)

NOTE: Numbers in parentheses are standard errors.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

The standard errors of the weighted estimates in Tables 3-7 and 3-8 were computed using a jackknife replication method of variance estimation that takes account of the complex sample design. Since the HD approach treats all the imputed data as if they were reported, the standard errors with this approach are underestimated. The estimates of standard error for the AC and CC analysis are valid estimates. In this study, the additional weighting adjustment for the CC analysis did not have substantial effect on the precision of the estimates.

Effects of Imputation on HLM Analysis

The sample design for the IEA Reading Literacy Study is a hierarchical one, with samples of schools, classes within schools, and students within classes. Using the hierarchical linear model (HLM), advantage can be taken of this hierarchical structure in the analysis to examine the effects of schools and classrooms, as well as background variables on student achievement (Bryk and Raudenbush 1992).

With HLM, variables measured at different levels (such as school, classroom, and student) are included in the model. In the analyses reported here, only two levels of variables are included. Level-1 variables are related to individual students. These variables are the same as those employed in the regression analyses reported in the previous section. Level-2 variables are related to classrooms. The only level-2 variable included in the HLM models is the proportion of minority students in the classroom. All but one of the level-1 variables are treated as fixed effect variables. The variable, minority status, is treated as a random effect variable. With these specifications, HLM controls for the between-classroom variance in student minority status and provides separate estimates of the effect of minority status on reading comprehension in each classroom.

The software used in this study for estimating the HLM (Bryk and Raudenbush 1989) can handle only complete datasets for a two-level model. For this reason, the HLM analyses were conducted only under the HD and CC estimation approaches. In these analyses, each student's score on the student-level predictor variables was converted into a deviation score from the mean of all students (grand mean centered). The random effect variable, minority status, was expressed as a deviation from the classroom mean (group mean centered). The proportion of minority students in each classroom was transformed into a deviation from the mean of the proportions of minority students across classrooms.

Table 3-11 shows the results from the HD and CC analyses of the HLM. The gamma coefficients from the HLM analysis are analogous to the regression coefficients from a regression analysis. In this analysis, the gamma coefficients for the proportion of minority in the class and minority status are similar for the two estimation approaches, -64.5 and -16.0, respectively, for the HD analysis, and -64.8 and -16.6 for the CC analysis. The gammas are, however, different for some of the fixed effect variables. In particular, as with the regression analyses, the coefficients are appreciably different for the category father absent from the household.

3.5 Discussion

Item nonresponse regularly occurs in survey data, but usually at a low level for most items (as is the case with the IEA Reading Literacy Study). Its presence complicates the analysis of the survey data. It is particularly problematic for multivariate analysis where low levels of item nonresponse for individual items can accumulate into a sizable fraction of the sample having missing responses for one or more items in the analysis.

Table 3-11. HLM analysis predicting the narrative performance scores for grade 4 students: Results using the HD imputation and CC analysis

Predictor variable	Hot-deck imputation		Complete cases	
	gamma	standard error	gamma	standard error
Base coefficient	556.1	2.0	564.0	2.0
Proportion minority in class[1]	-64.5	6.7	-64.8	7.0
Gender (1=female)	14.6	2.3	q	2.7
Age	-1.9	0.2	-1.8	0.2
Minority status[2] (1=black or Hispanic)	-16.0	3.9	-16.6	4.4
Father education - high school	5.9	4.7	7.7	5.8
Father education - some college	10.0	5.0	12.8	6.1
Father education - college	16.7	4.6	22.5	5.7
No father in household	21.1	7.7	32.6	10.6
Mother education - high school	13.2	4.8	14.0	6.1
Mother education - some college	13.3	5.0	13.3	6.4
Mother education - college	13.8	4.8	12.1	6.2
Family wealth index	6.0	1.3	5.4	1.5
Family composition	16.0	2.4	15.9	2.9
Extended family	-20.2	2.5	-22.8	3.0
Use of foreign language at home	-9.1	2.8	-12.2	3.3
Sample size		6,220		4,280

[1]Level-2 variable.

[2]Random effect variable.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

The CC analysis, which is the default procedure for handling missing data in most statistical packages, is widely used in practice. It is easy to implement but clearly inefficient. In the regression models examined in this paper, almost a third of the sampled students were discarded. The CC approach assumes that the complete cases are a random subsample of all cases (Little 1993), an assumption that is often unjustified in practice. In the current case, there is clear evidence that the students with one or more missing values on the predictors in the regression models differ from those with complete data in terms of reading performance, race/ethnicity of the student, type of community served by the school, region of the country, and control of school. As a result, the regression analyses conducted for the complete cases are likely to have produced biased results.

Three other methods for handling item nonresponse have been employed in this research—the AC approach, the HD imputation approach, and the EM algorithm. The three approaches yielded very similar results in the regression analyses conducted. The EM algorithm, which is available through the BMDP (Dixon 1988) and GAUSS packages (Aptech Systems 1988), has theoretical attractions (Little 1992). This algorithm has been found to be superior to the CC and AC analyses even when the underlying normality assumptions are violated (Azen, Van Guilder, and Hill 1989; Little 1988a). However, it is a computer-intensive procedure, and software for its use with a particular form of analysis may not be readily available. It was, for instance, not applied in this research for the weighted regression analyses using the survey weights, or for the HLM analysis, because no suitable software was available.

As compared with the CC approach, the AC approach has the attraction of making fuller use of the available data. In a simulation study, Kim and Curry (1977) found the AC approach to be superior to the CC approach with weakly correlated data. A limitation to the AC approach is that it may produce a covariance matrix that is not positive definite, an outcome that poses problems for model estimation (yielding indeterminate slopes in a regression analysis). This limitation is severe when the independent variables in a regression model are highly correlated (Little 1992). A furthe: problem with the AC approach is that it is not available in all statistical packages, and it may not be available for particular forms of analysis. It was not applied with the HLM analysis here because of a lack of available software.

Imputation is widely used to handle item nonresponses in survey research. A considerable amount of research has been conducted on alternative methods of imputation and their properties (see, for example, the three volume report of the National Research Council's Panel on Incomplete Data—Madow, Nisselson, and Olkin 1983; Madow, Olkin, and Rubin 1983; and Madow and Olkin 1983). By assigning values for all missing responses, the imputation approach creates a complete dataset that can be readily analyzed using complete data methods. In this respect it is like the CC approach. However, unlike the CC approach, it retains all the actual responses and it does not discard records with one or more missing values.

A limitation to imputation is that it can lead to an attenuation in covariances between some variables, thus distorting the results of multivariate analyses (Kalton and Kasprzyk 1986). This attenuation does not occur between a variable subject to imputation and the variables used as auxiliary variables in the imputation of that variable (e.g., the variables used to form the imputation classes with hot-deck imputation), but it does occur between the variable subject to imputation and other variables. For this reason, it is important to employ major variables associated with a variable subject to imputation as auxiliary variables in the imputation scheme. However, even when a variable is not used as an auxiliary variable in the imputation, the attenuation of its covariance with the variable subject to imputation is small, provided that the level of item nonresponse is low, as is the case in the IEA Reading Literacy Study. The regression coefficients in the analyses reported show no sign of such attenuation. It appears that the IEA Reading Literacy Study imputed dataset can be safely analyzed without concern for an appreciable attenuation of covariances.

Another concern with imputation is the effect on the standard errors of survey estimates. In essence, the hot-deck imputation used in this study duplicates some of the values from respondents to substitute for the missing data from nonrespondents. Therefore, treating the HD imputed dataset as complete responses is likely to overstate the precision of the survey estimates. One approach to variance estimation with an imputed set is to employ multiple imputations, completing the dataset several (say, 3 to 5) times and estimating the overall variance of a survey estimate from a combination of the average within-dataset and between-dataset variance components (Rubin 1987). This approach was not adopted here because of the added complexity involved. Other approaches to variance estimation with imputed datasets are under development (Lee, Rancourt, and Särndal 1991; Rao and Shao 1992; Tollefson and Fuller 1992; Fay 1991, 1992), but they are not yet available for general applications. As a result, there is no ideal solution currently available for variance estimation with the IEA Reading Literacy Study dataset. However, given the low levels of item nonresponse, the standard errors computed by treating the imputed values as reported values should overstate the precision of the survey estimates to only a slight extent.

In conclusion, this study shows that for the U.S. component of the IEA Reading Literacy Study, data analysis using the HD imputed data yielded similar results to those produced by the AC and EM methods of handling the missing data. Since analysis with the HD imputed data is the simplest to implement, it appears to be the best option for most analyses of the IEA Reading Literacy Study data. It should, however, be noted that an analyst of the IEA Reading Literacy Study dataset is not restricted to the HD approach. Since flags identifying the imputed values are provided in the dataset, the imputed values can readily be deleted and an alternative approach for handling the missing data can then be employed.

References

Aptech Systems. (1988). *GAUSS programming language.* Kent, WA:Author.

Azen, S.P., Van Guilder, M., and Hill, M.A. (1989). Estimation of parameters and missing values under a regression model with non-normally distributed and non-randomiy incomplete data. *Statistics in Medicine*, 8, 217-228.

Bryk, A.S., and Raudenbush, S.W. (1989). *An introduction to HLM: Computer program and user's guide.* Chicago: University of Chicago.

Bryk, A.S., and Raudenbush, S.W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Advanced Quantitative Techniques in the Social Science Series. Thousand Oaks, CA: Sage Publications.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Ser. B, 39, 1-38.

Dixon, W.J. (ed.) (1988). *BMDP statistical software manual.* Vol. 2. Berkeley, CA: University of California Press.

Elley, W.B. (1992). *How in the world do students read?* The Hague: The International Association for the Evaluation of Educational Achievement.

Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings, 1991 Annual Research Conference*, 429-440, Washington, DC: U.S. Department of Commerce, Bureau of the Census.

Fay, R.E. (1992). When are inferences from multiple imputation valid? *1992 Proceedings of the Section on Survey Research Methods*, 227-232, Alexandria, VA: American Statistical Association.

Jinn, J.H., and Sedransk, J. (1987). Effect on secondary data analysis of different imputation methods. *Proceedings of the Third Annual Census Bureau Research Conference*, 509-530. Washington, DC: U.S. Department of Commerce, Bureau of the Census.

Jinn, J.H., and Sedransk, J. (1989a). Effect on secondary data analysis of common imputation methods. *Sociological Methodology*, 19, 213-241.

Jinn, J.H., and Sedransk, J. (1989b). Effect on secondary data analysis of the use of imputed values: The case where missing data are not missing at random. *Proceedings on the Section on Survey Research Methods*, 51-61. Alexandria, VA: American Statistical Association.

Jinn, J.H., Sedransk, J., and Wang, R. (1991). The use of imputed values in secondary data analysis. *Proceedings of the Seventh Annual Census Bureau Research Conference*, 483-499. Washington, DC: U.S. Department of Commerce, Bureau of the Census.

Kalton, G. (1983). *Compensating for missing survey data.* Research Report Series. Ann Arbor, MI: The University of Michigan, Survey Research Center, Institution for Social Research.

Kalton, G., and Kasprzyk, D. (1982). Imputing for missing survey responses. *1982 Proceedings of the Section on Survey Research Methods*, 22-31. Washington, DC: American Statistical Association.

Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.

Kim, J.O., and Curry, J. (1977). Treatment of missing data in multivariate analysis. *Sociological Methods and Research*, 6, 215-240.

Lee, H., Rancourt, E., and Särndal, C. (1991). Experiments with variance estimation from survey data with imputed values. *1991 Proceedings of the Section on Survey Research Methods*, 690-695. Alexandria, VA: American Statistical Association.

Little, R.J.A. (1988a). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83, 1198-1202.

Little, R.J.A. (1988b). ROBMLE user's notes. Unpublished.

Little, R.J.A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*, 87, 1227-1237.

Little, R.J.A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421), 125-134

Little, R.J.A., and Rubin, D.B. (1987). *Statistical analysis with missing data.* New York: John Wiley.

Madow, W.G., Nisselson, H., and Olkin, I. (eds.) (1983). *Incomplete data in sample surveys. Vol. 1: Report and case studies.* New York: Academic Press.

Madow, W.G., and Olkin, I. (eds.) (1983). *Incomplete data in sample surveys. Vol. 3: Proceedings of a symposium.* New York: Academic Press.

Madow, W.G., Olkin, I., and Rubin, D.B. (eds.) (1983). *Incomplete data in sample surveys. Vol. 2: Theory and bibliographies.* New York: Academic Press.

Magidson, J. (1989). *SPSS/PC + CHAID manual.* Belmont, MA: Statistical Innovations, Inc.

Rao, J.N.K., and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, 79(4), 811-22.

Rubin, D.B. (1987). Multiple imputation for nonresponse in surveys. New York: John Wiley.

Santos, R.L. (1981). Effects of imputation on regression coefficients. *1981 Proceedings of the Section on Survey Research Methods*, 140-145. Washington, DC: American Statistical Association.

Tollefson, M., and Fuller, W.A. (1992). Variance estimation for samples with random imputation. *Proceedings of the Section on Survey Research Methods*, 758-763. Alexandria, VA: American Statistical Association

Wang, R., Sedransk, J., and Jinn, J.H. (1992). Secondary data analysis when there are missing observations. *Journal of the American Statistical Association*, 87, 952-961.

Appendix 1

Methods of Data Analysis with Missing Data

Hot-Deck Imputed Cases

Data imputed through hot-deck imputation are included as complete data and the complete-case analysis algorithm is applied.

Complete-Cases (or Listwise Deletion) Algorithm

The complete-case (CC) algorithm uses cases that are complete in all variables to estimate the mean vector and covariance matrix. If we write $x_i = (x_{i1}, x_{i2}, \ldots, x_{i\,(p+1)})'$ where $x_{i(p+1)} = y_i$, and x_{ij} and x_{ik} are any two components of x_i, then for $j,\ k = 1, \ldots, (p+1)$,

$$\bar{x}_j \equiv \sum_{i=1}^{n} x_{ij}\ \lambda_i \ /\ n_c$$

$$\bar{x}_k \equiv \sum_{i=1}^{n} x_{ik}\ \lambda_i \ /\ n_c$$

and

$$c\hat{o}v(x_j, x_k) = \sum_{i=1}^{n} \lambda_i\ (x_{ij} - \bar{x}_j)\ (x_{ik} - \bar{x}_k)\ /\ (n_c - 1)$$

where $\lambda_i = 1$ (or w_i for weighted analysis) if x_i is complete, otherwise it is equal to 0, and $n_c = \sum_{i=1}^{n} \lambda_i$ is the number of complete cases.

Available-Case (or Pairwise-Deletion) Algorithm

The available-case (AC) or pairwise-deletion algorithm estimates the mean of each variable using all observations for that variable and covariances between pairs of variables using only cases complete in both variables. Thus for $j,\ k = 1, \ldots, p+1$,

$$\bar{x}_j = \sum_{i=1}^{n} x_{ij}\ \lambda_{i(jj)} \ /\ n_{jj}\ ,$$

$$\bar{x}_k = \sum_{i=1}^{n} x_{ik}\ \lambda_{i(kk)} \ /\ n_{kk}\ ,$$

and

$$c\hat{o}v(x_j,x_k) = \sum_{i=1}^{n} \lambda_{i(jk)} (x_{ij}-\bar{x}_j) (x_{ik}-\bar{x}_k) / (n_{jk}-1)$$

where $\lambda_{i(jk)} = 1$ if x_{ij} and x_{ik} are present, otherwise it is equal to 0, and $n_{jk}=\sum_{i=1}^{n} \lambda_{i(jk)}$ is the number of cases with both x_j and x_k present.

The EM Algorithm

Following Little and Rubin's notation,[1] let us assume that $X_i \overset{iid}{\sim} N(\mu,\Lambda)$, with $\mu = (\mu_i, ... \mu_{(p+1)})$ and covariance matrix $\Sigma = (s_{jk})$. We write $X = (X_{obs}, X_{mis})$, where X represents a random sample of size n on $(X_1, ..., X_{p+1})$, X_{obs} is the set of reported values, and X_{mis} the missing data. We write $X_{obs} = (x_{obs,1}, x_{obs,2}, ..., x_{obs,n})$, where $x_{obs,i}$ represents the set of variables reported for observation i, where $i = 1, ..., n$. At the tth iteration, let $\theta^{(t)} = (\mu^{(t)}, \Sigma^{(t)})$ denote current estimates of the parameters.

The E step of the algorithm consists in calculating:

$$E(\sum_{i=1}^{n} x_{ij} \,|X_{obs},\theta^{(t)}) = \sum_{i=1}^{n} x_{ij}^{(t)}, \quad j=1,...,K$$

$$E(\sum_{i=1}^{n} x_{ij}x_{ik} \,|X_{obs},\theta^{(t)}) = \sum_{i=1}^{n} (x_{ij}^{(t)}x_{ik}^{(t)}) + c_{jki}^{(t)}, \quad j,k=1,...,K,$$

where $x_{ij}^{(t)} = x_{ij}$, if x_{ij} is reported, and $E(x_{ij}|x_{obs,i}, \theta^{(t)})$, if x_{ij} is missing; and $c_{jki}^{(t)} = 0$ if x_{ij} or x_{ik} are reported, and $\mathrm{Cov}(x_{ij}, x_{ik} \,|x_{obs,i}\, \theta^{(t)})$, if x_{ij} or x_{ik} are missing. This means that missing values x_{ij} are replaced by the conditional mean of x_{ij} given the set of values, $x_{obs,i}$ reported for that observaticn. These conditional means and nonzero conditional covariances can be found from the current parameter estimates by sweeping the augmented covariance matrix so that the variables $x_{obs,i}$ are predictors in the regression equation and the remaining variables are outcome variables.

In the M step, the new estimates $\theta^{(t+1)}$ of the parameters are estimated from the estimated complete-data sufficient statistic. Assuming that the hypothetical complete data X belong to the regular exponential family, the sufficient statistics are:

$$\mu_j^{(t+1)} = n^{-1}\sum_{i=1}^{n} x_{ij}^{(t)}, \quad j=1,...,(p+1)$$

$$\sigma_{jk}^{(t+1)} = n^{-1}\sum_{i=1}^{n} [(y_{ij}^{(t)} - \mu_j^{(t+1)})(y_{ik}^{(t)} - \mu_k^{(t+1)}) + c_{jki}^{(t)}], \quad j,k=1,...,(p+1)$$

[1] R.J.A. Little and D.B. Rubin. *Statistical Analyses with Missing Data* (New York: John Wiley, 1987).

Appendix 2
Scores Used in the IEA Reading Literacy Test

The Rasch scaling method was used to create an international scale that has a mean of 500 and a standard deviation of 100. Students who scored close to the international mean score of 500 were typically those who responded correctly to items that were of intermediate difficulty. For instance, they responded correctly to items that required processes like the following:

Narrative scale

- Can read a story about a shark that befriends a family of sardines and say why the shark was swimming alone.
- Can read a short fable about an elephant that was bothering a family of birds and say how the mother bird got the elephant to go away.

Expository scale

- Can read a short passage about quicksand and respond correctly to a question that asks how to recognize quicksand.
- Can read a description of the walrus and say how long it lives as stated in the passage.

Document scale

- Can read a simple map and identify the place south of point x.
- Can study a school timetable and work out which was the third lesson on Thursday.

Students who earned scores over 600 were able to respond correctly to very difficult items requiring the ability to read long complex stories or complicated figures and to make inferences about major themes, the motives of characters, or unusual relationships in the information given.

Students who scored below 400 had very limited reading ability. Typically they could respond correctly only on short simple passages where the items required limited processing or the answer was clearly stated in the passage.[2]

[2] W.B. Elley. *How in the World Do Students Read?* (The Hague: The International Associates for the Evaluation of Educational Achievement, 1992.)

4 Assessing the Dimensionality of the IEA Reading Literacy Data

Nadir Atash

4.1 Introduction

The definition of reading literacy adopted by the IEA International Steering Committee (ISC) specified that reading occurs in different contexts (e.g., school, home) and for different purposes (e.g., pleasure, homework). This definition implied that a reader would interact differently with different text tynes. In view of the context-based definition of reading literacy, a total test score was not derived for each student. Rather, based on a student's responses to items designated for each Reading Literacy Test domain and using the one-parameter logistic (Rasch) model (Wright and Stone 1979), test scores were estimated for each of the three literacy domains.

The Rasch model is the simplest of the Item Response Theory (IRT) class of models because it uses only one parameter to describe the item characteristic curve: the difficulty parameter. Specifically, the probability that subject i gets item j correct (P_{ij}) can be expressed as follows:

$$P_{ij} = P\ (x_{ij} = 1 \mid \theta_i) = \frac{1}{1 + e\left[-D\bar{a}\ (\theta_i - b_j)\right]}$$

where θ_i is the proficiency parameter for person i, $\bar{a}$ is the common level of item discrimination, D is a scaling factor, and b_j is the difficulty parameter for item j.

The summarization of students' performance within each reading literacy domain assumed unidimensionality within each domain. A critical assumption of measurement models, both classical and IRT, is that a set of items forming an instrument all measure one attribute in common. Given that this assumption is valid, it makes sense to interpret a total test score that is derived from all items contained in the instrument. If this assumption is false, interpretation of a single score, such as the total test score, may be severely limited—it is difficult to interpret a total test score from a set of items measuring different attributes. Since unidimensionality is a critical assumption of the Rasch scaling, it is necessary to assess whether or not the unidimensionality assumption is tenable.

The designation of items into each one of the three reading literacy domains involved lengthy discussions among National Research Coordinators (NRCs) and the ISC. These discussions centered on two questions: (1) Does the theoretical framework for defining reading literacy support three distinct reading literacy domains? and (2) Which items should be classified into each reading literacy domain? Using the U.S. national item response data, these two questions will be addressed in this chapter. Thus, in addition to testing the unidimensionality assumption of the U.S. item response data within each reading literacy domain, the underlying structure of the reading literacy domains will be investigated. Our aim here is to determine to what extent the data support the hypothesized structure (i.e., three domains). Before presenting the data, however, we will define dimensionality and describe various methods to assess it.

4.2 Defining Dimensionality

Studying dimensionality is one aspect of gathering validity evidence for specific uses of a test. As defined in the *Standards for Educational and Psychological Testing* (American Educational Research Association et al. 1985, 9), validity refers to

> *...the appropriateness, meaningfulness, and usefulness of the specific inference made from test scores...A variety of inferences may be made from scores produced by a given test, and there are many ways of accumulating evidence to support any particular inference. Validity, however, is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the scores. The inferences regarding specific uses of a test are validated, not the test itself.*

Consistent with this definition, Cronbach (1989) argues that, "validation of an instrument calls for an integration of many types of evidence. The varieties of investigation are not alternatives any one of which would be adequate." Because dimensionality can be viewed as one aspect of validity, analyzing dimensionality should be viewed within this broad definition of validity.

Lord (1980, 21) stated that "There is a great need for a statistical significance test for the dimensionality of a set of test items." While there is consensus about the importance of testing dimensionality, there appears to be some confusion about how to define it and, even more so, about how to test it. This lack of agreement is in part due to confusion between defining dimensionality and methods to assess it.

Hattie (1984, 50) defines unidimensionality as the "existence of one latent trait underlying the data." This definition is broad and can be operationalized differently. Hattie (1985, 140) has provided three alternative definitions of unidimensionality that may be considered as various types of operationalization of the general definition provided above.

Definition # 1: A set of items can be said to be unidimensional when it is possible to find a vector of values $\phi = (\phi_i)$ such that the probability of correctly answering an item g is $\pi_{ig} = f_g(\phi_\mathrm{i})$ and local independence holds for each value of ϕ.

Definition # 2: A perfectly unidimensional test is a function of the amount by which a set of item responses deviates from the ideal scale pattern.

Definition # 3: Consider two examinees designated as "A" and "B." Assuming that A's score on the test is greater than B's score, then A has more of *some* ability than B. If this ability is the same ability for all individual As and Bs who may be selected, then the test is unidimensional.

At first glance the three definitions appear to be quite different. Under a closer examination, however, the three definitions are found to be closely linked—in fact, the aforementioned general definition forms the logical basis for all three operational definitions. It should be pointed out, however, that because each operational definition emphasizes a different aspect of unidimensionality, the methodology to assess unidimensionality derived from each definition could be extremely different. Hattie (1984) has reviewed over 80 indices for determining unidimensionality. Most of these indices can be linked to one or more of the definitions cited above, although some of the indices lack a theoretical rationale.

Strictly speaking, the term "unidimensionality of a set of items" is misleading. The basis for studying unidimensionality is the interactions of examinees with a set of items, not the items themselves. "Theoretical or empirical studies of dimensionality that involve statistical/ psychometric techniques involve item-response data resulting from the examinee-item interaction and not the dimensionality of items as entities separate from examinees" (Carlson and Jirele 1993). In this chapter, for the sake of brevity, we use the phrase "unidimensionality of Reading Literacy data" or "unidimensionality." However, it should be clear that our reference to unidimensionality is always with the understanding that the database for studying it has resulted from the examinee-item interaction in a specific population.

4.3 Why is Assessing Dimensionality Important?

Assessing dimensionality is important for three major reasons. The first reason relates to the psychological interpretation of test scores. McNemar (1964, 268) states:

> *Measurement implies that one characteristic at a time is being quantified. The scores on an attitude scale are most meaningful when it is known that only one continuum is involved. Only then can it be claimed that two individuals with the same score or rank can be quantitatively and, within limits, qualitatively similar in their attitude toward a given issue. As an example, suppose a test of liberalism consists of two general sorts of items, one concerned with economic and the other with religious issues. Two individuals could thus arrive at the same numerical score by quite different routes. Now it may be true that economic and religious liberalism are correlated but unless highly correlated, the meaning of scores based on such a composite is questionable.*

Second, aside from the perspective of interpreting test scores, unidimensionality is important because the mathematical basis on which most measurement models are derived assumes unidimensionality. Finally, assessing dimensionality is also important from a theoretical perspective. Examining the underlying structure of the IEA Reading Literacy Test data may further our understanding of reading literacy itself and how it should be taught within our schools.

The assumption of unidimensionality, however, does not imply that other factors (e.g., motivation, anxiety) do not have an impact on test performance. What is assumed, however, is that the trait or ability under consideration is a dominant factor in explaining examinees' test performance. Zwick (1987, 246) contends that "In practice, the assumption of unidimensionality, required for the application of conventional IRT models, will always be violated to some degree." Hattie (1985, 147) says "It is more meaningful to ask the degree to which a set of items departs from unidimensionality than to ask whether a set of items is unidimensional."

4.4 Methods for Assessing Dimensionality

Earlier we indicated that studying dimensionality is one aspect of gathering validity evidence for specific uses of the IEA Reading Literacy Tests. It is with this broader context in mind that in this chapter we will summarize evidence to assist the reader in formulating an integrated evaluative judgment regarding the dimensionality of IEA Reading Literacy Study item-response data.

A variety of methods were available for assessing dimensionality of Reading Literacy item-response data. Clearly, it was neither desirable nor practical to apply each and every method to analyze dimensionality of the IEA Reading Literacy Study item-response data. Some of the methods were deemed to be not relevant, while some methods were known to be problematic. For the purpose of studying dimensionality of the item-response data, the following types of evidence were collected:

- Evidence based on reliability;
- Evidence based on principal components;
- Evidence based on factor analysis; and
- Evidence based on Item Response Theory.

Evidence Based on Reliability

Coefficient alpha, the internal consistency index, has been widely used to assess dimensionality (Hattie 1985). Cronbach (1951) has shown that coefficient alpha is a lower bound to the proportion of test variance attributable to common factors among the test items. Because a high value of alpha, according to Cronbach, may be indicative of a high first-factor saturation, the implication has been that alpha relates to dimensionality. Green, Lissitz, and Mulirk (1977) have argued that though high internal consistency, as indicated by a high value of alpha, resulted when a general factor was present, this did not rule out obtaining a high value of alpha when a general factor was not present.

A more serious limitation of alpha as a method for assessing dimensionality relates to the fact that alpha is dependent on the length of the test, whereas, conceptually, the dimensionality of a set of item scores should be independent of the length of the test. Because the mean item intercorrelations, which show the homogeneity of items within the test, are not dependent on test length, they may be used to assess dimensionality. Cronbach (1951) noted that a low mean item intercorrelation could denote a nonhomogeneous test and recommended that when the mean correlation is low, a careful examination of the item intercorrelations may show whether a test could be broken into more homogeneous subtests. Fur : more, Armor (1974) has suggested that inspecting the inter-item correlations for patterns of low or negative correlations would provide useful information regarding dimensionality.

The mean item-test correlations, which also indicate the homogeneity of test items and are not dependent on test length, may also be used to assess dimensionality. Point-biserial and biserial correlations are two alternative indices representing the correlation between the total test score (continuous variable) and a dichotomous item score. Similar to the mean inter-item correlations, however, the mean item-test correlations may be problematic. Thus, inspecting the item-test correlations for low or negative correlations would also provide useful information regarding dimensionality.

Evidence Based on Principal Components

Principal component analysis (PCA) and factor analysis (FA) traditionally have been used to investigate the dimensionality of responses to a set of items. PCA is a method of transforming a given set of variables into a new set of composite variables (principal components) such that the composite(s) extract maximum variance from the original set of variables.

The first principal component may be viewed as the single best summary of linear relationships exhibited in the data. Since the first principal component explains the maximum variance, then this variance, expressed as the percentage of total variance, has been used as an index of unidimensionality. "The implication is that the larger the amount of variance explained by the first component, the closer the set of items is to being unidimensional" (Hattie 1985, 146).

The question may be raised, "How high should the variance explained by the first principal component be to indicate unidimensionality?" Reckase (1979) recommended that the first component should account for at least 20 percent of the variance. Others (e.g., Carmines and Zeller 1979) have recommended that at least 40 percent of the total variance should be accounted for by the first principal component to indicate unidimensionality.

Another problem with using this index as a method for assessing dimensionality relates to the fact that a multidimensional set of item responses may explain a higher variance on the first component than does a unidimensional set of item responses. Thus, Lumsden (1957, 1961) asserted that the ratio of the first and second eigenvalues (i.e., variance explained by the first and second components) may provide a reasonable index of unidimensionality. However, because this index does not have a fixed maximum value, Divgi (1980) recommended that the difference between the first and second eigenvalues divided by the difference between the second and third eigenvalues may provide a more reasonable index for assessing dimensionality. Divgi's Index will be very high if the difference between the second and third eigenvalues is very small. To overcome this problem, we have proposed a new index to assess unidimensionality. Our proposed index AI (i.e., Atash's Index) is defined as follows:

$$AI = [(r_{kk} / p_1) + (p_2/p_1) + (p_3/p_1)] / 3$$

where r_{kk} is the reliability coefficient, p_1 is the proportion of variance explained by the first eigenvalue, p_2 is the proportion of variance explained by the second eigenvalue, and p_3 is the proportion of variance explained by the third eigenvalue.

The logical basis of AI is that for a unidimensional set of items:

1. The first eigenvalue should be large, approaching the reliability of the test (i.e., $r_{kk} \geq p_1$).

2. The second and third eigenvalues should be small relative to the first eigenvalue.

The above formula indicates that as p_1 approaches r_{kk}, AI will be small, thereby indicating unidimensionality.[1] On the other hand, if p_1 is small compared to r_{kk}, AI will be substantially greater than one, indicating multidimensionality.[2]

[1] As P_1 approaches r_{kk}, P_2 and P_3 will be small relative to P_1.

[2] When P_1 is small compared to r_{kk}, most likely P_2 and/or P_3 will not be small relative to P_1.

The sum of squared residual correlations, after removing the first component, also has been used as an index of unidimensionality. If the one-component model fits the data well, the residual correlations (i.e., the difference between observed correlations and correlations implied by the model) would be small. The root mean square off-diagonal elements of the residual correlations was used as a summary statistic to assess the unidimensionality of the IEA Reading Literacy Test items.

Evidence Based on Factor Analysis

It was stated earlier that factor analysis has been traditionally used to investigate the dimensionality of responses to a set of items. Linear factor analysis of dichotomously scored items in general do not produce satisfactory results (Carrol 1945; Drasgow and Parsons 1983). "In applying a linear factor analysis model, we are hypothesizing that dichotomous variables are linear combinations of continuous latent variables with infinite range, a mathematical impossibility" (Zwick 1987, 246-247).

Two promising alternatives to conventional factor analysis are factor analysis of item parcels (Cook and Eignor 1984) and full-information factor analysis (Bock and Aitkin 1981; Bock, Gibbons, and Muraki 1985). Factor analysis of item parcels was achieved by grouping items relating to the same passage in one subtest and then applying conventional factor analysis to the subtest scores. The number of items within each passage ranged from two to seven, with the majority of passages having four or five items. The conventional factor analysis of these subtest scores avoided the problems encountered in factor analyzing the dichotomously scored items. However, because of the low reliabilities of the subtest scores, the problems in estimating commonalities may persist. More importantly, due to low subtest score reliabilities, the correlations among the subtest scores may be attenuated, thereby affecting the results of the factor analysis of subtest scores.

Bock and Aitkin (1981) developed a method of factor analysis, based directly on Item Response Theory, that does not require estimation of inter-item correlation coefficients. "Because the Bock-Aitkin approach uses as data the frequencies of all distinct item response vectors, it is called 'full-information' item factor analysis" (Bock, Gibbons, and Muraki 1985, 262). The authors state (277-278):

> *Implementation of item factor analysis by marginal maximum likelihood estimation overcomes many of the problems that attend factor analysis of tetrachoric correlation coefficients: It avoids the problem of indeterminate tetrachoric coefficients of extremely easy or difficult items, it readily accommodates effects of guessing and of omitted or not-reached items, and it provides a likelihood ratio test of the statistical significance of additional factors.*

The full-information factor analysis was implemented using the TESTFACT computer program (Wilson, Wood, and Gibbons 1991). TESTFACT requires as input the fixed values of the **c** parameter in the three-parameter IRT model. By fixing the **c** parameter to 0, in effect the two-parameter IRT model was applied to the IEA Reading Literacy Study data. The TESTFACT program generated chi-square values indicating the fit of the data to the model. The difference in chi-square values between models of different dimensions was used to assess dimensionality of the Reading Literacy Study data.

Evidence Based on Item Response Theory (IRT)

The IEA Reading Literacy Test data, which consisted of dichotomously scored item responses, were scaled using the Rasch model (one-parameter IRT). "One of the major advantages of the Latent

Trait Models (e.g., the Rasch model) often cited is that there are many indices of how adequately the data 'fits' the model" (Hattie 1985, 152). Wright and Panchapakesan (1969, 25) asserted that "if a given set of items fit the (Rasch) model, this is evidence that they refer to a unidimensional ability, that they form a conformable set." Thus, one of the most useful tests of the unidimensionality assumption in the context of Rasch model is the test of fit to the model that is part of the calibration process. Specifically, item-fit statistics provided as part of the calibration were used to assess the dimensionality of the IEA Reading Literacy Study item responses.

4.5 Results

Reliability

Table 4-1 presents the coefficient alpha and the number of items for each domain and for the total test. Since the number of test items varies by domain, one cannot readily compare the reliability coefficients across domains or populations. To facilitate such comparisons, Table 4-1 also includes estimated coefficient alphas (using the Spearman-Brown prophecy formula) for a test with 82 test items—the largest number of test items (i.e., grade 9 test).

Table 4-1. Coefficient alpha and the number of items for each domain and for the total test

Domain	Grade 4			Grade 9		
	Number of items	Alpha		Number of items	Alpha	
		Observed	Adjusted*		Observed	Adjusted*
Narrative	20	0.857	0.961	26	0.875	0.957
Expository	19	0.766	0.934	24	0.846	0.949
Document	21	0.733	0.915	32	0.791	0.907
Total test	60	0.916	0.937	82	0.932	0.932

* Adjusted alpha coefficient reflects the estimated reliability of the test for a test with 82 items.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

The unadjusted coefficients indicate that for both populations, narrative test items are the most homogeneous (i.e., highest alpha), followed by expository test items, while the document test items are the least homogeneous (i.e., lowest alpha). Table 4-1 also indicates that the unadjusted reliability coefficients for the total test are greater than the reliabilities for the domains. This is not surprising given the difference in the number of test items. A comparison of adjusted coefficients shows that the reliability for the narrative domain is higher (for both grades) than the reliability of the total test. Further, for both grades, the adjusted reliability for the document domain is lower than the reliability for the total test, while for the expository domain the adjusted domain reliability and the total test reliabilities are similar.

What can we say about the observed coefficient alpha? If we adopt the rule of thumb that coefficient alpha greater than 0.80 is high, between 0.75 and 0.80 is moderate, and less than 0.75 is low, it can be concluded that (a) coefficient alpha for narrative items (both populations) and expository items (grade 9) is high; (b) coefficient alpha for expository items (grade 4) and document items (grade 9) is moderate, and (c) coefficient alpha for document items (grade 4) is low.

It was stated previously that a high coefficient alpha does not necessarily mean that a general factor is present, since high alpha can be obtained even though a general factor does not exist. Item intercorrelations may provide additional information regarding dimensionality. Further, a comparison of intercorrelations of test items within a domain (e.g., intercorrelation of narrative test items) with correlations of test items across domains (e.g., correlation of narrative items with expository items) may also provide useful information regarding dimensionality. If within-domain intercorrelations are substantially larger than across-domain correlations, this may be indicative of more homogeneity within the domain as compared to homogeneity of the entire test items. Table 4-2 presents the distributional characteristics of within-domain intercorrelations of test items, and Table 4-3 presents the distributional characteristics of across-domain intercorrelations of test items.

Table 4-2 indicates that for grade 4 the median within-domain test item intercorrelations are 0.234, 0.147, and 0.115 for the narrative, expository, and document domains, respectively. For grade 9, the corresponding figures are 0.195, 0.171, and 0.102. Based on the average within-domain intercorrelations, for both grades narrative test items are the most homogeneous (i.e., highest mean and median), followed by expository test items, while the document test items are the least homogeneous. Table 4-3 indicates that for both grades the median across-domain correlations involving the narrative items (i.e., narrative with expository and narrative with document domains) are lower than the median intercorrelations for the narrative domain test items shown in Table 4-2. For example, for grade 4, the median correlations are 0.171 and 0.133, respectively, for narrative with expository and narrative with document test items. Both of these median correlations are lower than the median intercorrelation for the narrative test items (i.e., 0.234). For both grades the median within-domain correlations for expository test items lie between the across-domain test item intercorrelations involving the expository test items. For grade 4, the median within-domain correlations for document test items lie between the across-domain test item intercorrelations involving the document test items, whereas for grade 9, the median within-domain correlations for document test items are smaller than the across-domain test item intercorrelations involving the document test items. This pattern of intercorrelations is suggestive of a lower degree of homogeneity for the expository and document scales.

Table 4-2. Distributional characteristics of within-domain test item intercorrelations

Domain	Mean	Median	Lowest	Highest
Grade 4				
Narrative	0.236	0.234	0.127	0.414
Expository	0.151	0.147	0.023	0.418
Document	0.123	0.115	0.026	0.385
Grade 9				
Narrative	0.210	0.195	0.025	0.604
Expository	0.177	0.171	0.063	0.376
Document	0.103	0.102	-0.002	0.643

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 4-3. Distributional characteristics of cross-domain test item intercorrelations

Domain	Mean	Median	Lowest	Highest
Grade 4				
Narrative with Expository	0.170	0.171	0.041	0.331
Narrative with Document	0.130	0.133	0.038	0.253
Expository with Document	0.108	0.106	0.014	0.219
Grade 9				
Narrative with Expository	0.178	0.173	0.041	0.389
Narrative with Document	0.116	0.118	-0.015	0.241
Expository with Document	0.114	0.115	0.002	0.235

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Correlation of test items with domain and test scores may provide additional information concerning the dimensionality of the IEA Reading Literacy Test data. The domain-item correlations (i.e., point-biserial correlations between the domain scores and item scores) and test-item correlations (i.e., point-biserial correlations between the total test score and item scores) are presented for each grade in Tables 4-4 and 4-5.

Table 4-4. Grade 4 item correlations with domain total score

Item	Point-biserial correlation	
	Within domain	Total test
NARRATIVE		
Bird1	0.579	0.545
Bird2	0.424	0.353
Bird3	0.514	0.470
Bird4	0.569	0.511
Bird5	0.428	0.396
Dog2	0.505	0.458
Dog3	0.501	0.463
Dog4	0.543	0.500
Dog5	0.537	0.507
Dog6	0.546	0.505
Shark1	0.522	0.477
Shark2	0.464	0.414
Shark3	0.514	0.463
Shark4	0.539	0.499
Shark5	0.564	0.501
Grandpa1	0.529	0.495
Grandpa3	0.585	0.551
Grandpa4	0.529	0.495
Grandpa5	0.599	0.568
Grandpa6	0.456	0.420
Mean	0.522	0.479
EXPOSITORY		
Card1	0.281	0.292
Card2	0.187	0.192
Walrus1	0.368	0.325
Walrus2	0.400	0.365
Walrus3	0.524	0.484
Walrus4	0.508	0.458
Walrus5	0.460	0.404
Walrus6	0.494	0.454
Sand2	0.489	0.513
Sand3	0.408	0.432
Item	**With domain**	**Total test**
EXPOSITORY (continued)		
Marmot1	0.493	0.452
Marmot2	0.427	0.356
Marmot3	0.420	0.351
Marmot4	0.454	0.406
Trees1	0.570	0.507
Trees2	0.447	0.339
Trees3	0.516	0.423
Trees4	0.387	0.266
Trees5	0.476	0.377
Mean	0.437	0.389
DOCUMENT		
Island1	0.311	0.257
Island2	0.246	0.193
Island4	0.325	0.238
Maria1	0.335	0.262
Maria2	0.428	0.321
Maria3	0.421	0.370
Bottle1	0.318	0.257
Bottle2	0.515	0.449
Bottle3	0.325	0.271
Bottle4	0.376	0.330
Bus2	0.549	0.470
Bus3	0.464	0.392
Bus4	0.472	0.351
Content1	0.302	0.258
Content3	0.379	0.353
Temp1	0.415	0.334
Temp2	0.414	0.323
Temp3	0.439	0.356
Temp4	0.430	0.377
Temp5	0.465	0.417
Mean	0.399	0.330
Mean (all items)	0.452	0.398

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 4-5. Grade 9 item correlations with domain total score

Item	Point-biserial correlation	
	Within domain	Total test
NARRATIVE		
Fox2	0.364	0.367
Fox3	0.431	0.408
Fox4	0.308	0.302
Fox5	0.190	0.183
Mute1	0.496	0.475
Mute2	0.550	0.535
Mute3	0.574	0.553
Mute4	0.370	0.333
Mute5	0.537	0.511
Shark2	0.335	0.317
Shark3	0.401	0.396
Shark4	0.372	0.369
Shark5	0.412	0.401
Reveng1	0.624	0.577
Reveng2	0.461	0.416
Reveng3	0.561	0.517
Reveng4	0.557	0.522
Reveng5	0.525	0.497
Reveng6	0.497	0.488
Reveng7	0.560	0.535
Angel1	0.572	0.504
Angel2	0.717	0.533
Angel3	0.592	0.496
Angel5	0.583	0.478
Angel6	0.602	0.508
Angel7	0.554	0.460
Mean	0.490	0.449
EXPOSITORY		
Marmot1	0.409	0.378
Marmot2	0.434	0.397
Marmot3	0.343	0.300
Marmot4	0.429	0.409
Laser1	0.350	0.328
Laser2	0.522	0.507
Laser3	0.441	0.409
Laser4	0.490	0.431
Laser5	0.474	0.428
Laser6	0.554	0.525
Liter1	0.504	0.460
Liter3	0.479	0.437
Liter4	0.563	0.545
Parac1	0.439	0.405
Parac2	0.332	0.307
Parac3	0.351	0.318
EXPOSITORY (continued)		
Parac5	0.382	0.339
Parac6	0.459	0.417
Smoke1	0.488	0.451
Smoke2	0.507	0.452
Smoke3	0.563	0.520
Smoke4	0.488	0.444
Smoke5	0.565	0.558
Smoke6	0.503	0.456
Mean	0.461	0.426
DOCUMENT		
Card1	0.215	0.164
Card3	0.165	0.116
Card4	0.167	0.106
Card5	0.323	0.266
Card6	0.330	0.292
Card7	0.176	0.144
Resourc1	0.343	0.265
Resourc2	0.483	0.427
Resourc3	0.464	0.401
Job1	0.377	0.356
Job2	0.344	0.336
Lynx1	0.254	0.205
Lynx2	0.401	0.356
Lynx3	0.36	0.336
Bus1	0.359	0.303
Bus2	0.428	0.365
Bus3	0.432	0.371
Direct1	0.438	0.361
Direct2	0.505	0.430
Direct3	0.465	0.387
Weather1	0.407	0.340
Weather2	0.329	0.266
Weather3	0.436	0.363
Weather4	0.412	0.350
Temp1	0.281	0.228
Temp2	0.422	0.354
Temp3	0.366	0.310
Temp4	0.357	0.298
Temp5	0.290	0.248
Aspirol1	0.384	0.368
Aspirol2	0.396	0.381
Aspirol3	0.502	0.455
Mean	0.363	0.311
Mean (all items)	0.432	0.388

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

BEST COPY AVAILABLE

Tables 4-4 and 4-5 indicate that the domain-item correlations are highest for the narrative test items. The mean correlation is 0.522 and 0.490 for grades 4 and 9, respectively. For the document domain, the domain-item correlations are the lowest for both populations--the mean correlation is 0.399 and 0.363 for grades 4 and 9, respectively. For both grades the mean domain-item correlations for the expository domain are higher than the mean correlations for the document domain and lower than the mean correlations for the narrative domain. One item for grade 4 (*card2*) and four items for grade 9 (*fox5, card3, card4, and card7*) show relatively lower correlations--0.187, 0.190, 0.165, 0.167, and 0.176, respectively.

Tables 4-4 and 4-5 also indicate that for both populations the test-item intercorrelations are generally lower than the domain-item correlations: with the exception of five items (*card1*, *card2*, *sand2*, and *sand3* for grade 4 and *fox2* for grade 9), all domain-item correlations are higher than the test-item correlations. On the average, the differences are about 0.05, indicating that the items within each domain are more homogeneous than the entire set of test items.

Tables 4-4 and 4-5 also indicate that the domain-item correlations are generally high. With the exception of three items for grade 4 and eight items for grade 9, all items have correlations that are higher than 0.30, which is generally considered acceptable. In each of the exceptions, some type of ambiguity may account for this low correlation with the domain scores. For example, the two expository items in grade 4 that have correlations less than 0.30 are both associated with the same passage. Reading specialists in the United States had difficulty in determining whether this passage should be classified as a document or an expository text. In the case of the two specific items in question, there was some question as to whether the reader had to process text or understand the format of a postcard and correctly identify the answer based on its position in the address. In the case of *island*, the question is somewhat vague. In the case of *fox*, one of the distractors could easily be considered a correct answer. In the other two remaining cases, one might attribute the low correlation to problems in the test item construction.

Principal Components

The eigenvalues and percent variance explained by the first three principal components for the IEA Reading Literacy Study data are shown in Table 4-6. The eigenvalues represent the amount of total variance in the data a given component explains. For example, the total variance accounted for by the first principal component is 5.531 for grade 4 narrative test items. Because the percent variance represents the proportion of the total variance explained by a given component, for both grades the variance explained by the first principal component is highest for the narrative domain (27.6 and 24.7 percent for grades 4 and 9, respectively) and lowest for the document domain (16.8 and 14.1 percent, respectively). Based on Reckase's rule of thumb, the narrative and expository items for both grades meet the unidimensionality criterion, whereas the document items fall short of the 20 percent criterion.

Table 4-7 includes the Lumsden, Divgi, and Atash Indices for both populations. Table 4-7 indicates some of the problems associated with the Lumsden and Divgi Indices:

1. For grade 4, the total test has a higher Lumsden Index than all of the three test domains;

2. For grade 9, the expository and document domains have a higher Divgi Index than the narrative and the total test; and

3. The Lumsden and Divgi Indices are not always in agreement.

Table 4-6. **Eigenvalue and percent variance explained by the first three factors for each reading literacy domain and the total test, by grade**

Factor	Grade 4					
	Eigenvalue			Percent variance		
	Principal component	Factor analysis	Full-information FA	Principal component	Factor analysis	Full-information FA
	Narrative					
1st Factor	5.531	4.832	8.926	27.6	85.1	44.6
2nd Factor	1.197	0.528	1.241	6.0	9.3	6.2
3rd Factor	0.990	0.316	0.932	4.9	5.6	4.6
	Expository					
1st Factor	3.865	3.097	6.588	20.3	71.6	34.7
2nd Factor	1.520	0.749	1.687	8.0	17.3	8.9
3rd Factor	1.233	0.479	1.299	6.5	11.1	6.8
	Document					
1st Factor	3.545	2.783	6.685	16.8	70.0	31.8
2nd Factor	1.462	0.687	1.556	7.0	17.3	7.4
3rd Factor	1.158	0.512	1.260	5.5	12.9	6.0
	Total Test					
1st Factor	10.192	9.413	20.917	17.0	82.6	31.7
2nd Factor	1.969	1.160	2.757	3.3	10.2	4.2
3rd Factor	1.595	0.829	2.165	2.7	7.3	3.2

Factor	Grade 9					
	Eigenvalue			Percent variance		
	Principal component	Factor analysis	Full-information FA	Principal component	Factor analysis	Full-information FA
	Narrative					
1st Factor	6.439	5.800	10.751	24.7	72.7	41.4
2nd Factor	1.990	1.427	2.302	7.7	17.9	8.9
3rd Factor	1.453	0.748	1.469	5.6	9.4	5.7
	Expository					
1st Factor	5.196	4.442	8.470	21.7	80.7	35.3
2nd Factor	1.303	0.615	1.456	5.4	11.2	6.0
3rd Factor	1.182	0.447	1.193	4.9	8.1	5.0
	Document					
1st Factor	4.364	3.592	8.397	14.1	70.8	26.2
2nd Factor	1.446	0.960	2.116	4.7	18.9	6.6
3rd Factor	1.371	0.521	1.532	4.4	10.3	4.8
	Total Test					
1st Factor	13.097	12.339	24.060	16.1	80.4	29.3
2nd Factor	2.477	1.874	3.415	3.1	12.2	4.2
3rd Factor	1.909	1.132	2.613	2.4	7.4	3.2

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 4-7. Lumsden, Divgi, and Atash Indices for each reading literacy domain and the total test, by grade

Domain	Grade 4			Grade 9		
	Lumsden Index	Divgi Index	Atash Index	Lumsden Index	Divgi Index	Atash Index
Narrative	4.6	19.7	1.2	3.2	8.1	1.4
Expository	2.5	8.2	1.5	4.0	32.6	1.4
Document	2.4	6.5	1.7	3.0	31.3	2.1
Total Test	5.2	22.8	1.9	5.2	18.6	2.0

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center foı Education Statistics, 1991.

Using a value of 1.5 or less as an indication of unidimensionality for the Atash Index, it can be concluded that for both grades the narrative and expository test items exhibit unidimensionality, while the document test items and the entire test items taken as a whole do not seem to exhibit unidimensionality.

The root mean square off-diagonal elements of the residual correlations are 0.0514, 0.0681, and 0.0586 for narrative, expository, and document items, respectively. The corresponding numbers for grade 9 are 0.0476, 0.0561, and 0.0463. This index shows that for both grades the residual correlations, on the average, are small.

Factor Analysis

It was stated earlier that because linear factor analysis of dichotomously scored items may not have produced satisfactory results, factor analysis of item parcels and full-information factor analysis were applied to the IEA Reading Literacy Study data. Table 4-8 presents the results of the factor analysis on parcels for grades 4 and 9. Table 4-8 shows that the percentage of variance attributed to the first factor is high for narrative (both grades) and expository (grade 9). The percentage of variance attributed to the first factor is low for grade 4 expository and document item parcels. For grade 4, the root mean squares of the residual correlations are lower than the corresponding numbers for grade 9. This difference in RMSs may be due to the difference in sample size for the two grades.

Table 4-8. First factor statistics based on item parcels

Domain	Number of parcels	Eigenvalue	Percent variance*	RMS residual correlations
		Grade 4		
Narrative	4	1.945	48.6	0.032
Expository	5	1.304	26.1	0.044
Document	6	1.496	24.9	0.027
		Grade 9		
Narrative	5	2.589	51.8	0.126
Expository	5	2.603	52.1	0.121
Document	7	2.605	37.2	0.104

RMS = root mean square.

*Percent variance is computed as the eigenvalue divided by the total variance. In designating total variance, we have considered the full-correlational matrix (with ones on the diagonal) instead of the reduced correlational matrix (with the commonalities on the diagonals).

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 4-9 presents the rotated factor loadings for one-, two-, and three-factor solutions for both grades. Because the factor loadings were obtained by analyzing all item parcels together, the underlying structure of the item parcels may be inferred from the rotated factor loadings. For grade 4, a two-factor solution can be meaningfully interpreted. The first factor has a high loading on *walrus* (E), *sand* (E), *marmot* (E), *trees* (E), *buses* (D), *temperature* (D), *bird* (N), *dog* (N), *shark* (N), and *grandpa* (N). With the exception of *buses* and *temperature*, all of these item parcels had been designated as either narrative or expository. Thus, the first factor can be labeled narrative-expository. Because the second factor has a high loading on *card* (E), *island* (D), *bottles* (D), *buses* (D), and *content* (D), it can be labeled document.

Table 4-9. Rotated factor loadings of item parcels for one-, two-, and three-factor solutions, by grade

Parcel	One-factor	Two-factor		Three-factor		
			Grade 4			
Card	0.331	0.080	0.582	0.045	0.271	0.289
Walrus	0.450	0.623	0.291	0.408	0.405	0.302
Sand	0.572	0.509	0.343	0.237	0.501	0.276
Marmot	0.561	0.703	0.008	0.582	0.204	0.157
Trees	0.510	0.633	0.029	0.490	0.198	0.169
Island	0.321	0.077	0.568	0.119	0.125	0.330
Maria	0.455	0.281	0.502	0.215	0.171	0.421
Bottles	0.529	0.340	0.546	0.239	0.194	0.515
Buses	0.562	0.470	0.391	0.330	0.175	0.444
Content	0.367	0.128	0.578	0.096	0.220	0.347
Temp	0.561	0.544	0.271	0.487	0.069	0.414
Bird	0.690	0.633	0.293	0.406	0.484	0.304
Dog	0.725	0.691	0.301	0.449	0.490	0.310
Shark	0.709	0.681	0.293	0.411	0.546	0.274
Grandpa	0.700	0.749	0.167	0.577	0.416	0.198
			Grade 9			
Card	-.362	0.334	0.174	0.146	0.271	0.214
Resource	0.498	0.462	0.239	0.216	0.463	0.187
Job	0.444	0.376	0.249	0.201	0.251	0.326
Lynx	0.462	0.383	0.267	0.240	0.329	0.230
Bus	0.477	0.453	0.218	0.191	0.420	0.220
Direct	0.450	0.392	0.240	0.225	0.404	0.147
Weather	0.504	0.493	0.217	0.193	0.512	0.178
Temperature	0.473	0.482	0.184	0.154	0.441	0.236
Aspirol	0.571	0.520	0.284	0.256	0.498	0.240
Fox	0.508	0.446	0.268	0.171	0.201	0.547
Mute	0.721	0.371	0.662	0.616	0.278	0.343
Shark	0.519	0.462	0.268	0.192	0.265	0.467
Revenge	0.765	0.723	0.613	0.554	0.334	0.428
Angel	0.573	0.234	0.593	0.582	0.234	0.159
Marmot	0.581	0.462	0.356	0.293	0.298	0.424
Laser	0.712	0.479	0.527	0.469	0.335	0.424
Literacy	0.612	0.291	0.588	0.556	0.245	0.244
Paracutin	0.574	0.452	0.355	0.279	0.241	0.495
Smoke	0.723	0.345	0.698	0.683	0.326	0.231

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

For grade 9 on the other hand, a three-factor solution can be meaningfully interpreted. The first factor has high loadings on *mute* (N), *revenge* (N), *angel* (N), *laser* (E), *literacy* (E), and *smoke* (E). The second factor has high loading on *resource* (D), *lynx* (D), *bus* (D), *direct* (D), *weather* (D), *temperature* (D), *aspirol* (D), *laser* (E), and *smoke* (E), while the third factor has high loading on *job* (D), *fox* (N), *shark* (N), *revenge* (N), *marmot* (E), *laser* (E), and *paracutin* (E). While factor two clearly can be labeled document, it is hard to distinguish between the first and third factors since both of these factors have high loadings on both narrative and expository parcels.

Table 4-10 presents the results of the full-information factor analysis for grades 4 and 9. For grade 4 the first factor extracted about 32, 26, and 24 percent of the total variance for the narrative, expository, and document domains, respectively. The corresponding numbers for grade 9 were 33, 25, and 21 percent. By comparison, the first factor for the total test (i.e., the entire test items disregarding domain designation) accounted for 27 and 23 percent of the total variance for grades 4 and 9, respectively. As compared to the first factor, the second and third factors generally accounted for much smaller percentages of the total variance for all domains for both grades.

Table 4-10. Full-information item factor analysis (three-factor solution)

Domain	Factor number	Percent variance	Chi-square change	Degrees of freedom
		Grade 4		
Narrative . . .	1	32.3	NA	NA
	2	3.8	207.9	19
	3	2.1	*	18
Expository	1	26.2	NA	NA
	2	6.1	1,018.9	18
	3	3.9	61.9	17
Document	1	24.4	NA	NA
	2	5.0	267.4	20
	3	3.6	153.0	19
Total test	1	26.8	NA	NA
	2	2.5	1,705.1	65
	3	1.2	926.8	64
		Grade 9		
Narrative	1	33.2	NA	NA
	2	7.1	1,819.0	25
	3	3.1	207.3	24
Expository	1	24.6	NA	NA
	2	4.4	157.5	23
	3	2.6	*	22
Document	1	20.5	NA	NA
	2	3.7	181.5	31
	3	3.1	431.7	30
Total test	1	23.4	NA	NA
	2	3.1	2,466.3	81
	3	0.8	829.5	80

NA = not applicable; * = small changes in chi-square value.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

The chi-square change for the improvement in fit by adding a second factor was 208 (degrees of freedom--df=19), 1,019 (df=18), and 267 (df=20) for grade 4 narrative, expository, and document domains, respectively. The corresponding numbers for grade 9 were 1,819 (df=25), 158 (df=23), and 182 (df=31). These values of chi-square need to be evaluated in light of the large sample sizes (in excess of 6,000 for grade 4 and 3,000 for grade 9) and the design effect of around 6 and 8 for grades 4 and 9, respectively. Therefore, corrections to the observed chi-square values that account for these two attributes were performed by transforming the chi-square values so that the actual sample size would function as if it were a simple random sample of 400 students.[3] The adjusted chi-square change for the improvement in fit by adding a second factor was 21.9 (df=19), 107.4 (df=18), and 28.1 (df=20) for grade 4 narrative, expository, and document domains, respectively. The corresponding numbers for grade 9 were 234.6 (df=25), 20.3 (df=23), and 23.4 (df=30). Therefore, the adjusted chi-square change for the improvement in fit by adding a second factor was not significant for grade 4 narrative and expository domains or grade 9 expository and document domains. For grade 4 expository domain and grade 9 narrative domain, however, the addition of the second factor significantly improved the fit. In both of these cases, the addition of a third factor did not significantly improve the fit.

Tables 4-11 and 4-12 present the unrotated factor loadings for a one-factor full-information factor solution for each domain for grades 4 and 9. Table 4-11 shows that except *trees4*, *temp3*, and *temp4*, all other items have factor loadings of 0.40 or higher. For grade 9 items, however, 2 narrative items (i.e., *fox5* and *mute4*), 2 expository items (i.e., *marmot3* and *parac2*), and 11 document items (i.e., *card3*, *card4*, *card5*, *card6*, *card7*, *resourc1*, *lynx1*, *lynx3*, *bus1*, *weather2*, and *temp1*) have factor loadings that are smaller than 0.40.

Table 4-11. Grade 4 factor loadings for one-factor full-information factor analysis

Item	Factor loading	Item	Factor loading
NARRATIVE		EXPOSITORY (continued)	
Bird1	0.669	Marmot1	0.527
Bird2	0.459	Marmot2	0.439
Bird3	0.570	Marmot3	0.429
Bird4	0.704	Marmot4	0.478
Bird5	0.699	Trees1	0.691
Dog2	0.572	Trees2	0.455
Dog3	0.614	Trees3	0.565
Dog4	0.662	Trees4	0.387
Dog5	0.611	Trees5	0.544
Dog6	0.729	DOCUMENT	
Shark1	0.725	Island1	0.517
Shark2	0.538	Island2	0.596
Shark3	0.648	Island4	0.614
Shark4	0.650	Maria1	0.638
Shark5	0.652	Maria2	0.602
Grandpa1	0.620	Maria3	0.687
Grandpa3	0.687	Bottle1	0.675
Grandpa4	0.629	Bottle2	0.576
Grandpa5	0.732	Bottle3	0.534
Grandpa6	0.498	Bottle4	0.457
EXPOSITORY		Bus1	0.631
Card1	0.582	Bus2	0.461
Card2	0.410	Bus3	0.434
Walrus1	0.617	Bus4	0.440
Walrus2	0.708	Content1	0.430
Walrus3	0.676	Content3	0.597
Walrus4	0.650	Temp1	0.442
Walrus5	0.557	Temp2	0.402
Walrus6	0.541	Temp3	0.333
Sand2	0.657	Temp4	0.343
Sand3	0.634	Temp5	0.423

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

[3] The target sample size for the IEA Reading Literacy Study was 400 (see Chapter 1 of this volume).

Table 4-12. Grade 9 factor loadings for one-factor full-information factor analysis

Item	Factor loading	Item	Factor loading
NARRATIVE		**EXPOSITORY (continued)**	
Fox2	0.553	Parac3	0.496
Fox3	0.486	Parac5	0.405
Fox4	0.472	Parac6	0.530
Fox5	0.216	Smoke1	0.554
Mute1	0.535	Smoke2	0.594
Mute2	0.604	Smoke3	0.664
Mute3	0.645	Smoke4	0.579
Mute4	0.376	Smoke5	0.705
Mute5	0.583	Smoke6	0.582
Shark2	0.554	**DOCUMENT**	
Shark3	0.646	Card1	0.491
Shark4	0.614	Card3	0.313
Shark5	0.556	Card4	0.232
Reveng1	0.719	Card5	0.353
Reveng2	0.490	Card6	0.329
Reveng3	0.635	Card7	0.255
Reveng4	0.623	Resourc1	0.379
Reveng5	0.530	Resourc2	0.585
Reveng6	0.560	Resourc3	0.536
Reveng7	0.676	Job1	0.439
Angel1	0.679	Job2	0.436
Angel2	0.759	Lynx1	0.263
Angel3	0.736	Lynx2	0.450
Angel5	0.731	Lynx3	0.368
Angel6	0.757	Bus1	0.374
Angel7	0.701	Bus2	0.480
EXPOSITORY		Bus3	0.475
Marmot1	0.534	Direct1	0.627
Marmot2	0.540	Direct2	0.702
Marmot3	0.351	Direct3	0.572
Marmot4	0.541	Weather1	0.574
Laser1	0.568	Weather2	0.369
Laser2	0.690	Weather3	0.576
Laser3	0.562	Weather4	0.454
Laser4	0.566	Temp1	0.383
Laser5	0.541	Temp2	0.509
Laser6	0.739	Temp3	0.408
Liter1	0.571	Temp4	0.419
Liter3	0.568	Temp5	0.457
Liter4	0.688	Aspirol1	0.671
Parac1	0.489	Aspirol2	0.567
Parac2	0.396	Aspirol3	0.587

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

BEST COPY AVAILABLE

To further investigate the underlying structure of the items, full-information factor analysis was applied to the entire set of test items for each population. Table 4-13 presents the factor loadings for one-, two-, and three-factor full-information factor analysis for grades 4 and 9. To enhance the interpretation of these factor ioadings, the output of the TESTFACT factor loadings were rotated by SAS using the Varimax option. Tables 4-14 and 4-15 present the rotated factor loadings for the three-factor full-information factor analysis for grades 4 and 9. Tables 4-14 and 4-15 indicate that for both grades the first factor can be labeled expository. The second factor for grade 4 and the third factor for grade 9 can be labeled document. The third factor for grade 4 and the second factor for grade 9 seem to be a "speed factor" because the items with high loadings on this factor (*grandpa* and *trees* items for grade 4 and *angel* items for grade 9) are at the end of the second testing session.

Evidence Based on Item Response Theory

The Rasch model assumes that the item characteristic curves take the form of parallel logistic distribution functions. Although one could plot and visually examine the curves for each item in the IEA Reading Literacy Study, there are better tests of fit to the model. These fit statistics, which take into account all departures from the assumptions of the model, are provided as part of the item calibration process. The output from the BIGSCALE program, a software package for conducting IRT analyses, provides all the necessary fit statistics (Tables 4-16 and 4-17). The column COUNT indicates the number of examinees correctly responding to an item, and the column SAMPLE refers to the number of examinees with valid responses for an item. Thus, by dividing COUNT by SAMPLE, the proportion of examinees correctly responding to an item (i.e., p-value) can be obtained.

The column CALIBRTN (calibration) indicates the item's estimated difficulty value on the logit scale. The logit values have been suffixed by "A" to indicate that item values have been anchored at the values obtained for the international calibration sample. The column ERROR represents the standard errors associated with estimating the item difficulty value.

Two types of fit statistics are provided—INFIT and OUTFIT. While both are measure of model fit—the degree to which the observed data agree with predicted values based on the model—the infit statistic is more sensitive to unexpected responses by people whose abilities are around the item's difficulty value. In contrast, the outfit statistic is more sensitive to responses by people whose abilities are some distance (on the logit scale) from the item difficulty value. MNSQ shows the mean-square infit (or outfit) statistics, with the expected value equal to 1. Values substantially less than 1 may indicate dependency in the data, while values substantially greater than 1 may indicate random error (noise).

The column DISPLACE represents the difference between the anchored value (based on a best fit of the data to the international calibration sample) and the item difficulty estimate resulting from a best fit of the data to the model based on the U.S. sample of students. The optimal fit for the international calibration sample may not necessarily produce item parameters that may be considered optimal for the U.S. sample. DISPLACE shows the departure from optimality for the U.S. sample relative to the international calibration sample.

Inspection of both of these fit statistics, which range from -25.9 to 25.7, reveals that they are generally within acceptable ranges when one considers the attributes of the sample design as well as the U.S. sample size. For example, in Table 4-16 the narrative domain data include fit statistics of 11.7 and 25.1, which appear to be quite large. However, to evaluate this statistic one must take into account the large sample size (in excess of 6,000) and the sampling design, which may also contribute to the inflation of this fit statistic. It is known that the design effect for estimating the grade mean is about 6 for grade 4 and about 8 for grade 9 (see Chapter 2 of this volume), and that the design effects for estimating

Table 4-13. **Factor loadings for one-, two-, and three-factor full-information factor analysis for grades 4 and 9**

Grade 4						
Item	One factor	Two factor		Three factor		
Bird1	0.640	0.620	-0.023	0.641	-0.085	-0.054
Bird2	0.400	0.378	-0.077	0.409	-0.192	-0.068
Bird3	0.543	0.522	-0.059	0.548	-0.131	-0.036
Bird4	0.643	0.619	-0.072	0.659	-0.250	-0.107
Bird5	0.648	0.612	-0.105	0.660	-0.239	-0.035
Dog2	0.530	0.519	0.05	0.538	-0.067	-0.078
Dog3	0.582	0.559	-0.006	0.587	-0.082	-0.015
Dog4	0.618	0.595	-0.022	0.624	-0.112	-0.035
Dog5	0.600	0.572	0.006	0.593	-0.033	-0.022
Dog6	0.682	0.666	-0.024	0.694	-0.077	-0.071
Shark1	0.673	0.648	-0.057	0.683	-0.207	-0.078
Shark2	0.496	0.479	0.005	0.503	-0.088	-0.069
Shark3	0.596	0.573	-0.021	0.604	-0.157	-0.069
Shark4	0.610	0.592	0.009	0.615	-0.058	-0.052
Grandpa1	0.584	0.580	0.186	0.589	0.070	-0.145
Grandpa3	0.652	0.647	0.235	0.649	0.112	-0.169
Grandpa4	0.588	0.594	0.245	0.590	0.169	-0.126
Grandpa5	0.695	0.708	0.309	0.709	0.180	-0.200
Grandpa6	0.473	0.465	0.122	0.470	0.070	-0.086
Card1	0.566	0.531	-0.096	0.568	-0.129	-0.046
Card2	0.402	0.373	-0.160	0.406	-0.047	0.072
Walrus1	0.497	0.499	0.085	0.502	0.169	0.104
Walrus2	0.605	0.603	0.104	0.608	0.154	0.063
Walrus3	0.597	0.581	0.003	0.603	-0.041	-0.006
Walrus4	0.564	0.545	0.037	0.568	-0.013	0.007
Walrus5	0.479	0.462	0.003	0.477	0.004	0.023
Walrus6	0.512	0.496	0.008	0.511	-0.028	-0.029
Sand2	0.655	0.632	-0.023	0.664	-0.121	-0.037
Sand3	0.628	0.604	-0.052	0.639	-0.148	-0.015
Marmot1	0.519	0.503	0.065	0.518	0.013	-0.055
Marmot2	0.418	0.409	0.054	0.419	0.005	-0.077
Marmot3	0.405	0.394	0.063	0.403	0.021	-0.050
Marmot4	0.474	0.456	0.060	0.468	0.015	-0.058
Trees1	0.590	0.606	0.362	0.584	0.502	-0.057
Trees2	0.376	0.380	0.254	0.358	0.346	-0.032
Trees3	0.482	0.487	0.301	0.466	0.383	-0.037
Trees4	0.296	0.307	0.301	0.283	0.348	-0.602
Trees5	0.425	0.441	0.334	0.402	0.575	0.019
Island1	0.365	0.326	-0.169	0.357	-0.046	0.094
Island2	0.338	0.290	-0.186	0.321	-0.594	0.039
Island4	0.273	0.251	-0.082	0.267	0.006	0.051
Maria1	0.302	0.281	-0.061	0.291	0.010	0.055
Maria2	0.347	0.314	-0.014	0.331	-0.012	0.111
Maria3	0.458	0.422	-0.130	0.448	-0.043	0.099
Bottle1	0.471	0.400	-0.244	0.451	0.003	0.208
Bottle2	0.534	0.503	-0.103	0.522	0.025	0.084
Bottle3	0.510	0.433	-0.294	0.495	-0.020	0.207
Bottle4	0.509	0.463	-0.151	0.500	0.028	0.144
Bus1	0.553	0.499	-0.263	0.538	-0.044	0.230
Bus2	0.523	0.491	-0.245	0.518	-0.024	0.236
Bus3	0.488	0.470	-0.165	0.478	-0.030	0.122
Bus4	0.385	0.358	-0.226	0.381	-0.031	0.217
Content1	0.389	0.352	-0.144	0.378	-0.002	0.092
Content3	0.569	0.528	-0.123	0.561	0.032	0.127
Temp1	0.388	0.364	-0.097	0.378	0.015	0.116
Temp2	0.375	0.352	-0.070	0.364	0.004	0.079
Temp3	0.397	0 375	-0.060	0.386	0.004	0.062
Temp4	0.362	0.346	-0.056	0.358	0.022	0.070
Temp5	0.528	0.501	-0.691	0.516	0.080	0.126

Table 4-13. Factor loadings for one-, two-, and three-factor full-information factor analysis for grades 4 and 9 (continued)

Grade 9						
Item	One factor	Two factor		Three factor		
Fox2	0.572	0.507	0.143	0.582	-0.191	-0.057
Fox3	0.479	0.430	0.133	0.490	0.188	-0.088
Fox4	0.459	0.417	0.072	0.458	0.045	-0.011
Fox5	0.229	0.199	0.081	0.222	0.068	0.016
Mute1	0.532	0.500	0.055	0.528	0.056	-0.054
Mute2	0.608	0.572	0.028	0.605	0.037	-0.106
Mute3	0.638	0.604	-0.013	0.645	0.008	-0.112
Mute4	0.369	0.346	-0.032	0.358	-0.031	-0.050
Mute5	0.576	0.547	0.015	0.571	0.036	0.128
Shark2	0.520	0.488	0.058	0.532	0.039	0.002
Shark3	0.621	0.577	0.101	0.626	0.093	0.066
Shark4 . . .	0.600	0.544	0.122	0.611	0.083	0.074
Shark5	0.542	0.491	0.109	0.551	0.117	0.021
Reveng1	0.679	0.640	0.042	0.687	0.065	0.138
Reveng2	0.470	0.438	0.063	0.485	0.107	0.141
Reveng3	0.598	0.558	0.055	0.604	0.089	0.137
Reveng4	0.602	0.555	0.062	0.604	0.084	0.107
Reveng5	0.562	0.532	0.037	0.566	0.035	0.046
Reveng6	0.554	0.529	0.021	0.558	0.023	0.014
Reveng7	0.649	0.616	0.012	0.655	0.037	0.038
Angel1	0.600	0.623	-0.353	0.580	-0.368	0.081
Angel2	0.646	0.693	-0.445	0.624	-0.500	0.056
Angel3	0.606	0.653	-0.486	0.573	-0.548	0.005
Angel5	0.585	0.652	-0.546	0.547	-0.630	0.022
Angel6	0.614	0.693	-0.531	0.585	-0.650	0.027
Angel7	0.564	0.647	-0.531	0.525	-0.645	0.039
Marmot1 . . .	0.467	0.440	0.064	0.470	0.051	0.027
Marmot2 . . .	0.473	0.425	0.139	0.476	0.142	0.006
Marmot3 . . .	0.325	0.290	0.095	0.318	0.088	0.012
Marmot4 . . .	0.492	0.455	0.103	0.497	0.108	0.011
Laser1	0.483	0.463	0.026	0.493	0.030	0.020
Laser2	0.622	0.574	0.108	0.630	0.140	0.041
Laser3	0.500	0.460	0.057	0.504	0.075	0.044
Laser4	0.491	0.458	0.098	0.498	0.105	0.024
Laser5	0.484	0.450	0.063	0.486	0.073	0.037
Laser6	0.652	0.614	0.068	0.664	0.087	0.010
Liter1	0.516	0.484	-0.024	0.504	-0.029	0.001
Liter3	0.502	0.487	-0.100	0.496	-0.090	0.058
Liter4	0.641	0.627	-0.111	0.636	-0.114	0.054
Parac1	0.454	0.408	0.107	0.465	0.146	0.099
Parac2	0.365	0.314	0.129	0.370	0.167	0.033
Parac3	0.439	0.389	0.110	0.451	0.142	0.076
Parac5	0.362	0.315	0.140	0.366	0.151	0.039
Parac6	0.479	0.428	0.118	0.475	0.142	0.059
Smoke1	0.502	0.483	-0.044	0.503	-0.047	0.049
Smoke2	0.525	0.501	0.013	0.535	0.036	0.081
Smoke3	0.602	0.587	-0.065	0.600	-0.063	0.061
Smoke4	0.519	0.513	-0.109	0.517	-0.101	0.072
Smoke5	0.653	0.637	-0.126	0.647	-0.152	0.019
Smoke6	0.534	0.499	0.028	0.532	0.057	0.050
Card1	0.389	0.312	0.144	0.358	0.119	0.117
Card3	0.213	0.193	0.017	0.180	-0.028	0.015
Card4	0.153	0.125	0.047	0.139	0.058	0.037
Card5	0.296	0.264	0.102	0.289	0.075	0.052
Card6	0.296	0.268	0.082	0.296	0.061	0.005

Table 4-13. Factor loadings for one-, two-, and three-factor full-information factor analysis for grades 4 and 9 (continued)

Grade 9						
Item	One factor	Two factor		Three factor		
Card7	0.216	0.174	0.146	0.210	0.149	0.019
Resourc1 . . .	0.281	0.245	0.104	0.265	0.038	0.087
Resourc2 . .	0.474	0.426	0.165	0.469	0.123	0.068
Resourc3 . . .	0.425	0.386	0.119	0.416	0.074	0.120
Job1	0.415	0.379	0.121	0.420	0.129	0.012
Job2	0.434	0.393	0.115	0.444	0.128	0.015
Lynx1	0.213	0.189	0.090	0.213	0.061	0.017
Lynx2	0.390	0.346	0.149	0.387	0.141	0.064
Lynx3	0.352	0.332	0.062	0.358	0.050	0.005
Bus1	0.321	0.290	0.066	0.316	0.028	0.089
Bus2	0.388	0.363	0.123	0.383	0.049	0.121
Bus3	0.385	0.348	0.136	0.379	0.074	0.136
Direct1	0.426	0.399	0.090	0.423	-0.050	0.315
Direct2	0.508	0.472	0.090	0.503	-0.048	0.344
Direct3 . . .	0.420	0.387	0.089	0.407	-0.006	0.230
Weather1 . . .	0.457	0.401	0.140	0.451	0.104	0.086
Weather2 .	0.302	0.272	0.053	0.284	0.009	0.094
Weather3 . . .	0.445	0.400	0.114	0.434	0.080	0.108
Weather4 . . .	0.371	0.336	0.103	0.364	0.088	0.090
Temp1	0.311	0.275	0.092	0.303	0.067	0.099
Temp2	0.406	0.366	0.137	0.401	0.091	0.106
Temp3	0.341	0.309	0.051	0.329	0.019	0.087
Temp4	0.343	0.305	0.119	0.336	0.073	0.081
Temp5	0.380	0.342	0.037	0.365	0.012	0.017
Aspirol1	0.649	0.617	0.038	0.675	0.045	0.033
Aspirol2	0.529	0.507	0.010	0.526	-0.047	0.100
Aspirol3	0.498	0.456	0.106	0.484	0.063	0.116

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 4-14. Rotated factor loadings for the three-factor full-information factor analysis for grade 4

Item	Factor 1	Factor 2	Factor 3
Bird1	0.464	0.298	0.355
Bird2	0.428	0.186	0.005
Bird3	0.426	0.341	0.120
Bird4	0.645	0.346	0.001
Bird5	0.545	0.416	0.072
Dog2	0.477	0.256	0.136
Dog3	0.454	0.365	0.136
Dog4	0.475	0.383	0.147
Dog5	0.412	0.356	0.226
Dog6	0.582	0.346	0.193
Shark1	0.617	0.349	0.099
Shark2	0.449	0.232	0.120
Shark3	0.561	0.301	0.068
Shark4	0.533	0.304	0.166
Shark5	0.619	0.222	0.087
Grandpa1	0.441	0.271	0.308
Grandpa3	0.524	0.224	0.387
Grandpa4	0.457	0.182	0.418
Grandpa5	0.537	0.254	0.473
Grandpa6	0.357	0.198	0.262
Card1	0.440	0.353	0.128
Card2	0.245	0.303	0.144
Walrus1	0.220	0.365	0.327
Walrus2	0.328	0.402	0.358
Walrus3	0.419	0.362	0.230
Walrus4	0.393	0.352	0.217
Walrus5	0.336	0.290	0.184
Walrus6	0.356	0.308	0.196
Sand2	0.503	0.407	0.163
Sand3	0.489	0.395	0.155
Marmot1	0.417	0.234	0.227
Marmot2	0.347	0.173	0.189
Marmot3	0.329	0.160	0.190
Marmot4	0.378	0.203	0.217
Trees1	0.254	0.226	0.693
Trees2	0.105	0.186	0.454
Trees3	0.170	0.250	0.524
Trees4	0.049	0.138	0.424
Trees5	0.025	0.205	0.674
Island1	0.199	0.289	0.121
Island2	0.451	0.279	-0.427
Island4	0.156	0.202	0.103
Maria1	0.164	0.222	0.109
Maria2	0.167	0.286	0.106
Maria3	0.257	0.351	0.154
Bottle1	0.193	0.438	0.133
Bottle2	0.305	0.375	0.211
Bottle3	0.228	0.468	0.152
Bottle4	0.268	0.412	0.165
Bus1	0.244	0.508	0.163
Bus2	0.224	0.504	0.154
Bus3	0.266	0.385	0.161
Bus4	0.138	0.404	0.104
Content1	0.213	0.301	0.129
Content3	0.303	0.438	0.217
Temp1	0.188	0.323	0.141
Temp2	0.205	0.281	0.123
Temp3	0.238	0.283	0.138
Temp4	0.203	0.271	0.143
Temp5	0.288	0.417	0.175

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 4-15. Rotated factor loadings for the three-factor full-information factor analysis for grade 9

Item	Factor 1	Factor 2	Factor 3
Fox2	0.373	0.458	0.193
Fox3	0.479	0.074	0.226
Fox4	0.340	0.173	0.261
Fox5	0.171	0.041	0.151
Mute1	0.451	0.207	0.220
Mute2	0.498	0.262	0.260
Mute3	0.512	0.307	0.275
Mute4	0.307	0.179	0.115
Mute5	0.470	0.243	0.238
Shark2	0.384	0.215	0.297
Shark3	0.435	0.212	0.419
Shark4	0.417	0.212	0.406
Shark5	0.433	0.155	0.325
Reveng1	0.566	0.274	0.310
Reveng2	0.449	0.135	0.211
Reveng3	0.512	0.214	0.265
Reveng4	0.508	0.222	0.263
Reveng5	0.412	0.234	0.320
Reveng6	0.397	0.247	0.309
Reveng7	0.474	0.277	0.368
Angel1	0.286	0.634	0.151
Angel2	0.271	0.740	0.152
Angel3	0.179	0.708	0.206
Angel5	0.121	0.786	0.174
Angel6	0.089	0.890	0.199
Angel7	0.084	0.773	0.195
Marmot1	0.330	0.175	0.291
Marmot2	0.387	0.103	0.299
Marmot3	0.611	-0.638	0.363
Marmot4	0.429	0.070	0.312
Laser1	0.352	0.205	0.273
Laser2	0.497	0.175	0.373
Laser3	0.381	0.166	0.289
Laser4	0.398	0.131	0.298
Laser5	0.377	0.175	0.287
Laser6	0.496	0.233	0.379
Liter1	0.346	0.237	0.273
Liter3	0.361	0.333	0.170
Liter4	0.458	0.399	0.247
Parac1	0.448	0.100	0.207
Parac2	0.330	0.026	0.237
Parac3	0.430	0.099	0.194
Parac5	0.314	0.039	0.228
Parac6	0.451	0.113	0.210
Smoke1	0.404	0.245	0.191
Smoke2	0.449	0.229	0.222
Smoke3	0.430	0.380	0.225
Smoke4	0.375	0.342	0.181
Smoke5	0.352	0.480	0.329
Smoke6	0.451	0.207	0.220
Card1	0.232	0.054	0.319
Card3	0.113	0.083	0.114
Card4	0.100	0.010	0.121
Card5	0.207	0.062	0.215
Card6	0.228	0.088	0.184
Card7	0.199	-0.034	0.162
Resourc1	0.152	0.084	0.229
Resourc2	0.358	0.125	0.355
Resourc3	0.256	0.118	0.343
Job1	0.347	0.084	0.256
Job2	0.361	0.094	0.267
Lynx1	0.160	0.045	0.143
Lynx2	0.296	0.056	0.291
Lynx3	0.265	0.125	0.215
Bus1	0.183	0.117	0.254
Bus2	0.216	0.125	0.315
Bus3	0.213	0.105	0.335
Direct1	0.063	0.257	0.468
Direct2	0.150	0.205	0.549
Direct3	0.151	0.173	0.410
Weather1	0.303	0.117	0.339
Weather2	0.146	0.115	0.228
Weather3	0.269	0.123	0.340
Weather4	0.236	0.083	0.288
Temp1	0.180	0.071	0.259
Temp2	0.253	0.100	0.326
Temp3	0.185	0.130	0.257
Temp4	0.220	0.092	0.265
Temp5	0.249	0.164	0.220
Aspirol1	0.475	0.275	0.406
Aspirol2	0.309	0.241	0.370
Aspirol3	0.289	0.163	0.372

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 4-16. Grade 4 IRT item statistics

DOMAIN	NUM	COUNT	SAMPLE	CALIBRTN	ERROR	MNSQ	INFIT	MNSQ	OUTFIT	DISPLACE*
NARRATIVE										
Bird	1	3810	5812	.73A	.03	.9	-5.3	.9	-5.4	.27
	2	4088	5812	.02A	.03	1.2	11.7	1.5	12.3	-.14
	3	2711	5812	1.31A	.03	1.1	7.0	1.1	5.7	-.21
	4	4539	5812	-.73A	.04	1.1	3.1	1.0	-.2	-.37
	5	5399	5812	-1.96A	.05	.9	-3.0	.8	-2.8	
Dog	2	3742	5812	.95A	.03	1.1	3.6	1.1	3.3	.42
	3	4645	5812	-.17A	.03	.9	-6.0	.9	-3.9	.34
	4	4367	5812	.57A	.03	.9	-7.4	.8	-8.1	.68
	5	3850	5812	1.34A	.03	1.1	6.1	1.1	6.3	.90
	6	4861	5812	-.17A	.03	.7	-15.8	.6	-13.9	.61
Shark	1	5035	5812	-1.01A	.04	.8	-7.8	.6	-7.2	.14
	2	4263	5812	-.79A	.04	1.5	20.1	2.0	15.5	-.86
	3	4589	5812	-.71A	.04	1.1	4.7	1.0	.5	-.27
	4	4242	5812	-.37A	.04	1.1	6.3	1.1	2.7	-.37
	5	3689	5812	-.26A	.04	1.4	18.2	1.6	12.8	-.94
Grandpa	1	4006	5812	-.06A	.03	1.1	6.6	1.1	3.6	-.32
	3	3834	5812	.35A	.03	.9	-3.2	.9	-4.4	-.08
	4	4290	5812	.35A	.03	.9	-5.2	.9	-5.0	.41
	5	4318	5812	-.03A	.03	.9	-8.0	.7	-10.0	.08
	6	2674	5812	.64A	.03	1.4	25.1	1.7	23.0	-.96
EXPOSITORY										
Card	1	6076	6325	-1.92A	.05	.6	-11.6	.4	-9.9	.52
	2	6157	6325	-3.04A	.08	1.0	.4	1.2	1.1	
Walrus	1	5879	6325	-1.42A	.04	.7	-12.5	.6	-7.7	.41
	2	5952	6325	-1.46A	.05	.5	-17.7	.4	-12.1	.53
	3	4976	6325	-.43A	.03	.9	-6.0	.8	-6.5	
	4	4995	6325	-.46A	.03	.9	-4.9	.8	-4.8	-.04
	5	4958	6325	-.17A	.03	.9	-6.8	.9	-4.3	.20
	6	3694	6325	.68A	.03	1.0	.0	1.0	.6	-.13
Sand	2	5256	6334	-.83A	.04	1.0	-1.6	.8	-4.9	-.10
	3	5702	6334	-1.04A	.04	.7	-11.4	.5	-11.2	.38
Marmots	1	3398	6334	.63A	.03	1.1	5.2	1.1	2.9	-.44
	2	2574	6334	.88A	.03	1.2	15.1	1.2	12.4	-.85
	3	2467	6334	1.40A	.03	1.0	4.3	1.1	4.5	-.41
	4	2734	6334	1.36A	.03	1.0	.6	1.1	2.9	-.23
Trees	1	4663	6334	.27A	.03	.8	-17.1	.7	-14.0	.31
	2	3079	6334	1.34A	.03	1.0	3.4	1.1	4.5	
	3	3409	6334	1.45A	.03	1.0	-1.9	1.0	.1	.40
	4	2137	6334	1.81A	.03	1.1	4.2	1.1	5.7	-.28
	5	4526	6334	.95A	.03	1.1	-3.6	.9	-3.2	.81
DOCUMENT										
Island	1	5803	6302	-.51A	.04	.6	-17.3	.6	-11.2	.66
	2	6038	6302	-1.64A	.05	.7	-7.7	.7	-4.4	.38
	4	5112	6302	-.05A	.03	1.0	-.2	1.0	1.1	.16
Maria	1	5044	6302	-1.15A	.05	2.1	24.7	2.7	18.7	-1.41
	2	4047	6302	.04A	.03	1.4	22.6	1.7	17.7	-.92
	3	5266	6302	-.75A	.04	1.2	6.4	1.1	2.7	-.37
Bottle	1	6079	6316	-.97A	.04	.4	-16.4	.3	-16.3	.83
	2	4705	631	.73A	.03	.8	-14.7	.7	13.3	.45
	3	6095	6316	-1.55A	.05	.5	-25.9	.4	-10.1	.57
	4	5880	6316	-.36A	.04	.5	-5.9	.4	-18.5	.84
Buses	1	5737	6299	-1.10A	.05	.8	-9.8	.6	-6.9	.09
	2	4398	5262	.54A	.03	.9	-6.7	.8	-8.4	
	3	1712	6305	2.65A	.03	1.2	14.9	.8	-6.8	-.14
	4	2832	6295	.96A	.03	.9	-2.3	1.3	13.1	-.90
Contents	1	5882	6316	-1.42A	.05	.5	-21.7	1.0	-.4	.10
	3	5939	6316	-.76A	.04	1.3	12.2	.4	-15.8	.69
Temperature	1	4868	6316	-.40A	.04	1.0	1.0	1.4	7.9	-.54
	2	2161	6316	2.42A	.03	1.0	-.6	1.2	6.6	.04
	3	3004	6316	1.64A	.03	1.0	2.2	1.0	2.4	-.05
	4	3008	6316	1.50A	.03	1.0	-19.5	1.1	4.2	-.19
	5	5317	6316	.20A	.03	.7		.6	-15.0	.58

*Values close to 0 are left blank.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 4-17. Grade 9 IRT item statistics

DOMAIN	NUM	COUNT	SAMPLE	CALIBRTN	ERROR	MNSQ	INFIT	MNSQ	OUTFIT	DISPLACE*
NARRATIVE										
Fox	2	2915	3162	-1.59A	.07	.9	-2.4	.8	-2.0	.12
	3	2429	3162	.56A	.04	1.0	-.5	.9	-1.5	.66
	4	2923	3162	-.13A	.05	.8	-11.1	.6	-9.0	1.16
	5	2845	3162	-1.28A	.06	1.1	2.7	2.2	8.9	.11
Mute	1	1679	3162	1.22A	.04	1.1	4.6	1.1	3.5	-.07
	2	1811	3162	1.03A	.04	1.0	-.4	1.0	-.4	
	3	2180	3162	-.11A	.05	1.1	4.6	1.2	3.0	-.51
	4	1275	3162	1.88A	.04	1.3	13.2	1.5	12.8	-.12
	5	1948	3162	.59A	.04	1.1	2.5	1.1	3.0	-.22
Shark	2	2971	3173	-2.03A	.08	1.0	.7	.9	-.4	
	3	2952	3173	-1.66A	.07	.8	-4.0	.6	-4.2	.19
	4	2990	3173	-1.64A	.07	.7	-7.2	.6	-3.6	.38
	5	2807	3173	-1.38A	.06	1.1	2.3	1.2	2.2	-.18
Revenge	1	2255	3173	-.64A	.05	1.4	11.1	1.3	3.8	-1.04
	2	1691	3173	.88A	.04	1.2	9.9	1.3	8.6	-.41
	3	2168	3173	-.03A	.05	1.1	4.6	1.1	2.2	-.47
	4	1972	3173	.98A	.04	1.0	-2.4	.9	-1.9	.19
	5	2080	3173	.31A	.04	1.1	3.8	1.1	2.9	-.28
	6	2184	3173	.35A	.04	1.1	-.6	1.0	1.0	
	7	2505	3173	-.49A	.05	1.0	-1.5	.9	-1.5	-.18
Angels	1	1774	3173	1.02A	.04	1.0	-2.5	.9	-2.8	-.12
	2	2194	3173	.35A	.04	.9	-6.9	.7	-7.6	
	3	2193	3173	.49A	.04	.9	-6.4	.8	-6.9	.13
	5	2244	3173	.54A	.04	.9	-7.2	.7	-7 6	.27
	6	2317	3173	.39A	.04	.8	-9.5	.7	-8.9	.27
	7	2317	3173	.40A	.04	.9	-6.0	.8	-5.9	.28
E.POSITORY										
Marmots	1	2656	3177	-.60A	.05	.9	-4.4	.9	-2.4	.26
	2	2488	3177	-1.05A	.05	1.4	10.4	1.6	6.6	-.69
	3	2116	3177	-.18A	.05	1.4	15.0	1.8	14.7	-.50
	4	2593	3177	-.41A	.05	.9	-5.2	.8	-3.1	.28
Laser	1	2920	3177	-1.55A	.06	.8	-4.7	.8	-2.3	.23
	2	2557	3177	-.17A	.05	.8	-11.1	.6	-9.1	.41
	3	2507	3177	-.95A	.05	1.3	7.8	1.3	3.7	-.50
	4	2010	3177	1.01A	.04	1.0	.3	1.0	.7	.52
	5	1986	3177	.64A	.04	1.0	1.4	1.0	.8	.12
	6	2547	3177	-.97A	.04	1.1	2.3	.8	-2.7	-.41
	7	1796	3177	.79A	.05	1.0	1.1	1.1	1.8	-.05
	9	2202	3177	.39A	.04	1.0	-2.3	.9	-1.7	.25
	10	2263	3177	-.19A	.04	1.0	1.1	.9	-3.0	-.21
Paracutin	1	2193	3192	.22A	.05	1.1	-2.3	1.1	2.6	.05
	2	2653	3192	-1.03A	.04	1.2	-2.2	1.6	6.8	-.22
	3	2861	3192	-1.04A	.05	.8	3.0	.7	-5.0	.39
	5	2310	3192	.53A	.05	1.1	5.6	1.1	3.5	.55
	6	2297	3192	.03A	.04	1.0	-6.8	1.0	-.3	.05
Smoke	1	1772	3192	.95A	.04	1.0	2.9	1.0	1.5	.05
	2	1482	3192	1.33A	.04	1.0	-.1	1.0	.5	-0.5
	3	1524	3192	.97A	.04	.9	1.5	.9	-2.8	-.34
	4	2109	3192	-.10A	.04	1.2	-1.2	1.2	4.2	-.45
	5	2345	3192	-.06A	.04	.9	-3.3	.8	-6.3	.06
	6	1341	3192	1.43A	.04	1.0	7.0	1.0	.9	-.19
DOCUMENT										
Card	1	3254	3308	-.71A	.06	.3	-21.8	.2	-15.7	1.08
	3	3254	3308	-1.81A	.09	.4	-10.0	.6	-3.5	.71
	4	3148	3308	-1.72A	.08	1.1	.9	1.6	3.8	
	5	2662	3307	-.18A	.05	1.2	5.9	1.4	6.1	-.25
	6	2010	3310	.71A	.04	1.3	14.0	1.5	12.5	-.55
	7	3137	3309	-1.49A	.08	1.0	-.7	1.2	1.4	.13
Resources	1	2480	3301	-.81A	.06	2.3	22.6	2.9	16.5	-1.79
	2	2542	3306	1.19A	.04	.8	-11.2	.8	-9.6	.79
	3	1854	3305	2.23A	.04	1.1	6.7	1.2	5.8	.79
Job	1	2602	3310	.07A	.05	1.1	2.6	1.0	.7	-.12
	2	2856	3310	-.36A	.05	1.0	-1.0	.9	-1.4	.08
Lynx	1	2714	3310	-.23A	.05	1.2	6.3	1.5	7.5	-.17
	2	1568	3310	1.93A	.04	1.0	1.0	1.1	2.9	.04
	3	1878	3310	1.55A	.04	1.1	4.3	1.1	4.5	.12
Bus Schedule	1	1961	3322	.14A	.05	1.8	25.7	2.2	19.7	-1.44
	2	2329	3322	.66A	.04	1.0	.0	1.0	-.5	-.06
	3	1838	3321	1.25A	.04	1.0	2.7	1.0	2.7	-.25
Directions	1	2744	3296	-.32A	.05	1.1	2.1	.9	-1.5	-.21
	2	2742	3321	.06A	.05	.8	-7.6	.6	-8.4	.17
	3	2366	3321	.66A	.04	.9	-3.2	.9	-2.8	
Weather	1	2948	3320	-1.35A	.07	1.6	8.4	1.3	2.8	-.82
	2	2669	3312	.36A	.04	.9	-2.5	1.0	-.3	.29
	3	2851	3322	-.80A	.06	1.2	5.0	1.1	1.1	-.45
	4	1719	3322	1.19A	.04	1.1	6.7	1.2	5.8	-.49
Temperature	1	3016	3322	-.95A	.06	1.0	.5	1.0	.4	
	2	2685	3322	.53A	.04	.8	-9.7	.8	-6.9	.46
	3	2565	3322	.46A	.04	1.0	-1.6	1.0	-.1	.18
	4	2722	3322	-.03A	.05	1.0	.0	1.0	.2	
	5	3146	3322	-1.34A	.07	.8	-4.1	.8	-2.5	.25
Aspirol	1	3179	3322	-1.94A	.09	1.0	-.4	.6	-3.2	
	2	2988	3322	-.76A	.06	.9	-2.7	.8	-3.5	.09
	3	1895	3322	1.82A	.04	.9	-4.3	.9	-3.9-	.41

*Values close to 0 are left blank.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

regression coefficients are typically around 2 (see Chapter 1 of this volume). The design effects for the Rasch fit statistics have not been estimated directly. For this report, we used a value of 6 to approximate the design effects of the Rasch fit statistics.

Therefore, corrections that account for these two attributes would yield adjusted fit statistics of 1.23 and 2.65.[4] The first would typically be considered of an acceptable magnitude, while the second would not. Further examination of Table 4-16 for grade 4 reveals that after taking into consideration the sampling attributes, two items on the narrative scale and six items on the document scale would have fit statistics that are considered high (i.e., adjusted fit greater than 2.0). For grade 9 (Table 4-17), three items on the document scale are considered high. Overall, however, it is reasonable to conclude that the data seem to adequately fit the one-parameter Rasch model.

4.6 Summary and Conclusions

In this chapter evidence regarding the dimensionality of the IEA Reading Literacy Test scores were presented and discussed. The central analytical theme revolved around testing the unidimensionality assumption for the U.S. item-response data within each reading literacy domain. Additionally, the underlying structure of the reading literacy domain was investigated. The aim here was to determine to what extent the data supported the hypothesized structure of three domains.

Realizing that in practice the assumption of unidimensionality will always be violated to some degree, the unidimensionality of the IEA Reading Literacy Study response data were investigated. The central question driving the investigation was, Is there a dominant factor explaining responses to the IEA Reading Literacy Test items? or, in other words, Can we ascertain the degree to which the IEA Reading Literacy item responses depart from unidimensionality? Evidence relating to a variety of methods was presented for assessing the dimensionality of reading literacy item response data. The evidence presented indicated overwhelmingly that overall the assumption of the unidimensionality was met for each reading literacy domain. However, for both grades, the narrative domain departed less from unidimensionality than the other two domains. The extent of departure from unidimensionality for the document domain was the highest for both grades.

Turning our attention to the second major question of the study (Can we ascertain the underlying structure of the IEA Reading Literacy item-response data?), overall the evidence indicated that the data did not support the three hypothesized domains. In particular, the distinction between the narrative and expository domains was not supported by the data. For both populations, the data, however, supported two substantive factors (i.e., narrative-expository and document) and a third factor that seemed to be related to "speed."

The emergence of a speed factor in both grades may have far reaching implications for test design and interpretation. The question, How can we interpret student performance on items appearing at the end of a testing session? does seem to warrant further investigation. When summary scores (e.g., IRT

[4] To account for the above mentioned sampling attributes, we wished to transform the fit statistics so that the actual sample size would function as if it were a simple random sample of 400 students. To do this, we assumed that the fit statistic is inversely proportional to the square root of the sample size and is directly proportional to the square root of the design effect, which for this example was assumed to be 6 and 8 for grades 4 and 9, respectively.

For example, to transform the observed fit statistic of 17.5. we performed the following calculations:

$$\textit{Adjusted Fit} = \textit{Observed Fit} \times (\sqrt{400/6000/6}$$

$$= (11.7)(0.2582)/(2.449)$$
$$= 1.23$$

scale scores) are used to repɔrt student performance, the issue cf speed may not be that important. However, educators increasingly are becoming interested in obtaining richer descriptions of student performance on subtests, thereby bringing the issue of speed into more prominence: if items on a particular subtest are located toward the end of a testing session, how could we report student performance on that subtest given that the location of these items in the test may have adversely affected student performance on these items?

The issue of speed also has implication for establishing proficiency levels—another test interpretation activity gaining more popularity among educators. The methods to establish proficiency levels heavily rely on item performance. Thus, if student performance on items appearing at the end of a testing session is adversely affected, the application of these methods becomes problematic. If resources permit, by counterbalancing the order of test items across different forms, we may be able to obtain unbiased estimates of student performance on subtests and test items.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Armor, D.J. (1974). Theta reliability and factor scaling. In H.L. Costner (ed.), *Sociological methodology*, (17-50). San Francisco, CA: Jossey-Bass.

Bock, R.D., and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters. Application of an EM algorithm. *Psychometrika*, 46, 443-459.

Bock, R.D., Gibbons, R.D., and Muraki, E. (1985). Full-information factor analysis. *Applied Psychological Measurement*, 12, 261-280.

Carlson, J.E., and Jirele. (1993). Dimensionality of NAEP scales that incorporate polytomously scored items. Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Carmines, E.G., and Zeller, R.A. (1979). *Reliability and validity assessment.* Beverly Hills, CA: Sage.

Carrol, J.B. (1945). The effect of difficulty and classic success on correlation between items or between tests. *Psychometrika*, 10, 1-19.

Cook, L.L., and Eignor, D.R. (1984). Assessing the dimensionality of NAEP reading test items: Confirmatory factor analysis of item parcel data. Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.

Cronbach, L.J. (1989). Construct validation after thirty years. In R.L. Linn (ed.), *Intelligence: Measurement, theory, and public policy -- Proceedings of a symposium in honor of Lloyd G. Humphreys*, 147-171. Chicago: University of Illinois Press.

Divgi, D.R. (1980). Dimensionality of binary items: Use of mixed model. Paper presented at the annual meeting of the National Council on Measurement in Education, Boston.

Drasgow, F., and Parsons, C.K. (1983). Application of unidimensional items response theory models to multidimensional data. *Applied Psychological Measurement*, 7, 189-199.

Green, S.B., Lissitz, R.W., and Mulirk, S.A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement*, 37, 827-838.

Hattie, J.A. (1984). An empirical study of various indices for determining unidimensionality. *Multivariate Behavioral Research*, 19, 49-78.

Hattie, J.A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9, 139-164.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Reading, MA: Addison-Wesley.

Lumsden, J. (1961). The construction of unidimensional tests. *Psychological Bulletin*, 58, 122-133.

Lumsden, J. (1957). A factorial approach to unidimensionality. *Australian Journal of Psychology*, 9, 105-111.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer and H.I. Braun (eds.), *Test validity*, 33-45. Hillsdale, NJ: Erlbaum.

McNemar, Q. (1964). Opinion-attitude methodology. *Psychological Bulletin*, 43, 289-347.

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests. Results and implications. *Journal of Educational Statistics*, 4, 207-230.

Wilson, D.T., Wooc, R., and Gibbons, R. (1991). *TESTFACT*. Chicago, IL: Scientific Software.

Wright, B.D., and Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 29, 23-48.

Wright, B.D., and Stone, M.H. (1979). *Best test design: Rasch measurement*. Chicago: MESA Press.

Zwick, R. (1987). Assessment of dimensionality of year 15 read g data. In A.E. Beaton (ed.). *The NAEP 1983-84 technical report*, 245-284. Princeton, NJ: Educational Testing Service.

5 Exploring the Possibilities of Constructed-Response Items

Barbara Kapinus and Nadir Atash

The purpose of this chapter is to discuss some of the issues surrounding the use of constructed-response items as they compare to multiple-choice items in large-scale assessments. In the current climate of assessment in the United States, we are witnessing a strong abandonment of the multiple-choice format in favor of the constructed-response format. However, it is our position that both have a purpose and a place when they are well developed, well conceived, and properly used. However, while there is a great deal of research to guide the development and use of multiple-choice items, there is not nearly so much support when using constructed-response items. Through the examples drawn from our experiences with the IEA Reading Literacy Study, with additional explorations, and with very small-scale observations and interviews of students, we wish to demonstrate some of the necessary criteria to consider when developing and scoring constructed-response items.

5.1 Background

The IEA Reading Literacy Tests were developed collaboratively by the IEA International Steering Committee (ISC) and the National Research Coordinators (NRCs). Given the diversity of the group of people involved in designing the reading tests, it is natural to expect disagreements about the form of the assessment instruments. In an effort to enhance coverage of the reading literacy assessment domains (thereby favorably affecting consequential validity of the tests), a group of individuals involved in developing the tests, including the U.S. NRC, argued for inclusion of some constructed-response items in the IEA Reading Literacy Tests. These items, however, were included on an exploratory basis (i.e., constructed-response items would not be included in the scaling). The inclusion of the constructed-response items reflected evolving theory in both reading and assessment in the United States.

> *The scientific paradigm that undergirds standardized and virtually all criterion-referenced tests, which has been in the process of breakdown for the last two decades, has reached a critical stage. Standardized and criterion-referenced tests, rooted in an anachronistic paradigm, are a major barrier to the restructuring of the nation's schools. As we enter the last decade of the twentieth century, it is becoming apparent, at least to those outside the testing and measurement establishment, that the assumptions intrinsic to the technology of standardized and most criterion-referenced tests are untenable. Out of the ashes of this paradigm, from the many varied and imperfect efforts underway to solve the practical problems of assessing educational achievement,*

is slowly emerging a new paradigm, one based on a set of foundational assumptions that are in sharp contrast to those that underlay the current paradigm (Berlak et al. 1992).

This quote captures the context in the United States within which the items for the IEA in reading were forged. An effort was made to incorporate newer theories of reading and to move beyond present assessment paradigms, as much as possible, given the need for international consensus on the assessment. To that end, several open-ended items were incorporated into the assessment. This section sets out some of the background in both reading and assessment that supported the use of open-ended items in the United States.

5.2 Reading Theory

Since the 1970s, theories of reading in the United States have moved from a behaviorist orientation to a cognitive view. The behaviorist model regarded reading as a set of sequential skills acquired through drill and practice on small pieces of text (e.g., letters and words). The more recent cognitive or schema theoretic perspe.'ive views reading as a complex cognitive activity wherein readers use knowledge about language, text, and the context of the reading situation and personal background knowledge of the topic to build an understanding of the text through a complex interactive process. Meaning is "constructed" using information from the text, information brought to the reading, and cues from the context in which reading takes place. This shift in theory is attributable to the work of a number of people including Chomsky and Fillmore, who were linguists; Bransford and Franks, who were cognitive researchers; and Rummelhart, Stein, Glenn, Thorndyke, Meyer, Halliday, and Hasan, who examined text structures (Pearson 1984). This cognitive model of reading has been elaborated and extended through research, including that of Anderson (1984), Anderson and Pearson (1984), Collins, Brown, and Larkin (1979), and Paris, Lipson, and Wixson (1983). Some of the most recent models of reading emphasize readers' responses to text and view the proficient reader as capable of responding to text in a variety of rich and complex ways, from forming a general understanding to responding from personal or critical perspectives (Rosenblatt 1938, 1978; Langer 1990a, 1990b).

As a result of the changes in theory, educators in the United States are increasingly coming to regard reading as a process of "developing" meaning from text, not simply "getting the meaning." The reading framework for the 1992 National Assessment of Educational Progress (NAEP) states that "reading for meaning involves a dynamic, complex interaction among three elements: The reader, the text, and the context" (National Assessment Governing Board 1992, 10). The NAEP framework asserts that good readers cannot merely "form an understanding of what they read," but they also "extend, elaborate, and critically judge its meaning" (p. 9). Several states, including Maryland, Pennsylvania, and Michigan, use similar definitions of reading to guide the development of curriculum and assessments.

While there are many views of reading in the United States—and some of them are the sources of heated debate—the prevailing view of reading has moved away from the notion of small, discrete subskills that must be mastered in a hierarchical manner to "get" the meaning that resides in the text. The present concept of reading is that of a contextually driven, purposeful, strategic use of skills and background knowledge in order to construct an understanding of and response to the text. It is essentially a generative process rather than a passive receiving of meaning. In order to assess students' proficiency in reading as it is defined here, it is necessary to provide students with some opportunities to generate responses that indicate an ability to integrate their own individual background knowledge with text ideas to construct actively understandings and responses that are possibly divergent.

5.3 Problems with Traditional Large-Scale Reading Assessments

This generative aspect of reading is difficult to capture in multiple-choice items, even those items that require complex thinking. At best, multiple-choice items allow students to identify an understanding, inference, or judgment that someone else has generated. If we are to assess the ability of students to build their own understandings of the text and spontaneously extend and support those understandings, it would seem to be a good idea to explore the use of constructed-response items.

Perhaps more importantly, the question of instructional relevance needs to be taken into consideration. In the past, the use of large-scale assessment tools, consisting solely of multiple-choice items, has led to undesirable instructional practice such as spending unreasonable amounts of time on teaching how to respond to multiple-choice items or emphasis on specific, decontextualized skills in meaningless drills.

In response to these abuses of assessment and the frustration that educators experience when student achievement measures disregard many of the proficiencies developed and emphasized in classrooms, there is a growing movement in the United States to make assessment more congruent with good instruction (Linn, Baker, and Dunbar 1991; Valencia, McGinley, and Pearson 1990) and real-world tasks (Wiggins 1992).

Given that assessment has come to influence instruction, and thus the form of a large-scale assessment can have an effect on the form and substance of instruction (Mislevy 1991), the exclusion of important student behaviors may have undesirable curricular implications. This has led to a careful consideration of the consequential validity of tests in the United States. Messick (1988) asserts that if an assessment causes educators to adopt more effective approaches to instruction, it is considered to demonstrate positive consequential validity. The heavy reliance on multiple-choice items, with the consequence of excluding important student behaviors from the assessment domain, has raised concerns about the consequential validity of many current, large-scale reading assessments.

5.4 Issues Associated with the Use of Constructed-Response Reading Items

We believe that the use of constructed-response items on large-scale reading assessments would appear to tap the reading process more completely and to reflect good instruction and real-life reading responses more faithfully than relying completely on multiple-choice items. However, constructed-response items, partly because they have seldom been used, are associated with several issues. These include reliability of scoring, possible bias for certain cultural minorities, and potential contamination of the reading construct assessed by a reliance on writing fluency. In addition, there is a question of just what constructed-response items are tapping. While we suspect that constructed-response items assess different aspects of reading from multiple-choice items, we do not have sufficient information about the processes that students use in responding to constructed-response items to be sure exactly what we are tapping with those items.

Both the pilot test and the main study allowed members of the U.S. Steering Committee to conduct a series of explorations to better understand how constructed-response items, as compared to multiple-choice items, work in assessing reading and to consider some of the issues related to the use and scoring of constructed-response items. The explorations were guided by the following questions:

1. What types of information do constructed-response items provide that is not evident from multiple-choice items?

2. What types of processes and strategies do readers seem to be using in responding to multiple-choice items as compared to constructed-response items?

3. How does the use of different scoring guides affect scores and the information gathered from constructed-response items?

4. To what extent is writing a confounding factor in the scores of constructed-response items?

5. What is the relationship between scores on multiple-choice test items and scores on constructed-response items?

6. What can be said about the psychometric qualities, (e.g., reliability) of the constructed-response items?

The overall question that these explorations inform is whether constructed-response items are worth the time required to answer them and the time and money required for scoring.

Exploration 1: What types of information do constructed-response items provide that is not evident from multiple-choice items?

The first exploration arose from the need to understand how to create an equivalent U.S. test item to one that was proposed internationally. The item was associated with an excerpt from the children's book *Pippi Longstocking*. The multiple-choice item under consideration was

"What kind of person was Pippi?"

a. Kind to strangers
b. Shy with adults
c. Cheeky
d. Cooperative

With respect to this item and its related passage, the following concerns were raised:

- The correct response (i.e., Cheeky), may not be familiar to the U.S. students.

- The passage may be more familiar to certain types of students who may have read the book, *Pippi Longstocking*, or who may have seen the movie.

- Due to cultural factors, as well as comprehension problems, certain types of students in the U.S. may fail to capture the humor in the passage that was essential in understanding what kind of person Pippi was.

For the pilot test, four forms of each test had been developed. These forms were intended to be spiraled within each classroom selected. The *Pippi* passage appeared on all forms, so we could alter one form and still have the requested number of responses for the International Coordinating Center (ICC). In the U.S., about 1,080 grade 4 students responded to one of the four forms (about 270 students responded to each test form.) On one of the four pilot test forms, the question was presented as an open-ended item. The intention was that by comparing the 270 students' responses to the open-ended item to their responses to the multiple-choice items, the aforementioned concerns could be addressed. Responses to the open-ended form of the item were analyzed to address the following issues:

1. Did the student show evidence of having read the passage or is the response based on previous knowledge (e.g., student had seen the movie prior to the assessment)?

2. Was the student aware of the humor in the passage?

3. Did the student's response capture the notion of "cheeky"?

4. Was the student's response similar to any of the distractors on the multiple-choice form of the item?

5. What were some plausible answers other than "cheeky"?

In addition, to obtain more indepth information about strategies and processes student use to respond to these items, a followup study was conducted with 14 grade 4 students who were asked to read the passage from *Pippi Longstocking* and answer the question about Pippi either in open-ended or multiple-choice format. Students were then asked to describe how they figured out the answer to the question (see the protocol in Appendix 1 to this chapter).

This exploration addressed the question of whether constructed-response items provide information about students' reading that cannot be obtained from multiple-choice items. Of the 242 responses to the open-ended version of the item that were scorable or complete, 200 clearly indicated that the students had not previously read the entire story. Only six responses clearly indicated that the students had read the entire story, either by saying so or by including information from the story not in the passage. Additionally, about 30 responses gave some evidence that the student might have read the story before, but the evidence was not conclusive. These responses usually included information that might have come from the original story or from inferences that the student had made. Six responses simply did not give any indication of whether the student had or had not read the story previously.

In examining whether students would spontaneously describe Pippi in terms that were equivalent to cheeky, the study found that few students offered a response that captured the idea of cheeky. Of the 242 responses, about 28 responses meant something similar to cheeky, for example, "smarty" and "brat." Only one of the responses was similar to one of the distracters on the multiple-choice version of the item, that is, kind. Other plausible responses included "adventurous," "confusing," "weird," "funny," "liar," "playful," "independent," "silly," "orphan," and "pest." With respect to the question of students' awareness of the humor in the passage, only 47 of the responses gave clear evidence of awareness of the humor in the passage.

These patterns in students' responses led to hypotheses that some students were using information from having read the story previously to answer the item, and indeed it was difficult for students to answer the item in its open-ended form based solely on the passage. In addition, it seemed that students might be able to answer the multiple-choice form of the item correctly simply by eliminating the distracters--even though they did not know what cheeky meant. Consequently, a followup interview of 14 students was conducted and the following observations were made:

- Several of the students indicated that they used information other than that presented in the passage in order to answer the item, for example, remembering the book or the movie.

- Some of the students who were given the multiple-choice form of the item chose cheeky for the answer, but when they were asked what cheeky meant, they could not give a definition or an example. These students indicated that they had chosen the response because the other responses did not make sense.

- Observations of students' behaviors in choosing the correct response suggested that students answering the multiple-choice item were doing more interacting with the item and distracters than with the passage. That is, they were not looking back at the passage but rather focusing on the test item.

Conclusions

The findings in this exploration suggest that both multiple-choice items and constructed-response items can be of limited use when the passage is highly familiar to some of the students. In addition, the excerpt did not really provide enough information for students who had not previously read the story of Pippi to be able to provide a reasonable answer to the test item.

The responses to the open-ended version of the item gave indications that some students lacked comprehension of the humor in the passage or they were using information from outside the passage that the multiple-choice form of the item did not provide. Indeed, it was the responses to the open-ended version of the item that showed the problems with the passage that supported the ultimate decision not to use that passage in the actual study.

The interviews of students about their responses to the item in open-ended or multiple-choice format suggested that students engaged in a process of eliminating distracters on the multiple-choice form of the item, as compared to thinking about the passage on the open-ended version.

Exploration 2: Are the processes used in answering a constructed-response item different from those used in answering a multiple-choice item?

The second exploration grew out of a need to consider whether students were indeed using different processes for answering the open-ended items than the multiple-choice items. The following questions guided the exploration:

- Do students employ different strategies when answering open-ended items as compared to multiple-choice items?

- Do multiple-choice and open-ended items both promote students' reinspection of the passage in order to provide a response?

- Do students seem aware of the strategies they use in answering the items and can they articulate them?

Thirty-eight grade 4 students from two suburban classrooms and 36 grade 9 students from four classes in a suburban high school were asked to read one of the two passages that contained an open-ended item on the IEA Reading Literacy Test (main study), and then to answer the open-ended item. The two grade 4 passages were *grandpa* and *walrus* and the open-ended items were "Why did the parents decide to ask Grandpa back to the table?" and "What problem would the walrus have if it lost its eye teeth?" The grade 9 passages were *a shark makes friends* and *a woman learns to read*, and the open-ended items were "If the writer had to make this story longer, what do you think the pilchard would do next?" and "What do you think would be the disadvantages in your country for an adult who could not read or write?"

For this exploration, these items were presented in either an open-ended or a multiple-choice format. Each student received one passage with an open-ended item and one passage with a multiple-choice item. The order for passage and item type was counterbalanced. Each student met individually with one of the two researchers. The students were asked to read the first passage and write the answer to the first item. When they had answered the item, they were asked how they had determined the answer and whether they had looked back at the passage in order ιo respond. The process was then repeated for the second passage.

This exploration provided the following observations related to the question of possible differences in the processes students used for answering open-ended and multiple-choice items. First, on the basis of student responses, it may be concluded that for grade 4 students the question related to the *grandpa* passage was harder than the question on the *walrus* passage. Only one student had the wrong answer to the multiple-choice item on the *walrus* passage, while five students had wrong answers to the multiple-choice item on the *grandpa* passage. Thus, format alone might not account for differences in student response strategies. Table 5-1 presents the distribution of scores on the open-ended items for each of the two passages. As shown on the *grandpa* passage, one student received a score of "4," no one scored "3," 13 scored "2," and three students scored "1." On the *walrus* passage, 16 students scored "4," two scored "3," one scored "2," and no one scored below "2."

Table 5-1. Distribution of scores for two grade 4 passages: Number of students with score

Passage	Score				
	1	2	3	4	Total
Grandpa	3 (17.7)	13 (76.5)	0 (0.0)	1 (5.9)	17 (100)
Walrus	0 (0.0)	1 (5.3)	2 (10.5)	16 (84.2)	19 (100)

NOTE: Numbers in parentheses are percentages. Percentages may not add to 100 due to rounding.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

At the grade 9 level, all students selected the correct response to the multiple-choice items. Table 5-2 shows the distribution of scores on the open-ended items for each of the two passages for grade 9 students. As shown, no student received a score below "2" on the open-ended items; 12 students received a score of "4" on the open-ended item related to *shark*, as compared to 9 students for the passage *literacy*.

Table 5-2. Distribution of scores for two grade 9 passage: Number of students with score

Passage	Score				
	1	2	3	4	Total
Literacy	0 (0.0)	1 (6.7)	5 (33.3)	9 (60.0)	15 (100)
Shark	0 (0.0)	1 (5.3)	6 (31.6)	12 (63.2)	19 (100)

NOTE: Numbers in parentheses are percentages. Percentages may not add to 100 due to rounding.

SOURCE: IEA Reading Literacy Study, U.S. National Study Data, National Center for Education Statistics, 1991.

In their overall reporting of strategies for answering both multiple-choice and open-ended questions, grade 4 students tended to talk about the story or the information in the passage rather than any specific strategy behaviors. Grade 9 students, on the other hand, reported using a process of elimination of distracters as their approach for answering the multiple-choice item on the expository passage *literacy*. This pattern did not hold for the passage *shark*, where no students reported focusing on the distracters in order to answer the multiple-choice item.

For both grade 4 and grade 9 students, few students reported looking back at the passage in order to answer the multiple-choice items. For grade 4 students, four and five students reported looking back at the passage in order to answer the *walrus* and *grandpa* items, respectively. For grade 9, only two students reported looking back at the passage in the process of answering the multiple-choice items.

For both grade 4 and grade 9 students, there was far more looking back at the passage in order to answer the open-ended questions. For grade 4 students, 17 and 11 students reported looking back at the passage in order to answer the open-ended questions for the *walrus* and *grandpa* passages, respectively. For grade 9 students, eight and seven students reported looking back at the passage in order to answer the open-ended items for *shark* and *literacy*, respectively.

The evidence collected in this exploration reveals the following findings.

- In response to multiple-choice items, students generally did not look back at the passage.

- In response to the open-enaed items, although a large proportion of students did not look back at the passage to respond to the items, far more students reported looking back at the passage than did so for the multiple-choice items.

- When asked how they determined the answers to the items, students, especially in grade 4, generally talked about the content of the passage rather than strategies such as looking back, thinking about the passage, or drawing conclusions.

- When answering open-ended items, students had a tendency to interact with the passage rather than the item, while they interacted more with the item and distracters when answering multiple-choice items.

An interesting observation in this exploration was that some students who reported looking back at the passage actually were not observed to do so. When a researcher finally asked one of the students, "Did you really look back at the story?" the student replied that she "looked back in my mind." This prompted the researchers to consider addressing this strategy more directly in a subsequent exploration. It did suggest that a passage might be revisited mentally if not through an actual physical review.

Conclusions

Constructed-response items seem to elicit the types of reading behaviors that are valued in the current theories of reading and reading instruction. Students tended to look back and think back on the passage more in answering the open-ended form of the questions than in responding to the multiple-choice format. This reinforces the notion and instructional practice that promotes the purpose of reading as building one's own understanding of the text rather than using the text to guess someone else's answer to a question.

This exploration also demonstrated that open-ended items, as well as multiple-choice items, can be too easy, providing little information about students' reading proficiency. However, when open-ended items do not give much information, the cost is far higher both in the time students spend on the items and in the time and effort required to score those items as compared to that for multiple-choice items.

Exploration 3: To what extent can scoring guides determine the types of information about students' reading provided by open-ended items?

The third exploration compared two different versions of the scoring guides to determine how the differences in scoring guides supported the observation and reporting of rich information and certain attributes of reading. The original scoring guides (Exhibit 5-1), based on the IEA recommendations for scoring the open-ended items, gave credit for plausible answers that were not necessarily passage based. In addition, most of the open-ended items were being scored either right or wrong with no partial credit or levels of correctness. On the basis of such scoring guides, the open-ended items may not be offering much more information about students' performance than the multiple-choice items. Furthermore, in the cases of text-independent correct responses, these items may provide information that is not valid (i.e., student response is not based on an interaction of the student and the passage and thus is not a demonstration of reading proficiency).

Exhibit 5-1. Original scoring guides for the passages for *grandpa* and *blue whale*

A. Why did the parents ask grandpa back to the table?

9 = no response

1 = gives an unacceptable response
gives response that does not include reason for parents' change in attitude

2 = gives an acceptable response
"They realized they had been selfish."
"They were embarrassed after watching their son."
"They put themselves in his place and realized how hard it was for him."
"They learned from their son's activity what it could be like to be an old person."

B. What might be some ways scientists could study blue whales?

9 = no response

1 = unacceptable response
"There are fish in my school."

2 = gives one way
"Follow them around."
"Capture a blue whale."

3 = gives two or more distinctive ways; gives one way with some elaboration.
"Scientists could put radios on whales and then follow them around."

SOURCE: IEA Reading Literacy Study. U.S. National Study data, National Center for Education Statistics, 1991.

The new refined scoring guides were developed similar to the 1992 NAEP progress scoring guides. First, a generic rubric was developed (see Appendix 2 to this chapter) and specific item scoring guides were developed for each passage (Exhibit 5-2 shows the refined scoring guide for the passage *blue*

Exhibit 5-2. Refined scoring guide for *blue whale*

New Scoring Guide for Blue Whale

General Scoring Rubric

What might be some ways scientists could study blue whales?

In answering this question students should make use of the information about the whales' size making this task difficult and should also include information from beyond the text.

Scoring Rationale

The question requires readers to build on the notion that studying blue whales is difficult because they are so large and containing them is difficult. A complete response would require readers to use their own background knowledge together with the above notion from the text. Responses should be plausible in the light of what the passage says about blue whales. References to other specific text information about blue whales such as what they eat or where they live are important components of extensive responses to this question.

9 = No response

0 = Off Task "They are interesting."

1 = Unsatisfactory Responses that are unrelated to the actual question, incomplete, or incorrect. "Look closely at them and check them a lot." "Their weight, width, if they are harmful."

2 = Partial Responses that show an incomplete understanding of the passage by offering suggestions that contradict passage information that tells that the size of blue whales limits the possibilities of putting the animals in a cage. Partial responses also include suggestions that are not really related to passage information. "Dissect one and look at its insides." "Keep them in custody so they can observe the whales and their instincts."

3 = Essential These responses indicate a comprehension of information in the passage related to the question by offering suggestions related to the blue whales' size or habits. While these answers are correct, they are not really elaborated or supported. "Make a specific place to store them, or go out into the ocean and study them in their natural environment." "They could lure the whales into a small bay, and close it off for a while, and they could study them there."

4 = Extensive These responses indicate an extended understanding of the passage by offering suggestions that are clearly linked to text information and are supported or explained. These responses contain information that is not only related to the text but is internally consistent. "Some ways scientists could study blue whales are going out to the ocean and observe and monitor them for several days at a time. They could also put trackers on them and find them every year." "Scientists could study them in their own habitat by diving down in wet suits. Maybe, if possible, scientists could build a huge aquarium fit for a whale to study."

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

whale). Four open-ended items in the pilot test were scored on the basis of the new scoring guides (new rubric). A comparison of students' scores based on the original and the refined scoring guides addressed the question, "how does the use of different scoring guides affect scores and information gathered from constructed-response items?"

Figures 5-1 to 5-4 contain the comparisons of students' scores on the two guides for four open-ended items from four different passages. Figure 5-1 depicts the distribution of scores (based on the original scoring guide) for each category of the scores (based on the refined scoring guide) for the open-ended item on the *grandpa* passage. All students receiving a score of "0" (i.e., lowest score possible) based on the refined scoring guide had received a score of "1" (lowest score possible) based on the original scoring guide. (It should be pointed out that the meaning of "0" and "1" may differ relative to the two scoring guides.) All students who received a score of "4" (i.e., highest score possible) based on the refined scoring guide also had received a score of "2" (highest score possible) based on the original scoring guide. However, not all the students receiving a "2" based on the original guide received a "4" based on the refined guide. A similar pattern of relationships can be observed for the other three open-ended questions from the other passages.

As shown in the figures, the refined scoring guides seem to be more stringent in what is demanded as adequate indication of comprehension (i.e., fewer students received the maximum score under the refined guide). Based on the refined scoring guides, the discrimination among students' responses at the upper end of the score distribution seems to have been expanded. Certainly the wording of the refined guides focuses potential reporting on a greater range of aspects of reading.

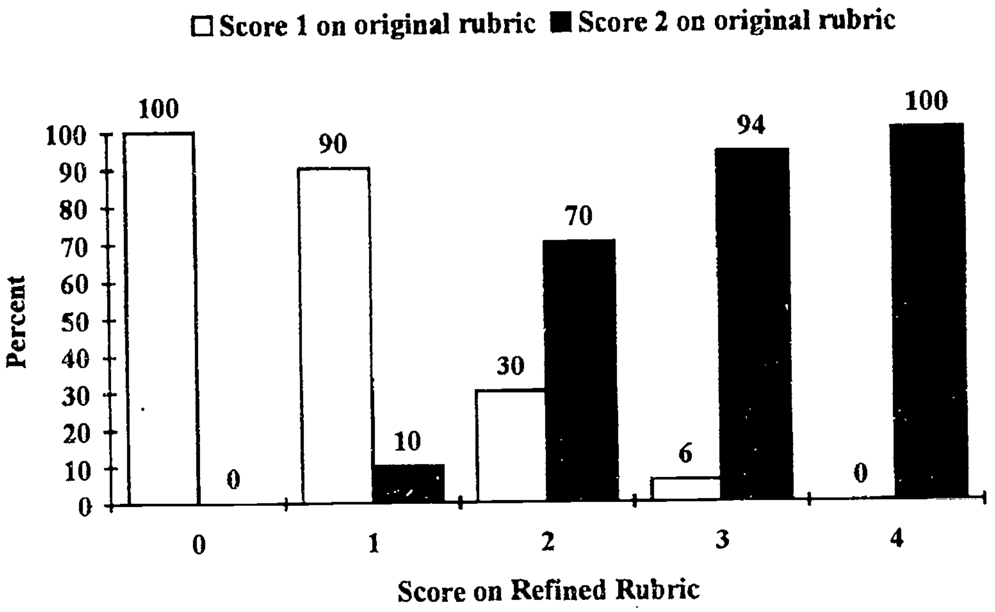
Conclusions

The comparisons of the two scoring guides and students' scores on those guides suggest that scores can vary depending on the degree to which text-based information is demanded, and that demanding text-based information increases the potential to score the desired abilities. In addition, it seems that care must be taken in developing scoring guides in order to assure that they are truly focused on reading behaviors and not on background knowledge. Finally, scoring guides must be constructed to capture the potential richness of responses to open-ended items. This suggests more than a three-point full-credit, partial-credit, no-credit scoring guide.

Exploration 4: Are open-ended items biased in favor of those who can write fluently?

Reading experts have raised the question of whether the responses to open-ended reading items measure writing fluency or reading comprehension. Using random samples from the pilot study, 36 student responses from grade 4 and 34 student responses from grade 9 were analyzed as a preliminary step. The number of words written by each student in response to an open-ended item were counted and recorded. The relationship between the score on the open-ended item and the length of response was determined. For the main study, this exploration was repeated using a larger sample size. Specifically, a random sample of 365 grade 4 students and 389 grade 9 students was analyzed using procedures similar to the one used with the pilot test. Three alternative approaches were used to determine the relationship between the length of response and the score on the open-ended item. First, the correlation between the two variables was computed. Second, the pattern of numbers was examined to determine whether some high score responses had very few words and, conversely, whether some low score responses were extremely long. Third, box-whisker plots were constructed to examine the distribution of length of responses by score categories on the open-ended items.

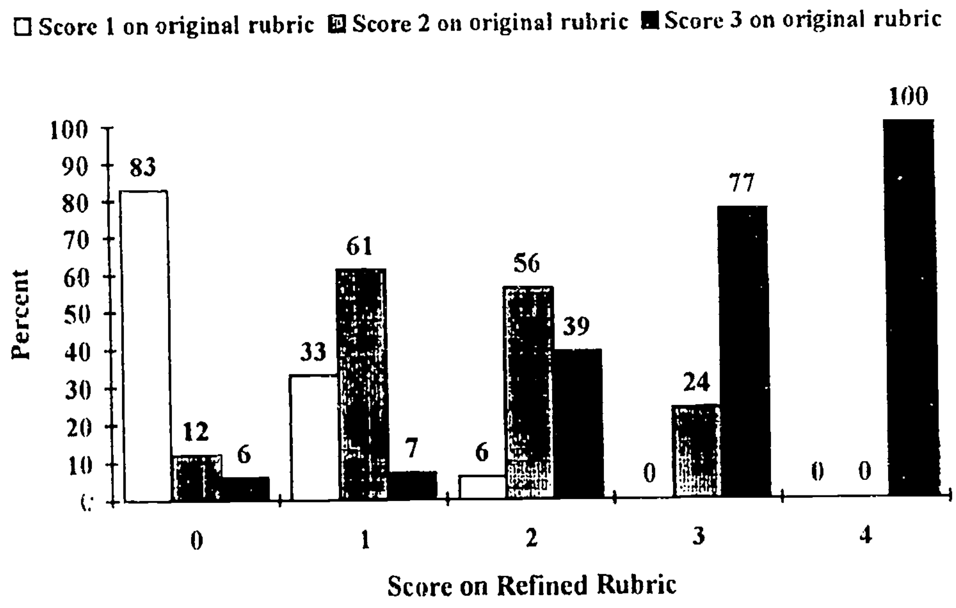
Figure 5-1. Comparison of refined and original scoring guides, grade 4 passage *grandpa*



NOTE: Figure shows that among students with a score of 1 on the refined rubic, 90 percent had a score of 1, and 10 percent had a score of 2 on the original rubic.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

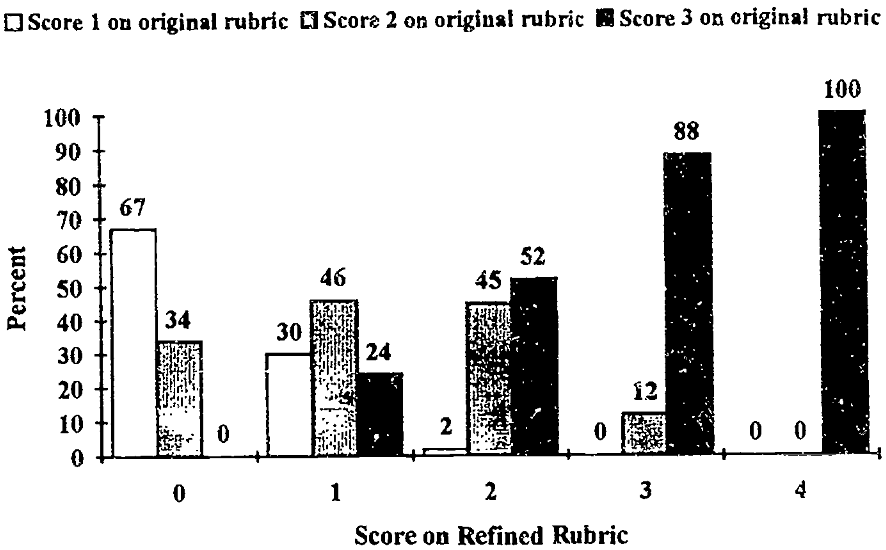
Figure 5-2. Comparison of refined and original scoring guides, grade 4 passage *blue whale*



NOTE: Figure shows that among students with a score of 0 on the refined rubic, 83 percent had a score of 1, 12 percent had a score of 2, and 6 percent had a score of 3 on the original rubic. Percentages may not add to 100 due to rounding.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

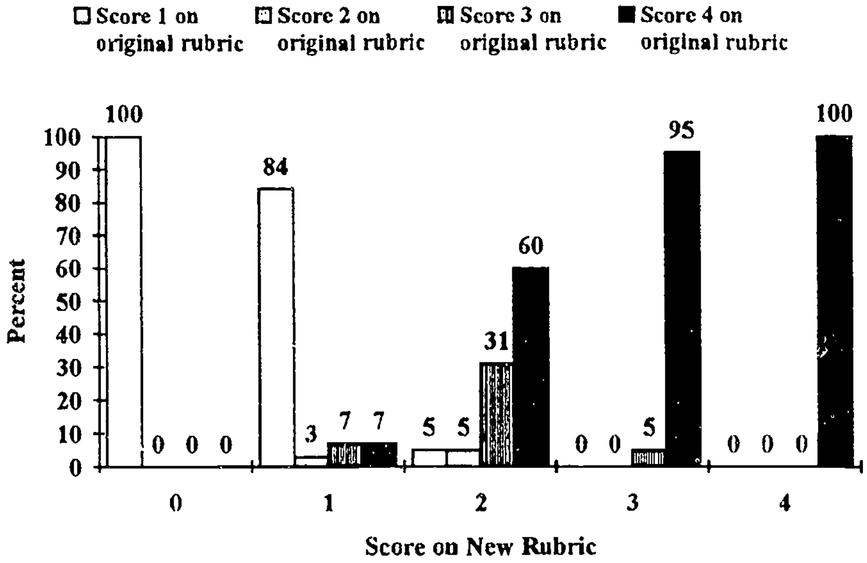
Figure 5-3. Comparison of refined and original scoring guides, grade 9 passage *whale*



NOTE: Figure shows that among students with a score of 1 on the refined rubic, 30 percent had a score of 1, 46 percent had a score of 2, and 24 percent had a score of 3 on the original rubic. Percentages may not add to 100 due to rounding.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Figure 5-4. Comparison of refined and original rubrics, grade 9 passage *lynx*



NOTE: Figure shows that among students with a score of 1 on the refined rubic, 84 percent had a score of 1, 3 percent had a score of 2, 7 percent had a score of 3, and the remaining 7 percent had a score of 4 on the original rubic. Percentages may not add to 100 due to rounding.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991

Table 5-3 presents the correlation between the length of response and the score on the open-ended items on all the analyses. With the exception of the grade 9 pilot test, these correlations are significantly different from 0. However, at best, only about 38 percent of the variation among lengths of responses and scores on open-ended item is in common.

Table 5-3. Correlation between response length and the response score on the open-ended items

Population	N	Correlation	P
Grade 4			
Pilot test passage	36	0.554	0.0001
Main test passage #1	405	0.614	0.0001
Main test passage #2	405	0.545	0.0001
Grade 9			
Pilot test passage	34	0.203	0.2494
Main test passage #1	365	0.588	0.0001
Main test passage #2	389	0.459	0.0001

N = sample size; P = statistical significance of correlation.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

The patterns of responses were examined to determine whether some high scoring responses had very few words and, conversely, whether some low scoring responses were extremely long. Tables 5-4 to 5-9 present cross-tabulations of response length by score on the open-ended items. For these tables, the categories of response length were defined as follows: 1 = 19 or fewer words written; 2 = 20 to 40 words written; and 3 = more than 40 words written.

Table 5-4. Relationship between response length and response score for grade 4 pilot test passage

Score	Response length category			
	1-19 words	20-40 words	41+ words	Total
1	3	0	0	3
2	6	6	2	14
3	2	0	0	2
4	2	7	8	17
Total	13	13	10	36

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 5-5. Relationship between response length and response score for grade 4 main study passage #1

Score	Response length category			
	1-19 words	20-40 words	41+ words	Total
1	21	6	2	29
2	54	28	4	86
3	50	67	10	127
4	8	66	66	140
Total	133	167	82	382

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 5-6. Relationship between response length and response score for grade 4 main study passage #2

Score	Response length category			
	1-19 words	20-40 words	41+ words	Total
0	32	0	0	32
1	107	29	7	143
2	53	22	3	78
3	39	35	7	81
4	8	24	16	48
Total	239	110	33	382

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 5-7. Relationship between response length and response score for grade 9 pilot test passage

Score	Response length category			
	1-19 words	20-40 words	41+ words	Total
1	0	0	0	0
2	0	0	2	2
3	3	6	2	11
4	0	6	15	21
Total	3	12	19	34

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 5-8. Relationship between response length and response score for grade 9 main study passage #1

Score	Response length category			
	1-19 words	20-40 words	41+ words	Total
0	16	1	0	17
1	6	3	1	10
2	18	57	32	107
3	16	52	85	153
4	0	13	65	78
Total	56	126	183	365

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 5-9. Relationship between response length and response score for grade 9 main study passage #2

Score	Response length category			
	1-19 words	20-40 words	41+ words	Total
0	9	2	0	11
1	15	4	1	20
2	26	18	12	56
3	54	75	11	140
4	13	90	59	162
Total	117	189	83	389

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

As these tables indicate, in general, longer responses tend to have higher scores. However, not all long responses have received high scores. For example, for the grade 4 passage 1 (main study) two students who had written more than 40 words had scored only "1" on the open-ended item. On the other hand, eight students who had written less than 20 words received a score of "4" (the highest possible score).

The same pattern of relationships was observed on the basis of the "box and whisker" plots showing the distribution of length of responses by score categories of the open-ended items (Figures 5-5 to 5-8). These figures indicate, in general, that as the mean number of words increases, the score on the open-ended items also increases. However, there seems to be a large overlap in the distribution of number of words written, particularly between adjacent categories of open-ended scores.

The following are examples of responses that reflect exceptions to the relationship between the length of the responses and the score.

Why did the parents decide to ask Grandpa back to the table?

Response A: *They let him eat at the table because the boy made him a cheap wooden bowl. After they heard that he was making a bowl for Grandpa. They felt sorry and started crying. After that they let him eat at the table.*

This response is long, but the answer is simply incorrect. The response received a score of "2."

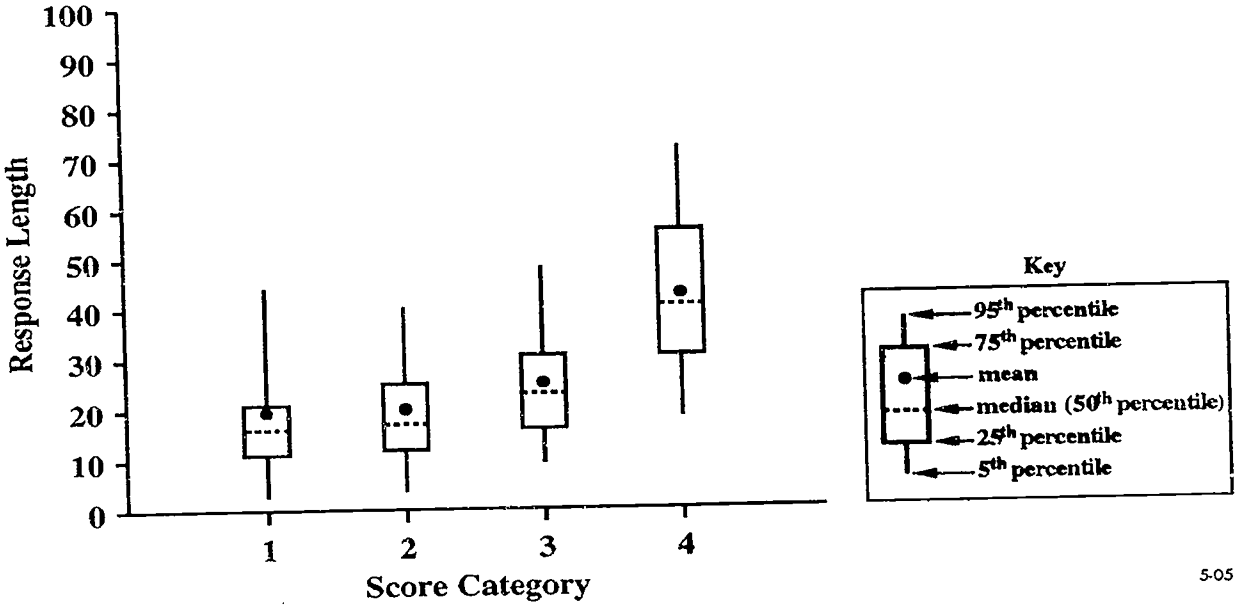
Response B: *When the father and mother get older they are going to shake a lot too.*

This response, while brief, captures an important inferred connection and received a score of "4."

Conclusions

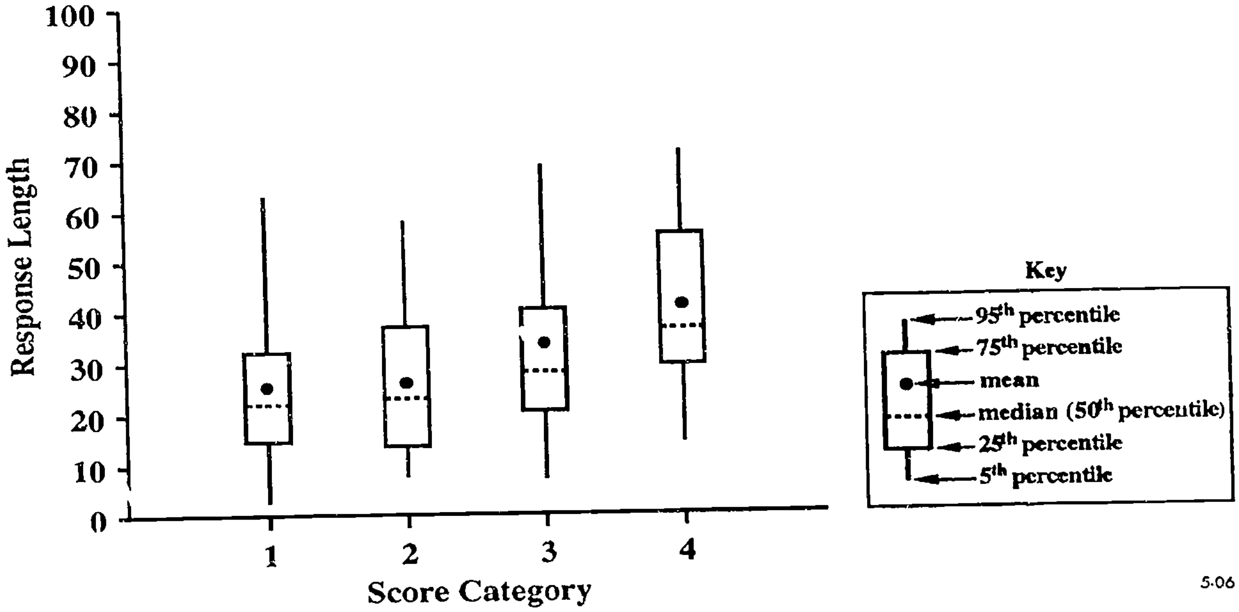
This exploration examined the relationship between writing fluency and performance on the open-ended items. The findings indicate that there is indeed a strong relationship between sheer quantity of writing on the responses and the scores. This is not surprising for two reasons. First, there is a relationship between reading and writing; good writers are often good readers (Tierney and Shanahan 1991). An understanding and command of written language is necessary for achievement in both areas. In addition, it is only reasonable to believe that students who offer more information are more likely to have complete, thorough responses, although the low scores on some lengthy responses demonstrated that wordiness alone did not guarantee success. What was also interesting, however, was the occurrence of high scores on brief responses. Students sometimes demonstrated clear understanding and sound interpretation in a very few words. The careful construction of scoring guides is one way of working to limit the confounding of reading performance with writing fluency.

Figure 5-5. Distribution of response length, by category of open-ended score for grade 4 main study passage #1



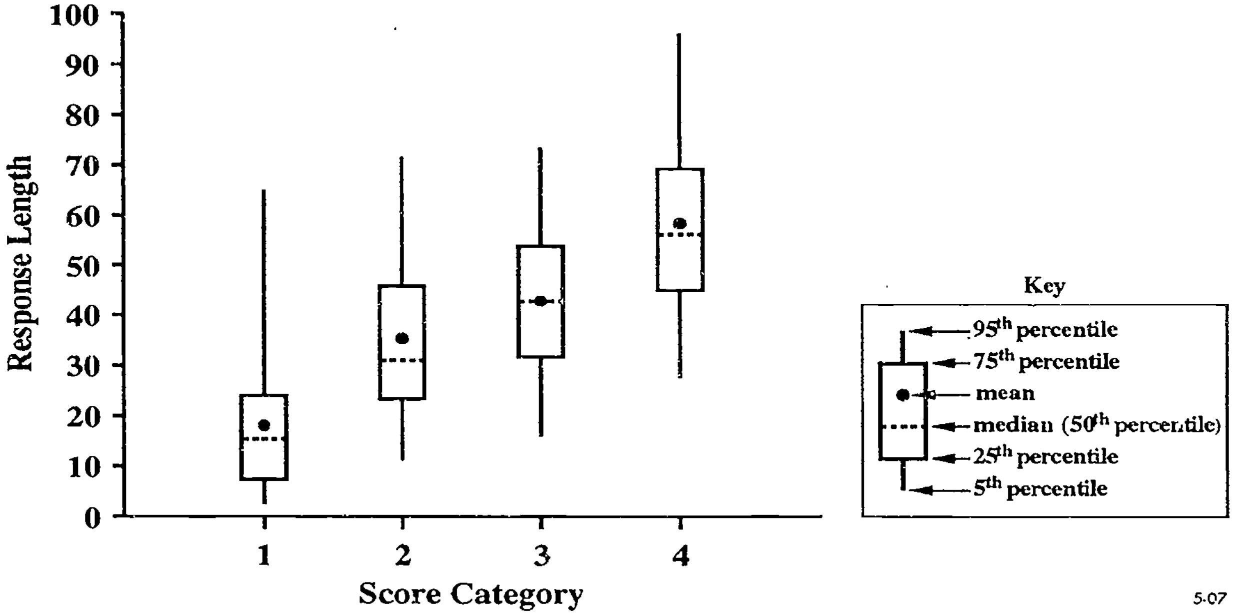
SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Figure 5-6. Distribution of response length, by category of open-ended score for grade 4 main study passage #2



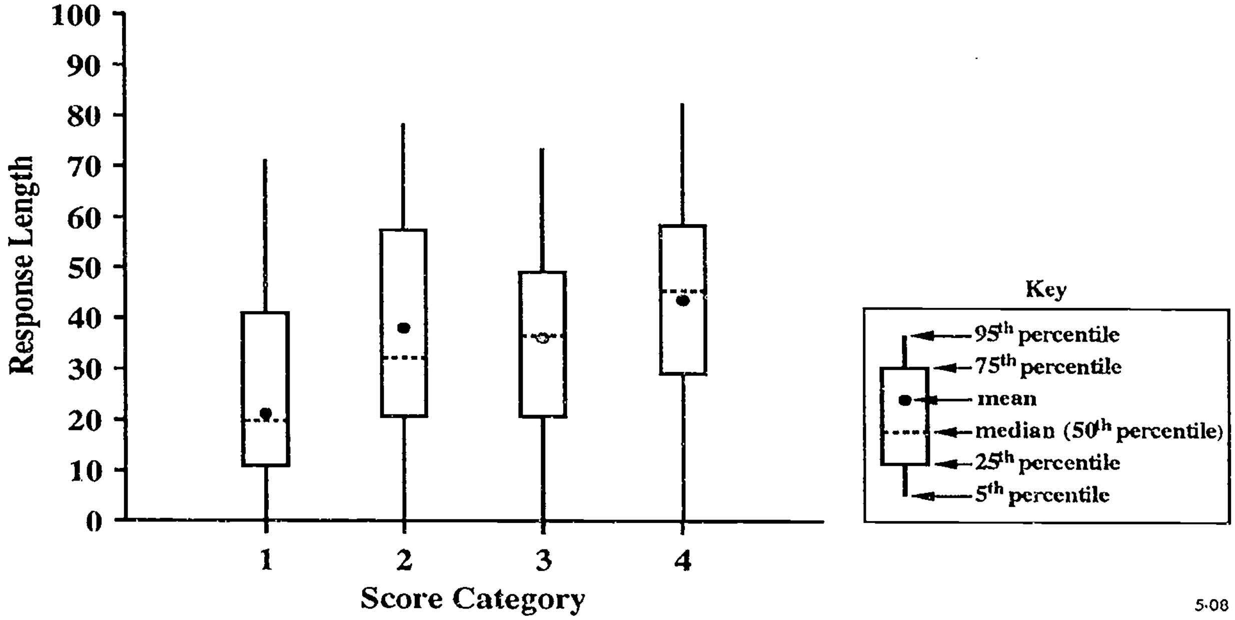
SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Figure 5-7. Distribution of response length, by category of open-ended score for grade 9 main study passage #1



SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Figure 5-8. Distribution of response length, by category of open-ended score for grade 9 main study passage #2



SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Exploration 5: What is the relationship between scores on multiple-choice test items and scores on constructed-response items?

Open-ended items are more expensive than multiple-choice items, both in terms of the time students must take to respond to them and the time and cost of scoring. In order to justify the extra expenditures, it is necessary to understand what additional information, if any, we find out by using the open-ended items as opposed to multiple-choice items.

Using the one-parameter IRT (Rasch) methodology, scale scores were obtained for each student in the three reading literacy domains. Since each student's score on the multiple-choice items was also available (both for the pilot and main studies), the relationship between scores on open-ended items and scales scores (derived from multiple-choice test items) could easily be studied. Two alternative approaches were used to determine the relationship between scores on open-ended items and scale scores. First, the correlation between scale scores and scores on open-ended items was computed. Second, a table that shows the mean and standard deviation of scale scores within each category of score on an open-ended item was constructed.

Table 5-10 presents mean scale scores for each IEA Reading Literacy Study domain by ratings of constructed responses. As this table indicates, the mean scale scores for students with high scores on the constructed-response items are substantially higher than the mean scale scores for students with low scores on the constructed-response items. The patterns of increase in mean scales scores are similar across the two populations.

Table 5-10. Mean scale scores and standard deviations, by ratings of constructed-response items

Test passage	Item rating	Number of students	Narrative		Expository		Document	
			Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Grade 4								
Walrus	2	435	466.1	85.7	474.2	61.7	489.2	68.3
	3	1,266	510.7	85.6	505.1	66.6	521.5	69.3
	4	2,139	566.7	85.7	548.7	68.8	559.8	74.7
	5	2,032	603.7	78.9	581.5	68.9	583.7	76.8
Grandpa	2	2,220	502.8	82.6	509.1	66.3	521.0	71.2
	3	1,265	574.7	76.9	553.5	69.0	561.1	77.7
	4	1,334	615.1	69.5	581.3	66.1	585.3	72.9
	5	746	634.2	66.4	597.5	69.3	598.0	73.0
Grade 9								
Literacy	2	10	415.5	46.2	404.8	59.3	400.9	64.5
	3	120	448.5	74.7	445.7	90.2	447.9	73.8
	4	968	523.7	87.7	524.6	94.0	519.1	75.1
	5	1,847	576.6	89.9	579.9	99.7	553.2	77.5
Shark	1	156	452.5	94.7	449.0	94.6	468.6	81.0
	2	573	508.1	87.2	504.2	91.3	507.5	75.1
	3	1,089	549.3	91.6	547.9	101.7	532.8	78.0
	4	1,387	569.4	92.6	576.4	101.5	552.5	77.9

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

To summarize the strength of the relationship between reading literacy scale scores and students' responses to the open-ended items, the coefficient of determination (R^2) was computed (Table 5-11). As shown, the proportion of variance in reading literacy scale scores explained by students' responses to the open-ended items ranges from 14 percent to 33 percent for grade 4, and from 8 percent to 13 percent for grade 9. Further, for both populations the proportion of variance in document scales accounted for by the criterion variable is lower than the corresponding number for the other two domains. Across all domains, the proportion of variance accounted for by the open-ended items is larger for grade 4 than grade 9, although it should be kept in mind that the passages were not the same across the two grades.

Table 5-11. Coefficient of determination (R^2) between reading literacy scale scores and students' ratings of constructed responses to two open-ended items

Domain	Grade 4		Grade 9	
	Walrus	Grandpa	Literacy	Shark
Narrative	0.211	0.330	0.128	0.104
Expository	0.206	0.214	0.118	0.109
Document	0.136	0.151	0.102	0.77

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Conclusions

The relationship between scores of multiple-choice items and the constructed-response items were correlated as indicated by the data in Table 5-10. However, the proportion of variance in reading scale scores explained by the constructed-response items was low. These findings suggest that the two types of items measure different but related aspects of the domain of reading. The exploration indicates that constructed-response items do tap aspects of reading achievement that are not assessed by multiple-choice items. This finding supports the notion that the use of a combination of constructed-response and multiple-choice items might give a more complete sampling of the domain than multiple-choice items alone.

Exploration 6: What can be said about the psychometric qualities of the constructed-response items?

The objective of the use of the constructed-response items on the IEA was not to use those items alone as estimates of students' reading proficiency, but to study how those items worked as compared to the multiple-choice items and whether they contributed additional information on student reading literacy to the overall study. Each student in both populations responded to only two constructed-response items that certainly by themselves may not yield reliable estimates of reading achievement. Furthermore, each of the two constructed-response items pertained to a different type of text (i.e., one narrative and one expository), and therefore scores on the two items could not really be combined to o .tain a more reliable estimate of students' reading literacy abilities. However, a consideration of the usefulness of the constructed-response items must address the reliability of those items. Although we did not obtain a direct estimate of the reliability of the constructed-responses used in the IEA Reading Literacy Study, we did estimate a lower-bound reliability as follows:

- First, we estimated the correlation between the multiple-choice items and constructed-response items under the condition that the two measures were perfectly reliable. Using the divergent and convèrgent validity studies as our guide, we estimated that a correlation of 0.707 (i.e., 50 percent common variance) between the multiple-choice items and constructed-response items would be reasonable provided both variables (i.e., scores on multiple-choice and constructed-response items) were measured without any error (i.e., reliability = 1.0).

- Next, using the "correction for attenuation" formula, we estimated the reliability of constructed-response items.[1] Table 5-12 presents the observed correlation between the multiple-choice and constructed-response items, reliability for the multiple-choice items, as well as the estimated reliability for the constructed items.

Table 5-12 shows that the estimated reliability for the constructed-response items for grade 4 is substantially higher than the estimated reliabilities for grade 9. The difference in estimated reliabilities could be due to one or more of the following factors: (1) the type of responses elicited by the items; (2) student motivation to respond to the constructed-ıesponse items; and (3) scoring guides used to score the student responses.

Table 5-12. Correlation between multiple-choice and constructed-response items, reliability of multiple-choice items, and estimated reliability of constructed-response items

Test passage	Correlation with multiple-choice items	Reliability of multiple-choice items	Estimated reliability of constructed-response items
Grade 4			
Walrus	0.459	0.766	0.550
Grandpa	0.574	0.857	0.769
Grade 9			
Literacy	0.344	0.846	0.280
Shark	0.322	0.875	0.237

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Conclusions

The reliabilities for scores on the constructed-response items varied across grade levels and items. On the constructed-response item for the fourth grade *grandpa* passage, the estimated reliability was close to that for the multiple-choice items on both of the fourth grade passages with constructed-response items. The reliability for the other fourth grade constructed-response item on the *walrus* passage was lower but not unreasonable. The reliabilities for the ninth grade items was very low, but the motivation of these students to write a careful, thorough response could be questioned. The scoring guides and the items themselves might have been problematic, also. This was an initial, limited effort to use constructed-response items on this assessment. In other large-scale assessments, reliabilities of constructed-response item score have been substantially higher. For example, on the 1992 NAEP, the overall percentages of agreement between readers of items were 89 percent for grade 4, 85 percent for grade 8, and 88 percent for grade 12 (Mullis, Campbell, and Farstrup 1993). The differences in reliability of scores on the two

[1] The reliability of the constructed-response items were estimated using the correction for attenuation formula:

$$r'_{xy} = \frac{r_{xy}}{\sqrt{r_{xx}}\sqrt{r_{yy}}}$$

where r'_{xy} is the corrected (i.e., unattenuated) correlation between variables x and y; r_{xy} is the observed correlation between variables x and y; r_{xx} is the reliability of variable x; and r_{yy} is the reliability of variable y. To estimate the reliability of constructed-response items (i.e., r_{yy}) we assumed $r'_{xy} = .707$. For example, the observed correlation between the constructed-response items and the multiple-choice items in the *Grandpa* passage for grade 4 was 0.574. Thus,

$$\sqrt{r'_{yy}} = \frac{r_{xy}}{r'_{xy}\sqrt{r_{xx}}} = \frac{0.574}{0.707\sqrt{0.857}} = 0.877$$

$$r'_{yy} = 0.769$$

assessments might reflect the differences in development and use of constructed-response items for the assessments. Given additional time and attention, it is likely that reliabilities for constructed response items on IEA would increase.

5.5 Discussion

Although the studies summarized here represent only some initial explorations of the processes and strategies students use when answering constructed-response items, they do begin to provide some insights related to the questions posed at the beginning of this chapter—insights that should be considered when deciding whether to use constructed-response items on large-scale assessments and when developing those items.

Question 1: What types of information do constructed-response items provide that is not evident from multiple-choice items?

The findings of Exploration 1 suggest that students can and do answer multiple-choice items by interacting with the item and distracters more than with the passage. Responses to the open-ended form of the item gave insight into students' ability to go beyond a literal understanding of the passage, for example, perceiving the humor. While it was not clear that students interacted more with the passage on the open-ended items than on the multiple-choice items, it was evident that students in some cases were drawing on experiences outside of the assessment situation and the stimulus passage to answer the item.

Although it was not part of the original question, one finding was strong evidence suggesting problems with using excerpts of potentially very familiar text on a reading assessment. Students' responses to the passage sometimes reflected recall of a previous reading of the book or the movie rather than the comprehension of the specific passage used on the assessment.

Question 2: What types of processes and strategies do readers seem to be using in responding to multiple-choice items as compared to constructed-response items?

The findings of Exploration 2 suggest that open-ended items promote more interaction with the passage, at least in the form of looking back at it. The findings also imply that students, especially grade 4 students, are not very strategic about answering either multiple-choice or open-ended items. The high percentage of correct responses for both groups of students implies that the items were extremely easy and might not have elicited much in the way of strategic approaches to responding.

The implication for developing open-ended items is that those items should be thought provoking and should ask only questions that could not be posed just as effectively and perhaps more efficiently in a multiple-choice format. In addition, open-ended questions might be framed to demand that students revisit the passage to find support for their inferences and conclusions. For example, the item, "If the writer were to make this story longer, what do you think the pilchard would do next?" could have an additional direction, "Support your opinion with evidence from the story."

Question 3: How does the use of different scoring guides affect scores and the information gathered from constructed-response items?

The findings of Exploration 3 indicate that different scoring guides can emphasize various aspects of an answer and consequently affect scores. If guides are not constructed carefully, scores might not reflect a measure of the construct. The early guides often gave students credit for a plausible answer even if there was no indication that the student actually read the passage. This was, in the case of the question, "What do you think would be the disadvantages in your country for an adult who could not read or write?" also a function of the question itself. Questions must demand text based responses if they are to be scored for evidence of building an understanding of the text itself.

In addition, if guides do not explicitly address important aspects of reading, such as supporting inferences, there can be no reporting on those aspects in assessment results. For example, if the relation of text information to background knowledge is an important aspect of reading, it should be explicitly addressed in scoring responses to reading.

Question 4: To what extent is writing a confounding factor in the scores of constructed-response items?

The analyses conducted for Exploration 4 indicated that while a strong relationship existed between the length of responses and scores, that relationship was not always present. As students are asked more frequently in their classroom work to construct their own written responses to reading questions, it is likely that they will become more fluent both in their writing and in their ability to respond to open-ended questions. That fluency can eliminate some of the concern about the writing contamination of reading scores.

Question 5: What is the relationship between scores on multiple-choice test items and scores on constructed-response items?

The analyses conducted for Exploration 5 to ascertain the relationship between scores on multiple-choice test items and scores on constructed-response items showed that while there is a significant relationship between the two variables, nevertheless, based on coefficient of determination, the variance in common between the two variables was at best 33 percent. While some of the variation not common between the two measures (i.e., unique variation) may be due to measurement error (i.e., measurement error tends to attenuate the true relationship between the two measures), it appears that the two variables are measuring different aspects of reading proficiency.

Question 6: What can be said about the psychometric qualities (e.g., reliability) of the constructed-response items?

The finding of Exploration 6 indicated that the estimated reliability of the constructed-response items was lower than the corresponding reliabilities for the multiple-choice test items. This is not surprising given the substantial difference in the number of items between the constructed-response and multiple-choice items. Because the objective of the use of the constructed-response items for the IEA Reading Literacy Study was not to use those items as estimates of students' reading proficiency, but to study how those items worked as compared to the multiple-choice items and whether they contributed additional information on student reading literacy to the overall study, the relatively low reliabilities for the constructed-response items do not affect the interpretation of study results.

It should be pointed out that the potential sources of systematic and random error are much greater in administering and scoring constructed-response items than the multiple-choice test items. One major source of error relates to the consistency of scoring the constructed-response items. In particular, ambiguities in the scoring guide may add random error to the scores of the constructed-response items. Thus, it is imperative to develop scoring guides that are appropriate for the type of student responses elicited by the item.

The design of the present study did not allow estimation of various components of the error model for scoring and analyzing constructed-response items. Further investigations are needed (e.g., generalizability studies) to ascertain the error components of the scores based on constructed-response items under various conditions.

Summary

The overall question that all of these explorations address is whether constructed-response items offer sufficient benefits to justify the added expenditure of time and money necessary both to answer them and to score them. The explorations indicate that students use different strategies in responding to the constructed-response items. Sometimes those strategies involve more direct interaction with the text than multiple-choice items that sometimes require only elimination of distracters.

While there remain concerns about reliable scoring and the possible contamination of reading scores by writing proficiency, the explorations indicate that these are not factors that should inhibit the use of constructed-response items. It is possible that as we become better at framing as well as scoring constructed-response items, the reliability will increase. As students become more accustomed to responding to open-ended items, their writing proficiency will likely be less important.

The ultimate concern, especially in the United States, is that reading assessments tap, as closely as possible, the constructive processes of building understandings of text. Multiple-choice items, no matter how well crafted, have limited capacity for tapping this generative aspect of reading that is being increasingly emphasized in our educational goals, our curriculum, our instruction, and the other testing conducted in this country.

References

Anderson, R.C. (1984). Role of the reader's schema in comprehension, learning, and memory. In R.C. Anderson, J. Osborn, and R.J. Tierney (eds.), *Learning to read in American schools.* Hillsdale, NJ: Erlbaum.

Anderson, R.C., and Pearson, P.D. (1984). A schema-theoretic view of basic processes in reading comprehension. In P.D. Pearson (ed.), *Handbook of reading research*, 255-292. New York: Longman.

Berlak, H., Newmann, F.M., Adams, E., Archbald, D.A., Burgess, T., Raven, J., and Romberg, T.A. (1992). *Toward a new science of educational testing and assessment.* Albany, NY: State University of New York Press.

Collins, A., Brown, J.S., and Larkin, K.M. (1979). Inference in text understanding. In R.J. Spiro, B.C. Bruce, and W.F. Brewer (eds.), *Theoretical issues in reading comprehension.* Hillsdale, NJ: Erlbaum.

Langer, J.A. (1990a). The process of understanding: Reading for literary and informative purposes. *Research in the Teaching of English*, 24, 229-260.

Langer, J.A. (1990b). Understanding literature. *Language Arts*, 67, 812-816.

Linn, R.L., Baker, E.L., and Dunbar, S.B. (1991). Complex performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20, 8: 15-21.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer and N.I. Braun (eds.), *Test validity*, 33-45. Hillsdale, NJ: Erlbaum.

Mislevy, R.J. (1991). *A framework for studying differences between multiple-choice and free response test items.* Research Report. Princeton, NJ: Educational Testing Service.

Mullis, I.V.S., Campbell, J.R.., and Farstrup, A.E. (1993). *NAEP 1992 reading report card for the nation and the states.* Washington, DC: U.S. Department of Education, National Center for Education Statistics.

National Assessment Governing Board (1992). *Reading framework for the 1992 National Assessment of Educational Progress.* Washington, DC: U.S. Government Printing Office.

Paris, S., Lipson, M., and Wixson, K. (1983). Becoming a strategic reader. *Contemporary Educational Psychology*, 8, 293-316.

Pearson, P.D. (1984). Twenty years of research in reading comprehension. In T.E. Raphael (ed.), *The contexts of school-based literacy*, 43-62. New York: Random House.

Rosenblatt, L. (1983). *Literature as exploration.* 4th Ed. New York: Modern Language Association.

Rosenblatt, L. (1978). *The reader, the text, and the poem.* Cambridge, MA: Harvard University Press.

Tierney, R.J., and Shanahan, T. (1991). Research on the reading-writing relationship: Interactions, transactions, and outcomes. In R. Barr, M.L. Kamil, P. Mosenthal, and P.D. Pearson (eds.), *Handbook of reading research.* Vol. 2, 246-280. New York: Longman.

Valencia, S.W., McGinley, W., and Pearson, P.D. (1990). Assessing reading and writing. In G.G. Duffy (ed.), *Reading in the middle school*, 124-146. Newark, DE: International Reading Association.

Wiggins, G. (1992). Creating tests worth taking. *Educational Leadership*, 49, 8: 26-33.

Appendix 1

Protocol for Exploration One Interview

I am trying to find out more about how students read and answer questions about their reading. The information that I find will help me make better reading tests and even better reading lessons.

I am going to ask you to read a short passage and to answer one question about the passage.

When you finish, I will ask you some questions about how you read and answered the question.

Here is the passage. It comes from a longer book. I'd like you to read it carefully and answer the question at the end. Take your time and let me know when you are finished.

Finish

1. Now I'd like you to tell me how you figured out the answer to the question.
   - Prompt 1:Did you look back at the story?
   - Prompt 2:What else did you think about when you answered the question?
2. Have you ever read or heard the story of Pippi? Seen the movie? Did you think about what you read, heard, or saw when you answered the question?
3. Is the question like the questions you usually answer about reading? If not, how is it different?

Thank you very much for your help. You gave me a lot of good ideas that will be a big help to me.

Appendix 2

IEA Generic Rubric for Open-Ended Questions

9	No Response	
0	Off Task	Responses not related to the question.
1	Unsatisfactory	These responses indicate miscomprehension of the question or the passage. They often contain incomplete, incorrect, or fragmented information.
2	Partial	These responses demonstrate only some comprehension. They give information on only one part or aspect of a question or do not anchor the response in the text. When elaboration is required, these responses only give text information.
3	Essential	These responses demonstrate adequate comprehension. Although they contain essential information, either there are few specific references to the text or there is little elaboration.
4	Extensive	These responses demonstrate rich comprehension. They contain complete, relevant information that is internally consistent and related. They also contain specific references to the text and, where called for, elaboration based on background knowledge.

6 Interpreting the IEA Reading Literacy Scales

Irwin S. Kirsch and Peter B. Mosenthal

6.1 Introduction

It is widely recognized that an inevitable degree of uncertainty exists when attempting to transform research into action plans. This uncertainty creates a necessary gap between policy research and policy formation. It also raises questions about a suitable role for policy research. In a classic volume published some 40 years ago, Lerner and Lasswell (1951) argued that the appropriate role for policy research is not to define policy but rather to establish a body of evidence from which informed judgments can be made.

Some 35 years later, Messick (1987, 158) refined and extended this thinking to the role of large-scale assessments as a form of policy research. He noted that if "large-scale assessments are to function effectively as policy research — to provide empirically grounded interpretations to inform policy judgment — a number of key features must be exhibited." Among these are *relevance*, or the provision for measuring diverse background and program information to illuminate context effects and treatment or process differences; *comparability*, or the capacity to provide data or measures that are commensurable across time periods and across subpopulations of interest; and *interpretability*, or the ability to provide evidence that will enhance the understanding and interpretation of what is being measured.

Through the use of Item Response Theory (IRT), as well as other sampling and design procedures, significant contributions have been made in the use of large-scale assessments in terms of the relevance and comparability criteria that Messick identified (Beaton 1987, 1988; Mislevy 1985, 1991). While efforts have been made to address the interpretability issue in large-scale assessments (Carroll 1993), these attempts have generally lagged behind the more fully developed psychometric and methodological procedures.

Recently, to inform the discussions on international competitiveness and educational achievement, the IEA conducted an international assessment of reading literacy. The purpose of this assessment was to profile the literacy abilities of representative samples of 9-year-old and 14-year-old students. To address the issue of relevance, extensive student and teacher background questionnaires were developed and administered in each participating country and in each classroom. To address the issue of comparability, complex sampling and scaling procedures were implemented to ensure that appropriate comparisons could be made between countries and among major subpopulations of interest within a country.

While the criteria of relevance and comparability have been addressed in the design and implementation of the IEA Reading Literacy Study, less attention has been focused on the criteria of interpretability (i.e., what the scores of this assessment mean). Meaning is usually made by comparing mean scores across subpopulations of interest, by examining performance on individual tasks, or by assuming it is inherent in the label that is used to organize a set of tasks — such as reading comprehension, critical thinking, or problem solving. To address the issue of interpretability in large-scale assessments, the authors have developed a paradigm for anchoring and interpreting tasks along a scale that was initially employed in the design and interpretation of the National Assessment of Educational Progress (NAEP) Young Adult Literacy Assessment (Kirsch and Jungeblut 1986) and subsequently refined (Kirsch and Mosenthal 1990; Kirsch, Jungeblut, and Campbell 1992).

The resulting information attempts to guide a reader's interpretation of performance both within and across domains of tasks by identifying and evaluating a set of variables that underlie performance on those domains. Typically, a domain of tasks over which performance is modeled is referred to as a scale. Depending on the type of psychometric model used, tasks are displayed along a scale based on one or more characteristics such as difficulty, guessing, or discriminating power. In addition to estimating item characteristics, individual or group proficiency is also estimated across a particular set of tasks. Through these analyses, it becomes possible to estimate the percentages of various groups or individuals who perform at selected levels on a scale, as well as to construct a task map displaying where items are located along a particular scale.

Once tasks are displayed along a particular scale, one is compelled to ask why particular tasks cluster together at various points and what characteristics seem to distinguish easy from moderate tasks and moderate from difficult tasks. In general, the work conducted to date with the adult assessments reveals that while literacy is not a single skill suited to all types of tasks and materials, neither is it an infinite number of skills each associated with a different type of material and purpose for reading. Rather, there appears to be an ordered set of information-processing skills and strategies that get called into play to successfully perform the range of tasks falling along a particular scale. Through a deeper understanding of the variables that contribute to this ordering, we understand better what a scale measures and what a demonstrated level of proficiency may mean. Moreover, this type of information provides a framework for refining and extending a scale. With this information, one can begin to ask questions about what characteristics should be included in a test, how these characteristics should be manipulated across a scale, and whether these characteristics have differential impact on performance across subgroups of interest.

To date, our work has focused on domains of open-ended simulation tasks that have been administered to nationally representative samples of adults across the United States. The question remains as to whether the framework used in that research applies to school-based, multiple-choice items. The conduct of the IEA Reading Literacy assessment provides an opportunity to address this question.

The IEA Reading Literacy Study was conducted during the 1990-91 school year in some 32 countries around the world (Elley 1992). Nationally representative samples of 9-year-old and 14-year-old students were directed to read and respond to a broad range of materials over two testing periods. This chapter defines and evaluates a set of variables that could be used to understand the constructs being measured in the IEA survey and to compare and contrast these characteristics across grades and subpopulations of interest.

To illuminate the characteristics that underlie performance on the IEA scales, this chapter is organized as follows. Section 6.2 describes the procedures that were used to characterize the IEA materials and tasks. Included here are brief descriptions of narrative, expository, and document materials that were included in this survey. Next, the procedures used to analyze task complexity — that is, the set of material and process variables — are described as they apply generally across the three scales. Section 6.3 characterizes and evaluates the fourth and ninth grade narrative scales. The variables affecting difficulty for each narrative scale are identified and exemplary tasks are used to illustrate how these variables act, either alone or in combination with each other, to influence difficulty. These variables are then evaluated using both zero-order correlation and regression analyses. The various dimensions of these variables are then compared and contrasted across the fourth and ninth grade scales for the total, white, and minority populations. Section 6.4 focuses on the expository scales, and Section 6.5 focuses on the document scales. Finally, the concluding section of the paper summarizes the results and reconsiders the importance of interpretability, particularly as it relates to improving test design and use.

It is important to note that the analyses reported in this chapter have been restricted to students in the United States because we have no experience with, or knowledge of, the generalizability of our models to other languages or cultures. Moreover, because Item Response Theory assumes that item parameters are invariant among subpopulations of interest, we have chosen to use nationally weighted *p*-values (or percent correct) as our dependent measure so that we may evaluate performance among subpopulations of interest.

6.2 Procedures

The conceptual framework used in the IEA Reading Literacy Study included narrative, expository, and document materials in an attempt to reflect the diverse range of printed and written information that students around the world are expected to learn to read and use. Before characterizing the tasks and describing the analyses that were conducted, a brief description of the three types of stimulus materials is provided.

Nature of the Stimulus Materials

In the elementary grades, one of the principle vehicles for teaching reading is narratives, or stories. For the most part, the purpose of narratives is to entertain rather than to provide descriptions of events that actually transpired. As such, narratives tend to portray imag ıary or possible worlds and events that may include fictitious characters with human-like characteristics. In most cases, narratives tend to be organized into a series of episodes including setting, initiating event, goal definition, attempt, block, outcome, and resolution.

The setting describes a particular time and place and introduces the reader to the story's major characters. The initiating event is something that prompts a change in state in a character's thinking or actions. Often this initiating event causes the character to define or redefine its goal or purpose for behaving. This then leads to a series of attempts that may or may not be blocked by a variety of forces internal or external to the story's central character. At the completion of the attempts, there is some sort of outcome and often resolution wherein the character reflects upon his or her attempt to achieve a particular goal.

As students move from the elementary grades into middle grades and high school, there is an increased emphasis on having them learn to read expository prose. Exposition, for the most part, tends to be organized topically rather than episodically. Topics tend to include definitions and/or descriptions of phenomena. In some instances, such descriptions may include comparison and contrast of a phenomenon's state. In others, such descriptions may include characterization of steps that make up some process or procedure. In the IEA study, exposition focused on such topics as *the walrus, what is* *quicksand?, marmots, how to read the age of a tree, the promise of laser, a woman learns to read, paracutin (or the unusual formation of a volcano),* and *smoke.*

In addition to learning to read narratives and exposition, students and adults are expected also to be able to read and process information found in documents. This aspect of reading literacy has been receiving increasing attention in large-scale assessments in recent years. Included among the array of printed materials that are called documents are forms, tables, charts, graphs, advertisements, indexes, tables of contents, and maps.

The most basic form of documents is a *simple list*, composed of a label and a set of items that all share one or more characteristics identified by the label (e.g., "Food to Buy" followed by a list of items that are foods and all which are to be purchased). A slightly more complicated type of document is *combined lists*, which consist of two or more simple lists. These simple lists are typically concatenated such that one list represents a list of persons, places, or things, and one or more of the other lists provide modifying information (such as amounts or attributes). A typical example of such a document is a flight schedule that identifies the flight number of the plane, gate, city, and times of arrival and departure.

An even more complicated type of document includes *intersected lists*. This type of document is made up of three simple lists. One of these lists, (i.e., the intersected list) shares information with each of the two intersecting lists, one arrayed as a column along the left side of the intersected list and one arrayed as a row along the top of the intersected list. The classic example of such a list is the TV schedule in which the list of shows constitutes the intersected list, the list of channels constitutes the column intersecting list, and the list of times constitutes the row intersecting list.

Finally, the document reflecting the most complex structure is called a *nested list*. Nested lists consist of four or more simple lists whereby the labels of two of the simple lists are the same for each pair of lists. An example of such a document might include the amount of oil exports from the United States and Canada in the years 1990 and 1991. In this case, the labels related to countries would be repeated by year for the two years. Certain bus schedules, including arrival and departure times, also represent this type of document structure.

Just as tables can be classified in terms of the preceding categories, so can graphs and maps. Sometimes the information contained in a pie, bar, or line graph can be represented as a combined list, with one list representing persons or things and the other list representing amounts. In some cases, however, bar and line graphs may also represent intersected or nested lists. Maps typically represent a form of intersected list where locative information (for example, rivers, roads, towns, and parks) is arrayed in degrees latitude and longitude along a given axis.

Defining the Readability and Process Variables

In an attempt to characterize task complexity, we chose to examine two types of variables: readability and process variables. This perspective takes into account the fact that performance on any given literacy task depends to a large degree on what is read (i.e., material) and what the reader is asked to do with that material (i.e., question/directive) (Kirsch, Jungeblut, and Campbell 1992). In the following sections, we first identify the procedures for operationalizing material complexity. We then identify and define the process variables as they apply generally across the two grades and the three scales.

Readability Variables

Although a variety of procedures have been developed to characterize the structural complexity of prose materials, many — if not most — of these procedures are specific to a particular rhetorical type (e.g., narrative versus exposition) or genre type (e.g., mystery versus suspense versus fable). The limitation with such specific procedural analyses, of course, is that they do not allow for the comparison of structural complexity across different types of rhetorical structures and genres.

To permit such comparisons, psychologists have traditionally used a variety of readability formulas that focus on common lexical and sentential characteristics (Klare 1984). The advantage of these formulas is that they can be applied to prose materials of varying length such that they enable comparisons to be made on a variety of linguistic dimensions, including number of words in a passage, average number of words per sentence, average number of syllables per word, average number of sentences per 100 words, and overall readability (which typically combines two or more of the preceding variables). In research, instruction, testing, and assessment, the use of such readability indexes have consistently provided the baseline measures that permit comparison of structural complexity across all types of prose (Fry 1981). In many of these instances, readability is the variable that is manipulated such that it is one of the principal determinants of task difficulty (College Entrance Examination Board 1982). In other instances, it is reported primarily as a baseline measure against which the structural complexity of materials in other situations can be measured (Carver 1983).

To compare the structural complexity among the fourth and ninth grade narrative and expository scales, all passages were analyzed using the Fry (1977) readability formula. In addition to estimating overall readability, each passage was analyzed in terms of its number of words, number of syllables per sentence, and number of sentences per 100 words, since as Klare (1984) has noted, materials may represent complexity in one but not all of these dimensions. While this may not necessarily result in a substantially increased readability score (based on the aggregate interaction of two or more linguistic variables), this may, on the other hand, still influence overall task difficulty. This typically appears to be the case where an increase in syllables per word may significantly increase difficulty but not be reflected given the way readability tables are specified.

Unfortunately, readability formulas are not appropriate for analyzing the structural complexity of documents. This is because readability formulas require strings of serially connected words and sentences, whereas documents tend to represent information in either matrix, graphic, or pictorial formats. To estimate structural complexity among the document stimuli, two measures were used: (1) number of items and (2) type of document structure. To determine the number of specifics (or basic units of a document), documents were first divided into simple lists (Mosenthal and Kirsch 1989). These lists consist of a series of exemplars, or items that belong to a common class of elements (e.g., locations on a map, a list of resources, a list of job vacancies, a list of times when a bus leaves a specific location for downtown). Each class is said to consist of items that can be described by a common label (e.g., cities, low temperatures, and weather). Once a document had been divided into its simple lists, the number of items in the list were totaled as a measure of the length and density of a given document (Kirsch and Mosenthal 1990).

A second measure of document complexity used in this study was based on the structural complexity of the document (Mosenthal and Kirsch 1991). Documents organized as simple lists, comprising only a single label and a related list of items, were scored a "1" for structural simplicity. Examples of documents that received this score were the grade 9 documents *job vacancies* and *directions*. Surprisingly, no simple lists were used in the grade 4 document scale.

Documents organized as combined lists consisted of two or more simple lists in which one list tended to act as a list of subjects and the remaining list(s) served to modify this list. An example of such a document was the fourth grade *table of contents*, in which the list of titles served as a list of subjects and the lists of authors and pages served as lists of modifying information. Combined list documents were scored "2" for complexity. Note that basic graphs, such as the fourth grade *empty bottles*, were scored "2," as their underlying structure is similar to that of combined lists.

Documents organized as intersected lists consisted of three simple lists. One of these three lists (i.e., the intersected list) contained information that related to one of the intersecting lists as a row and a second intersecting list as a column. An example of this type of document was the fourth grade document *temperature*. Intersected list documents were scored "3" for complexity. Because of the way latitude and longitude are often specified, maps were included as representing an intersected list structure.

Documents organized as nested lists consisted of four or more simple lists in which identical labels were repeated under differing labels. An example of a document with such a structure was the ninth grade *bus schedule* in which the labels "Leaves Weston," "Leaves Trump," "Leaves Monument," and "Leaves Hilltop," were repeated both under the label "Inward - To City" and "Outward-From City." Nested-list documents were scored "4" for complexity. Graphs in which identical labels were nested under differing labels were also classified as representing nested documents.

Process Variables

In analyzing the IEA tasks, the authors hypothesized that the variables found to underlie difficulty on the adult literacy scales would similarly influence task difficulty on the narrative, expository, and document scales of the IEA Reading Literacy Tests. These three variables included type of information, type of match, and plausibility of distracting information.

The variable *type of information* represents how concrete or abstract the information in a question is relative to a passage or document. The more concrete the requested information is, the easier the task is judged to be; conversely, the more abstract the requested information is, the harder the task is thought to be. In our analysis cf the IEA tasks, this variable was scored on a five-point scale, with "1" representing the easiest dimension of this variable and "5" representing to the most difficult. (See Appendix 1 to this chapter for the specific scoring rules for this variable.) For instance, questions asking respondents to identify a person, animal, or thing (i.e., imaginable nouns) were said to request highly concrete information and received a score of "1." Questions asking respondents to identify goals, conditions, or purposes were said to request more abstract information. These tasks were thought to be more difficult to complete and were given a score of "3." Questions that tended to require respondents to identify an "equivalent" were judged to be the most abstract and were assigned a score of "5." Equivalence, in this case, tended to be an unfamiliar term or phrase for which respondents had to provide a definition or interpretation, or a predicating condition that had to be inferred from text.

A second process variable examined in this study was *type of match*. This variable dealt with the degree of difficulty associated with matching information in a question to information in a passage or document to information provided by multiple-choice alternatives. (It should be noted that many of the document tasks did not use a multiple-choice format.) This variable also was defined as ranging in difficulty from 1 to 5, with "1" again said to represent the easiest condition. (See Appendix 1 for the rules characterizing this variable.) Overall, four types of matching strategies were identified: locate, cycle, integrate, and generate. Type of match was judged to be easiest when there was a literal or synonymous relation between the information contained in the question, the text, and the correct response listed among the choices. As one or more of these relations required greater matching, inferencing, or integration of information, tasks were estimated to increase in difficulty. The most difficult type of match was said to be one in which the reader had to generate the appropriate interpretive framework to relate the information provided in the question, stimulus material, and alternative choices.

A third process variable examined was *plausibility of distracting information*. This variable dealt with the degree of difficulty associated with selecting the correct answer from among a list of multiple-choice answers. This variable, too, ranged in difficulty as defined by a five-point scale. (See Appendix 1 for the scoring rules characterizing this variable.) Tasks were said to be the easiest when no distracters were present in the list of alternative choices. That is, none of the distracters in the list were stated either literally or synonymously in the text. Tasks were judged to become more difficult as the number of distracters increased, as the distracter shared more features with the correct choice, and as the distracters were placed in closer proximity in the text where the correct response appeared. For instance, tasks tended to be moderately difficult when one or more of the distracters appeared as invited inferences that met some but not all of the conditions established in the question and that appeared in a paragraph other than the one containing the correct response. Tasks were judged as being the most difficult in terms of this variable when two or more distracters partially satisfied the conditions specified in the question and appeared in the same paragraph as the correct answer.

In the sections that follow, the material and process variables defined here are summarized and evaluated for each of the IEA scales. Similarities and differences among the scales are highlighted for the fourth and ninth graders as well as between the total, white, and minority populations.

6.3 The IEA Narrative Scales

Table 6-1 compares the grade 4 and grade 9 narrative scales with respect to summary statistics on a selected set of material and process variables. The four stories used to construct the grade 4 scale had an average readability level of 2.5 and ranged from grade 1 to grade 4. Thus, no passage contained within this scale required students to read beyond their current grade in school. These four passages had an average of 444 words and 47 sentences. The shortest story contained 292 words, and the longest had 703. The full set of information for these variables is provided for each grade in Appendix 2 to this chapter.

Table 6-1. Selected summary statistics comparing the grade 4 and grade 9 narrative scales

Variable	Grade 4		Grade 9	
	Mean	Range	Mean	Range
Readability	2.5	(1-4)	4.2	(1-6)
# of words	443.5	(292-703)	669.8	(422-1143)
# of sentences	47.3	(25-88)	54.8	(28-95)
# of words/sentence . . .	10.0	(8-12)	15.5	(10-15)
Percent correct	74.8	(52-94)	75.0	(44-96)
TOI	3.2	(1-5)	3.6	(2-5)
TOM	2.5	(1-4)	3.1	(1-5)
POD	2.3	(1-4)	2.8	(1-5)

TOI = type of information; TOM = type of match; POD = plausibility of distracting information.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Unlike the grade 4 scale, none of the five stories used on the grade 9 narrative scale were judged to be at grade level. In fact, the five stories had an average readability of only 4.2, and none of the five stories exceeded a sixth grade level. While the stories were several years below grade level, on average, they were somewhat longer and more complex than those used in the grade 4 scale. The grade 9 passages had an average of 670 words and 55 sentences. The shortest passage had 422 words while the longest had 1,143.

The 22 tasks making up the grade 4 narrative scale ranged from 94 to 52 percent correct with an average difficulty level of about 75. This range and average level of difficulty were very comparable to that shown for the grade 9 scale where the 29 tasks ranged from 44 to 96 percent correct with an average *p*-value of 75 percent. In comparison to the three process variables, the tasks on the grade 9 scale had a tendency to be more difficult and to have a broader range of scores with respect to the three process variables, but the mean values were not statistically different from those shown among the grade 4 tasks (Table 6-1).

Characterizing Narrative Tasks

Grade 4 Scale. Readability variables used in this study did not appear to influence task difficulty on the grade 4 narrative scale. For example, among the four hardest tasks or those having the lowest *p*-values, three had first and second grade readability levels. Conversely, among the five easiest tasks, four had third and fourth grade readability levels (see Appendix 2).

In contrast to the readability variables, tasks did tend to distribute themselves from easy to difficult in terms of their type of information, type of match, and plausibility of distracting information. To illustrate this, consider the grade 4 narrative *the bird and the elephant* shown in Exhibit 6-1.

This passage contained the fewest words (i.e., 292) and sentences (i.e., 25) of the set used in this fourth grade assessment. It had the second largest number of syllables per 100 words (i.e., 122) and sentences per 100 words (i.e., 9.6). It was rated as having a third grade readability level. Interestingly, the five survey items involving this passage covered the full range of difficulty on this scale, i.e., from 94 to 52 percent correct.

The easiest question and related set of choices are as follows:

The story ends happily because________________

A. The elephant died.

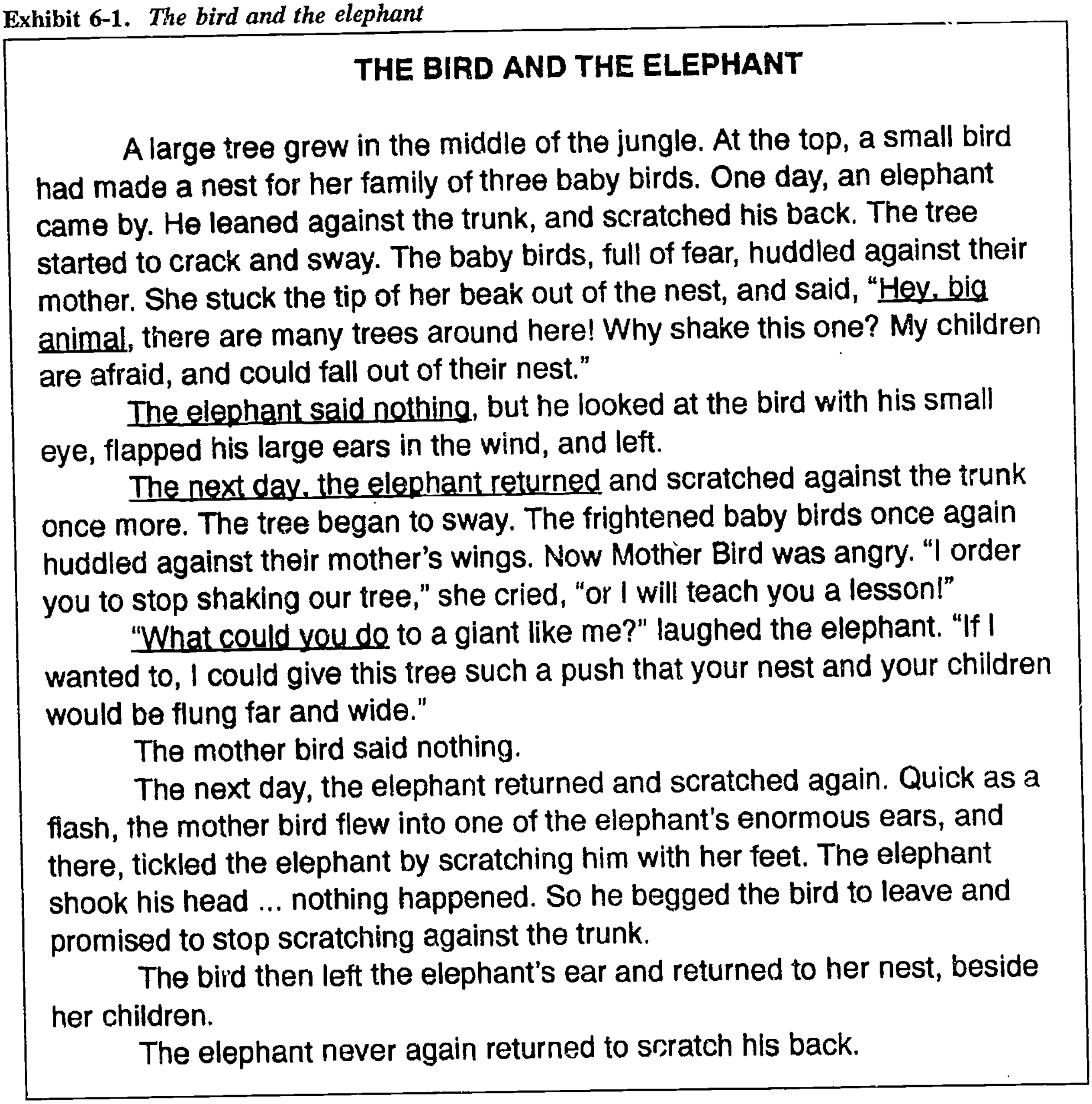
B. The elephant did not come back.

C. The tree was strong enough.

D. The birds learned to fly.

To complete the task, respondents had to recognize that the type of information requested was a "happy" event (with "happy" being an attribute). Moreover, respondents were directed to look at the end of the passage where they had to make a synonymous match between the text statement "The elephant never again returned to scratch his back," and the correct choice "the elephant did not come back." Finally, there was no distracting information, since none of the distracters were explicitly mentioned in the text; i.e., nowhere in the text did it say that the elephant died or the birds learned to fly.

Exhibit 6-1. *The bird and the elephant*



THE BIRD AND THE ELEPHANT

A large tree grew in the middle of the jungle. At the top, a small bird had made a nest for her family of three baby birds. One day, an elephant came by. He leaned against the trunk, and scratched his back. The tree started to crack and sway. The baby birds, full of fear, huddled against their mother. She stuck the tip of her beak out of the nest, and said, "Hey, big animal, there are many trees around here! Why shake this one? My children are afraid, and could fall out of their nest."

The elephant said nothing, but he looked at the bird with his small eye, flapped his large ears in the wind, and left.

The next day, the elephant returned and scratched against the trunk once more. The tree began to sway. The frightened baby birds once again huddled against their mother's wings. Now Mother Bird was angry. "I order you to stop shaking our tree," she cried, "or I will teach you a lesson!"

"What could you do to a giant like me?" laughed the elephant. "If I wanted to, I could give this tree such a push that your nest and your children would be flung far and wide."

The mother bird said nothing.

The next day, the elephant returned and scratched again. Quick as a flash, the mother bird flew into one of the elephant's enormous ears, and there, tickled the elephant by scratching him with her feet. The elephant shook his head ... nothing happened. So he begged the bird to leave and promised to stop scratching against the trunk.

The bird then left the elephant's ear and returned to her nest, beside her children.

The elephant never again returned to scratch his back.

A slightly more difficult question related to *the bird and the elephant* was:

What did the mother bird do to stop the elephant from returning to that tree?

A. She ordered him to stop.

B. She scratched his back.

C. She tickled his ear.

D. She stuck her beak into him.

The task based on this question and set of choices had a *p*-value of 81. To complete the survey item, respondents had to identify a specific action. For type of match, respondents were judged to infer that the elephant stopped scratching since the elephant could not stop the bird from tickling its ear. Finally, distracter A represented a low-level of distracting information and was scored a 2, since the mother bird actually did order the elephant to stop shaking her tree in the first paragraph.

An even more difficult task included the following item and choices:

What does the passage tell us?

A. When you're strong, you can bother others.

B. Elephants shouldn't shake trees.

C. The weak can sometimes overcome the strong.

D. Always face danger head-on.

The task associated with this item and choices had a *p*-value of 69. To complete this task, respondents had to identify a theme (this type of information was rated difficult). For type of match, respondents were given no clues as to what constituted the correct answer, as none of the information in the item matched the information stated in the text. Rather, respondents had to integrate the information to identify the correct answer, i.e., "The weak can sometimes overcome the strong." As a result, this task received a high score for type of match.

In addition to this rather difficult level of match, this task was judged to have a moderate level of distracting information. This is because distracters A, B, and D above represent plausible invited inferences (but not plausible moral themes) based on the narrative.

The most difficult task related to the passage above included the following item and choices:

Which sentence in the story tells us that the elephant thinks he is the strongest? It starts with these underlined words:

A. Hey, big animal,. . .

B. The elephant said nothing,. . .

C. The next day, the elephant returned. . .

D. What could you do. . .

This task had a *p*-value of 53. To complete this task, respondents had to identify a statement that meant the equivalent of the elephant thinking that he was the strongest. (As noted above, equivalents were scored as representing the most difficult type-of-information level.) For type of match, respondents had to infer that the elephant thinks he is the strongest. This is suggested by the elephant stating, "What could you do to a giant like me?" Also note that, in terms of plausibility of distracting information, choice A represents distracting information in the sense that, by acknowledging the elephant as "hey, big animal," the bird is suggesting that the elephant is indeed the strongest This choice, however, is based on the bird's thinking and not the elephant's. Another plausible distracter was C. One might infer that by returning the next day, the elephant considered himself stronger than the bird because he was able to torment the bird without the bird being able to respond.

Grade 9 Scale. Unlike the grade 4 students' narrative tasks where none of the readability variables seemed to discriminate among task difficulty, here two variables seemed to be strongly associated with performance: number of syllables per 100 words and readability level itself. Number of syllables per 100 words appeared to progress linearly from 115 for the easiest tasks to 127 for the most difficult. Similarly, readability level ranged from first grade level for the easiest three tasks to fourth, fifth, and sixth grade levels for the most difficult tasks.

In addition, the tasks used to assess grade 9 students tended to distribute themselves primarily in terms of two of the three process variables: type of match and plausibility of distracting information. Among the grade 9 tasks, type of requested information tended to include many *why* questions addressing goal, purpose, and cause. As such, this variable did not appear to discriminate much among the 29 tasks. To illustrate the manner in which these variables interacted, consider the passage *killing the fox* (Exhibit 6-2) and its related tasks. This passage contains the fewest words (i.e., 422) and sentences (i.e., 28) of the set used in the grade 9 assessment. It had 15 words per sentence and 118 syllables per 100 words. This passage had a fifth grade readability level (only one passage had a higher readability level for this scale). Despite the relatively high readability level, the survey items relating to this passage ranged from 93 to 73 percent correct (see Appendix 2.)

One of the easiest items on the grade 9 narrative scale applied to this passage. This item and its accompanying choices were as follows:

Why did the author shoot the fox?

A. He wanted to punish the fox.

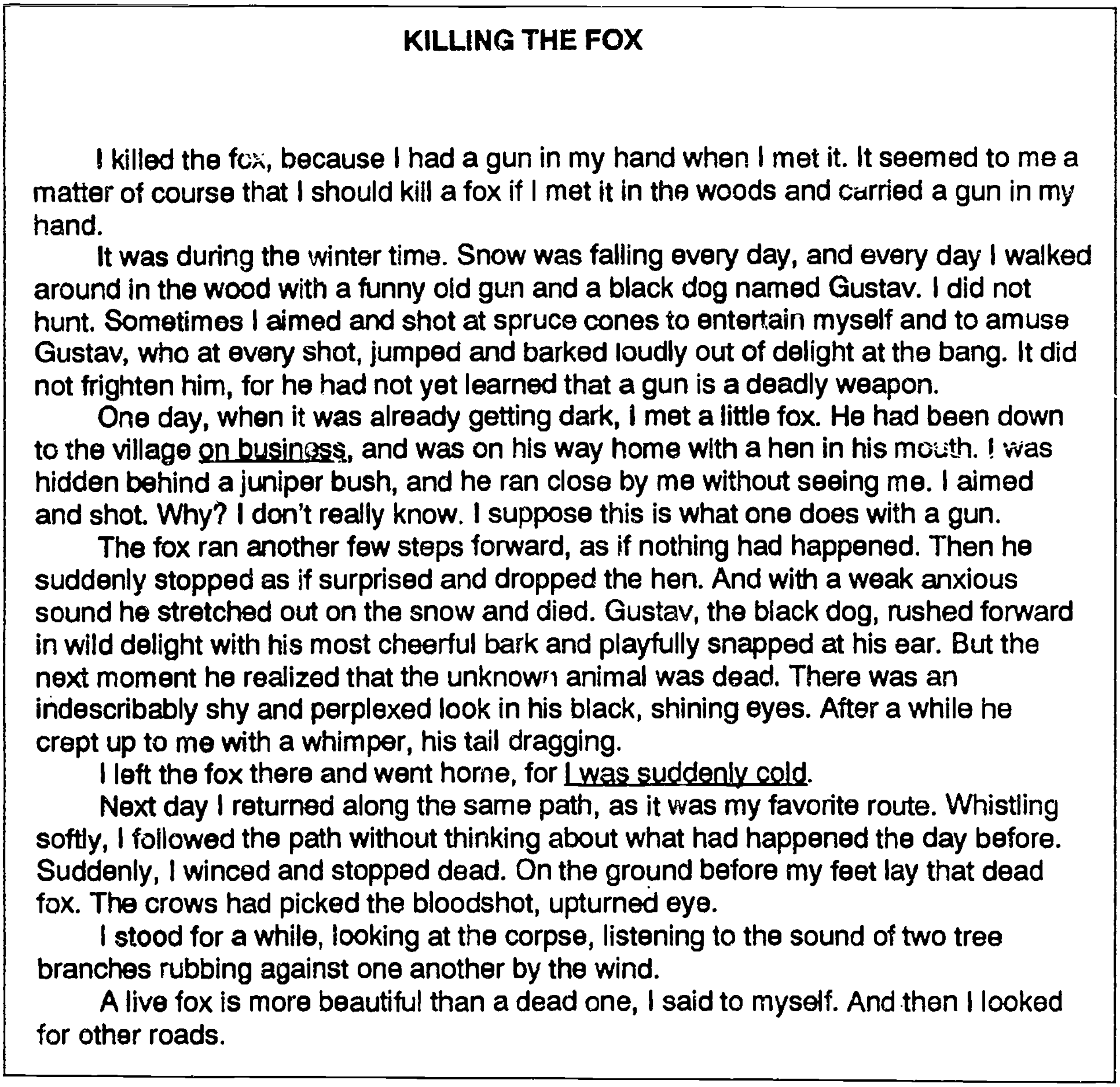
B. He was an experienced and skillful hunter.

C. He did it without thinking.

D. He was frightened by the fox.

The task associated with this survey item and choices had a *p*-value of 93. To complete this task, respondents had to identify why the author shot the fox. Respondents were able to locate the appropriate part of the text by making a literal match between the word "shot" in the question and the phrase "I aimed and shot" in the third paragraph of the passage. At this point, respondents had to make the synonymous match between the text statement "Why? I don't really know, I suppose this is what one does with a gun" and the correct choice "He did it without thinking." Also, this item was judged to have no plausible distracters, since none of the distracters were explicitly mentioned in the text.

Exhibit 6-2. *Killing the fox*



KILLING THE FOX

I killed the fox, because I had a gun in my hand when I met it. It seemed to me a matter of course that I should kill a fox if I met it in the woods and carried a gun in my hand.

It was during the winter time. Snow was falling every day, and every day I walked around in the wood with a funny old gun and a black dog named Gustav. I did not hunt. Sometimes I aimed and shot at spruce cones to entertain myself and to amuse Gustav, who at every shot, jumped and barked loudly out of delight at the bang. It did not frighten him, for he had not yet learned that a gun is a deadly weapon.

One day, when it was already getting dark, I met a little fox. He had been down to the village on business, and was on his way home with a hen in his mouth. I was hidden behind a juniper bush, and he ran close by me without seeing me. I aimed and shot. Why? I don't really know. I suppose this is what one does with a gun.

The fox ran another few steps forward, as if nothing had happened. Then he suddenly stopped as if surprised and dropped the hen. And with a weak anxious sound he stretched out on the snow and died. Gustav, the black dog, rushed forward in wild delight with his most cheerful bark and playfully snapped at his ear. But the next moment he realized that the unknown animal was dead. There was an indescribably shy and perplexed look in his black, shining eyes. After a while he crept up to me with a whimper, his tail dragging.

I left the fox there and went home, for I was suddenly cold.

Next day I returned along the same path, as it was my favorite route. Whistling softly, I followed the path without thinking about what had happened the day before. Suddenly, I winced and stopped dead. On the ground before my feet lay that dead fox. The crows had picked the bloodshot, upturned eye.

I stood for a while, looking at the corpse, listening to the sound of two tree branches rubbing against one another by the wind.

A live fox is more beautiful than a dead one, I said to myself. And then I looked for other roads.

A more difficult item and set of choices were as follows:

What did Gustav do when he understood that the fox was dead?

A. He ran and hid himself.

B. He crawled up to his master.

C. He took another road.

D. He snapped at the ear of the fox.

The task associated with this question had a *p*-value of 70. To complete this task, respondents had to identify the result associated with a specific condition. To match information between the question, the text, and the document, respondents first had to match "when he (Gustav) understood that the fox was dead," stated in the question, to the statement in the text, "he (Gustav) realized that the unknown animal was dead." To make this match, an anaphoric inference had to be made between "unknown animal" and "fox." To complete the match, respondents had to make a synonymous match between "he (Gustav) crept up to me with a whimper" in the text to "He crawled up to his master" from among the alternatives.

What made this task more difficult was the fact that, unlike the earlier example, this task involved distracting information that was judged to be difficult. One of the choices was "He (Gustav) snapped at the ear of the fox." In the passage above, the text states just before Gustav discovers that the fox was dead, the dog ". . . rushed forward in wild delight . . . and playfully snapped at his (the fox's) ear." This makes an excellent distracter because this information occurs in the same sequence of events just prior to the dog crawling up to his master. As such, it appears that plausibility of distracting information or type of information is driving the difficulty of this item more than type of match.

A passage generally associated with more difficult tasks on the grade 9 narrative scale was *mute*. This passage of moderate length consisted of 605 words and 53 sentences. It had, on average, 11 words per sentence, 127 syllables per 100 words, 7.5 sentences per 100 words and, like *the fox*, represented a fifth grade readability level. The survey items associated with *mute* ranged in difficulty from 44 to 71 percent correct.

At one point in the passage, we encounter the lines:

> "Come on chaps! Let's get to work! If we finish just one day late we'll get a penalty that will hurt!"

A moderately difficult question applied to this part of this passage included the following statement and choices:

The workers who came to build the school were in a hurry because

A. They wanted to return home quickly.

B. They would have to pay a fine if they didn't finish on time.

C. They would be rewarded if they finished before the expected date.

D. They didn't like working in front of the gaping villagers.

The task associated with this question and choices had a *p*-value of 71. To complete this task, respondents had to identify a cause underlying the workers' hurry to build the school quickly. To match information between the question, the text, and the document, respondents first had to match "workers who came to build the school were in a hurry" with the statement "Come on chaps! Let's get to work!" using a low-level inference. Respondents had to then make a second low-level inference to relate the requested information in the choice (i.e., "they would have to pay a fine if they didn't finish on time") to the corresponding information in the text (i.e., "If we finish just one day late we'll get a penalty that will hurt!").

In terms of plausibility of distracting information, we find that the distracters A, C, and D all represent plausible invited inferences relative to the text. In other words, there is nothing in the text that suggests such statements are not true and, moreover, parts of the choices (e.g., the villagers did, in fact, gape at the workers) actually reflect states or actions that occur in the text. As such, this distracting information was scored as moderately difficult.

At another point in the passage, we encounter the lines:

> One day, good old Cosme, who was the mute of the village, came running uphill through the path that led to South Nutsville. He panted, his big body staggering with the effort, his round face reddened, his shining bald head dripping with sweat, and screaming.
>
> Screaming? But he was mute. Yes, he was the mute man of the village. Nevertheless, he was a chatter-box. He was an engaging and communicative fellow, with a chattering and cheerful nature, who was always starting conversations with whomever was about. But he had the bad fortune to be mute.
>
> It had to be him! With so many people in the world who hardly talk and for whom hardly anything would change if they were to be mute... But no, it had to happen to him. Confusions of our chromosomes decide, before we are born, how we will be, from tip to toe. Two chromosomes that didn't get along well must have fallen to his lot. Indeed, one of them probably said, "He will be a great babbler." And the adjacent one said, "He will be mute."

A more difficult question and related set of choices applied to this section of the narrative included the following:

What did the writer mean by the statement that old Cosme's chromosomes didn't get along well (line 14)?

A. Fate had made a communicative man mute.

B. Old Cosme's chromosomes were mute.

C. Old Cosme's parents didn't get along well when discussing their son's fate.

D. The lonely life in the village made old Cosme mute.

The task associated with this question and choices had a *p*-value of 60. To complete this task, respondents had to identify an equivalent that required them to provide an interpretation (or definition) of a statement made by the author. To match information between the question, the text, and the document, respondents had only to make a literal match to relate first the information in the question to information in line 14 in the text. As such, locating the information in the text was not the difficult part. What was difficult, however, was that, once line 14 had been identified, respondents had to then make a high-level inference to construct an interpretation of the author's statement (i.e., that "two chromosomes that didn't get along well" really meant the same as "Fate had made a communicative man mute.")

In terms of plausibility of distracting information, we find that choice B (i.e., "Old Cosme's chromosomes were mute") was a good distracter in that it represented an invited inference based on information contained in the same paragraph in which line 14 occurred. As noted in Appendix 1 to this chapter, when such invited inferences tended to occur in the same paragraph as the answer, this made for a rather difficult distracter (scored "4" out of 5 for difficulty).

Finally, toward the end of the passage, we encounter the lines:

> From that moment, the peaceful village was shaken by the quivers and vibrations of the earth works, in particular, during the first days, when the monstrous machines snapped at the ground. Right beside the old school the ground was leveled. They dug out the rocks in order to place a circular platform and then left. In time, many trucks loaded with huge concrete beams and queer pieces arrived, while the whole village watched, unable to believe their eyes. The teacher, most surprised of all, exclaimed, "But it isn't a repair job. It's a completely different building....!"
>
> And they all commented, very intrigued:
>
> "But it's round! Will it be a baseball stadium?"
>
> "It will be huge! And we are so few!"
>
> "Have you seen it? It has no stairs! Just curving slopes!"
>
> "It doesn't have a single window!"
>
> And finally the teacher dared to show his surprise to the foreman. The foreman simply answered with a shrug, "Designs are designs."
>
> In the village, everybody was continually astonished. They spoke of nothing else, especially the mute.

The most difficult item on the grade 9 narrative scale related to this section of *mute* that came toward the end of the passage. This question and its accompanying choices were as follows:

The people of South Nutsville were surprised because

A. They thought giants were coming to attack them.

B. They had never seen machines before.

C. They didn't know anything about a building job on the school.

D. The work on the school was different from what they expected.

The task associated with this question and choices had a *p*-value of 44. To complete this task, respondents again had to identify a cause (in this case the cause underlying the surprise of the people of South Nutsville). This reflected a rather difficult level of type of information, receiving a score of "4." To match information between the question, the text, and the document, respondents had to integrate across several lines of text to identify the section of the passage specifically related to the item. Having done this, respondents then had to create a high-level inference to determine that the cause of the people's surprise stemmed from the fact that "the work on the school was different from what they expected" (i.e., choice D).

In terms of plausibility of distracting information, we again find a distracter located in the same paragraph as the answer. Note, in the above section of the passage, it states "But it isn't a repair job.

It's a completely different building!" Distracter C reflects an invited based on this statement that "they (i.e., the people of South Nutsville) didn't know anything about a building job on the school." Given this type of distracter, plausibility of distracting information was scored quite high for this question (i.e., a "4" out of 5).

Evaluating the Contribution of Variables to Task Difficulty

Grade 4 Analyses. To examine the extent to which the readability and process variables contributed to task difficulty on the narrative scale for the grade 4 total population, and for the white and minority subpopulations, zero-order correlations between the six readability variables and the three process variables were computed. The results of this correlation matrix are shown in Table 6-2.

As can be seen from this table, the three variables that correlated the highest with total, white, and minority *p*-values were the three process variables: type of match (i.e., -.75, -.73, and -.75, respectively), plausibility of distracting information (i.e., -.68, -.69, and-.66), and type of information (i.e., -.58, -.57, and -.58). Two of the six readability variables correlated with total, white, and minority *p*-values ___ these included number of words (i.e., .31, .34, and .27), and number of sentences (i.e., .29, .32, and .25).

As to be expected, the intercorrelations among readability variables were extremely high. For example, number of words correlated .99 with number of sentences and -.88 with words per sentence. The correlation between number of words per sentence and the number of sentences was also high (i.e., -.90), as was the correlation between syllables per 100 words and overall readability level (i.e., .87). Syllables per 100 words also correlated rather highly with the number of words (i.e., .75) and with the number of sentences (i.e., .77).

Table 6-2. Intercorrelations for grade 4 students between process and readability variables, and narrative task difficulty (represented by *p*-values) for total, white, and minority populations

	p-value	1	2	3	4	5	6	7	8	9	10	11
1	Total											
2	White	.99										
3	Minority	.98	.96									
	Readability variables											
4	No. of words	.31	.34	.27								
5	No. of sentences	.29	.32	.25	.99							
6	Words per sentence	-.18	-.23	-.10	-.88	-.90						
7	Syllables per 100 words	.28	.27	.29	.75	.77	-.44					
8	Sentences per 100 words	.16	.14	.20	-.07	-.18	.29	-.24				
9	Readability level	.12	.12	.11	.52	.52	-.37	.87	-.67			
	Process variables											
10	Type of information	-.58	-.57	-.58	-.25	-.29	.20	-.42	.40	-.52		
11	Type of match	-.75	-.73	-.75	-.05	.00	-.04	.09	-.47	.31	.46	
12	Plausibility of distracting information	-.68	-.69	-.66	-.20	-.14	.02	-.08	-.56	.23	.13	.69

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Some of the intercorrelations between readability and process variables were also significant. For example, type of information correlated -.42 with syllables per 100 words, .40 with sentences per 100 words, and -.52 with overall readability level. Sentences per 100 words correlated -.47 with type of match and -.56 with plausibility of distracting information.

Finally, the correlations were relatively high between type of match and plausibility of distracting information (i.e., .69) and bet-veen type of match and type of information (i.e., .46).

Regressions were next run to determine the relative strength among the readability and process variables in predicting task difficulty for total, white, and minority populations along the narrative scale. Because of the high ratio of variables to number of tasks, several rules were applied to help minimize overinterpretation of the data:

- Only those variables that were significantly correlated with p-values were included in the regression;

- Only variables that added significantly to the model were left in the regression; and

- Only those variables whose simple correlation with the dependent variable had the same sign as their beta weight were included in the regression.

While the first two rules seem rather apparent, some explanation is needed for the third. Typically, a partial regression weight whose sign is inconsistent with its zero-order correlation is a suppr^essor variable. Suppressor variables tend to be difficult to interpret and, more importantly, tend not to be replicable across samples. Applying these rules, three regressions were run. The results shown in Table 6-3 indicate that two process variables remained in the regression equation: type of information and plausibility of distracting information.

Table 6-3. Raw betas and *t*-ratios representing the regression of select process variables against total, white, and minority *p*-values for grade 4 narrative tasks

Process variable	Total (df=19)		White (df=19)		Minority (df=19)	
	Beta	*t*-ratio	Beta	*t*-ratio	Beta	*t*-ratio
Type of information	-5.00	-4.04**	-4.55	-3.89**	-5.89	3.84**
Plausibility of distracting information . .	-6.86	-4.98**	-5.52	-5.01**	-7.81	-4.58**
R^2 = ...		71%		71%		65%

**p < .01.; df = degrees of freedom.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

These two process variables accounte ' for 71 percent of the variance for the total and white populations, and 67 percent of the variance for the minority population. For each group, plausibility of distracting information received the largest standardized regression weight indicating its overall importance in the model. These results suggest that difficulty along this grade 4 scale is best accounted for by the distracters found among the multiple-choice items and by the type of information asked for in the survey items.

Grade 9 Analyses. To examine the extent to which readability and process variables contributed to task difficulty on the grade 9 narrative scale, a correlation matrix was computed between the six readability variables and the three process variables (Table 6-4). As can be seen from this table, the variables that had the highest correlation with total, white, and minority p-values were the two process variables, type of match (i.e., -.85, -.84, and -.85, respectively) and plausibility of distracting information (i.e., -.75, -.76, and -.74). The variables syllables per 100 words (i.e., -.63, -.58, and -.71) and overall readability level (i.e., -.54, -.50, and -.62) also showed a rather strong relationship with percent correct values.

Table 6-4. Intercorrelations for grade 9 students between process and readability variables, and narrative task difficulty (represented by *p*-values) for total, white, and minority populations

	p-value	1	2	3	4	5	6	7	8	9	10	11
1	Total											
2	White	.99										
3	Minority	.98	.96									
	Readability variables											
4	No. of words	-.39	-.34	-.49								
5	No. of sentences	-.34	-.29	-.43	.97							
6	Words per sentence	-.02	-.03	-.05	-.09	-.32						
7	Syllables per 100 words	-.63	-.58	-.71	.82	.82	-.16					
8	Sentences per 100 words	.44	.43	.50	-.26	-.07	-.75	-.45				
9	Readability level	-.54	-.50	-.62	.60	.48	.42	.80	-.89			
	Process variables											
10	Type of information	-.27	-.27	-.21	-.30	-.19	-.45	-.06	.32	-.29		
11	Type of match	-.85	-.84	-.85	.59	.53	.05	.71	-.46	.61	.20	
12	Plausibility of distracting information	-.75	-.76	-.74	.15	.05	.32	.24	-.43	.35	.12	.53

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

As was seen on the grade 4 scale, the intercorrelations among the readability variables were quite high. Number of words and number of sentences correlated .97. Syllables per 100 words correlated .82 with number of words as well as with number of sentences. In additicn, overall readability level correlated .80 with syllables per 100 words and -.89 with sentences per 100 words.

Table 6-4 also reveals moderate to strong relations between one of the process variables and several of the readability variables. Type of match correlated .61 with overall readability level, .71 with syllables per 100 words, .59 with number of words, and .53 with number of sentences.

Unlike the grade 4 narrative scale, the intercorrelations among the process variables were relatively low on the grade 9 scale — that is, below .20. The one exception was the correlation of .53 between type of match and plausibility of distracting information.

Based on these correlations, three regressions were run to determine the relative strengths of the readability and process variables in predicting difficulty on the grade 9 narrative scale for the total, white, and minority populations (Table 6-5).

Table 6-5. Raw betas and *t*-ratios representing the regression of select process variables against total, white, and minority *p*-values for grade 9 narrative tasks

Process variable	Total (df=26)		White (df=26)		Minority (df=25)	
	Beta	*t*-ratio	Beta	*t*-ratio	Beta	*t*-ratio
Type of match	-6.95	-7.02**	-6.33	-6.52**	-4.94	-3.35**
Plausibility of distracting information . .	-4.37	-4.63**	-4.30	-4.73**	-6.15	-6.06**
Syllables per 100 words	-	-	-	-	-1.00	-3.63**
R^2 = ...	85%		84%		90%	

***p*<.01; df = degrees of freedom.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Here we see that type of match and plausibility of distracting information significantly predicted difficulty for the total, white, and minority populations. Type of match is the most important predictor, followed by plausibility of distracting information for the total and white populations. Among the minority population, plausibility of distracting information was the best predictor. In addition, one of the readability variables added significantly to the model for the minority population. Together, these variables accounted for between 84 and 90 percent of the variance.

Summary

A set of variables has been included in the analyses which, in previous research with adults, has been shown to influence task difficulty. These variables, in essence, represent different aspects of a task where a task is defined as the stimulus material plus the question or directive asked over that material. The readability variables reflect the overall length and complexity of the stimulus material. The three process variables were developed to reflect various aspects of the interaction between the material and the level of processing needed to successfully respond to the survey item.

What these analyses reveal is that readability was not a significant predictor of overall task difficulty for either of the two narrative scales. Only one of the readability variables — syllables per 100 words — entered into the grade 9 model and this was only for the minority population. These results probably reflect more on the passages that were selected for this assessment than on the importance of readability as a predictor of task difficulty. It is worth emphasizing that none of the grade 4 passages were rated above a fourth grade level. And, while there was more variability among the grade 9 passages, none of these were rated above the sixth grade level.

The best predictors of the grade 4 and grade 9 narrative scales were the process variables. Interestingly, the variables that entered into the grade 4 and grade 9 regression models were somewhat different. At the grade 4 level, the most salient predictor was plausibility of distracting information followed by type of information. Tasks at the grade 9 level also were affected by the plausibility of distracting information. At this level, however, plausibility of distracting information was not the most salient characteristic. Type of match replaced type of information in the regression model and received the largest beta weight.

In light of these observations, consider the overall characteristics of the two narrative scales. Tasks did not tend to vary along either of the scales with respect to readability. Tasks that were easy tended to require students to locate literal or synonymous information that was relatively concrete. In addition, there were few if any choices that served as plausible distracters. Tasks became more difficult on the grade 4 scale as survey items required students to respond to more abstract information or to distinguish among more plausible distracters. At the grade 9 level, tasks that were moderately difficult tended to represent either a difficult type of match or plausibility of distracting information, but not both. Finally, tasks that were most difficult were judged to be high with respect to each of the process variables.

6.4 The IEA Expository Scales

Table 6-6 compares the grade 4 and grade 9 expository scales with respect to summary statistics on a selected set of material and process variables. The five passages used to construct the grade 4 scale had an average readability level of 4.4 and ranged from first to sixth grade. Thus, unlike the grade 4 narrative scale, the expository passages required students to read both at and slightly above their grade in school. In fact, the average readability of the grade 4 expository passages was almost two grade levels above the average grade 4 narrative story used in the IEA assessment. These five expository passages had an average of 204 words and 12 sentences. The shortest passage contained 56 words and the longest had 389. While they had higher readability levels, these five expository passages were less than half as long as the passages used on the narrative scale. The full set of information associated with these variables is provided for each grade in Appendix 2 to this chapter.

Table 6-6. Selected summary statistics comparing the grade 4 and grade 9 expository scales

Variable	Grade 4		Grade 9	
	Mean	Range	Mean	Range
Readability	4.4	(1-6)	6.0	(4-9)
# of words	204.2	(56-389)	398.8	(228-830)
# of sentences	12.2	(7-20)	22.8	(11-48)
# of words/sentence . . .	15.4	(8-21)	18.2	(14-23)
Percent correct	68.1	(34-97)	72.2	(44-96)
TOI	2.5	(1-5)	3.3	(1-5)
TOM	2.1	(1-5)	2.7	(1-5)
POD	2.9	(1-5)	3.3	(1-5)

TOI = type of information; TOM = type of match; POD = plausibility of distracting information.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

At the grade 9 level, only one of the five passages was rated to be at grade level; the other four ranged between fourth and sixth grade. In fact, the five passages had an average readability of only 6.0. While the grade 9 passages were, on average, several years below grade level, they were somewhat longer and more complex than those used on the grade 4 exposition scale. For example, the average passage used for grade 9 was almost twice as long as the average expository passage used for grade 4 — 399 words compared to 204 words. As was the case with the passages at grade 4, those at grade 9 were significantly shorter than the grade 9 passages used in the narrative scale. The grade 9 narratives averaged 670 words compared to the 399 words used in the average expository passage.

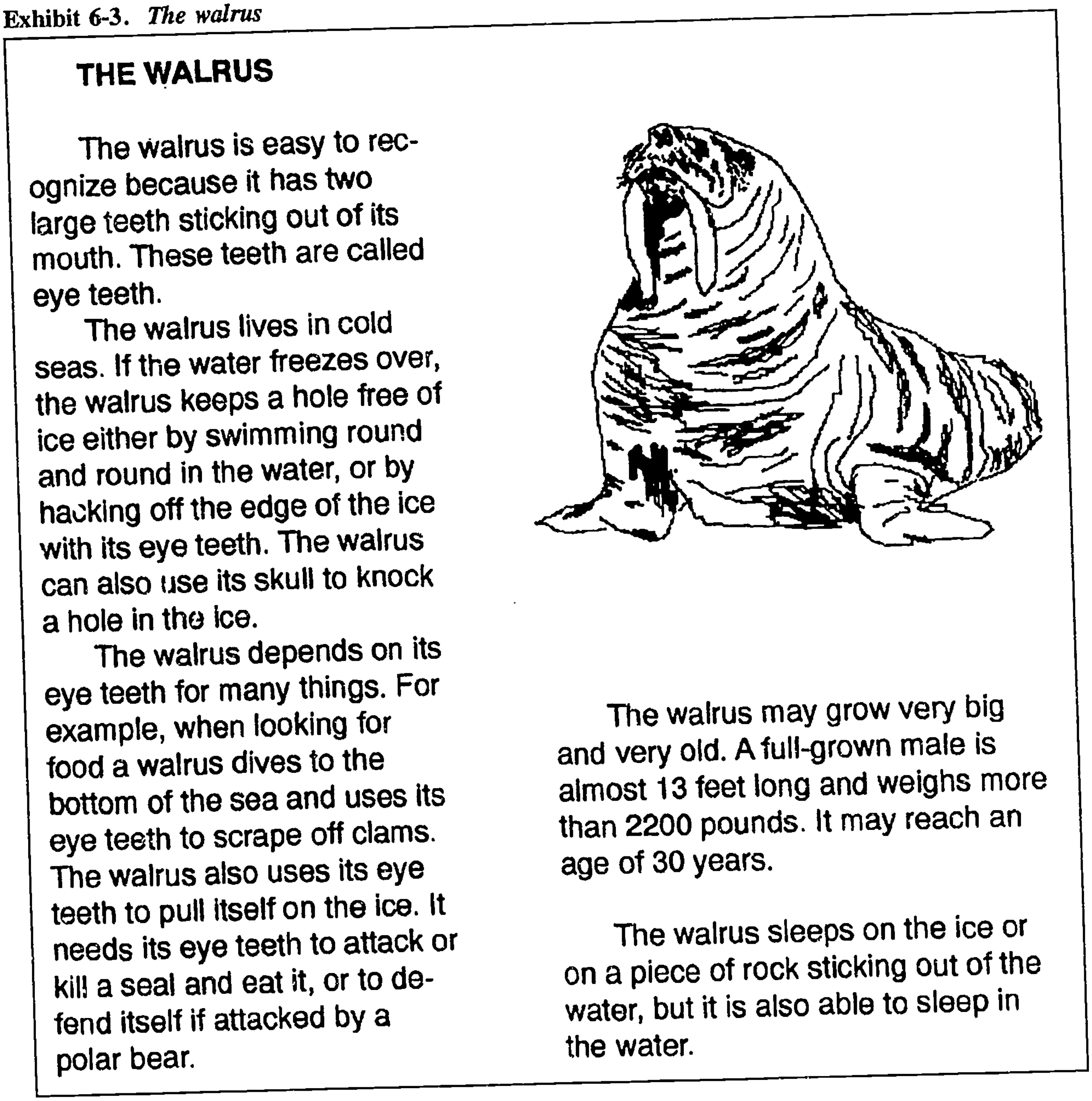
The 21 tasks asked over the five grade 4 expository passages ranged from 97 to 34 percent correct and had an average difficulty level of about 68 percent. As was the case among the narrative tasks, the 26 grade 9 expository tasks were very comparable in terms of range and average difficulty level to the grade 4 tasks (Table 6-6). These items ranged from 44 to 96 percent correct and had an average difficulty of 72 percent. Table 6-6 also shows that the three process variables ranged from "1" (easiest) to "5" (most difficult) on both the grade 4 and grade 9 expository scales. In addition, the tasks on the grade 9 scale had a tendency to be more difficult with respect to these three variables than those used in the grade 4 assessment.

Characterizing Expository Tasks

Grade 4 Scale. Unlike the grade 4 narratives, tasks on the grade 4 expository scale tended to distribute themselves from easy to difficult based on several of the six readability variables. In glancing at the ordering of tasks on this scale from easy to difficult (see Appendix 2), it appears that number of words and number of sentences were reasonably good predictors of how difficult a task is likely to be; that is, harder tasks tend to be associated with relatively longer passages.

In addition, and as was the case for tasks on the grade 4 narrative scale, tasks on the grade 4 expository scale also tended to distribute themselves from easy to difficult in terms of type of information, type of match, and plausibility of distracting information. To illustrate this, consider the passage *the walrus* (Exhibit 6-3).

Exhibit 6-3. *The walrus*



THE WALRUS

The walrus is easy to recognize because it has two large teeth sticking out of its mouth. These teeth are called eye teeth.

The walrus lives in cold seas. If the water freezes over, the walrus keeps a hole free of ice either by swimming round and round in the water, or by hacking off the edge of the ice with its eye teeth. The walrus can also use its skull to knock a hole in the ice.

The walrus depends on its eye teeth for many things. For example, when looking for food a walrus dives to the bottom of the sea and uses its eye teeth to scrape off clams. The walrus also uses its eye teeth to pull itself on the ice. It needs its eye teeth to attack or kill a seal and eat it, or to defend itself if attacked by a polar bear.

The walrus may grow very big and very old. A full-grown male is almost 13 feet long and weighs more than 2200 pounds. It may reach an age of 30 years.

The walrus sleeps on the ice or on a piece of rock sticking out of the water, but it is also able to sleep in the water.

This passage consisted of 207 words and 13 sentences. It had a readability level of sixth grade, which was the highest level of any passage on this scale. The survey items related to this passage ranged in *p*-values from 94 to 60.

The easiest task included the following item and choices:

How long can a walrus live?

A. 2 years

B. 4 years

C. 30 years

D. 100 years

To complete this item, respondents had to identify a specific number of years, which was relatively concrete in terms of type of information requested. The type of match involved making a literal match between "30 years" in paragraph four and "30 years" in the choices. While other numbers are mentioned in the text, none of these are given as alternatives in the question. Therefore, this question was rated easy in terms of each of the three process variables.

A task of comparable difficulty included the following question and choices:

Where does the walrus live?

A. In very cold water

B. In tropical countries

C. On the bottom of lakes

D. In cold forest country

This task had a *p*-value of 93. Note that this task, like the preceding one, requested relatively concrete information (i.e., location). Moreover, it involved a literal match between the question and the text and between the text and the choices. To complete the tasks, respondents had merely to match "walrus live" in the question to "walrus lives" in the first sentence of the second paragraph. Next, respondents had only to match "cold seas" in the text to "in very cold water" in the choices. Finally, as none of the distracters appeared as location information in the text, there was no distracting information relative to this question.

A more difficult task related to the passage above included the following question and choices:

We can tell that the walrus has to protect itself from

A. Seals

B. Bears

C. Eagles

D. Lions

This task had a *p*-value of 79. To complete this item, respondents had to identify an animal as type of information (which is highly concrete information). However, type of match is no longer literal in that respondents had to make a low-level inference to relate "walrus has to protect itself" in the question to "It (i.e., 'walrus') needs its eye teeth . . . to defend itself if attacked . . ." in the text. The factor, however, which contributes most to this task's difficulty is plausibility of distracting information. Note that, while seal does not satisfy the conditions as being an animal that attacks walruses (but rather the opposite), the word "seal" appears in the same sentence as the correct answer, polar bear. Moreover, seal is mentioned in terms of attack, thus making seals rather difficult distracting information between the text and the choices.

Another comparable task in the 70 to 79 percent correct range included the following item and choices:

What does a walrus do when it wants to get up on the ice?

A. It jumps up.

B. It cries for help.

C. It uses its eye teeth.

D. It uses its skull.

This task also had a *p*-value of 79. To complete this task, respondents had to identify a relatively concrete action. Respondents had to make a low-level inference relating "get up on the ice" in the question with "pull itself on the ice" in the text. This task again reflected a moderate level of distracting information in that a walrus could conceivably use its skull "to knock a hole in the ice" and then climb onto the ice. However, this answer was incorrect, as it was not what walruses do ' nder all (or most) conditions in getting up on the ice.

Finally, the most difficult task related to the walrus passage included the following question and choices:

How does the walrus get its food?

A. It catches fish with its eye teeth.

B. It scrapes clams off the bottom of the sea.

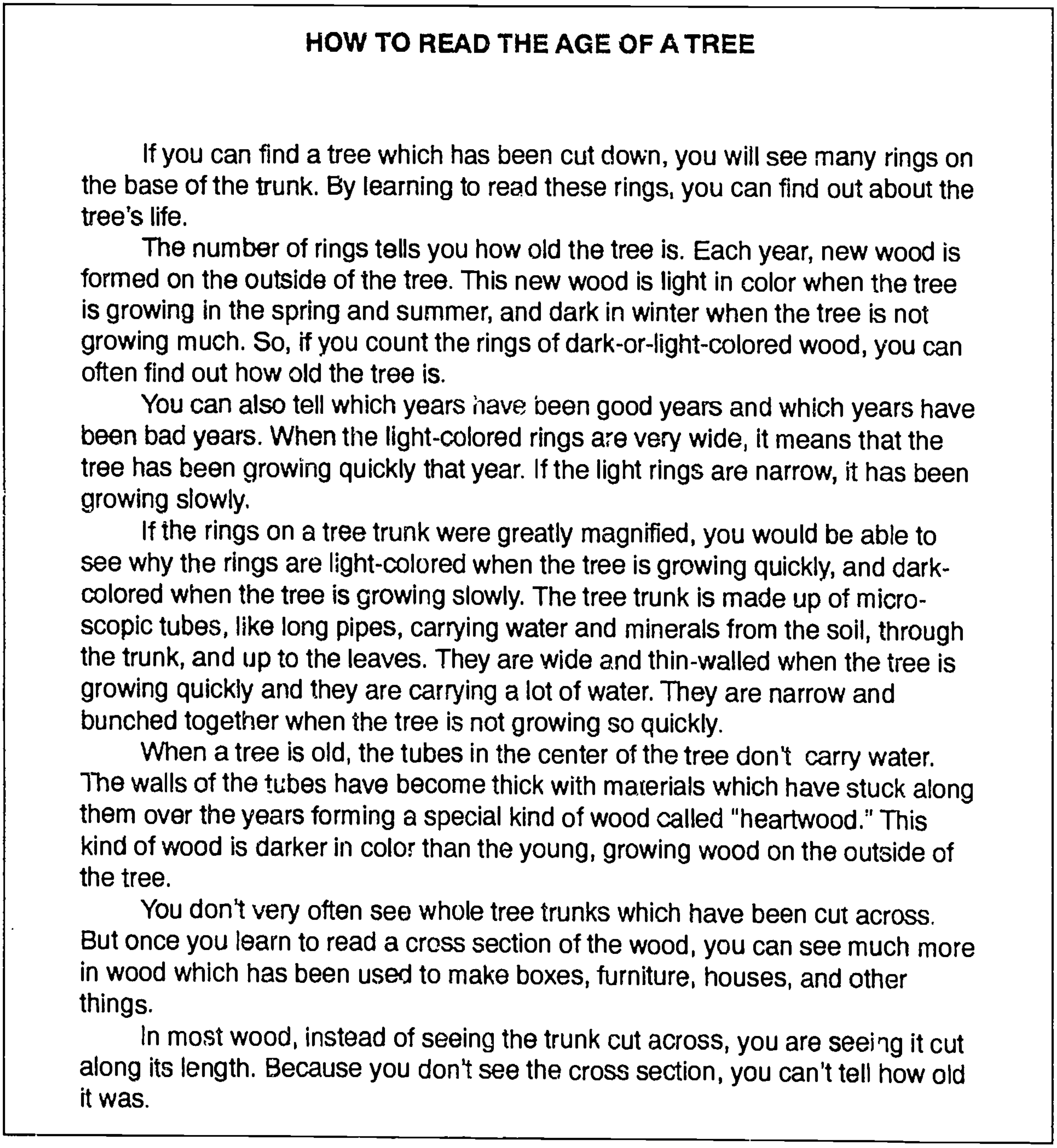
C. It knocks a hole in the ice with its skull.

D. It attacks polar bears.

This task had a *p*-value of 60. To complete this task, respondents had to identify a manner that represented a moderate level of difficulty for type of information. Type of match for this task was actually quite easy; it involved making a literal match between "food" in the item and "food" in the text. In fact, it is the only time food is mentioned. Note, however, that a statement among the choices is "It catches fish with its eye teeth." This distracting information represents a highly plausible invited inference since, in the same paragraph as the answer, there is a statement that walruses use their eye teeth to catch seals and eat them. Couldn't walruses also catch and eat fish using their eye teeth?

To illustrate tasks representing a greater degree of difficulty, consider those that applied to the passage *how to read the age of a tree* (Exhibit 6-4). This passage contained 389 words and 20 sentences and it had a fifth grade readability level. Although this passage did not represent the highest passage readability level on this scale, it was associated with the two hardest survey items the scale. Overall, the items ranged from 74 to 34 percent correct.

Exhibit 6-4. ***How to read the age of a tree***



HOW TO READ THE AGE OF A TREE

If you can find a tree which has been cut down, you will see many rings on the base of the trunk. By learning to read these rings, you can find out about the tree's life.

The number of rings tells you how old the tree is. Each year, new wood is formed on the outside of the tree. This new wood is light in color when the tree is growing in the spring and summer, and dark in winter when the tree is not growing much. So, if you count the rings of dark-or-light-colored wood, you can often find out how old the tree is.

You can also tell which years have been good years and which years have been bad years. When the light-colored rings are very wide, it means that the tree has been growing quickly that year. If the light rings are narrow, it has been growing slowly.

If the rings on a tree trunk were greatly magnified, you would be able to see why the rings are light-colored when the tree is growing quickly, and dark-colored when the tree is growing slowly. The tree trunk is made up of microscopic tubes, like long pipes, carrying water and minerals from the soil, through the trunk, and up to the leaves. They are wide and thin-walled when the tree is growing quickly and they are carrying a lot of water. They are narrow and bunched together when the tree is not growing so quickly.

When a tree is old, the tubes in the center of the tree don't carry water. The walls of the tubes have become thick with materials which have stuck along them over the years forming a special kind of wood called "heartwood." This kind of wood is darker in color than the young, growing wood on the outside of the tree.

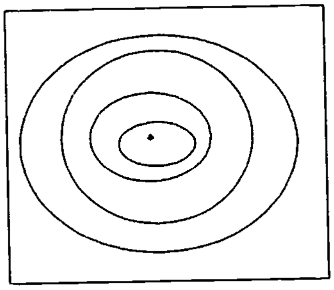
You don't very often see whole tree trunks which have been cut across. But once you learn to read a cross section of the wood, you can see much more in wood which has been used to make boxes, furniture, houses, and other things.

In most wood, instead of seeing the trunk cut across, you are seeing it cut along its length. Because you don't see the cross section, you can't tell how old it was.

One of the tasks based on this passage had a *p*-value of 35 percent and consisted of the item and choices shown in Exhibit 6-5.

Exhibit 6-5. Box 1

In the cross section of the tree trunk shown in Box 1, all the rings are wide and about the same width. This shows that the tree



BOX 1

A. Grew quickly all its life.

B. Grew slowly all its life.

C. Grew quickly when it was young and more slowly later.

D. Grew slowly when it was young and more quickly later.

To complete this item, respondents first had to identify an equivalent relating the characteristic growth patterns in Box 1 to a linguistic interpretation. Moreover, in terms of type of match, respondents found this to be a difficult task because the cross-section of the tree shown in Box 1 presents different information than what is stated in the question itself. According to the question, all the rings of the tree are said to be wide and about the same width. In Box 1, the rings of the tree are not all wide nor are they all the same width; the inner rings are more narrow than the outer rings.

To arrive at the correct answer for this task, readers had to generate the inference frame that the information in the question takes precedence over information in the drawing. Otherwise, in relating the information shown in Box 1 to the information in the text noted above, readers would infer that, because the inner rings are narrow, the tree began growing slowly. Because the outer rings are wider in the illustration, the tree then grew more quickly. According to this interpretation, the following distracter choice would, in fact, be correct:

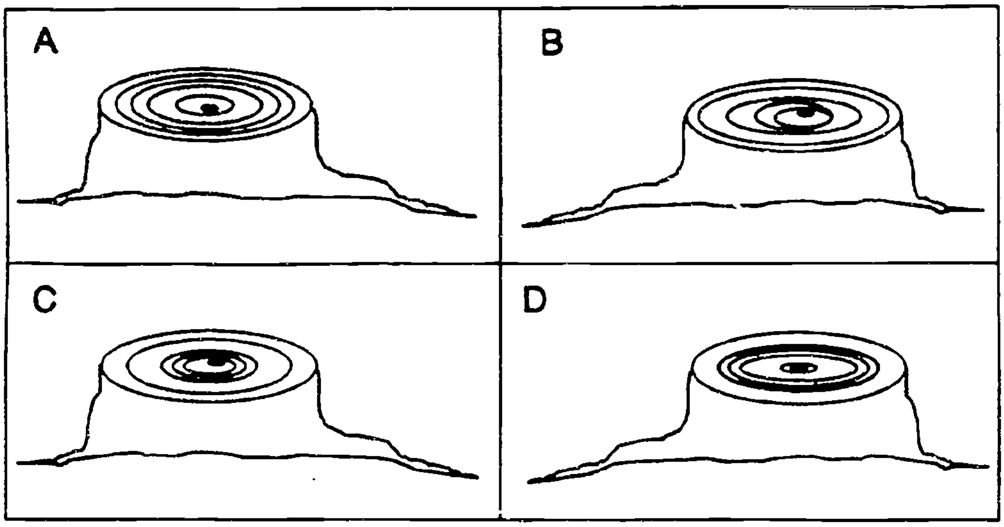
"grew slowly when it was young and more quickly later."

As such, not only is type of match extremely difficult for this task but so, too, is the plausibility of distracting information, since there is a distracter that corresponds with the pattern represented in the figure.

Another difficult task (*p*-value = 34) associated with this passage had the following question and choices shown in Exhibit 6-6.

Exhibit 6-6. *Tree trunks*

41. In a country which has a dry climate, it rains heavily every third year. Which drawing shows a tree trunk from this country?



Again, to complete this task, respondents had to relate linguistic information in the question (which suggests a climate condition in which it is dry for 2 years and rains heavily every third year) to a pictorial presentation of a tree having a corresponding ring pattern. As with other tasks related to this passage, type of match and plausibility of distracting information were difficult since different choices could be interpreted as possibly satisfying the condition of the question.

Grade 9 Scale. Tasks also appeared to distribute themselves from easy to difficult along the grade 9 expository scale in terms of readability; as the readability level of a passage increased, so did the difficulty of the tasks associated with this passage. Tasks also distributed themselves along the scale in terms of the three processing variables: type of information, type of match, and plausibility of distracting information. To illustrate this, consider the passage *a woman learns to read* (Exhibit 6-7).

The tasks associated with this passage tended to be of average difficulty; their *p*-values ranged from 60 to 73. The question and choices associated with the easiest task were as follows:

Mrs. Okashi often protests loudly if she is

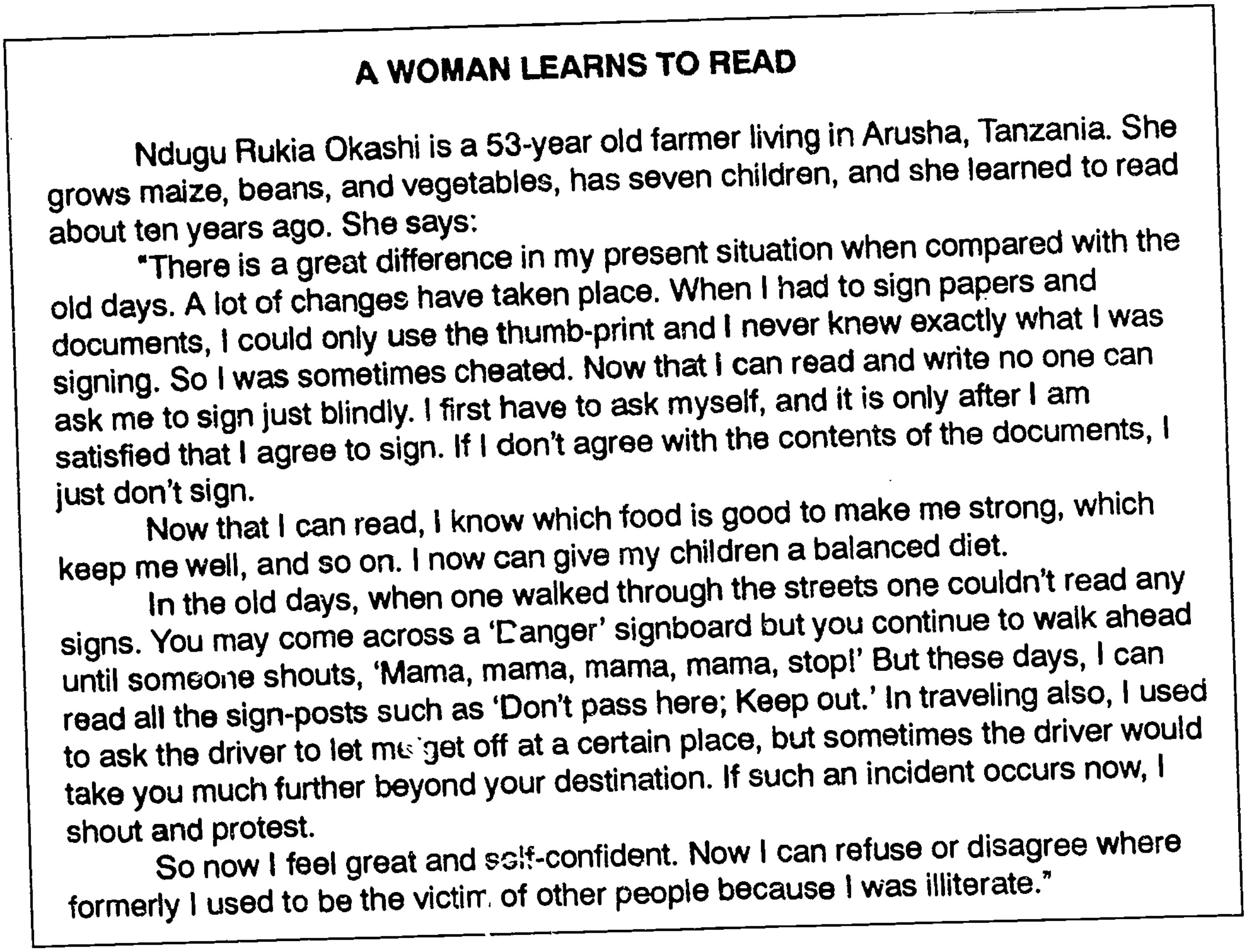
A. Charged too much by drivers.

B. Taken too far by drivers.

C. Mistakenly gets into danger.

D. Misled by public signs.

Exhibit 6-7. *A woman learns to read*



A WOMAN LEARNS TO READ

Ndugu Rukia Okashi is a 53-year old farmer living in Arusha, Tanzania. She grows maize, beans, and vegetables, has seven children, and she learned to read about ten years ago. She says:

"There is a great difference in my present situation when compared with the old days. A lot of changes have taken place. When I had to sign papers and documents, I could only use the thumb-print and I never knew exactly what I was signing. So I was sometimes cheated. Now that I can read and write no one can ask me to sign just blindly. I first have to ask myself, and it is only after I am satisfied that I agree to sign. If I don't agree with the contents of the documents, I just don't sign.

Now that I can read, I know which food is good to make me strong, which keep me well, and so on. I now can give my children a balanced diet.

In the old days, when one walked through the streets one couldn't read any signs. You may come across a 'Danger' signboard but you continue to walk ahead until someone shouts, 'Mama, mama, mama, mama, stop!' But these days, I can read all the sign-posts such as 'Don't pass here; Keep out.' In traveling also, I used to ask the driver to let me get off at a certain place, but sometimes the driver would take you much further beyond your destination. If such an incident occurs now, I shout and protest.

So now I feel great and self-confident. Now I can refuse or disagree where formerly I used to be the victim of other people because I was illiterate."

To complete this survey item, respondents had to select a condition — i.e., a type of information that was rated moderate in abstraction. In addition, once respondents had matched on "protests" in the question to "protest" in the last sentence of paragraph four, they then had to make a low-level inference to relate the phrase, " . . . sometimes the driver would take you much further beyond your destination" in the text, to the choice "taken too far by drivers." Finally, respondents had to avoid the highly distracting information that related the description of Ms. Okashi getting into danger (located in the same paragraph as the answer) to the choice "mistakenly gets into danger."

The most difficult task associated with this passage was the following:

Which of these phrases best expresses the underlying theme of this passage?

A. The benefits of becoming literate.

B. The way in which one person became literate.

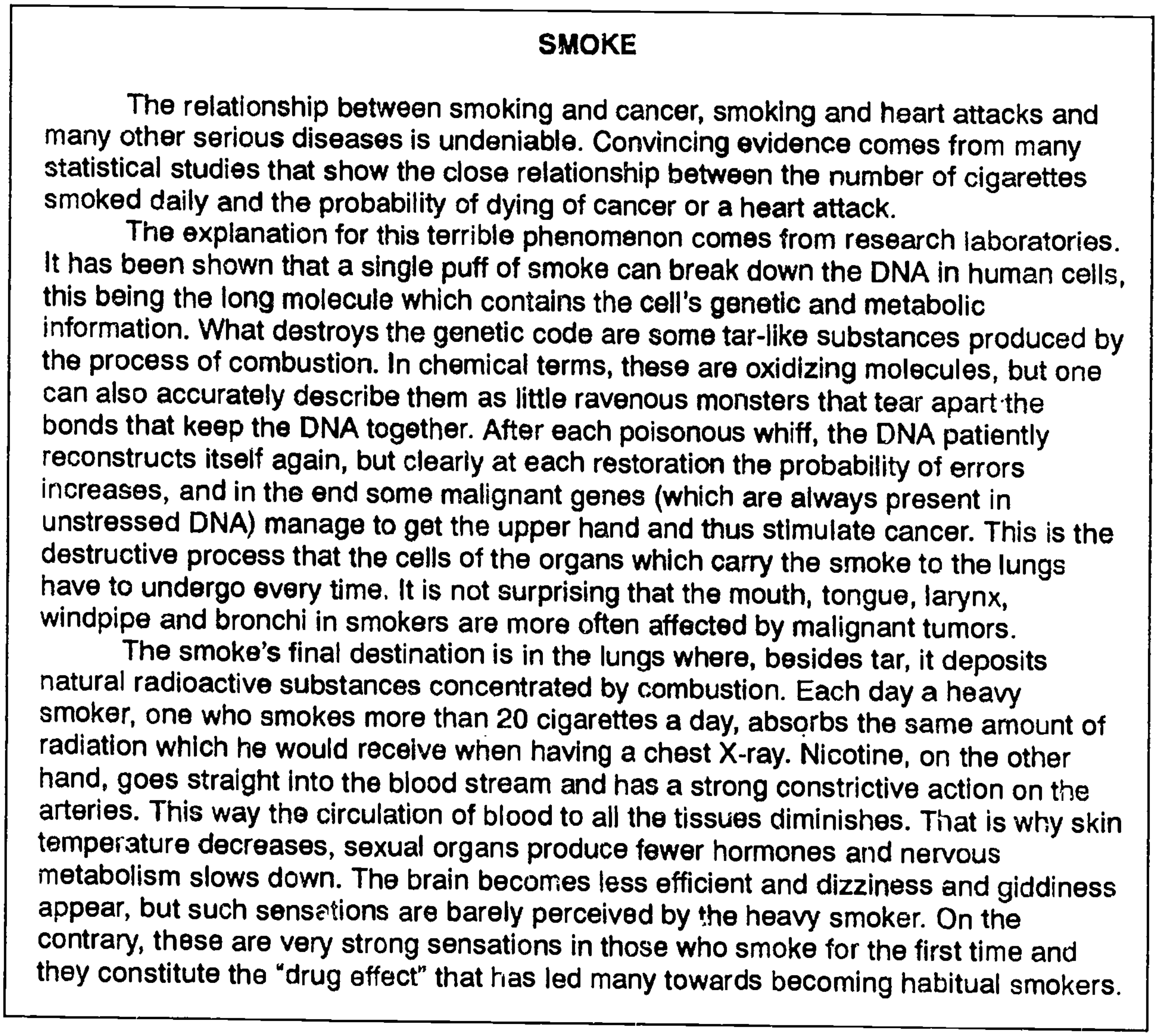
C. The problems of being an illiterate Tanzanian farmer.

D. The difficulties of coping in a literate world.

This task had a *p*-value of 60. To complete this item, respondents had to recognize a theme; a type of information that was rated as rather abstract. Moreover, type of match for this task was extremely difficult, as none of the words or phrases in the item keyed readers into a particular part of the text. Rather, to relate the question to the text and the overall text to the choices, respondents had to integrate information across the text. Finally, the distracter, "difficulties of coping in a literate world," is a highly plausible invited inference since much of the text does discuss this. Thus, this task was difficult in terms of all three process variables.

Another passage having a set of more difficult tasks was *smoke* (Exhibit 6-8). This passage had four items related to it that were the most difficult on the grade 9 expository scale. *Smoke* consisted of 368 words and 16 sentences. It had a readability level of ninth grade, which was the highest grade level of any passage used on this scale. The questions ranged from 75 to 46 percent correct.

Exhibit 6-8. *Smoke*



SMOKE

The relationship between smoking and cancer, smoking and heart attacks and many other serious diseases is undeniable. Convincing evidence comes from many statistical studies that show the close relationship between the number of cigarettes smoked daily and the probability of dying of cancer or a heart attack.

The explanation for this terrible phenomenon comes from research laboratories. It has been shown that a single puff of smoke can break down the DNA in human cells, this being the long molecule which contains the cell's genetic and metabolic information. What destroys the genetic code are some tar-like substances produced by the process of combustion. In chemical terms, these are oxidizing molecules, but one can also accurately describe them as little ravenous monsters that tear apart the bonds that keep the DNA together. After each poisonous whiff, the DNA patiently reconstructs itself again, but clearly at each restoration the probability of errors increases, and in the end some malignant genes (which are always present in unstressed DNA) manage to get the upper hand and thus stimulate cancer. This is the destructive process that the cells of the organs which carry the smoke to the lungs have to undergo every time. It is not surprising that the mouth, tongue, larynx, windpipe and bronchi in smokers are more often affected by malignant tumors.

The smoke's final destination is in the lungs where, besides tar, it deposits natural radioactive substances concentrated by combustion. Each day a heavy smoker, one who smokes more than 20 cigarettes a day, absorbs the same amount of radiation which he would receive when having a chest X-ray. Nicotine, on the other hand, goes straight into the blood stream and has a strong constrictive action on the arteries. This way the circulation of blood to all the tissues diminishes. That is why skin temperature decreases, sexual organs produce fewer hormones and nervous metabolism slows down. The brain becomes less efficient and dizziness and giddiness appear, but such sensations are barely perceived by the heavy smoker. On the contrary, these are very strong sensations in those who smoke for the first time and they constitute the "drug effect" that has led many towards becoming habitual smokers.

One difficult task associated with this passage (*p*-value of 51) included the following question and choices:

Smoke is dangerous to the lungs because

A. Nicotine and tar accumulate there.

B. It causes a greater predisposition to cancer there.

C. Stronger bonds form between DNA and malignant genes.

D. Tar and radioactive substances are deposited there.

To complete this task, respondents had to determine a cause that had the effect of smoking being dangerous to the lungs. (As noted in Appendix 1, this type of information was scored "4" in terms of its difficulty.) Moreover, for type of match, respondents had to infer that tar and radioactive substances deposited in the lungs over time are dangerous. Having made this inference, students then had to match the phrase "The smoke's final destination is in the lungs where, besides tar, it deposits natural radioactive substances concentrated by combustion" to the choice "tar and radioactive substances are deposited there."

Note that this task was made more difficult given that information in choice A (i.e., "nicotine and tar accumulate there") also occurs in the same paragraph as the correct answer. This choice is wrong only given the fact that nicotine does not stay in the lungs but is passed into the blood stream where it does harm which, consequently, makes smoke dangerous.

Finally, the most difficult task on the grade 9 expository scale (with a *p*-value of 46) had the following question and choices:

Which of these phrases best indicates the writer's attitude toward smoking?

A. "Patiently reconstructs..."

B. "Ravenous monsters..."

C. "Constrictive action..."

D. "Habitual smokers..."

To complete this task, respondents had to again determine an equivalent relation between "writer's attitude" and "ravenous monsters. . ." Moreover, type of match was made extremely difficult in this task as respondents had to make two high-level inferences — one relating the question to the text and one relating the text to the choices. It is worth noting that the phrase "ravenous monsters" per se does not directly relate to the writer's attitude toward smoking; rather, it relates to "tar-like substances" or "oxidizing molecules." Moreover, each of the other distracters (including "habitual smokers") are terms present in the text. To select "ravenous monsters" over the distracters, readers had to generate the frame that the writer abhors smoking due to its destructive processes. Observing this, readers then had to select ravenous monsters over the other choices since this was the concept most closely related with destructive processes. Given that some of the other choices also relate to the destructive processes of smoking (however, without the prominence of choice B), they serve as excellent distracters, thus making plausibility of distracting information quite high for this task as well.

Evaluating the Contribution of Variables to Task Difficulty

Grade 4 Analyses. To examine the extent to which readability and process variables contributed to the difficulty of the tasks on the grade 4 expository scale, a correlation matrix was computed (Table 6-7).

As can be seen from Table 6-7, both readability and process variables were significantly correlated with total, white, and minority *p*-values. Plausibility of distracting information had the highest correlations for all three sets of *p*-values (i.e., -.88, -.87, and -.87, respectively). This was followed by type of match (i.e., -.74, -.75, and -.70) and number of words per sentence (i.e., -.79, -.77, and -.83). Moderate correlations also were found with type of information (i.e., -.66, -.66, and -.64), number of words (-.64, -.62, and -.67), and sentences per 100 words (i.e., .63, .61., and .68).

As with the narrative scales, there were high intercorrelations among the readability variables. Number of words and number of sentences correlated highest (.96). Sentences per 100 words correlated -.91 with words per sentence and -.91 with overall readability level. The readability level also correlated highly with words per sentence (.81) and with syllables per 100 words (.71).

Moderate to high intercorrelations also were found among the process variables. Type of information correlated .80 with type of match and .63 with plausibility of distracting information. In addition, type of match correlated .73 with plausibility of distracting information.

Table 6-7. Intercorrelations for grade 4 students between process and readability variables, and expository task difficulty (represented by *p*-values) for total, white, and minority populations

	p-value	1	2	3	4	5	6	7	8	9	10	11
1	Total											
2	White	.99										
3	Minority	.99	.98									
	Readability variables											
4	No. of words	-.64	-.62	-.67								
5	No. of sentences	-.47	-.45	-.42	.96							
6	Words per sentence	-.79	-.77	-.83	.74	.56						
7	Syllables per 100 words	.17	.17	.15	-.16	-.15	.17					
8	Sentences per 100 words	.63	.61	.68	-.70	-.60	-.91	-.40				
9	Readability level	-.45	-.44	-.50	.47	.47	.81	.71	-.91			
	Process variables											
10	Type of information	-.66	-.66	-.64	.69	.69	.45	-.27	-.45	.19		
11	Type of match	-.74	-.75	-.70	.72	.71	.51	-.37	-.43	.17	.80	
12	Plausibility of distracting information	-.88	-.87	-.87	.58	.58	.61	-.08	-.52	.39	.63	.73

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Based on these correlations, three regressions were run to determine the relative predictive strength of the readability and process variables on grade 4 task difficulty for the total population, white, and minority populations. The rules for including variables in each of the regressions were the same as they were for grade 4 and grade 9 narrative scales. The results of these analyses are shown in Table 6-8.

Table 6-8. **Raw betas and *t*-ratios representing the regression of select process variables against total, white, and minority *p*-values for grade 4 expository tasks**

Process variable	Total (df=18)		White (df=16)		Minority (df=15)	
	Beta	*t*-ratio	Beta	*t*-ratio	Beta	*t*-ratio
Words per sentence	-2.24	-3.85**	-2.15	-3.33**	-2.96	-5.01**
Plausibility of distracting information.	-11.26	-5.95**	-11.22	-5.57**	-11.28	-5.99**
R^2 = ...	87%		85%		90%	

**$p<.01$; df = degrees of freedom.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

The data in Table 6-8 show that one process and one readability variable combine to predict task difficulty along the grade 4 expository scale. Interestingly, plausibility of distracting information was the most important predictor of task difficulty among this set of tasks. This result was also found among the grade 4 tasks on the narrative scale. Perhaps because of the greater range in readability among the expository passages, words per sentence also contributed significantly to the regression model. In combination, these variables accounted for between 85 and 90 percent of the variance among the three populations of interest. No differences were noted among the variables predicting difficulty for the total, white, and minority populations.

Grade 9 Analyses. To examine the extent to which readability and process variables contributed to the difficulty on the grade 9 expository scale, a correlation matrix was computed between the six readability variables and the three process variables (Table 6-9).

Table 6-9. **Intercorrelations for grade 9 students between process and readability variables, and expository task difficulty (represented by *p*-values) for total, white, and minority populations**

	p-value	1	2	3	4	5	6	7	8	9	10	11
1	Total											
2	White	.99										
3	Minority	.98	.97									
	Readability variables											
4	No. of words	.13	.14	.03								
5	No. of sentences	.28	.28	.27	.97							
6	Words per sentence	-.46	-.48	-.46	-.06	-.32						
7	Syllables per 100 words	-.68	-.65	-.67	-.03	-.25	.87					
8	Sentences per 100 words	.58	.58	.57	.34	.55	-.93	-.91				
9	Readability level	-.65	-.64	-.67	-.15	-.38	.90	.99	-.96			
	Process variables											
10	Type of information	-.68	-.66	-.72	.05	-.02	.15	.38	-.23	.35		
11	Type of match	-.73	-.73	-.73	-.27	-.33	.07	.27	-.24	.27	.43	
12	Plausibility of distracting information	-.64	-.62	-.66	-.11	-.24	.46	.48	-.48	.50	.27	.38

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

As displayed in Table 6-9, both readability and process variables were significantly related to *p*-values for total, white, and minority populations. Overall readability level correlated -.65, -.64, and -.67 with total, white, and minority *p*-values, respectively. Other readability variables demonstrating a moderate level of association included syllables per 100 words (-.68, -.65, and -.67), sentences per 100 words (.58, .58, and .57), and words per sentence (-.46, -.48, and -.46). In addition, all three process variables had a comparable level of association with percent correct for each population. Type of match had the highest correlation (-.73) with all three groups.

There were a number of high intercorrelations among the readability variables. As with the grade 4 expository scale, many of the intercorrelations were .87 or higher.

In contrast, the intercorrelations between the three process variables were only modest: .43 between type of information and type of match; .27 between type of information and plausibility of distracting information; and .38 between type of match and plausibility of distracting information.

In addition, the intercorrelations between the readability and process variables ranged from low to somewhat moderate, with plausibility of distracting information correlating .46 with words per sentence, .48 with syllables per 100 words, -.48 with sentences per 100 words, and .50 with readability level.

Three regressions were run to evaluate the relative strength of the readability and process variables in predicting difficulty on the ninth grade expository scale. The same three rules outlined earlier in this chapter were applied, and the results are shown in Table 6-10.

The data in Table 6-10 show that all three process variables were significant predictors of task difficulty. In addition, overall readability level also contributed to difficulty on this scale. As was noted with the grade 4 analyses, there were no differences among the three populations in either the variables that entered into the regression models or in the relative size of the beta weights obtained.

Table 6-10. Raw betas and *t*-ratios representing the regression of select process variables against total, white, and minority *p*-values for grade 9 expository tasks

Process variable	Total (df=21)		White (df=21)		Minority (df=21)	
	Beta	*t*-ratio	Beta	*t*-ratio	Beta	*t*-ratio
Readability level	-2.02	-3.32**	-1.94	-3.04*	-2.37	-4.18**
Type of information	-3.39	-3.75**	-2.99	-3.15**	-4.75	-5.66**
Type of match	-4.56	-4.72**	-5.10	-4.57**	-5.06	-5.64**
Plausibility of distracting information. .	-2.64	-2.58**	-2.39	-2.21*	-3.56	-3.72**
R^2 = ...	88%		85%		93%	

* $p < .05$; **$p < .01$; df = degrees of freedom.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Summary

On each of the expository scales, both readability and process variables contributed to overall task difficulty for each of the three populations studied. Easy tasks tended to be associated with relatively uncomplicated passages. The questions asked over these texts involved concrete information, literal matching of information, and few, if any, plausible distracters. Not surprisingly, the most difficult tasks tended to be associated with longer, more complicated texts and tasks requiring more complicated levels of processing.

It is again interesting to note that the most important predictor at the grade 4 level on both the narrative and expository scales was plausibility of distracting information. At grade 9, this variable was replaced in importance on both scales by type of match. Of the four prose scales discussed so far, the grade 9 expository scale represented the broadest range of text and process variables. As a result, these regressions had the most variables enter and remain in the models. As a set, these variables accounted for the largest variance — from 85 to 93 percent. In contrast, the grade 4 narrative scale had the most restricted range and also the least amount of variance accounted for — 60 to 70 percent.

6.5 The IEA Document Scales

Table 6-11 compares the grade 4 and grade 9 document scales with respect to summary statistics on a selected set of material and process variables. The six documents included a map, three tables, a bar graph, and a simple bus schedule. In terms of document type, there were no simple lists and no nested lists found on this scale; three of the five documents were combined lists and two were intersected lists. In addition, the documents were relatively short, having an average of only 25 items. The shortest document contained 15 items and the longest, 40 items. In contrast, the nine documents used on the grade 9 scale were not only longer — 71 items versus 25 — than those found on the grade 4 scale, they covered a broader range of document types as well. These documents included a form, three simple lists, two tables, a bar graph, and a complex bus schedule. Thus, each of the four document types — from simple list through nested lists—were represented on the grade 4 scale.

The 23 tasks asked over the five grade 4 documents ranged from 29 to 97 percent correct and had an average difficulty level of about 76 percent. As was noted for the narrative and expository scales, Table 6-11 shows that the 34 grade 9 document tasks were very comparable in terms of average difficulty level to the grade 4 tasks. These items ranged from 48 to 98 percent correct and had an average difficulty of 78 percent. It is worth noting that the range of difficulty for the grade 9 tasks was somewhat narrower than that for the grade 4 tasks. As shown in Appendix 2 to this chapter, there were five tasks on the grade 4 scale that had *p*-values in the 20-49 range compared to only one task on the grade 9 document scale. Table 6-11 also shows that tasks on the grade 9 scale had a tendency to be rated more difficult with respect to these three variables than those used in the grade 4 assessment.

Table 6-11. Selected summary statistics comparing the grade 4 and grade 9 document scales

Variable	Grade 4		Grade 9	
	Mean	Range	Mean	Range
# of items	25.2	(15-40)	70.7	(16-154)
Type of document. . . .	2.5	(2-3)	2.4	(1-4)
Percent correct	76.0	(29-97)	77.9	(48-98)
TOI	1.5	(1-2)	2.0	(1-5)
TOM	1.9	(1-5)	2.6	(1-5)
POD	2.3	(2-5)	2.2	(1-5)

TOI = type of information; TOM = type of match; POD = plausibility of distracting information.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Characterizing Tasks Along the Document Scale

Grade 4 Scale. Tasks on the grade 4 document scale tended to distribute themselves from easy to difficult primarily in terms of the three process variables, although there also appeared to be a moderate relationship with overall length as reflected in the number of items in a document. To illustrate this, consider the document *empty bottles* (Exhibit 6-9).

This graph was organized as a combined list consisting of 18 items. The tasks associated with this document ranged from 93 to 75 percent correct. The easiest of these tasks included the following question and choices:

Which class got the prize for the most bottles?

A. Mr. Green's class

B. Mr. Mack's class

C. Miss Barber's class

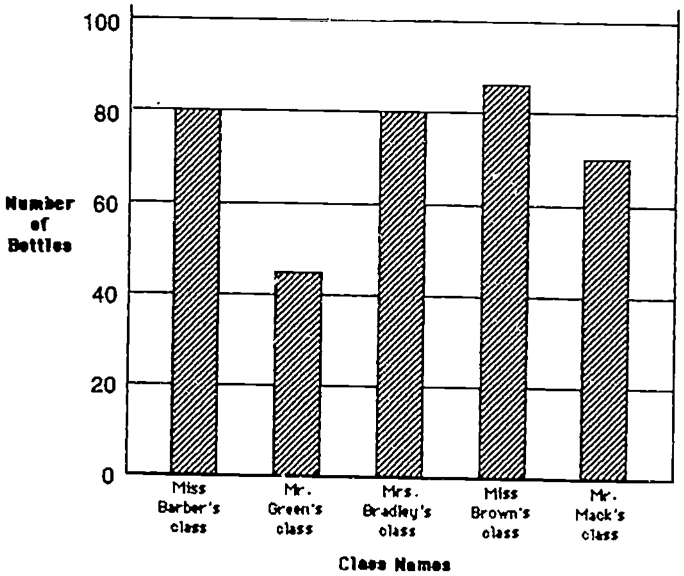
D. Miss Brown's class

To complete this item, respondents had to identify an attribute (i.e., a particular class). This task involved associating "most" in the question with the highest bar in the graph and then making a literal match between "Miss Brown's class" in the document and "Miss Brown's class" in the choices. As other classes collected bottles, there were some distracters. But since none of these distracters came close to the bar representing the number of empty bottles collected in Miss Brown's class, this task received a score of only "2" for plausibility of distracting information.

Exhibit 6-9. ***Empty bottles***

EMPTY BOTTLES

Mid-Town School had a bottle collection.
Children in each class brought empty bottles to school.
The principal made a bar graph of the number of bottles from five classes. Use the graph to answer questions 9-12.



A slightly more difficult task associated with this document included the following question and choices:

Which two classes collected exactly 80 bottles?

A. Miss Barber's class and Mrs. Bradley's class

B. Miss Barber's class and Mr. Mack's class

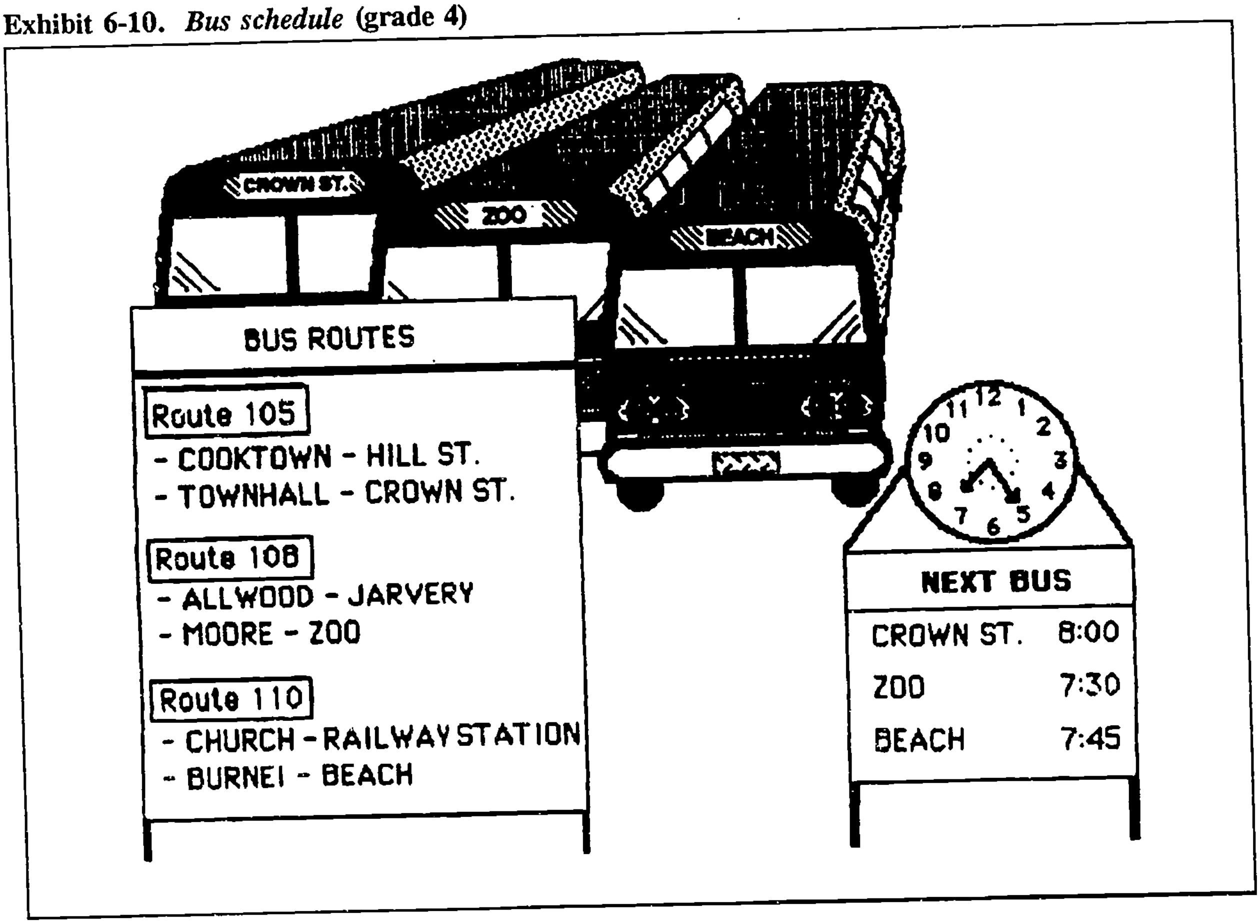
C. Miss Brown's class and Mrs. Bradley's class

D. Miss Brown's class and Mr. Mack's class

To answer this question, respondents had to again identify an attribute (i.e., two particular classes). The type of match this time, however, involved not a locate but a cycle. In short, respondents had to search the document based on the search criterion of "80 bottles." To complete the question, they had to match 80 twice in the document in order to identify Miss Barber's class and Mrs. Bradley's class (Choice A). Again, plausibility of distracting information was relatively low as no other classes closely approximated 80 bottles.

Tasks representing the other end of the difficulty continuum tended to be based on the *bus schedule* (Exhibit 6-10). This schedule was loosely organized as a combined list consisting of 27 items. Tasks associated with this document ranged from 91 to 29 percent correct.

Exhibit 6-10. *Bus schedule* **(grade 4)**



BUS ROUTES

Route 105
- COOKTOWN - HILL ST.
- TOWNHALL - CROWN ST.

Route 108
- ALLWOOD - JARVERY
- MOORE - ZOO

Route 110
- CHURCH - RAILWAY STATION
- BURNEI - BEACH

NEXT BUS

CROWN ST.	8:00
ZOO	7:30
BEACH	7:45

A moderately difficult task associated with this document (with a *p*-value of 70) included the following open-ended question:

Where do you think the bus stops first on Anne's way to the railway station?

To complete this question, respondents had to identify a location (i.e., a particular class). The type of match actually involved cycling whereby respondents first had to make a literal match between "railway station" in the question and "railway station" in the schedule. Having made this match, respondents then had to identify the first stop. This question is further made difficult because the label "Route 110" and other items such as "zoo" serve as significant distracting information, as they could be construed as being stops before the railway station.

A much more difficult task associated with the bus schedule (with a *p*-value of 46) included the following open-ended question:

What is the name of the place where buses stop just before the zoo?

Note that this question is quite similar to the preceding question in terms of type of information requested and type of match. However, what makes this question particularly difficult is the fact that the item zoo appears twice in the document (i.e., once in the list of bus routes and once in the list of destinations). In the event that students fail to identify "Next Bus" as representing departures, they were likely to identify "Crown St." as the correct answer. Thus, the item zoo appearing twice and the fact that Crown St. comes before zoo in the departure list makes for particularly difficult distracting information.

Finally, the most difficult survey item (with a *p*-value of 29) that related to the bus schedule (and was also the most difficult question appearing on the grade 4 document scale) included the following open-ended question:

How long will it be before the next bus leaves for the zoo?

To answer this question, respondents had to identify and calculate amount information. The type of match in this case involved cycling, whereby respondents first had to identify the time on the clock. Next they had to identify zoo in the list labeled Next Bus and the time associated with zoo (i.e., 7:30). Finally, respondents had to subtract the time of the clock from the time shown on the clock. In short, as the information on the clock had no literal referent in the question and since this match involved a mathematical operation, it was quite difficult. Also, as there is a time listed directly before zoo for the item Crown St., many respondents construed 8:00 as the time from which 7:30 was to be subtracted. As such, 8:00 represented rather difficult distracting information.

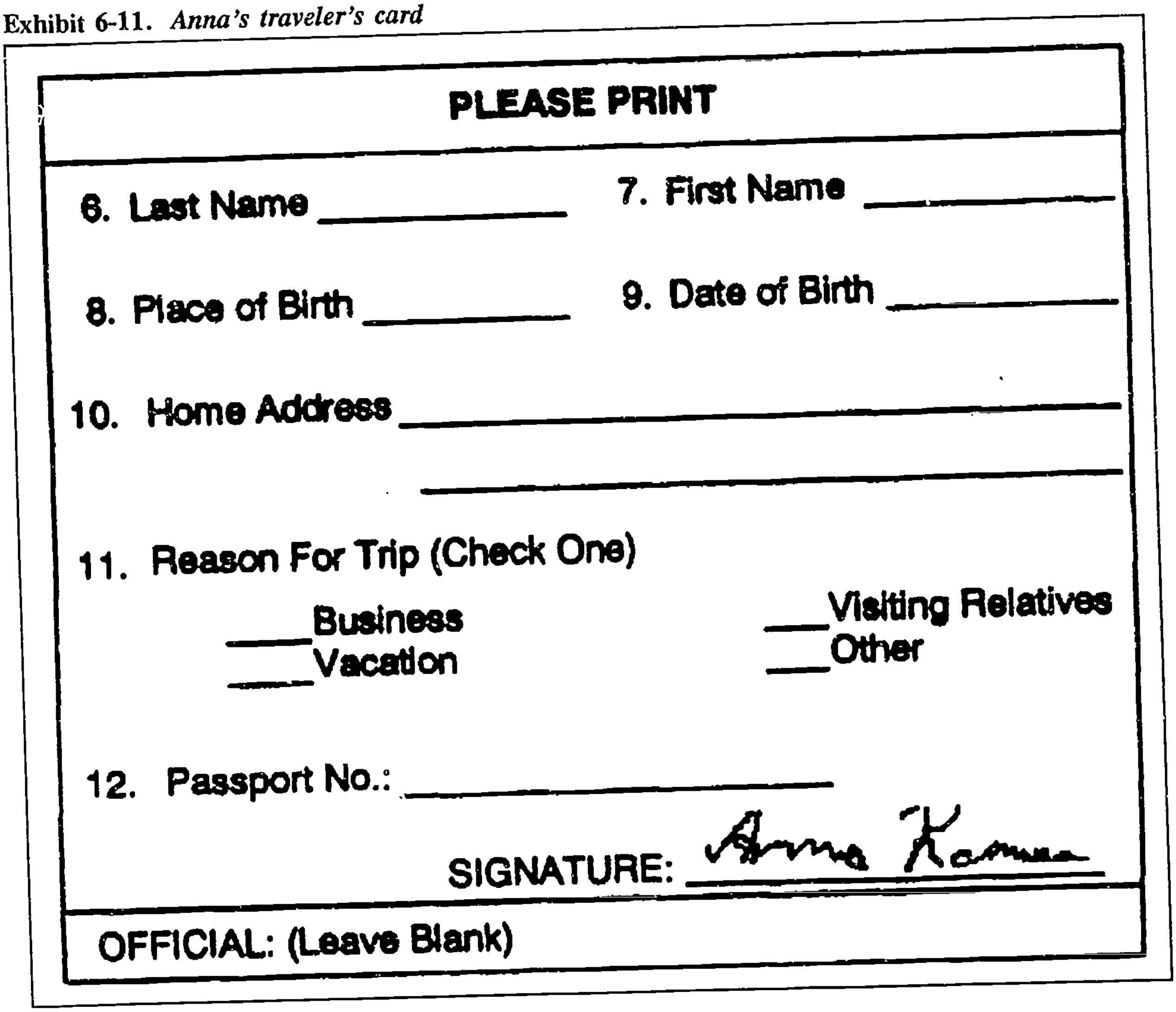
Overall, the documents used on the grade 4 scale were designed representing rather simple organizational structures having a relatively low number of items. In terms of process variables, the document tasks asked respondents to identify rather concrete information representing such things as persons, places, things, attributes, and amounts.

Where document tasks varied, however, was in terms of their type of match and plausibility of distracting information. Tasks that were easy required subjects to locate a single item based on one or two literal or synonymous feature matches with few distracters. Tasks became more complex when they required respondents to perform independent cycle matches. Tasks continued to increase in difficulty as they required respondents to perform dependent cycle tasks involving more difficult levels of distracting information. The most difficult task required respondents not only to cycle but also to perform a mathematical operation in the context of information representing a fairly high level of plausibility.

Grade 9 Scale. Tasks on the grade 9 document scale also tended to distribute themselves from easy to difficult, based on the three process variables. As was noted with the grade 4 document scale, there was a moderate degree of association with overall length of the document. To illustrate this, consider *Anna's traveler's card* show~~n in~~ Exhibit 6-11.

This form consisted of 15 items. The tasks associated with this document ranged from 98 percent correct to 61 percent correct. The easiest of these tasks (with a *p*-value of 98) required respondents to complete the label "Place of Birth" with the appropriate information provided in the question.

Exhibit 6-11. *Anna's traveler's card*



PLEASE PRINT

6. Last Name ____________ 7. First Name ____________

8. Place of Birth __________ 9. Date of Birth ___________

10. Home Address ______________________________
____

11. Reason For Trip (Check One)

____Business ___Visiting Relatives
____Vacation ___Other

12. Passport No.: ____________________

SIGNATURE: Anna Kama

OFFICIAL: (Leave Blank)

To do this, respondents had to identify a place (i.e., "Nadi"). The type of match involved making a synonymous match between "born in Nadi" and "Place of Birth." Aithough there are other places mentioned in the description of Anna's background, no other places of birth are mentioned or referred to. Consequently, this task was said to have a low level of plausibility of distracting information.

A much more difficult task applied to the above document required respondents to identify the reason for Anna's trip. (This task had a *p*-value of 61) Among the choices listed were "Business," "Vacation," "Visiting Relatives," and "Other." To answer this question, respondents had to identify a purpose (i.e., a moderately difficult type of information). The type of match in this instance required respondents to integrate information in the second paragraph of Anna's description and then make an inference that her purpose for traveling was business. Given that the category "Other" was available and that respondents had to infer that her travel was for business and not simply "to represent her country at the South Pacific Games in the high jump," Other tended to serve as a fairly high level distracter.

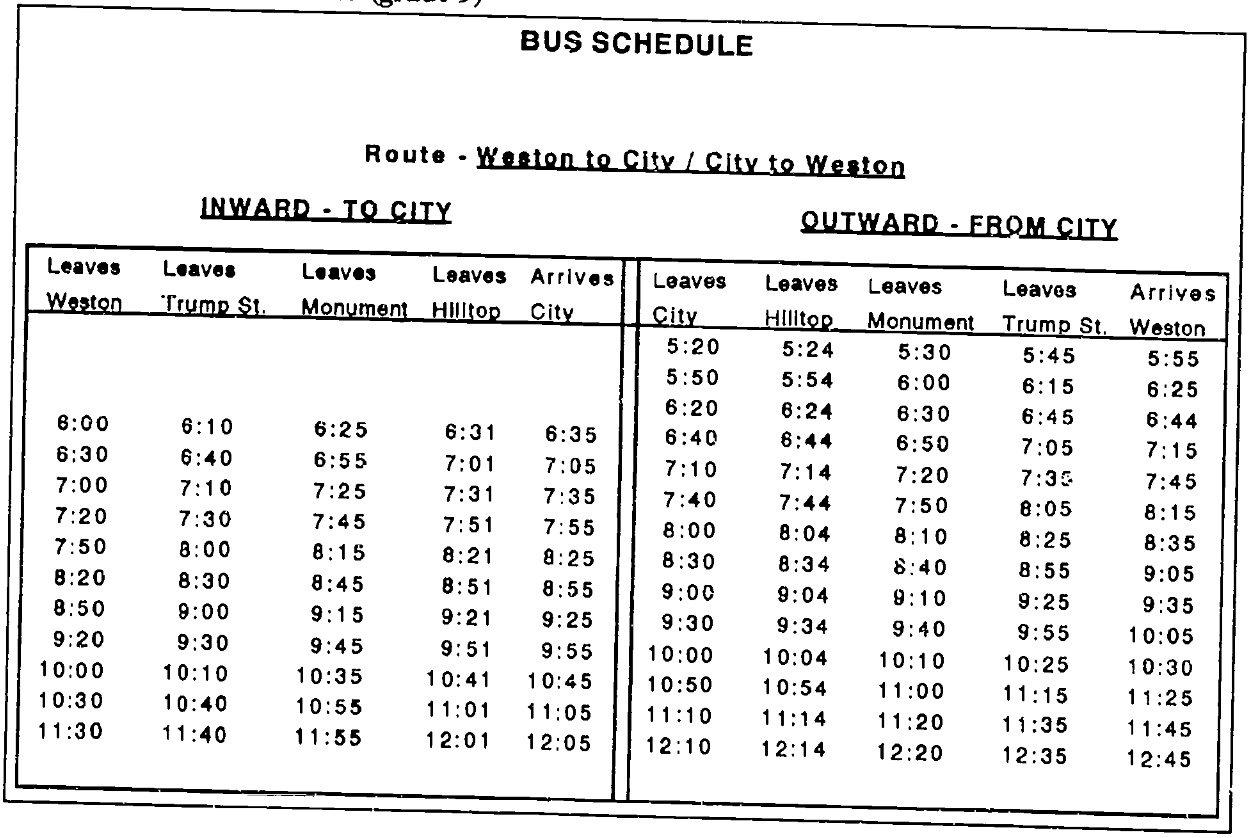
Another set of moderate tasks were associated with the *bus schedule* in Exhibit 6-12.

In contrast to the grade 4 bus schedule, which represented a combined list structure, the bus schedule appearing on the grade 9 document scale represented a nested list structure. Moreover, the grade 9 schedule consists of about five times as many items as the schedule for grade 4 (i.e., 140 versus 27). Tasks associated with this grade 9 schedule document ranged from 71 percent correct to 56 percent correct.

The easiest of these four open-ended questions was,

If you miss the 8:21 bus from Hilltop to City, what time would you arrive at City if you took the next bus?

Exhibit 6-12. *Bus schedule* (grade 9)



BUS SCHEDULE

Route - Weston to City / City to Weston

INWARD - TO CITY

Leaves Weston	Leaves Trump St.	Leaves Monument	Leaves Hilltop	Arrives City
6:00	6:10	6:25	6:31	6:35
6:30	6:40	6:55	7:01	7:05
7:00	7:10	7:25	7:31	7:35
7:20	7:30	7:45	7:51	7:55
7:50	8:00	8:15	8:21	8:25
8:20	8:30	8:45	8:51	8:55
8:50	9:00	9:15	9:21	9:25
9:20	9:30	9:45	9:51	9:55
10:00	10:10	10:35	10:41	10:45
10:30	10:40	10:55	11:01	11:05
11:30	11:40	11:55	12:01	12:05

OUTWARD - FROM CITY

Leaves City	Leaves Hilltop	Leaves Monument	Leaves Trump St.	Arrives Weston
5:20	5:24	5:30	5:45	5:55
5:50	5:54	6:00	6:15	6:25
6:20	6:24	6:30	6:45	6:44
6:40	6:44	6:50	7:05	7:15
7:10	7:14	7:20	7:35	7:45
7:40	7:44	7:50	8:05	8:15
8:00	8:04	8:10	8:25	8:35
8:30	8:34	8:40	8:55	9:05
9:00	9:04	9:10	9:25	9:35
9:30	9:34	9:40	9:55	10:05
10:00	10:04	10:10	10:25	10:30
10:50	10:54	11:00	11:15	11:25
11:10	11:14	11:20	11:35	11:45
12:10	12:14	12:20	12:35	12:45

To complete this question, respondents had to identify a time (a type of information representing a low level of difficulty). Note, however, that the type of match is rather difficult in that it requires respondents to make a three-feature match including "Leave," "Hilltop," and "8:21" and then search for a time based on the criterion of "next bus." Having accomplished this, respondents then had to make the two-feature match "Arrives" and "City" to identify the correct answer as "8:55." Although "Leaves Hilltop" is a label in the departure list, there are no times in this list that correspond to "8:21." Consequently, the level of distracting information associated with this task is not particularly high.

A more difficult task associated with the above bus schedule (with a *p*-value of 60 included the following open-ended question:

When does the <u>first</u> bus from Weston to City leave Monument each day?

To answer this question, respondents again had to identify a time. What made this question somewhat more difficult was the type of match required to answer it. To complete this item, respondents had to make a four-feature match between information in the question and information in the document. These matches were between "from" to "Leaves," "Weston" to "Weston," "To" to "Arrives," and "City" to "City." Having completed this, respondents then had to match on the criterion of the "first bus." Also note that because there is no specific time associated with any of the points of departure, and given that these points of departure are mentioned in the "from city" portion of the schedule, these points of departure represent a fairly high level of distracting information.

Finally, the most difficult item (with a *p*-value of 56) that related to the bus schedule included the following open-ended question:

Which is the <u>latest</u> bus you can catch from Monument to arrive at Weston before 11 o'clock?

To answer this question, respondents once again had to identify amount information. The type of match in this case also involved multiple feature matching. First, respondents had to match on "from" in the question to "Leave" in the document, as well as on "Monument" to "Monument," "arrive at" to "Arrive," and "Weston" to "Weston." Having completed this, respondents had to then find the latest bus in the list labeled "Arrives Weston" that arrived before "11 o'clock." In short, that represents quite a feat of feature matching. What makes this task somewhat more palatable, however, is the fact that in the event that respondents fail to extrapolate based on time information in Arrives Weston, they could still get this answer right for the wrong reason—namely, that "10:10" comes before "11:00" in the list Leaves Monument. As such, this task tended to represent a rather low level of distracting information.

Overall, the tasks on the grade 9 document scale represented a broader range of organizational structures than was noted on the grade 4 scale. On average, these documents contained almost five times as many items as the grade 4 documents. As with the grade 4 document scale, tasks on the grade 9 document scale tended to vary most in terms of their type of match and plausibility of distracting information. Tasks that were easy required subjects to locate a single item based on one or two literal or synonymous feature matches with few distracters. Tasks became more complex when they required respondents to perform independent cycle matches. Tasks continued to increase in difficulty as they required respondents to perform dependent cycle tasks involving more difficult levels of distracting information. The most difficult task required respondents not only to cycle but also to match on a large number of features between information in the survey item and in the document. This was often carried out in the presence of items representing a rather high level of distracting information.

Evaluating the Contribution of Variables to Task Difficulty

Grade 4 Analyses. To examine the extent to which structural and process variables contributed to the difficulty of the tasks on the grade 4 document scale for the total, white, and minority populations, a correlation matrix was computed (Table 6-12).

Table 6-12. Intercorrelations between process and structural variables, and document task difficulty (represented by *p*-values) for total, white, and minority grade 4 students

	p-value	1	2	3	4	5	6	7
1	Total							
2	White	.99						
3	Minority	.99	.98					
	Structural variables							
4	No. of items	-.37	-.38	-.35				
5	Type of document	-.19	-.19	-.21	.30			
	Process variables							
6	Type of information	-.41	-.41	-.38	.16	.05		
7	Type of match	-.76	-.75	-.78	.25	.12	.36	
8	Plausibility of distracting information	-.67	-.68	-.65	.24	-.25	.02	.34

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

As can be seen from this table, the variables having the highest correlation with total, white, and minority *p*-values were the process variables type of match (i.e., -.76, -.75, and -.78, respectively) and plausibility of distracting information (i.e., -.67, -.68, and -.65). Moderate correlations also were obtained with type of information (i.e., -.41, -.41, and -.38) and number of items (i.e., -.37, -.38, and -.35).

Regression Analyses. Based on these correlations, three separate regressions were run to determine the relative strength of the structural and process variables in predicting difficulty on the fourth grade document scale. In running these regressions, the same rules were followed that were identified earlier in this paper. The results of these regression analyses are shown in Table 6-13.

Table 6-13. Raw betas and *t*-ratios representing the regression of select process variables against total, white, and minority *p*-values for grade 4 documents tasks

Process variable	Total (df=19)		White (df=19)		Minority (df=20)	
	Beta	*t*-ratio	Beta	*t*-ratio	Beta	*t*-ratio
Type of match	-9.76	-4.54**	-8.84	-4.43**	-13.99	-5.54**
Plausibility of distracting information	-13.38	-4.67**	-12.74	-4.79**	-13.81	-3.83**
Type of Information	-8.60	-1.96*	-8.50	-2.09*	-	-
R^2 =	81%		81%		77%	

* $p < .05$; **$p < .01$; df = degrees of freedom.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Each of the three process variables contributed to the overall regression model for the total and white populations. Among minority grade 4 students, type of information did not remain in the model. Interestingly, plausibility of distracting information was again most salient among grade 4 students. In combination, these variables accounted for 77 to 81 percent of the variance in *p*-values.

Grade 9 Analyses. To examine the extent to which structural and process variables contributed to the difficulty of the tasks on the grade 9 document scale for the total, white, and minority populations, a correlation matrix was computed (Table 6-14).

Table 6-14. Intercorrelations between process and readability variables, and document task difficulty (represented by *p*-values) for total, white, and minority grade 9 students

	p-value	1	2	3	4	5	6	7
1	Total							
2	White	.99						
3	Minority	.98	.96					
	Structural variables							
4	No. of items	-.34	-.34	-.32				
5	Type of document	-.33	-.434	-.28	.30			
	Process variables							
6	Type of information	-.63	-.61	-.68	.09	.46		
7	Type of match	-.89	-.88	-.90	.34	.45	.75	
8	Plausibility of distracting information	-.64	-.65	-.59	.31	.15	.07	.45

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

As can be seen from this table, type of match had the highest correlation with total, white, and minority *p*-values (i.e., -.89, -.88, and -.90, respectively), followed by plausibility of distracting information (i.e., -.64, -.65, and -.59) and type of information (i.e., -.63, -.61, and -.68). Intercorrelations between structural and process variables tended to be low to moderate, with type of document correlating .46 with type of information, and .45 with type of match. The correlation between the two structural variables was somewhat low at .30. The intercorrelations among process variables ranged from low, between type of information and plausibility of distracting information (i.e., .07), to moderate between type of match and plausibility of distracting information (i.e., .45), to high between type of information and type of match (i.e., .75).

Based on these correlations, three regressions were run to evaluate the relative strength of the structural and process variables in predicting task difficulty along the grade 9 document scale. The results of these regressions are shown in Table 6-15.

In Table 6-15, we find that both type of match and plausibility of distracting information to be significant predictors of document task difficulty for the total, white, and minority populations.

Table 6-15. Raw betas and *t*-ratios representing the regression of selected process variables against total, white, and minority *p*-values for grade 9 document tasks

Process variable	Total (df=31)		White (df=31)		Minority (df=31)	
	Beta	*t*-ratio	Beta	*t*-ratio	Beta	*t*-ratio
Type of match	-8.79	-10.40**	-7.93	-9.65**	-11.53	-10.00**
Plausibility of distracting information	-3.80	-4.16**	-3.73	-4.20**	-3.74	-3.00**
R^2 =	87%		86%		85%	

**p < .01; df = degrees of freedom.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

These two variables accounted for 85 to 87 percent of the variance. Again, type of match received the largest beta weight indicating its relative importance for understanding factors accounting for difficulty on the grade 9 document scale.

Summary

Both the grade 4 and grade 9 document scales had similar characteristics in terms of the variables contributing to task difficulty. Although number of items correlated with overall difficulty, it did not remain in any of the regression analyses at either grade 4 or grade 9. In contrast, both plausibility of distracting information and type of match were significant predictors for both scales. In addition, type of information was also a significant predictor for the total and white populations for grade 4, but did not remain in the regression model for minority students.

As with the other two reading literacy scales, plausibility of distracting information was most salient at grade 4. While it still contributes to predicting overall difficulty at grade 9, type of match appears to be the more important predictor among grade 9 students.

6.6 Discussion

The purpose of this chapter was to describe and evaluate a set of variables that were hypothesized to underlie task difficulty among U.S. grade 4 and grade 9 students participating in the IEA Reading Literacy Study. To meet this objective, we have refined and extended a paradigm that was used with previous surveys of adult literacy (Kirsch and Jungeblut 1986; Kirsch and Mosenthal 1990; Kirsch, Jungeblut, and Campbell 1992).

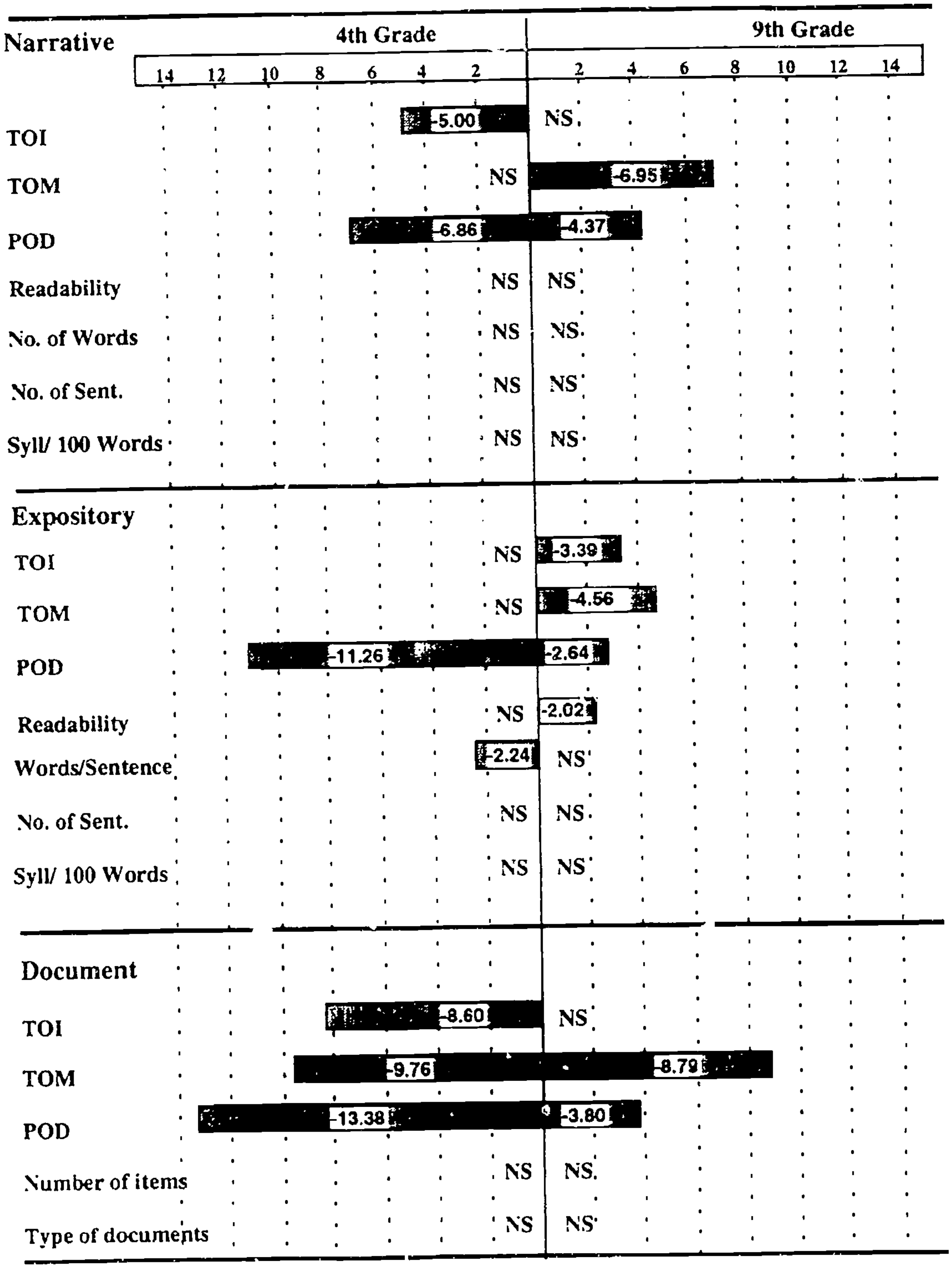
In applying this paradigm to the IEA Reading Literacy Study, tasks within a given scale were arrayed from easy to difficult. Next, they were characterized in terms of a set of variables that take into account both the nature of the material being read and what the reader is directed to do with this material. For the narrative and expository scales, material complexity was defined as (1) number of words in a passage; (2) number of sentences in a passage; (3) average number of words per sentence; (4) average number of syllables per 100 words; (5) average number of sentences per 100 words; and (6) overall readability level (Fry 1981; Klare 1984). For the document scale, complexity was defined in terms of the type of document structure represented and the number of items or specifics contained within the document.

Three variables were used to represent the type and level of processing associated with answering questions on the three scales. These included type of match, plausibility of distracting information, and type of information. These variables were designed to take into account various aspects of the interactions among questions, stimulus materials, and multiple-choice distracters.

To evaluate the relationships among these variables and task difficulty for the total, white, and minority populations, correlation matrices were computed and, based on the results obtained, regression models were constructed using a set of rules that help minimize overinterpretation of the findings. The results of the regression analyses for the total population are summarized in Figure 6-1. These results highlight important task characteristics for each literacy scale. These results can be compared and contrasted, across both grades and scales, to better understand what is being measured and what the various scores may mean.

Before interpreting the results shown in Figure 6-1, a word of caution is needed. Direct comparisons of regression weights can sometimes lead to an erroneous interpretation unless colinearity is examined. In regression analyses, one of two colinear variables can be found significant or nonsignificant as a function of minor changes in either variable, while the overall fit of the model as indicated by R^2 remains relatively unchanged. Such appears to be the case for the grade 4 narrative scale. Two variables (TOI and POD) receiving significant weights seem to have some colinearity with TOM.

Figure 6-1. Comparison of regression weights across grade levels for selected process and ⁻eadability variables by scales



TOI = type of information; TOM = type of match; POD = plausibility of distiacting information.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, Natio ıal Center for Education Statistics, 1991.

Hence, despite the fact that TOM had the highest zero-order correlation with p-values, it did not remain in the regression model. In contrast to the grade 4 results, analyses at grade 9 indicate thatTOI is not significantly related to p-values or the other two process variables. With this variable, colinearity is not a problem and the results are more directly interpretable.

This cautionary note notwithstanding, it is useful to examine the pattern of results shown in Figure 6-1. First, these data reveal that difficulty on the grade 4 and grade 9 literacy scales tended to be explained better by the process variables than by the readability or structure variables. Only on the expository scales do we find readability (i.e., average number of words at the grade 4 and overall readability level at grade 9) contributing significantly to task difficulty. What these results suggest is that the IEA Reading Literacy Study tended to be more of a measure of how well students were able to respond to different types of questions and distracters than how well they were able to read and understand a wide range of texts.

This does not mean that readability of text complexity is not an important aspect of the reading process. These results indicate that this aspect of reading was not well represented in the materials used in this assessment. It will be remembered, for example, that the average readability among the grade 4 narrative passages was 2.5, with no passages being above grade 4. Similarly, at grade 9 the average readability was grade 4.4, with no story rated above grade 6.

In examining the data shown in Figure 6-1, we also can see that plausibility of distracting information was a significant predictor on each of the six scales. By comparing across grades, we see that this variable was more salient for grade 4 students than for grade 9 students, suggesting that one of the most important aspects of the IEA survey was students' skill at being able to reduce uncertainty in light of distracting information presented to meet the criteria of the task. In contrast, type of match was more salient among the grade 9 scales receiving the largest beta weight in each regression model. These data suggest that at grade 9, tasks distinguished among students most in terms of whether they could match, cycle, integrate, or generate information based on the texts they were given to read. Type of information seemed less consistent in contributing to task difficulty across the two grades and three scales. At the grade 4 level, it contributed significantly to the narrative and document scales, while at grade 9, type of information predicted difficulty on the expository scale.

This pattern of results suggests that apart from plausibility of distracting information, the six scales seemed to tap into somewhat different processing dimensions. These findings reinforce the decision of the IEA to report results in terms of within-grade scoring rather than in terms of vertical scaling. Vertical scaling would have put both grades 4 and 9 onto common scales linked by common tasks. The underlying assumption would have been that each scale is basically measuring the same aspects of narrative reading.

These findings raise important questions and provide insights in interpreting the results of this assessment as well as in conceptualizing literacy in general. Why were certain dimensions of task difficulty more salient than others? Was this the intention of the survey designers or are the results more of a chance occurrence? Since test objectives are rarely assembled using a perspective such as the one outlined here, it is important to distinguish between the skills that designers intend to measure and what the test actually measures. Along this same line, it would be interesting to know whether the patterns of task characteristics observed in this survey are similar to other U.S. measures of reading literacy and whether or not they are similar to the instruments translated into other languages as part of the IEA assessment.

Another question that must be addressed is on what basis should we decide the principal dimensions of task difficulty in the future. In our opinion, this question is more important than any conclusions that we might offer in providing a final analysis of the IEA Reading Literacy Study. Clearly, this survey could have more systematically addressed the issue of readability across the three scales. We believe that using the paradigm described in this chapter, designers have a better understanding of factors that can be manipulated to affect difficulty along a scale in the future. This knowledge could be used to construct instruments that more systematically address dimensions believed to be important.

On a larger level, the analyses described in this chapter also address the issue of what constitutes literacy. The findings suggest that this question really has two components: an internal validity and an external validity. The internal validity component suggests that the variables that may affect performance may vary from one group of tasks to another as the saliency of task dimensions differ. However, despite differences between demonstrated proficiencies among groups of interest, the saliency of these variables should remain rather uniform across these populations. Similarly, the external validity component suggests that the constructs of literacy scales, again as defined by assessments and tests, may vary from assessment to assessment and from test to test. It should be noted, however, that comparability across assessments or tests should be based as much on an understanding of skills that are contributing to performance as on a statistical linking of distributions. This raises the issue of how we might proceed to build better interpretive bridges between various assessments and tests of literacy such that questions of policy might be more effectively addressed at all levels of education.

References

Beaton, A.E. (1987). *Implementing the new design: The NAEP 1983-84 technical report.* Princeton, NJ: Educational Testing Service.

Beaton, A.E. (1988). *Expanding the new design: The NAEP 1985-86 technical report.* Princeton, NJ: Educational Testing Service.

Carroll, J.B. (1993). Test theory and the behavioral scaling of test performance. In N. Frederikson (ed.), *Test theory for a new generation of tests.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Carver, R.P. (1983). Is reading rate constant or flexible? *Reading Research Quarterly*, 18, 190-215.

College Entrance Examination Board. (1982). *Degrees of reading power.* Readability Report, 1982-83 Academic Year. New York: The College Board.

Elley, W. (1992). *How in the world do students read?* The Hague: International Association for the Evaluation of Educational Achievement.

Fry, E. (1981). A partial reading model utilizing language unit size by frequency. In M.L. Kamil (ed.), *Directions in reading: Research and instruction.* Thirtieth Yearbook of the National Reading Conference, 103-107. Washington, DC: National Reading Conference.

Fry, E. (1977). Fry's readability graph: Clarifications, validity, and extension to level 17. *Journal of Reading,* 21, 242-252.

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

Kirsch, I.S., Jungeblut, A., and Campbell, A. (1992). *Beyond school doors: The literacy needs of job seekers served by the U.S. Department of Labor.* Princeton, NJ: Educational Testing Service.

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

Kirsch, I.S., and Jungeblut, A. (1986). *Literacy: Profiles of America's young adults.* NAEP Report No. 16-PL-01. Princeton, NJ: Educational Testing Service.

Klare, G.R. (1984). Readability. In P.D. Pearson, R. Barr, M. Kamil, and P. Mosenthal (eds.), *Handbook of reading research*, Vol. 1, 681-744. White Plains, NY: Longman.

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

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

Mislevy, R.J. (1991). Randomization-based inferences about latent variables from complex samples. *Psychometrika*, 56, 177-196.

Mislevy, R.J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association*, 80, 993-997.

Mosenthal, P.B., and Kirsch, I.S (1989). Lists: The building blocks of documents. *Journal of Reading,* 33, 58-60.

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

Mosenthal, P.B., and Kirsch, I.S. (1992). Understanding the constructs of prose, document, and quantitative literacy. Paper presented at the annual meeting of the National Reading Conference, San Antonio, Texas.

Mosenthal, P.B., and Kirsch, I.S. (1992). Types of document knowledge: From structures to strategies. *Journal of Reading*, 36, 64-67.

Appendix 1

Scoring Rules for Narrative and Expository Variables

Type of Information

Type of information requested refers to the nature of information that readers must identify to complete a narrative or expository question or directive. As Mosenthal and Kirsch[1] have noted, narratives and exposition consist of a rather restricted range of information types. These information types form a continuum of concreteness that was operationalized as follows for purposes of this analysis.

In sum, the scoring rules for type of information were as follows:

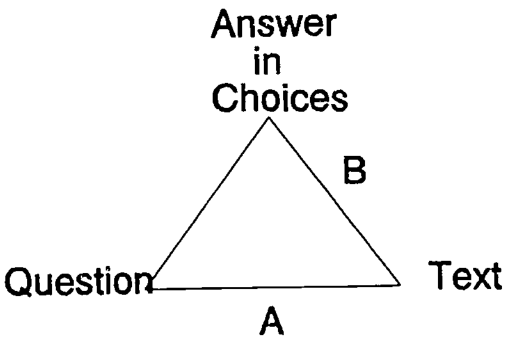
When the requested information is a:

- Person, animal, or thing, *score 1.*
- Amounts, time(s), attributes, actions, and locations, *score 2.*
- Manner, goal, purpose, condition, or predicate adjective, *score 3.*
- Cause, result, reason, evidence, or theme, *score 4.*
- Equivalent, *score 5.*

Type of Match

The variable *type of match* refers to the processes used to relate information in the question to information in the text to information in the choices. How this information is related is illustrated in the following figure.

The type of match triangle



The scoring rules for type of match (based on the figure) were as follows:

- When the relations between the question and text and between the text and the answer are both literal or synonymous, **score 1** for type of match.

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

- When the relation between the question and text or between the text and the answer requires a low text-based inference while the other requires literal or synonymous match, **score 2** for type of match.
- When the relation between the question and text and between the text and the answer both require a low text-based inference, **score 3** for type of match.
- When either the relation between the question and the text or between the text and the answer requires a high text-based inference, **score 4** for type of match.
- When the relation between the question, the text, and the answer requires the reader to generate the appropriate interpretive framework to relate the three, **score 5** for type of match.

Plausibility of Distracting Information

Plausibility of distracting information refers to whether or not an identifiable match exists between information in the question and the text, or between the text and the distracters, which makes it difficult for readers to identify the correct answer. There were five degrees of plausibility. In sum, the scoring rules for plausibility of distracting information were as follows:

In scoring for plausibility of distracting information:

- When there is no distracting information in the text, **score 1**;
- When distracters contain information that corresponds literally or synonymous to information in the text but not in the same paragraph as the answer, **score 2**.
- When distracters contain information that represents plausible invited inferences not based on information related to the paragraph in which the answer occurs, **score 3**.
- When one distracter in the choices contains information that is related to the information in the same paragraph as the answer, **score 4**.

  a) When two or more distracters in the choices contain information that is related to the information in the same paragraph as the answer, **score 5**; or

  b) When one or more distracters represent plausible inferences based on information outside the text, **score 5**.

Scoring for Document Variables

Type of match refers to the processes used to relate information in the question to information in the document. Unlike type of match in performing narrative and exposition tasks, type of match on the document scales did not appear to require making additional matches from the question to the document to the choices. Four types of document-matching strategies were identified: locate, cycle, integrate, and generate strategies. On average, these represented successively more difficult tasks. The rules for scoring document variables in terms of type of match were as follows:

Locate

- If match is 1 feature, literal or synonymous with 1 response, **score 1.**
- If match is 2 feature, literal or synonymous with 1 response, or 1 feature, literal or synonymous with 2-3 item response with number of responses specified in question or directive, or 1 feature, low text-based inference with 1 response, **score 2.**
- If match is 3 feature, literal or synonymous with 1 response, or 2 feature, literal or synonymous, with 2-3 item response with number of responses specified in question or directive, or 1 feature literal or synonymous with inferred mathematical operation, **score 3.**
- If match is 4 feature, literal or synonymous match with 1 response, or 2 feature, literal or synonymous match with 2-3 item response with number of responses not specified in the question, or 1 feature, high text-based inference with 1 response, **score 4.**
- If match is 4 feature literal or synonymous with conditional information, **score 5.**

Cycle

- If match involves a series of 1 feature, literal or synonymous independent matches, **score 2.**
- If match involves a series of 1 feature, literal or synonymous dependent matches, or 2 feature, literal or synonymous independent matches, or 1 feature, literal or synonymous indepenuent matches that include counting with 3 or more numbers, **score 3.**
- If match involves a series of 2 feature, literal or synonymous dependent matches, **score 4.**

Integrate

- If match involves 2 or more 1 feature matches that are compared or contrasted, or the integration of text information to answer document information, **score 3.**
- If match involves 2 or more 2 feature matches that are compared, **score 4.**
- If match involves 2 or more 2 feature matches that are contrasted, **score 5.**

Generate

- If match requires respondents to infer a causal pattern or trend, or make a unique inference based on prior knowledge or highly conditional information, **score 5.**

Plausibility of Distracting Information

Plausibility of distracting information. This variable has to do whether or not there are features from a question, or directives given, or requested information that appear in the document but, once matched or identified, do not yield the correct requested information. In sum, the rules for scoring plausibility of distracting information were as follows:

When *plausible distracters*:

- Do not appear for either given or requested information, ***score 1***.
- For either given or requested (but not both) appear in a node other than the answer node, ***score 2***.
- For both given and requested appear in different nodes other than the answer node, ***score 3***.
- For both given and requested, both appear in the same node other than the answer node, ***score 4***.
- For both given and requested appear in the same node as the answer, ***score 5***.

Appendix 2

Characteristics of Tasks Used on the IEA Reading/Literacy Study

Table 1. Characteristics of the variables used in the analyses of grade 4 narrative tasks

Task no.	Narrative	*p*-value total	Type of match	Plausibility of districting information	Type of information	Number of words	Number of sentences	Number of words per sentence	Syllables per 100 words	Number of sentences per 100 words	Readabilit grade leve
47	Bird . .	94	1	1	2	292	25	12	122	9.6	3rd
62	Dog . .	89	1	1	1	703	88	8	130	8.5	4th
4	Shark .	88	1	1	4	467	46	10	115	12.0	1st
67	Dog .	86	2	2	2	703	88	8	130	8.5	4th
64	Dog . . .	82	3	3	3	703	88	8	130	8.5	4th
46	Bird .	80	2	2	2	292	25	12	122	9.6	3rd
6	Shark . .	81	2	1	4	467	46	10	115	12.0	1st
65	Dog . . .	78	3	3	3	703	88	8	130	8.5	4th
39	Grandpa .	77	2	2	2	312	30	10	111	8.3	2nd
5	Shark . .	76	2	2	4	467	46	10	115	12.0	1st
7	Shark . .	76	1	1	4	467	46	10	115	12.0	1st
38	Grandpa .	77	3	2	4	312	30	10	111	8.3	2nd
44	Bird . . .	74	2	3	2	292	25	12	122	9.6	3rd
35	Grandpa .	73	2	4	2	312	30	10	111	8.3	2nd
66	Dog . . .	70	4	2	3	703	88	8	130	8.5	4th
43	Bird . . .	69	4	3	4	292	25	12	122	9.6	3rd
37	Grandpa .	70	3	3	4	312	30	10	111	8.3	2nd
63	Dog . . .	68	3	3	3	703	88	8	130	8.5	4th
8	Shark . .	67	2	2	4	467	46	10	115	12.0	1st
36	Grandpa .	66	3	3	4	312	30	10	111	8.3	2nd
45	Bird . .	52	4	3	5	292	25	12	122	9.6	3rd
40	Grandpa .	52	4	4	4	312	30	10	111	8.3	2nd

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 2. Characteristics of the variables used in the analyses of grade 9 narrative tasks

Task no.	Narrative	*p*-value total	Type of match	Plausibility of districting information	Type of information	Number of words	Number of sentences	Number of words per sentence	Syllables per 100 words	Number of sentences per 100 words	Readabil grade le
7	Shark . .	96	1	1	4	467	46	10	115	12.0	1st
8	Shark . .	94	1	2	4	467	46	10	115	12.0	1st
10	Shark . .	93	1	1	4	467	46	10	115	12.0	1st
2	Fox	93	1	1	3	422	28	15	118	6.0	5th
9	Shark . .	93	2	1	4	467	46	10	115	12.0	1st
4	Fox	93	2	2	2	422	28	15	118	6.0	5th
5	Fox	91	3	2	3	422	28	15	118	6.0	5th
11	Shark . .	89	2	2	4	467	46	10	115	12.0	1st
37	Magician	80	3	2	3	712	52	14	120	7.5	4th
3	Fox	79	2	4	4	422	28	15	118	6.0	5th
47	Angel . .	77	3	2	3	1,143	95	12	130	7.3	6th
1	Fox	75	2	4	3	422	28	15	118	6.0	5th
49	Angel . .	75	3	2	2	1,143	95	12	130	7.3	6th
50	Angel . .	75	4	3	3	1,143	95	12	130	7.3	6th
31	Magician	73	4	3	4	712	52	14	120	7.5	4th
48	Angel . .	73	4	2	2	1,143	95	12	130	7.3	6th
34	Mute . . .	71	3	3	4	605	53	11	127	7.5	5th
36	Magician	71	2	5	2	712	52	14	120	7.5	4th
45	Angel . .	71	4	3	4	1,143	95	12	130	7.3	6th
46	Angel . .	71	4	2	3	1,143	95	12	130	7.3	6th
33	Magician	70	4	3	4	712	52	14	120	7.5	4th
35	Magician	68	4	3	4	712	52	14	120	7.5	4th
34	Magician	65	4	5	4	712	52	14	120	7.5	4th
36	Mute . . .	64	4	2	5	605	53	11	127	7.5	5th
33	Mute . . .	60	4	4	5	605	53	11	127	7.5	5th
44	Angel . .	59	5	5	4	1,143	95	12	130	7.3	6th
32	Magician	57	4	5	3	712	52	14	120	7.5	4th
32	Mute . . .	56	4	4	5	605	53	11	127	7.5	5th
35	Mute . . .	44	5	4	4	605	53	11	127	7.5	5th

SOURCE: IEA Reading Literacy Study, U.S. Nati)nal Study data, National Center for Education Statistics, 1991.

Table 3. Characteristics of the variables used in the analyses of grade 4 expository tasks

Task no.	Narrative	*p*-value total	Type of match	Plausibility of districting information	Type of information	Number of words	Number of sentences	Number of words per sentence	Syllables per 100 words	Number of sentences per 100 words	Readabilit grade leve
42	Postcard .	97	1	1	1	56	7	8	112	12.5	1st
41	Postcard .	96	1	2	1	56	7	8	112	12.5	1st
56	Walrus . .	94	1	1	2	207	13	16	127	6.5	6th
55	Walrus . .	93	1	1	2	207	13	16	127	6.5	5th
3	Quicksand	90	1	2	2	141	11	13	119	7.4	4th
2	Quicksand	83	1	2	2	141	11	13	119	7.4	4th
57	Walrus . .	79	1	3	1	207	13	16	127	6.5	6th
58	Walrus . .	79	2	3	1	207	13	16	127	6.5	6th
59	Walrus . .	79	2	3	2	207	13	16	127	6.5	6th
1	Quicksand	79	3	2	3	141	11	13	119	7.4	4th
29	Trees . . .	74	2	2	3	389	20	19	116	5.8	5th
33	Trees . . .	72	3	3	2	389	20	19	116	5.8	5th
60	Walrus . .	59	1	4	3	207	13	16	127	6.5	6th
25	Marmot .	55	2	3	2	228	11	21	120	5.5	6th
31	Trees . . .	55	3	4	4	389	20	19	116	5.8	5th
30	Trees . . .	50	3	3	4	389	20	19	116	5.8	5th
28	Marmot .	45	2	4	2	228	11	21	120	5.5	6th
26	Marmot .	42	2	3	1	228	11	21	120	5.5	6th
27	Marmot .	41	3	4	4	228	11	21	120	5.5	6th
32	Trees . . .	35	5	5	5	389	20	19	116	5.8	5th
34	Trees . . .	34	5	5	5	389	20	19	116	5.8	5th

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 4. Characteristics of the variables used in the analyses of grade 9 expository tasks

Task no.	Narrative	*p*-value total	Type of match	Plausibility of distracting information	Type of information	Number of words	Number of sentences	Number of words per sentence	Syllables per 100 words	Number of sentences per 100 words	Readabili grade lev
26	Laser . . .	92	1	2	2	830	48	17	119	6.3	5th
3	Paracutin	90	2	1	3	270	20	14	109	6.3	4th
16	Marmot .	85	2	3	2	228	11	21	120	5.5	6th
2	Paracutin	84	2	3	1	270	20	14	109	6.3	4th
19	Marmot .	83	2	4	2	228	11	21	120	5.5	6th
27	Laser . . .	82	2	2	4	830	48	17	119	6.3	5th
31	Laser . . .	82	1	4	3	830	48	17	119	6.3	5th
17	Marmot .	80	2	3	1	228	11	21	120	5.5	6th
28	Laser . . .	80	2	2	1	830	48	17	119	6.3	5th
4	Paracutin	80	3	3	3	270	20	14	109	6.3	4th
42	Smoke . .	76	3	3	3	368	16	23	146	4.8	9th
5	Paracutin	74	3	3	3	270	20	14	109	6.3	4th
6	Paracutin	74	3	3	4	270	20	14	109	6.3	4th
40	Literacy .	74	2	4	3	298	19	16	123	6.0	6th
39	Literacy .	72	2	2	5	298	19	16	123	6.0	6th
1	Paracutin	71	4	3	3	270	20	14	109	6.3	4th
18	Marmot .	69	2	4	4	228	11	21	120	5.5	6th
41	Smoke . .	69	2	4	4	368	16	23	146	4.8	9th
29	Laser . . .	66	2	3	5	830	48	17	119	6.3	5th
30	Laser . . .	65	4	5	4	830	48	17	119	6.3	5th
38	Literacy .	64	4	3	4	298	14	16	123	6.0	6th
37	Literacy .	60	5	5	4	298	14	16	123	6.0	6th
38	Smoke . .	58	2	5	4	368	16	23	146	4.8	9th
39	Smoke . .	50	4	5	4	368	16	23	146	4.8	9th
40	Smoke . .	51	4	5	4	368	16	23	146	4.8	9th
43	Smoke . .	46	5	3	5	368	16	23	146	4.8	9th

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 5. Characteristics of the variables used in the analyses of grade 4 document tasks

Task no.	Document	*p*-value	Matching given information	Plausibility of distracting information	Type of requested information	Number of items	Document type
11	Bottles	97	1	2	1	18	2
18	Content	97	1	2	1	22	2
49	Island	96	1	2	1	15	3
9	Bottles	96	1	2	1	18	2
19	Content	94	1	2	2	22	2
12	Bottles	93	2	2	1	18	2
17	Content	93	1	2	2	22	2
48	Island	92	1	2	1	15	3
13	Buses	91	1	2	2	27	2
24	Temperature . .	85	1	2	1	29	3
54	Timetable	84	2	2	2	40	3
51	Island	82	2	2	2	15	3
52	Timetable	81	2	2	1	40	3
20	Temperature . .	78	1	2	2	29	3
50	Bottles	75	3	2	2	18	2
50	Island	70	3	2	1	15	3
14	Buses	70	3	3	1	27	2
53	Timetable	65	3	3	2	40	3
22	Temperature . .	49	3	2	2	29	3
23	Temperature . .	49	3	2	2	29	3
16	Buses	46	1	5	1	27	2
21	Temperature . .	36	3	3	2	29	3
15	Buses	29	5	4	2	27	2

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 6. Characteristics of the variables used in the analyses of grade 9 document tasks

Task no.	Document	*p*-value	Matching given information	Plausibility of distracting information	Type of requested information	Number of items	Document type
6	Card	98	1	1	1	16	2
8	Card	98	1	1	1	16	2
7	Card	97	1	1	1	16	2
28	Aspirol	96	1	2	2	106	2
27	Temperature . .	95	1	2	1	29	3
9	Card	95	1	1	1	16	2
12	Card	95	1	2	1	16	2
23	Temperature . .	91	1	2	2	29	3
29	Aspirol	90	1	2	1	106	2
19	Weather	89	2	2	1	154	2
21	Job	87	1	2	1	41	1
21	Weather	86	2	2	1	154	2
16	Directions	83	3	1	3	21	1
17	Directions . . .	83	3	1	3	21	1
23	Predator	82	3	3	2	68	4
26	Temperature . .	82	3	2	2	29	3
10	Card	81	1	1	1	16	2
20	Weather	81	3	2	2	154	2
24	Temperature . .	81	3	3	2	29	3
20	Job	79	2	2	1	41	1
25	Temperature . .	78	3	2	2	29	3
14	Resources	77	3	2	3	61	3
13	Resources	75	3	1	3	61	3
18	Directions	72	3	1	3	21	1
14	Bus	71	3	2	3	140	4
11	Card	61	3	4	3	16	2
13	Bus	60	4	4	2	140	4
25	Predator	58	5	2	4	68	4
30	Aspirol	58	3	5	2	106	2
15	Resources	57	4	5	3	61	3
15	Bus	56	5	2	3	140	4
22	Job	54	4	5	1	41	1
22	Weather	53	4	4	2	154	2
24	Predator	48	5	2	5	68	4

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

7 Creating a Measure of Reading Instruction

Marilyn R. Binkley, Linda M. Phillips, and Stephen P. Norris

7.1 Introduction

The very last sentence of *Becoming a Nation of Readers: The Report of the Commission on Reading*, proclaims in bright red ink that "America will become a nation of readers when verified practices of the best teachers in the best schools can be introduced throughout the country" (Anderson et al., 120). While this is clearly an ideal worth striving for, this statement presupposes that it is clear what the best practices might be in all cases. However, this very point has been debated in seemingly endless ways for centuries.

Similarly, how best to capture definitions and descriptions of best practices is also unresolved. While it appears reasonable to believe that it might be possible to observe successful teachers and describe what they have done, the question of how to validate empirically the success of the practice in other settings remains. One might collect survey data associated with a common or standard student outcome measure. Assuming that the survey questions are appropriate, one might expect to create models of effective instruction.

In principle, cross-national survey studies conducted by International Association for the Evaluation of Educational Achievement (IEA) and the International Assessment of Educational Progress (IAEP), as well as national surveys such as National Assessment of Educational Progress (NAEP), have taken this approach. These studies have tried to measure both the outcome of instruction and the instruction itself. The measures of instruction, along with other important descriptive variables, are then used to explain differences in outcomes across nations or subpopulations.

While this approach is currently accepted practice, the usefulness of the reported instructional data is questionable. Reports of associations between an isolated instructional practice and achievement, on an item-by-item or construct-by-construct basis, do not always provide sufficient insight into the context of instruction. Knowing that one nation or subpopulation did more or less of a particular thing is of limited value. Data users are frequently at a loss as to how to balance one finding against another and to reasonably construct an instructional program. Further, these reported associations often appear to contradict the research literature reported in journals. Rather, data users, curriculum specialists, and practitioners are more concerned with the mix of instructional practices—the combinations that work. Therefore, it might be more useful to step back and group items into more meaningful units for analysis. These units should, in principle, correspond to theories of instruction.

In this paper we suggest that there might be a more meaningful way to analyze and report on instructional variables. It is our position that, in ways that are analogous to the construction of a cognitive test, one might develop survey specifications—a blueprint or theoretic frame—for measuring instruction. The theoretic frame, in a large-scale national or cross-national survey, would be *all inclusive* in that it would capture the widest possible range of instructional practices. Then by placing items into the theoretic frame, we could begin to explore whether there were more or less effective instructional programs.

Using instructional data from the IEA Reading Literacy Study, we hope to demonstrate the strength of the procedures we would propose for future work. Along the way, we also will point out other design issues that need attention as well. We start by reporting on the association between three instructional items and achievement—the standard practice. In these three examples we hope to make clear some of the limitations of this approach. We contrast this approach with our strategy. We propose a theoretic frame, based on the reading research literature, into which we might place constructs derived from the data. These constructs were developed from groups of items drawn from the Teacher Questionnaire of the Reading Literacy Study, which were subjected to exploratory factor analyses. The conclusions that we can then draw are based on the intersection of the reported data and the theoretic frame. In this way, we marry exploratory statistical procedures with instructional research theory and consequently arrive at stronger interpretive understandings about the current state of instructional practice.

Because the theoretic frame we propose was developed after the survey had been conducted, and because the items were not specifically written to meet the criteria of definitions established in that frame, our discussion of findings is limited and tentative. Our point is to demonstrate a methodology that we believe might be more effective in designing future work and ultimately that could better inform practice.

In effect, this paper has two messages. One draws an interpretation of the IEA Reading Literacy Study data on reading instruction in the United States. The second, and for us the more crucial point, emphasizes the methodological issues that need further consideration in future studies so that findings can be based on a more rigorous and solid footing. In fact, the overall point of the paper does not depend on the particular interpretation of reading theory that we put forth. Rather, alternative interpretations of the data might be just as valid. However, we strongly believe that without the initial statement of a theoretic frame, one cannot be sure what was measured.

7.2 The Available Data

The IEA Reading Literacy Study provided an unusually good opportunity to work with data about instruction.[1] Within the United States, the glimpse into the "black box" that constitutes grade 4 reading instruction was provided by 190 items related to classroom reading instructional activities on the IEA Reading Literacy Study Teacher Questionnaire. Over 300 teachers from 167 schools responded. Thus, when used with the proper weights, this sample constitutes a nationally representative sample of grade 4 reading classes and their teachers in the United States.

Unlike other academic disciplines where a clearly specified list of topics to be covered at a particular grade level could be produced, or where particular instructional strategies could be associated with topics, reading instruction is much more amorphous. The 190 items represented views about issues in reading instruction, instructional activities that represent common practice, attributes of numerous instructional programs, and some standard policy stances relating to materials, time allocation, and class groupings.

[1] For a complete description of the IEA Reading Literacy Study, see Binkley and Rust (1994).

The 190 items did not necessarily represent the full range of possible instructional strategies. The items had been generated through a consensus process whereby representatives from each of the participating countries could add or delete items through open discussion. Thus, there might have been some undetected or unanticipated bias in the array of items. While each item tended to represent a discrete activity, any of these activities might be associated with one or more philosophical approaches to reading instruction. For example, many programs call for such activities as "comparing pictures and stories." Therefore, it would be difficult to attribute a single item to a particular philosophical approach.

7.3 Item-Level Analysis

For the purposes of providing examples of standard reporting practices with regard to instructional data, we have chosen three items from the Teacher Questionnaire that correspond closely to items included in NAEP. This makes it possible to compare data and to examine findings more generally.

The items were drawn from question 30 of the grade 4 Teacher Questionnaire (see the appendix to this chapter), which asked teachers to state how often their students were typically involved in specific reading activities. The activities ranged from learning letter-sound relationships, to making generalizations and inferences, to reading in other subject areas. The three items reported on ask about phonics instruction, the use of writing in response to reading, and the provision of time to read silently in class. The criteria for the selection of these items was based solely on the correspondence to NAEP. We look first at the frequency of learning letter-sound relationships and/or phonics.

Example 1: Frequency of Phonics Instruction

As concisely stated by Lundberg and Linnakyla (1992, 2), phonics most often refers to a stage-wise, objective-based strategy where specific decoding skills are taught with the aim of full mastery within the first 2 school years. The purpose of this instruction is to make certain that children understand the fundamental nature of the alphabetic principle and that they acquire ready familiarity with frequent words and with spelling patterns and their mapping to sounds. (Also see Adams 1990; Chall 1967; and Anderson et al. 1985.)

Within the context of the IEA Reading Literacy Study Teacher Questionnaire, teachers were asked not about their adherence to the more global theoretic positions most often associated with phonics instruction, but rather how frequently they specifically included instruction in letter-sound relationships in their classroom activities.

While the data reported in Table 7-1 suggest that the more frequently the teacher reports the use of phonics instruction, the lower the mean score of students for each of the three scales, there are no statistically significant differences among the means of groups of students whose teachers reported using this practice to varying extents for any scale. The level of statistical significance must be controlled for the fact that many comparisons are involved (six for each of the three scales), with the result that even the apparent differences between the extreme groups in each case are not significant.

This contrasts with the results of the 1992 NAEP reading assessment, in which grade 4 students of teachers who reported a heavy emphasis on the use of phonics had considerably lower mean reading achievement than those whose teachers reported a moderate emphasis. This second group, in turn, had slightly lower mean achievement than students of teachers who reported little or no emphasis on phonics.

Table 7-1. Class mean reading proficiency scores, by frequency of phonics instruction

Frequency	Percent	Narrative	Expository	Document
Almost never	22.5	564 (7.0)	550 (6.7)	562 (5.9)
About once or twice a month	15.5	556 (6.4)	538 (5.8)	553 (6.8)
About once or twice a week	35.0	557 (5.4)	542 (4.4)	551 (3.4)
Almost every day	27.0	550 (6.4)	535 (5.3)	545 (4.4)

NOTE: Numbers in parentheses are standard errors.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Although the IEA and NAEP items have very similar wordings, in the IEA Reading Literacy Study the question asks about a specific subset of activities, while the question in NAEP places phonics in a context where it represents an entire instructional approach. In contrast to the other instructional approaches (literature-based reading, integration of reading and writing, and whole language) considered in NAEP, phonics alone would tend to be limited only to beginning reading instruction[2] in its intent. It is the only approach included in that list that focuses solely on the decoding aspect of reading instruction. Therefore, we might expect that the NAEP data would show larger effects associated with this instructional stance than the IEA data, gathered via the questions that refer solely to specific letter-sound relationships.

In interpreting the data in both the IEA Reading Literacy Study and NAEP, we note that where grade 4 teachers report high levels of phonics instruction, students tend to have lower achievement scores. Given that this activity or instructional approach is recommended for beginning or delayed readers (Stahl 1992), and that it should be suspended after grade 2 if students demonstrate adequate abilities (Anderson et al. 1985), it seems that the students who are receiving phonics instruction may have entered grade 4 with lower reading abilities. Consequently, we can draw no conclusions about the efficacy of phonics instruction based on these data.

Example 2: Writing in Response to Reading

When reading is considered in the larger context of language usage or communication, the interrelationships between speaking and listening and reading and writing become more prominent. There has been a growing emphasis on tying reading and writing more closely together because of the natural ways in which they complement each other and call upon related cognitive capacities (Loban 1963; Durkin 1988; Moffett and Wagner 1983; Lewin 1992; Farr et al. 1991; Reid 1990; Clay 1985). Strategies are multiplying for having students respond in ways that more closely emulate what people more generally do when reading. Consequently, more and more children are being asked to write summaries, to keep a personal reading journal, or to write to a friend about a book and their reactions to it (McGinley and Madigan 1990).

Teachers were asked to report the frequency with which they ask their students to respond in writing to something they have read. As seen in Table 7-2, the data show that the group of students whose teachers reported that students almost never write in response to something that they have read is too small to draw any meaningful conclusions. There are no differences in mean achievement among students with teachers in the other three groups, who reported with varying frequency that they use this practice.

[2] This does not mean that advocates of phonics instruction believe that once children have learned phonics, reading instruction is complete. Rather, they then advocate continued reading instruction to facilitate comprehension.

Table 7-2. **Class mean reading proficiency scores by teacher-reported frequency of written responses to reading**

Frequency	Percent	Narrative		Expository		Document	
Almost never	0.1	497	(34.2)	500	(27.3)	521	(20.4)
About once or twice a month	18.2	559	(7.5)	543	(7.0)	559	(7.2)
About once or twice a week	48.8	558	(4.2)	543	(5.7)	551	(4.0)
About once a day	32.4	557	(5.9)	541	(5.2)	551	(5.4)

NOTE: Numbers in parentheses are standard errors. Percentages may not add to 100 due to rounding.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Similarly, the 1992 NAEP reading assessment showed no significant difference in mean reading achievement among grade 4 students whose teachers reported that they write almost every day in response to something they have read, or write at least once a week, or write less than weekly.

Can we conclude that writing in response to reading has no effect on reading achievement? In their extensive review of the literature related to the integration of reading and writing, Tierney and Shanahan (1991) would likely argue that the data are inconclusive at this time because, despite methodological advances in exploring this area, "the research on reading-writing relationships should be viewed as still in its infancy." Consequently, the instructional strategies that have been implemented to date might in fact be misguided or misused. Further, it is striking that an overwhelming majority of the research conducted in this area has focused on students of high school or college age. Therefore, one would wonder about the ability of grade 4 students to successfully use similar approaches. Alternatively, we have no measure as to whether beginning this type of instruction at this grade level results in higher achievement in successive grades. Once again, there is little indication from these data as to an appropriate policy decision.

Example 3: Silent Reading in Class

Numerous researchers have found that how much a child reads is highly associated with various measures of reading achievement (Anderson, Wilson, and Fielding 1988; Greaney 1980; Greaney and Hegarty 1984; Kirsch and Guthrie 1984; Krashen 1988; Heyns 1978). However, Thurlow et al. (1984) have pointed out that perhaps too little actual sustained reading may be occurring in school as a part of reading instruction. In the Reading Literacy Study, teachers were asked how frequently their students silently read in class.

The data in Table 7-3 show that the groups of students whose teachers reported that they read silently either almost never or once or twice a month are too small to draw any meaningful conclusions. There are no significant differences across the scales in the mean achievement of the students whose teachers reported silent reading once or twice a week and those whose teachers reported silent reading almost every day.

In comparison, the 1992 NAEP reading assessment showed that grade 4 students whose teachers reported that their students read silently almost every day had somewhat higher mean achievement than did students whose teachers reported that their students read silently at least once a week. The group of students whose teachers reported that they read silently less than weekly was very small.

There are two things that stand out when considering how to interpret these findings. First, in contrast to, or perhaps as a consequence of, the warning that not enough sustained reading was going on in classrooms during the early 1980s, we note that the IEA survey and NAEP surveys in the 1990s find about 98 percent of the teachers reporting silent reading in class at least once a week. Second, there does not seem to be much difference in performance associated with whether students read at least once or twice a week or almost every day.

Table 7-3. Class mean reading proficiency scores, by frequency of silent reading in class

Frequency	Percent	Narrative	Expository	Document
Almost never	0.1	484 (-)	469 (-)	498 (-)
About once or twice a month	1.8	561 (14.5)	559 (4.3)	556 (15.9)
About once or twice a week	12.8	547 (12.7)	535 (11.0)	546 (9.7)
Almost every day	85.4	560 (3.2)	544 (3.4)	554 (3.1)

(-) = Sample size is too small to computer standard error.

NOTE: Numbers in parentheses are standard errors. Percentages may not add to 100 due to rounding.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

The real question, however, is whether this use of instructional time has improved reading achievement. Pearson and Fielding (1991) were surprised that methods designed to increase the amount of uninterrupted reading children do in class had met with limited success. They reported on three methods—book floods (students are inundated with very easy access to numerous books), use of a classroom library, and sustained silent reading. The findings showed that book floods were most successful in settings where few books were available prior to the intervention (Elley and Mangubhai 1983; Ingham 1982; Holdaway 1979). The work of Morrow and Weinstein (1986) indicated that the voluntary free-time reading of grade 2 students could be increased if the classroom library was well stocked and if there were related book enjoyment activities. However, this did not seem to transfer to increased out-of-school reading. Sustained silent reading seemed most successful when it was accompanied by peer and teacher interaction about books (McCracken 1971; Cline and Kretke 1980; Collins 1980; Manning and Manning 1984). Once again, there is no clear indication of what an appropriate policy decision might be.

In considering each of these three examples, it is striking that each one by itself does not seem to enlighten the reader. The question and the data, without the surrounding information from a review of the pertinent literature, would often lead the reader to perhaps an inaccurate conclusion. When placed in the context of the literature, we are again left with more questions than we might like. We are left wondering if we have sufficient information on which to base a judgment, or whether we have asked the right question in the right way. Clearly, we could continue to review each of the remaining 187 instructional variables in the same way. However, how it might help to improve our understanding of instructional practices is unclear.

One alternative might be to ask teachers questions about instruction at a slightly more abstract level. We might ask, as NAEP did, about their emphasis on various instructional programs such as literature-based reading, integration of reading and writing, and whole language. However, without explanation and definition of what is intended by these terms, teachers are apt to respond with a more widely divergent array of activities than if asked to respond with regard to more specific activities.

If we continue to ask questions at the same level of specificity, is there some way we might group the questions and aggregate the responses so that we might get some measure of adherence to a particular program? In considering this issue, we wondered what a teacher might do with information regarding a particular instructional activity. When presented with a particular activity as a model of good instruction, teachers are likely to evaluate it against what they are already doing and against the theory of reading they have espoused. Although all teachers use most activities listed in the questionnaire, how they put them together and what aspects they emphasize are dependent on their own implicit theory of reading.

Can we then look at how teachers tend to group instructional activities and the correspondence with theoretic approaches to understanding reading and reading instruction? To do so, we must define theories of reading. It is important to note, however, that theories of reading are a subset of theories of learning. So, there is a clear connection to larger currents of thought about theories of learning, curriculum, and instruction.

7.4 Moving Beyond the Item by Establishing a Framework for Theories of Reading

In practice, teachers do not just focus their programs on a single, discrete activity. They tend to be somewhat more goal oriented and to follow or create a pattern intended to move the student to a greater accomplishment than could any single activity. The use of a combination of activities should, in principle, reflect an implicit theory of learning and perhaps a more explicit theory of reading.

Because theories of reading have been promulgated since Plato, they are quite numerous. Consequently, a way to characterize these theories systematically is needed. Logically, certain attributes are common to all theories of reading. We have identified seven global attributes that make it difficult to contrast theories of reading.

- Reading theories are evolutionary. No reading theory stands on its own or comes from nowhere. Each theory draws on previous conceptualizations and modifies these to suit particular ends.

- Reading theories are partial. No reading theory says all that can be said about reading. The complexity of reading is immense, and new insights into its nature are being made constantly.

- Each reading theory has a focal phenomenon. No reading theory attempts to do everything. Rather, each has a specific focus. For example, some theories concentrate on word recognition processes (Gough 1972, 1985; LaBerge and Samuels 1974, 1985; Stanovich 1991; Rumelhart 1985), while others concentrate on comprehension, almost to the exclusion of letter-level processes (Just and Carpenter 1985; Kintsch and van Dijk 1978).

- Reading theories with the same focal phenomenon take variant positions. For instance, within the group of theories that have the focal phenomenon of word recognition, the emphasis can vary. Some theories emphasize strict linear processing (e.g., Gough 1972, 1985); others emphasize interactive processing (e.g., Rumelhart 1985; Stanovich 1991; LaBerge and Samuels 1974, 1985).

- Versions of each reading theory range from moderate to extreme. Within each theory there are variant positions, with some proponents holding extreme positions and others holding moderate ones. For example, some theories maintain that reading is acquired naturally, and some proponents argue that teaching children to read any more than they are taught to speak is a cause of reading failure (Goelman, Oberg, and Smith 1984). On the other side are those who maintain that while learning to read is natural, some instruction can help to prevent reading failure (Applebee and Langer 1983).

- Reading theories are complementary. They tend to be different rather than contradictory. They tend not to contradict in many instances because they have different focal phenomena or because they concentrate on a different aspect of the same phenomenon. Hence, different theories tend to contribute to an understanding of different aspects of reading. Taken together, they provide a more comprehensive view than any of them taken singly.

- Not all reading theories look equally closely at reading. Some theories take a local, microscopic look; others a global, macroscopic one. Word-recognition theories generally tend to be local and microscopic in their examination of reading, while comprehension theories tend to be global and macroscopic.

The lack of a common focus makes it hard to contrast reading theories. However, there seem to be certain positions regarding distinctive attributes of reading theories that have had salience at different times. Our aim is not to challenge the different theories, nor to repeat what has been said and written before. Rather, our aim is to identify the distinctiveness of each of the dominant theories and to place it in a time sequence that would convey its evolutionary nature.

To classify the reading theories, we have drawn on Straw's categorization system that distinguishes among five periods based on three criteria: locus of meaning, nature of knowledge needed to be literate, and purpose of literacy (Straw 1989). However, we have extended his descriptors by focusing also on the attributes of theories of reading acquisition, instruction, and processes that would likely be associated with his periods. This is not to say that each theory fits neatly into only one of these categories or in any one period of time. As noted at the beginning of this discussion, given the evolutionary nature of reading theories and instructional practice, remnants of earlier periods and conceptions of reading continue to hold a very important place in both theory and practice.

The Progression of Reading Theories

The categorization system that we use divides conceptualizations of reading into just four periods. It combines Straw's first two periods—transmission and translation—into one that we label transmission, and maintains the remaining three—interaction, transaction, and social construction.[3]

Transmission. As defined by Straw, in the *transmission* period the meaning of text rests with the author, and the knowledge incorporated into a text by an author is to be reproduced by the reader. This conceptualization of reading supports conceptions of teaching and learning that hold up the teacher as the source of knowledge and the student as the recipient of that knowledge. The purpose of reading is to reproduce the author's intention.

Straw contrasts this with the *translation* period, where meaning lies in the text, and the text is seen as independent of its author. The reader is seen as a decoder of text, not of the authors' intentions. To decode text, the reader needs knowledge about reading and literature skills. Emphasis is placed on the entertainment value of text, as well as on the information found in it (Just and Carpenter 1980; Davis 1944; Gough 1972; LaBerge and Samuels 1974).

[3] While Straw designates specific dates for each period, we believe that for our purposes these designations are not important. Rather, we are more interested in the progression across time, and the way in which these periods correspond to notions of acquisition, instruction, and processes. Further, we believe that while many might disagree with his time periods, few would argue with the progression.

From our perspective we see the two as similar because in both instances the meaning of the text rests outside of the reader, and in both cases the reader is expected to reproduce what is someone else's meaning and knowledge as represented in the text.

One might expect stage models of reading acquisition to be associated with this period. Stage models assume that human development progresses through a series of qualitatively different stages, that the stages are hierarchically ordered, and that higher stages cannot be reached without going through the ones below (Chall 1983; Gough and Hillinger 1980; and to a lesser extent, Mason 1980). A singular similarity across stage theories—that an understanding of the alphabet is basic to reading acquisition (Juel 1991)—can help to explain the connection we see. Given that the alphabet is an abstraction that is removed from the basic understanding of words, it requires some intervention on the part of a teacher, parent, or guide to facilitate learning.

Given the notion of a hierarchy inherent in stage theories, it also seems reasonable to see reading acquisition as the accretion of subskills or components that together make up reading (Barrett 1968, Gray 1960). When reading is seen as a collection of skills, such as letter recognition, ability to make letter-sound correspondences, word recognition, finding the main idea, sequencing ideas, and making inferences, one is dependent on an expert to order those skills in some logical progression for learning. Most basal reading series are structured on such a model of reading acquisition.

This period is very prescriptive in its stance. A central tenet is that students must be taught to use the single system of language properly. The definitions and rules of this system form the basis for what is taught. Teaching phonics before children have a concept of reading is the epitome of the prescriptive approach to reading literacy instruction. For instance, some theorists argue that phonics instruction is the only way to begin to teach reading: first teach the letters, then the sounds of each letter, then the many phonic generalizations, and so on (Flesch 1955; Balmuth 1982).

Information-processing theories of reading, which compare human mental processes to the operation of a computer, are highly consistent with the stance of this period. According to many of these theories, information taken in by the senses is processed by a series of discrete processors. The output for one processor becomes the input for the next one in a linear series of steps. For instance, Gough proposes a comprehensive description of the reading process that begins with an eye fixation. Thence, the visual system produces an iconic image, which is matched against patterns for letter recognition. These patterns are then mapped onto lexical entries and stored for each word until they can be arranged into a larger unit of meaning. The process then starts from the beginning again with a fixation on another element of text (Gough 1985). Variations on this view would allow for information to be *chunked* into whole units (LaBerge and Samuels 1985), or might be somewhat more interactive (Ruddell and Speaker 1985; Rumelhart 1985).

This prescriptive/information processing tradition, characteristic of the transmission period, continues to exert a powerful influence on school language arts programs. Teachers continue to include grammar and phonics instruction in their programs; publishers supply grammar texts and phonics drill books; and computer software is offered based on the prescriptive traditions (Reinking and Bridwell-Bowles 1991).

Interaction. During the *interaction* period, meaning resides with readers and text. The theories of this period assume that three sources of knowledge are needed by readers: knowledge of authors and of text and personal experience. The good reader is the one whose background knowledge fits the text. These theories also assume that meaning is determinate. Reading is seen as a means whereby authors and readers can share knowledge and experience (Frye 1957; Goodman 1970; and Rumelhart 1977, 1985).

This period marks the beginning of a shift in views regarding instruction and cognitive processing. In contrast to the very prescriptive view previously held, we see the development of a psycholinguistic view where language is perceived as an instrument and the vernacular speech that children bring to school is seen as an adequate base for learning to read. Spoken language is seen as the overt performance of underlying, abstract abilities that involve phonological, syntactic, and semantic components of linguistic competence. The theme of building on those things the student already knows—linking the more formal language of school to the informal vernacular and the more disciplined academic understandings to the experientially acquired concepts already in place in the mind of the learner—fit the definition of interaction.

Learning to read is a matter of employing these components in the processing of meaning. Reading is much more than recoding visual symbols into their spoken equivalents. It involves readers in using their knowledge of oral language and their powers of conceptualization to derive meaning from print. The reader's knowledge of language includes familiarity with the syntactic order of linguistic elements and the semantic relationships among them. The reader's background experience with oral language is assumed to be a crucial factor in reading development.

Psycholinguistic approaches to reading instruction are based on the principle of continuity between home and school in the young child's experience and language. Beginning readers encounter written materials as part of their natural language development. They are encouraged to read them fluently in terms of their own language and meanings, rather than precisely and accurately in terms of what appears on the printed page as is required in the prescriptive approach. The graphic symbols are only part of the information that readers use; syntactic and semantic predictions supplement the visual display. These sources of information are available from the child's own linguistic competence acquired in the preschool years.

In the psycholinguistic view, language is a self-contained system to be acquired and refined by the individual. Psycholinguists are primarily concerned with the individual reader and how that reader establishes meaning for text. Of primary concern are the intrapersonal context, the background knowledge and skills that the reader brings to the task of interpreting a text, and individual differences in knowledge and skills. Consistent with this psycholinguistic view, we see the development of schema-theoretic views where individuals are believed to possess cognitive structures called "schemata" (Anderson and Pearson 1984). These schema consist of organized sets of concepts, and understanding a piece of text occurs when stimuli from the text are fitted into one of these structures.

Transaction. During the *transaction* period, constructing meaning is considered to be a generative act. The meaning of text is indeterminate and is constructed by readers while reading. In order to construct meaning, readers draw on a variety of knowledge sources including the text, knowledge of language, and experience. In contrast to the first two models described above, which are communicative, transactional theories assume that reading is more than the reception or processing of information in text. The reader generates meaning in response to text. The purpose of reading, in contrast to the communicative purpose of the previous models, is actualization (Rosenblatt 1978; Tompkins 1980; Harste, Burke, and Woodward, 1982; Straw 1989).

In this period, there are the beginnings of a number of major shifts in the stances taken by theorists. The stage models that prevailed in the two prior periods begin to be challenged by another conception—the nonstage model that assumes human development is continuous, and that reading does not require qualitatively different abilities for children and adults. So, what is required of a child to read a piece of text is the same as what is required of the adult; the difference is that the adult has a broader base of knowledge on which to draw in making an interpretation (Goodman and Goodman, 1979; Harste, Burke, and Woodward 1982; Smith 1973). This shift is necessary if one believes in having meaning produced or generated by the reader. Without an array of basic thinking processes, knowledge could not be generated.

Consequently, reading acquisition is no longer seen as necessarily based on formal, well-structured, sequential instruction. Theorists in this period maintain that reading acquisition is a natural activity analogous to learning to speak one's native language. Children learn to speak naturally, without formal instruction, when reared in the context of other speakers of the language. And learning to read, just like learning to speak and to walk, emerges early in life from children's experiences with spoken and written language (Goodman 1986; Harste and Woodward 1989; Kastler, Rosen, and Hoffman 1987; Pearson 1985). Children learn to read earlier in the context of more diverse oral language use (Snow and Perlman 1985) and through more active engagement with written language (Cullinan 1989; Strickland and Morrow 1989; Sulzby 1985). Even within this group of theorists, however, there is often the acknowledgment that children profit from help (Ehri 1987; Goodman 1986; Harste and Woodward 1989).

During the *transaction* period, the psycholinguistic views expand to include a somewhat more sociolinguistic position. From this larger perspective, language cannot be separated from its social context and reading is viewed not only as a set of cognitive processes, but also as social and linguistic processes (Wells 1986). As a social process, reading is used to establish, structure, and maintain social relationships among people. As a linguistic process, reading is used to communicate intentions and meaning between authors and readers (Olsen, Torrance, and Hildyard 1985).

Both the psycholinguistic and sociolinguistic theories of the period lead to experiential learning or the achievement of linguistic abilities through engagement in language use. Children are encouraged and allowed to learn to read by reading for purposes that are personally meaningful. School reading programs provide opportunities for reflective appraisal of these communications (Moffett 1983). Traditional, prescriptive information about language, such as the rules governing the relationships between words called nouns and other words called verbs, may provide some useful tools of appraisal. To this reflective repertoire could be added powerful tools of appraisal in the form of sociolinguistic understandings about such factors as the effects of certain kinds of audiences, situations, and purposes on meaning.

As opposed to the subskill view that characterized earlier periods, we see the emergence of a holistic view that maintains that reading is more than the sum of its parts and involves more than a collection of skills (Goodman 1986; Harste and Woodward 1989). Every reading act, according to holistic theories, requires the integration of skill, background knowledge, purpose and intention, and attitudes. Consequently, the characteristics of the reader and the text cannot be analyzed separately, as assumed by earlier reading theories. Reading emerges in the transaction between readers and text. In contrast to earlier interactive models, which assume that the text and the reader are separable entities, both readers and text are seen as aspects of a total event according to transactional theories of reading (Beach and Hynds 1991).

Learning to read is seen as a process of being socialized into the uses of written language. There is a renewed interest in the home as a setting in which some children become literate and from which schools can learn how to establish settings that are more effective for general literacy teaching (Harste, Burke, and Woodward 1982).

Social Construction. In the newly emerging period called *social construction*, knowledge is socially patterned and conditioned. The locus of meaning is in the social context, not with any person or object. As in the transaction period, the focus is on the construction of meaning—not by a single author or reader, but rather by society as a whole (Vygotsky 1978; Hunt 1990; Hynds 1990).

Here the primary concern is with the interpersonal context, the organization of reading events, the interaction of participants, the influences of the interaction on the processes of reading, and how the reading influences the interaction of the participants. Language neither develops as an autonomous system nor is it used as one, since language is a personal, social, and cultural phenomenon (Guthrie and Greaney 1991). It is not learned as a system and then put into use; it is learned as its functions are learned. Effective language use involves much more than learning words, pronunciation, spelling, and grammar. It involves sensitivity to audience factors such as social status and conventions such as turn-taking in conversation. The pragmatic dimension of language is the central focus in this newly emerging, very sociolinguistic view.

We arrive then at the definition of four dominant theory systems:

- **Transmission,** where the meaning of the text lies outside the reader who is expected to reproduce it, where teaching is based on a prescriptive view of language, instruction is hierarchical and subskill in nature, and processing is done in linear fashion.

- **Interaction,** where the meaning of the text resides with both the text and the reader who is expected to have some background knowledge that fits the text, and where we see an interaction between the vernacular language of the student and the more formal language of school and text.

- **Transaction,** where meaning is generated by the reader while reading, where a reader of any age is expected to read in the same manner, albeit with differing levels of knowledge on which to base an interpretation, and in which the reading act is clearly considered to be holistic in nature and is tightly integrated into the socialization associated with active language use.

- **Social Construction,** where all knowledge is socially patterned and constructed not by individuals, but rather in a group context.

7.5 Turning Groups of Items into Meaningful Constructs

Having established definitions of four dominant theory systems, we turn back to the IEA Reading Literacy Study data to determine how the two might fit together. While in principle, items might have been written or might be grouped on the basis of a theoretic stance, the questionnaire was not developed in a manner that explicitly reflected theoretic stances. Among the set of instructional items, a number of particular items could be classified as representative of more than one school of thought. For example, very few reading theorists would argue against having students do such things as "compare pictures and stories" or "understand why they are reading." Most, if not all, would consider these activities a natural part of reading and thinking. However, theorists would argue about how one might arrive at or structure such activities. Consequently, a number of items could be expected to be associated with a number of theoretic stances.

Within the questionnaire, there were blocks of items that had the same response scales and were grouped together as a single "question" because they had a common theme that tied the block of items to the literature on reading. For example, in question 53 on the grade 4 Teacher Questionnaire (see the appendix to this chapter), the 30 subsumed items are all generally tied to teaching practices. However, the group of items tap more than one aspect of instructional practice. Some of the items are statements related to student- or teacher-directedness (e.g., "Students have a choice in what they will do"), while others relate more closely to the content of reading instruction (e.g., "Specific skills are taught at certain times").

Therefore, to establish both a theoretic (i.e., based on reading theory) and an empirical basis for our groupings, we engaged in exploratory factor analyses to get at the latent structure of these items. As a general strategy, a principal factor solution was obtained and, in the first instance, factors with eigenvalues greater than 1 were rotated to an oblique solution. In subsequent analyses, factors were rotated until a solution was obtained that exhibited good simple structure and whose factors could be assign.d meaning consistent with a theory of reading. Factor scores were then estimated to provide measures of the latent variables identified.

To illustrate what we have done, we focus on four of these omnibus questions. The first is a measure of what teachers believe about reading instruction (question 43). The second focuses on what they do when teaching reading (question 53). The third focuses on what they have students do (question 30). The fourth looks at what they assess (question 46). See the appendix to this chapter for a copy of these questions.

What Teachers Believe About Reading Instruction. In question 43, teachers were asked to indicate their level of agreement with statem.nts about issues in reading instruction. This question provides a glimpse into teachers' beliefs about reading theory and how instruction should be organized.

As seen in Table 7-4, based both on an empirical rule of thumb where we considered factors with eigenvalues greater than 1 and a theoretic stance where the group of items in the first factor contrast with those in the second, we defined two factors from this question block. The first factor, labeled *sequenced instruction*, is characterized by sequencing, mastery of prior levels before moving on, accuracy, and heavy teacher direction. While this stance is likely to be consistent with what phonics advocates might suggest, it is broader than just phonics. Although never specifically stated, one might read into this factor a belief in developmental stages that are carefully orchestrated by either the materials or the teacher. Sequence also may be related to beliefs about the logic of the subject matter moving from simple to more complex.

Although a number of the items in this factor are not specifically unique to transmission, and some are not exclusively related only to the period, the items loading in this factor mostly characterize the theoretic stance underlying *transmission*. "Accuracy" is representative of reproduction of an author's or text's message or knowledge. The necessity for correctness can easily be associated with a rule-driven or prescribed notion of language use. The controlled movement across graded sets of materials can be related to the idea of a hierarchy and stages of development. All of these attributes are characteristic of the transmission period.

In considering the distribution of teachers' responses to the items in this factor, the general picture that emerges is that, at a minimum, 60 percent of the teachers appear to disagree with beliefs that are consistent with this factor. However, there are four items where this pattern is not as strong. Two items are related to the use of sequenced materials in class. Here teachers seem to be more evenly divided in their beliefs. Teachers also seem to be strongly supportive of providing feedback and monitoring student progress.

In contrast, the second factor, *extensive exposure to reading*, is characterized by students' active involvement in frequent extended reading, both at school and at home. There is little mention of teacher direction in this factor. It is characterized most by its focus on what the student does. Here are elements of whole language approaches, with students being given a more central role in constructing meaning.

Similarly, there is mention of the integration of reading and writing where students are encouraged to read texts they themselves have written.

Table 7-4. What teachers believe about reading instruction

Factor loading	Item	Disagree	Agree
Factor 1 -- *Sequenced Instruction*		Percent*	
0.58	Reading learning materials should be carefully sequenced in terms of language structures and vocabulary	44	41
0.56	Most of what a student reads should be assessed	60	22
0.56	Every mistake a student makes in reading aloud should be corrected at once	82	12
0.55	Teachers should carefully follow the sequence of the textbook	72	14
0.55	Teachers should always group students according to their reading ability	84	13
0.54	All students' comprehension assignments should be carefully marked to provide them with feedback	23	67
0.52	Students should not start a new book until they have finished the last	69	17
0.46	When my students read to me, I expect them to read every word accurately	65	27
0.46	Class sets of graded reading material should be used as the basis for the reading program	37	32
0.45	Students should learn most of their new words from lessons designed to enhance their vocabulary	57	27
0.39	Teachers should keep careful records of every student's reading progress	7	84
0.32	A word recognition test is sufficient for assessing students' reading levels	90	3
0 31	Students who can't understand what they read haven't been taught proper comprehension skills	66	10
0.24	9-year-olds should not have access to books they will read in the next year at school	76	10
Factor 2 -- *Extensive Exposure to Reading*			
0.51	Students should take a book home to read every day	13	76
0.41	Every day students should be read to by the teacher from a story book	11	86
0.40	Students should always understand what they are reading	21	58
0.39	All students should enjoy reading	10	82
0.38	Students should be encouraged to read texts they have written	1	95
0.32	Students should always understand why they are reading	12	74
0.30	Most students improve their reading best by extensive reading on their own	11	75

* Percentages may not add to 100 because response category "uncertain" has not been included.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Once again, although we see that all the items clustered in this factor need not solely be tied to a single particular period, we find that the underlying theme of the items in this factor would appear to be most closely associated with either the interaction or transaction periods. The movement between school and home, and between reading and writing, represent an integration between the more formal language of school and the vernacular that would be associated with either a psycholinguistic or sociolinguistic stance characteristic of these periods. These views are further developed by the statements of enjoyment and extensive independent reading.

Teachers appear to strongly support the beliefs espoused in this factor. More than 74 percent of the teachers agree with all but one of the items. In that item, *students should always understand what they are reading,* teachers seem to be permitting students a bit more latitude, and perhaps leaving more room for students to be challenged by working at constructing meaning more interactively.

What Teachers Do. In question 53, teachers were asked how often they used specified teaching practices in their classes. The items reflect a teacher's views and behavior with regard to who controls learning. What is at issue across these items is the degree of autonomy that students are given.

Across all the items in each of the three factors, there is an underlying assumption that the teacher is orchestrating instruction (Table 7-5). The teacher is creating an environment in which students are expected to learn certain things—both content and process. Within this structured environment there is, however, a broad range in which instruction and learning can flourish.

Table 7-5. What teachers do

Factor loading	Item	Rarely	Frequently
		Percent	
Factor 1 -- *Student Centered*			
0.72	Students are given the opportunity to consider what they think they have learned, as well as their perception of their strengths and weaknesses	70	30
0.72	Students are given the opportunity to assess their own progress	76	24
0.70	Students are encouraged to compare their written texts with the reading selection	88	12
0.69	Students are encouraged to use the reading selection as a source for ideas when writing their texts	61	39
0.65	Students are given the opportunity to provide input on how they will be assessed	92	8
0.60	Students are given the opportunity to work on a variety of different projects	67	33
0.59	Students establish their own purposes and goals	85	15
0.54	Students are given the opportunity to discuss various possible themes for the selection	72	28
0.54	Students are encouraged to compare their written texts with other students' written texts	81	19
0.50	Students decide how they will approach their texts	90	10
0.40	Students have a choice in what they will do	84	16
0.34	Students are given feedback by the teacher on the themes or main ideas of the selections they read	54	46
Factor 2 -- *Materials Directed*			
0.73	Students are given guided practice with skills	34	66
0.63	Specific skills are taught at certain times	35	65
0.53	Students are expected to follow the activities outlined in the lesson the teacher has planned	15	85
0.45	Students are invited to consider how skills apply to what they have written	59	41
0.39	Students are told what they have learned and have yet to learn	54	46
0.38	Students are directed to answer a set of the teacher's questions	55	45
0.29	Students are given teacher feedback on how they compare with other students	87	13
Factor 3 -- *Shared Direction*			
0.61	Students receive feedback from the teacher on their ideas	17	83
0.57	Students are informed as to the purposes of lessons	15	85
0.51	Students deal with issues and topics related to their own experiences	52	48
0.43	Students are directed to proceed based upon set guidelines	23	77
0.43	Students share their ideas with each other	43	57
0.41	Stu nts are told how what they know relates to a topic	49	51
0.27	Students are assigned specific topics to study	62	38

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Based both on the empirical test and the theoretical contrasts across sets of items, three distinct patterns emerge.

Factor one, *student centered,* stresses student independence. Students are asked and encouraged to consider and decide how they are doing, what they are doing, and how they will do it. This does not imply anarchy. Rather, within a structured environment, students are given every opportunity to organize themselves and the materials they use to construct meaning.

The seeming autonomy of the student, who is the reader, very closely emulates the terms of the *transaction* or *social construction* periods. However, the group nature of the social construction period is not represented in any of the items. Within this factor, we see the marked movement away from a hierarchical or staged stance. Students establish their purposes, decide how to approach their tasks, and work interactively between the text and their own ideas. These attributes are all characteristic of the *transaction* period.

An inspection of the distribution of teachers' responses seems to indicate that, at most, only about a third of the teachers surveyed are likely to strongly support extensive use of student-centered teaching strategies. It seems reasonable to conclude that, for the most part, teachers are still likely to be making most of the decisions regarding instruction and are probably providing direct instruction.

The items in factor two, *materials directed*, represent the other end of the continuum. Here students are directed as to what to do in a specified sequence. The teacher carefully maps out what will be done in accordance with a highly structured and ordered sense of progression. The theme of this set of items is that the expert or source of knowledge orchestrating instruction and meaning rests outside of the student This view of reading would be most closely aligned with the *transmission* period.

As indicated by responses to the first three items in this factor, two-thirds or more of the teachers surveyed indicate that students are expected to work frequently on activities that are skills oriented and orchestrated in specific ways by the teacher or the materials they have been assigned. Teachers are more evenly divided with regard to their use of the remaining strategies included in this factor, with the exception of the last listed item, which few teachers do frequently.

The items in the third factor, *shared direction*, represent a give and take between teachers and students. Teachers provide a high level of direction and feedback, but students are expected to generate ideas, to share with one another, and to relate what they are learning to their own experiences. What underlies this collection of items is the sense that students are given a great deal of latitude while they work within a prescribed structure.

Despite the single item that might indicate the possibility of theories of social construction coming into play (i.e., students share their ideas with each other), the factor as a whole seems more oriented toward the give and take between the teacher who is modeling desired behaviors through the feedback and structure of lessons and the students' use of their own knowledge. This reliance on and integration between the students' own knowledge and the structure provided by either text or teacher underlies the theories associated with the *interaction* period.

Although the distribution of teacher responses seems to vary a great deal across the items within this factor, close inspection of the items reveals an inherent logic consistent with the notion of shared direction. Teachers who believe in and practice behaviors that are consistent with an authoritative, facilitating approach are more likely to provide students with feedback and are less likely to assign specific topics. In principle, there seems to be reasonably high acceptance of this perspective among teachers.

In general, it seems safe to conclude that the majority of teachers do not regularly use practices that put the student at the center and with the most control. Rather, the teachers surveyed seem to most favor teaching practices associated with shared direction or that are materials directed.

What Teachers Have Students Do. In question 30, teachers were asked how frequently they have students do certain reading activities. In contrast to the last question, where the focus was on descriptions of teacher behaviors, this question looks at the kinds of assignments and activities teachers expect students to complete. Again, based both on an empirical rule of thumb, using only factors with an eigenvalue greater than 1, and on theoretic contrasts, three factors emerge from this question, as shown in Table 7-6.

Table 7-6. What teachers have students do

Factor loading	Item	Rarely	Frequently
		Percent	
Factor 1 -- *Schema-based activities*			
0.76	Making predictions during reading	16	84
0.71	Making generalizations and inferences	15	55
0.67	Relating experiences to reading	21	79
0.65	Orally summarizing their reading	31	69
0.63	Looking for the theme or message	25	75
0.42	Studying the style or structure of a text	60	40
Factor 2 -- *Integrated language arts activities*			
0.63	Listening to students reading aloud to small groups or pairs	33	67
0.59	Discussion of books read by students	63	37
0.56	Dramatizing stories	95	5
0.46	Drawing in response to reading	72	28
0.45	Diagramming story content	82	18
0.43	Writing in response to reading	23	77
0.42	Reading other students' writing	60	40
0.38	Student leading discussion about passage	70	30
0.35	Reading plays or dramas	97	3
0.30	Comparing pictures and stories	45	55
Factor 3 -- *Skills-based activities*			
0.81	Learning letter-sound relationships	41	59
0.65	Word attack skills	23	77
0.37	Learning new vocabulary from texts	8	92
0.35	Answering reading comprehension exercises in writing	9	91
0.35	Playing reading games (e.g., forming sentences from jumbled words)	82	18

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

In factor one, *schema-based activities*, students focus on the organization and interrelated aspects of text. They move back and forth from the detail to the overarching theme to make predictions and generalizations. They use what they know from experience and about the structure of text.

The instructional activities in this factor closely mirror the definition of the *interaction period.* The period focuses on reliance on background knowledge of the reader, which serves as a context for understanding. In activities such as *making predictions, relating experiences to reading,* and *looking for the theme or message,* students are calling forth the appropriate schemata for organizing the information gathered from the text.

For all but two of the items included in this factor, over 70 percent of the teachers report frequently having students do these things. In looking at the items, it is clear that they represent very common practices associated with a directed reading lesson and have been suggested and included in teaching manuals for years. With regard to the two remaining items, *making generalizations and inferences* and *studying the style or structure of a text,* if one believed in a hierarchy of skills these would be likely to be considered beyond the range of a grade 4 student. Therefore, it is not surprising that fewer teachers reported frequent use of these activities.

In factor two, labeled *integrated language arts activities,* the emphasis is on bringing all communication modes together. Students listen and discuss, read and write, and respond through other symbolic modes (drama, art).

The items grouped in this factor share an underlying theme, which is closely tied to the sociolinguistic theories characteristic of the *transaction period* where reading and language more generally are situated in a social context. The heavy reliance on discussion, dramatization, and writing of text seems to indicate an emphasis on a more experiential approach to learning.

That there is a great deal of variability in the frequency with which teachers report using the instructional activities in this group is to be expected given the nature of these items. Having students dramatize stories or read plays or dramas is quite time consuming and possibly results in little added benefit given the heavy time commitment. Even if the teacher were committed to this type of approach, we would expect such differences among the items. However, in looking at those items teachers report using frequently, we note that they need not be associated with this type of program. Students are often asked to read aloud for diagnostic purposes. Students in any class frequently write something in response to reading. And, it is not uncommon to have teachers in any subject area draw students' attention to the accompanying pictures or diagrams in order to make comparisons with the text. Given the dispersion of response rates, one would be very hard pressed to make any statement about teachers' commitment to this approach as a whole.

In the third factor, *skills-based activities*, the emphasis is on what is literally in the text. It is a very bottom-up orientation focusing on letters, words, sentences, and text-based understanding. This factor could most be associated with the *transmission period*, where the teacher or the text organizes tasks to be accomplished that become increasingly more difficult and call forth increasingly more complex coordinated skills.

The teachers surveyed seem to use the instructional activities included in this factor very frequently. That only 58 percent report frequently teaching letter-sound relationships is not surprising, because these are teachers of grade 4 students who, in principle, should have moved beyond this particular type of activity. Similarly, playing reading games would also be most likely to be associated with earlier grades—preschool, kindergarten, and grades 1 and 2.

What Teachers Test. In question 46, teachers were asked how frequently they assessed certain aspects of reading (Table 7-7).

In their assessments teachers appear to emphasize three different concepts. As seen in factor one, *contextualized reading*, teachers are testing the entire process. The basics of decoding and vocabulary are given less emphasis in this factor than relating reading to what the student knows. The second factor, *reading skills*, focuses entirely on the basic subskills of reading--decoding, phonics. The third factor, *literal understanding,* maintains a heavy, text-based, bottom-up orientation. Teachers are focusing on what is specifically in the text.

One would be hard pressed to associate the assessment emphases with particular periods. Each has somewhat overlapping elements. For example, word recognition and vocabulary are very closely related, although the former is more strictly a decoding activity while the later represents some level of understanding. It seems reasonable, however, to say that *contextualized reading*, due to its more inclusive nature, would more likely to be associated with either the *interaction* or *transaction* periods. *Reading skills* implies a more subskill approach and an analytic organization of instruction, which would require someone outside the learner to organize. Consequently, a case could be made that this would best match the *transmission* period. The third factor, *literal understanding,* given the progression from word to sentence to text, appears to have elements of a hierarchy that are prevalent in both the *transmission* and *interaction* periods.

Table 7-7. What teachers test

Factor loading	Item	Rarely	Frequently
		Percent	
Factor 1 -- *Contextualized reading*			
0.85	Use of background knowledge	11	89
0.72	Literary appreciation	20	80
0.62	Amount of reading	16	84
0.51	Vocabulary	5	95
0.51	Decoding	16	84
Factor 2 -- *Reading skills*			
0.99	Phonic skills	21	79
0.42	Reading study skills	10	90
Factor 3 -- *Literal understanding*			
0.64	Word recognition	12	88
0.56	Text comprehension	1	99
0.50	Sentence understanding	2	98

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991

What is most striking about this group of factors and the distribution of teacher responses to the items in each factor is that the teachers surveyed report frequently assessing everything, irrespective of the content implied in the factor, and perhaps irrespective of what they might be teaching.

7.6 Joining Data and Theory

The illustrations of how the items group within each question block raise the question of whether we can arrive at some systematic, integrated view of what is going on within classes. Do teachers in fact integrate their beliefs, their practices, the activities they have students do, and their assessment practices in ways that represent particular reading theories? The factor analyses of the four omnibus questions resulted in 11 distinct factors that could be associated with certain theoretical stances. Mapping those factors to the four reading periods, based on reading theory, might provide a tentative answer to our question.

Teachers' beliefs grouped into two factors: sequenced instruction and extensive exposure to reading. The first contains items that specifically related to a highly prescriptive, sequential, teacher-as-source-of-knowledge view of reading. These are characteristic of the transmission period. In contrast, the extensive exposure factor puts reading in a very natural acquisition mode with children doing extensive reading on their own. There is very little teacher intervention, and students are encouraged to read a great deal by themselves. These beliefs would be most typical of the transaction period.

With regard to what teachers do (e.g., how they provide instruction), the concern is most with where the locus of meaning resides. The first factor, student centered, puts the learner/reader in control. Statements of comparing their own writing with reading selections represent the generative act of constructing meaning, which is characteristic of the transaction period. Readers and students are expected to discuss various themes and to espouse a variety of positions regarding the meaning of text. The second factor, materials directed, places the student in the position of recipient. All actions are organized by the teacher in specific patterns. This is very prescriptive, and students are carefully guided through a series of skills. This position is most consistent with the transmission period. The third factor, shared direction, contains items indicative of schema theory and psycholinguistic models—students are expected to deal with issues related to their experience and proceed according to set guidelines. There is a structure to what they do that is imposed from outside but within which they exercise freedom—sharing their ideas. This give and take fits best with the interaction period.

In considering what teachers have students do, there were also three factors. The first, schema-based activities, involved drawing background knowledge of both content and text structure together to construct meaning. These activities are characteristic of the interactive period. The second factor, integrated language arts activities, includes activities that involve an integration of symbolic forms, an interaction with the text, and peers to go beyond the text. While it is probably most characteristic of the transaction period, it also would fit within the definition of the interaction period. The third factor, skills-based activities, only includes items that focus on small units of text—words, sentences. The very literal nature of these items places the factor in either the transmission or interaction period.

In looking at what teachers test, we note that two factors, skills and literal understanding, are very highly interrelated. Both focus on a bottom-up approach and are prescriptive and subskill in nature. Both would be associated with transmission. Literal understanding might also be associated with the interaction period if one assumes that it is the first step in developing an understanding of text. In contrast, the first factor, contextualized reading, represents a much broader view of the reading process, incorporating aspects of schema theory (use of background knowledge) and natural approaches to acquisition (amount of reading). This type of testing would be consistent with the interaction or transaction periods.

Table 7-8 serves as a summary of the placement of each of these 11 factors based on the logical relationship between the items and reading theory. What we see from this summary table is that the factors regarding beliefs, teacher behaviors, student activities, and testing practices can be aligned with the dominant theories of reading. However, there are two issues to be considered. First, were all the theories represented equitably? Second, what, if anything, is this configuration of data and the direction of teacher responses telling us about the state of the art of instruction?

In Table 7-9, we have duplicated Table 7-8, but have replaced the asterisk (*) with the number of items associated with each factor. In addition, we have limited any single factor to a single theoretic period, emphasizing the earliest period with which it would be associated. This data display makes it clear that based on the factor analyses of the responses of American teachers, there seems to be an imbalance across the theories in terms of how well they are represented. For example, none of the factors seem particularly representative of theories of social construction. Overall, the number of items associated with transmission is about 50 percent greater than those associated with transaction.

At least two possible reasons for this imbalance come to mind. First, one might consider whether the items include an adequate and representative sample of the beliefs, behaviors, and activities that might be associated with each period. Second, one might wonder if American teachers have had sufficient exposure to each theoretic stance so that they might be explicitly implementing and/or emphasizing instructional practices that might be associated with particular theoretic stances.

Both of these interpretations have possible implications for future study. In the first case we might argue that items should be developed to reflect a specified framework. The items would then be more evenly distributed and would be associated with each theoretic stance before the data were collected. This presents a challenge for survey developers: can the items be written so that they relate to one and only one theoretic stance? The experience in this study suggests this might be difficult. The items seem more easily interpretable as a group than independently. Individual items might have more than one interpretation and as a consequence could be associated with more than one theoretic stance. However, once grouped into a factor it seems easier to infer the meaning as a respondent associated with the individual item.

Table 7-8. Relating data to theory

Factor		Transmission	Interaction	Transaction	Social
What Teachers Believe . . .	Sequenced instruction	*			
	Extensive exposure			*	
What Teachers Do	Materials directed	*			
	Shared direction		*		
	Student centered			*	
What Teachers Have Students Do	Skills-based activities	*			
	Schema-based activities		*		
	Integrated language arts		*	*	
What Teachers Test	Skills assessment	*			
	Literal understanding	*	*		
	Contextualized reading		*	*	

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 7-9. Relating the number of data items to theory

Fctor		Transmission	Interaction	Transaction	Social
What Teachers Believe	Sequenced instruction	14			
	Extensive exposure			7	
What Teachers Do	Materials directed	7			
	Shared direction		7		
	Student centered			12	
What Teachers Have Students Do	Skills-based activities	5			
	Schema-based activities		6		
	Integrated language arts		10		
What Teachers Test	Skills	2			
	Literal understanding	3			
	Contextualized reading		5		

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

The second hypothesized interpretation regarding a measurement of implementation changes the question asked. It implies that these items would be used repeatedly over time and that we would track whether there were changes in the combinations of items clustered within factors and in the number of teachers employing certain combinations of instructional strategies. When placed in juxtaposition with information about modes of teacher training, this might provide insight into how to efficiently change instructional practice.

For the sake of argument, assuming that there had been adequate representation of each of the theoretic stances, could we then determine something about how teachers organize and implement instruction?

7.7 Do Teachers Organize Instruction According to an Implicit Theory of Reading?

In principle, teachers might be expected to align their beliefs about instruction, their actions, what they have students do, and what they test according to a consistent theory of either reading or learning. As we noted in the discussion above, the emerging factors could theoretically be associated with particular theoretic stances. To test whether teachers would organize instruction according to these theoretic stances, we conducted a second-order factor analysis.

The second-order factor analysis resulted in three relatively meaningful factors—two that distinguish between two schools of thought in instruction, and a third that captures all testing. Table 7-10 displays the second-order factor loadings. The first two factors that emerge are derived from the item blocks on beliefs, practices, and activities. The first of the two, *interaction emphasis,* seems to be most related to notions of reading and learning as an interaction between the teacher or author and the student. In contrast, the second factor, *transmission emphasis,* is associated with reading and learning theories based on a notion of the transmission of knowledge from the teacher or author to the student.

Table 7-10. Theory and practice combined

Factor loading	Primary factor name*
Second Order Factor 1 -- *Interaction emphasis*	
0.73	Integrated language arts
0.72	Schema-based activities
0.72	Student-centered teacher behaviors
0.64	Shared-direction teacher behaviors
Second Order Factor 2 -- *Transmission emphasis*	
0.69	Materials-directed teacher behaviors
0.37	Sequenced instruction teacher beliefs
Second Order Factor 3 -- *Assessment*	
0 76	Contextualized reading assessment
0.75	Text-based understanding assessment
0.63	Skills assessment
0.52	Skills-based activities

*The 11th identified factor, extensive exposure, did not load on any of the second-order factors and is therefore not included in this table.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Interaction Emphasis

An interaction emphasis may be characterized as having the meaning of the text reside with both the text and the reader, who is expected to have some background knowledge that fits the text (Straw 1989). There is also an expectation that there will be an interaction between the vernacular language of the student and the more formal language of school and text.

Based solely on a theoretic stance, three of the first-order factors in this construct would be associated with interaction. There is a consistent theme of the integration of reading and writing, of student and author knowledge, and of the shared decision making between the student and the teacher. The fourth first-order factor, student-centered teacher behaviors, might be seen as differing from the others and, from a theoretic perspective, might be more closely linked to transaction. Although it is linked in the factor analysis, teachers do not report frequent use of these behaviors.

Transmission Emphasis

Instruction and reading theories that can be grouped under the heading of transmission may be characterized as placing the meaning of the text outside the reader who is expected to reproduce it (Straw 1989), organizing teaching according to a prescriptive view of language (Balmuth 1982), providing instruction that is hierarchical and subskill in nature (Barrett 1968; Gray 1960), and processing that is done in a linear fashion (Gough 1985).

The two first-order factors that fall into this category are strongly prescriptive and demand a high level of accuracy consistent with a view of language usage that is correct, and that is known by the teacher and the authors of texts and materials. Thus, they fit together well, both theoretically and empirically.

Assessment Emphasis

Second-order factor 3 brings all the questions on assessment back together, including *skills-based activities*. This is not particularly surprising because these activities are often workbook or worksheet pages that a teacher would be likely to grade and are not too different in kind from what teachers would use for a skills assessment. Despite the fact that there are three possible emphases, assessment seems to run together. A teacher who tests a great deal is likely to test everything frequently.

What then is this telling us about how teachers organize instruction? First, we see that only two of the four defined theoretic stances are represented. The two that are represented have been well documented and disseminated. Theories related to *transmission* have been in mainstream practice for generations. Theories associated with interaction have been well documented since the early 1970s. As such, it is not surprising to see that both theoretic stances appear to be heavily entrenched. In contrast, the transaction theories began to take center stage during the mid- to late-1980s. Consequently, given the age of the teaching staff and their distance from undergraduate training, it is not surprising that this stance is not well represented. Social construction theories are now just coming into the research literature, and consequently it may be premature to expect teachers to base their instruction on this stance.

Second, if we also consider response patterns, we can say something about how consistently teachers organize their actions in ways that are consistent with their beliefs. However, we must do so with a high level of tentativeness. Given that assessment emphasis formed a separate factor, one is likely to conclude that teachers are not necessarily basing their testing on their beliefs or instructional strategies. At a minimum, 80 percent of the teachers report administering tests of every kind frequently (see Table 7-7). Irrespective of their beliefs and practices, teachers seem to test everything frequently.

When we consider the responses associated with factors related to instructional emphases, different pictures emerge. We consider first the items and response rates associated with a *transmission* emphasis. As noted before, two first-order factors were associated with this construct. The first was derived from the question asking whether teachers agree or disagree with particular issues regarding reading instruction. The second asked about the frequency with which teachers used certain instructional strategies. One would expect that if teachers aligned their instructional practice with their beliefs, teachers would agree with those statements about issues in reading that corresponded with related activities they used frequently. Or, conversely, they would rarely use those activities with which they disagreed.

Given the factor loadings listed in Table 7-10 for these first-order factors, it appears that there is a relatively strong tendency for teachers to align their beliefs and practices within a theoretic stance. However, as noted in Table 7-11, where we have reproduced the parts of Tables 7-4 and 7-5 that included the first-order factors that formed *transmission* emphasis, there is a level of incongruity between the responses to the first-order factors and consistent representation of theory. While a large proportion of teachers disagree with many of the items listed in this factor, many teachers report frequently using activities that represent this approach to instruction.

There are a number of potential explanations for this lack of congruity. First, although teachers may be more likely to disagree with the theoretic stance of *transmission*, they only may have access to texts and teaching materials that fit that stance. Alternatively, given the movement in the field as reflected in journals and teacher-oriented magazines, as well as the emphasis of most inservice courses, reporting this view may be considered the socially accepted thing to do. Another possible explanation may be that many teachers who no longer strongly believe in these theories are constrained by policies enacted at the school, district, or state levels. Whatever, we seem to be picking up some level of disequilibrium.

When we consider the response pattern associated with an *interaction* emphasis, we are confronted by yet another picture (Table 7-12). The first-order factors are all reports of frequencies. Two factors describe what teachers do, and the other two are reports on what teachers have students do. Based on the factor loadings, as reported in Table 7-10 where each of the loadings is at least .64, it might be reasonable to assume that there is a strong alignment between the teaching behaviors and instructional activities included in this second-order factor.

However, the response distributions may be telling a somewhat different story. With regard to what teachers do, there seems to be a clear divide. Teachers report rarely using those practices associated with a student-centered approach, but they seem to indicate more frequently using teaching approaches that represent shared direction. This is also in keeping with the historic progression of theory and the fact that this generally is an *interaction* factor. However, it would seem that teachers with a high factor score on the second-order factor would be likely to be moving into the newer teaching strategies.

With regard to what teachers have students do, we note that teachers report frequently assigning activities associated with schema-based theories. In contrast, teachers' reports of use of the activities associated with integrated language arts activities vary greatly from item to item. That there is such wide disparity between frequent and rare use of activities is not surprising when one carefully examines the items and considers how each activity would fit into instruction. For example, the two items that less than 5 percent of teachers report using frequently, dramatizing stories and reading plays or dramas, in fact would and should play a comparatively small role in the overall instructional program because they are so time-consuming. In contrast "writing in response to reading," which is a more generalized and significant part of the language arts curriculum and would relate to a larger variety of topics that might be covered, is used frequently by 77 percent of teachers. What this points to is the need to establish a more logically uniform response scale as opposed to the absolute scale that is currently used.

Table 7-11. Transmission emphasis

Factor loading	Item	Disagree	Percent
		Percent*	
Factor 1 -- *Sequenced Instruction*			
0.58	Reading learning materials should be carefully sequenced in terms of language structures and vocabulary	44	41
0.56	Most of what a student reads should be assessed	60	22
0.56	Every mistake a student makes in reading aloud should be corrected at once	82	12
0.55	Teachers should carefully follow the sequence of the textbook	72	14
0.55	Teachers should always group students according to their reading ability	84	13
0.54	All students' comprehension assignments should be carefully marked to provide them with feedback	23	67
0.52	Students should not start a new book until they have finished the last	69	17
0.46	When my students read to me, I expect them to read every word accurately	65	27
0.46	Class sets of graded reading material should be used as the basis for the reading program	37	32
0.45	Students should learn most of their new words from lessons designed to enhance their vocabulary	57	27
0.39	Teachers should keep careful records of every student's reading progress	7	84
0.32	A word recognition test is sufficient for assessing students' reading levels	90	3
0.31	Students who can't understand what they read haven't been taught proper comprehension skills	66	10
0.24	9-year-olds should not have access to books they will read in the next year at school	76	10
Factor 2 -- *Materials Directed*		Rarely	Frequently
0.73	Students are given guided practice with skills	34	66
0.63	Specific skills are taught at certain times	35	65
0.53	Students are expected to follow the activities outlined in the lesson the teacher has planned	15	85
0.45	Students are invited to consider how skills apply to what they have written	59	41
0.39	Students are told what they have learned and have yet to learn	54	46
0.38	Students are directed to answer a set of the teacher's questions	55	45
0.29	Students are given teacher feedback on how they compare with other students	86	14

*For factor 1, percentages may not add to 100 because the response category "uncertain" has not been included.

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

Table 7-12. -- Interaction emphasis

Factor loadinj:	Item	Rarely	Frequently
Factor 1 -- *What teachers do: Student centered*		Percent	
0.72	Students are given the opportunity to consider what they think they have learned, as well as their perception of their strengths and weaknesse:i	70	31
0.72	Students are given the opportunity to assess their own progress	76	24
0.70	Students are encouraged to compare their written texts with the reading selection	88	12
0.69	Students are encouraged to use the reading selection as a source for ideas when writing their texts	61	39
0.65	Students are given the opportunity to provide input on how they will be assessed	92	8
0.60	Students are given the opportunity to work on a variety of different projects	67	33
0.59	Students establish their own purposes and goals	85	15
0.54	Students are given the opportunity to discuss various possible themes for the selection	72	28
0.54	Students are encouraged to compare their written texts with other student's written texts	81	19
0.50	Students decide how they will approach their texts	90	10
0.40	Students have a choice in what they will do	84	16
0.34	Students are given feedback by the teacher on the themes or main ideas of the selections they read	54	46
Factor 3 -- *What teachers do: Shared direction*			
0.61	Students receive feedback from the teacher on their ideas	17	83
0.57	Students are informed as to the purposes of lessons	15	85
0.51	Students deal with issues and topics related to their own experiences	52	48
0.43	Students are directed to proceed based upon set guidelines	23	77
0.43	Students share their ideas with each other	43	57
0.41	Students are told how what they know relates to a topic	49	51
0.27	Students are assigned specific topics to study	62	38
Factor 1 -- *What teachers have students do: Schema-based activities*			
0.76	Making predictions during reading	16	84
0.71	Making generalizations and inferences	15	55
0.67	Relating experiences to reading	21	79
0.65	Orally summarizing their reading	31	69
0.63	Looking for the theme or message	25	75
0.42	Studying the style or structure of a text	60	40
Factor 2 -- *What teachers have students do: Integrated language arts activities*			
0.63	Listening to students reading aloud to small groups or pairs	33	67
0.59	Discussion of books read by students	63	37
0.56	Dramatizing stories	95	5
0.46	Drawing in response to reading	72	28
0.45	Diagramming story content	82	18
0.43	Writing in response to reading	23	77
0.42	Reading other students' writing	60	40
0.38	Student leading discussion about passage	70	30
0.35	Reading plays or dramas	97	3
0.30	Comparing pictures and stories	45	55

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

It is possible to both theoretically and empirically tie these factors together into a cohesive whole. The only outlier, as previously mentioned, is the student-centered first-order factor. Theoretically this is a bit out in front of the theory represented in the other three factors.However, it is not inconsistent with the stance. If these items were to be administered again in 3 to 5 years, within the United States one might predict a reshuffling of the first-order factors such that shared direction and schema-based activities would group together and would be distinct from student-centered and integrated language arts activities. This assumes that reading theory is truly evolutionary instead of revolutionary.

7.8 Conclusions and Recommendations

Throughout the discussion we have emphasized the need for revising and improving the design and analysis of surveys related to instruction. Essentially, we stressed the importance of beginning with a theoretic frame and designing the measurement instrument so that we might do more confirmatory analyses. We have used the data from the IEA Reading Literacy Study as a case study. We have manipulated and explored that data to illustrate a methodological point.

The findings from this data suggest that the proposed methodology would allow us to get a sense of the degree of implementation of the various theoretic stances. The data seem to tell us that, in general, a growing number of teachers are beginning to move away from the theoretic stance associated with a *transmission* emphasis. However, many are still assigning student activities that would represent the tasks of that period. Others have adopted an *interactive* emphasis. They frequently use instructional strategies that represent that mode of interaction characteristic of that period. Similarly, they have incorporated the notion of schema theory into their assignments for students.

To consider these findings conclusive, however, it is important to stress the necessity for improving the instrument design. Three suggestions follow. First, it seems imperative that we begin with a theoretic frame that can serve as the blueprint or specifications for item development. In this way we can more systematically sample the domain of instruction. In this paper we have suggested a possible frame that identified four periods—*transmission, interaction, transaction*, and *social construction*. While we have defined these periods in terms of reading theory most specifically, they are based on more generalized theories of instruction and learning. Consequently, we see that they also might be used more generally in relation to other curriculum areas.

Second, we would advocate that items be written to correspond specifically with each period. This, however, is likely to be very difficult. Teachers would have difficulty in responding to the global terms. On the other hand, including a wide range of activities so that there is the probability that many would relate to each period is important.

Third, item response categories should be designed so they easily facilitate logical comparisons. The weights given to the responses "frequently" and "rarely" should be based not on absolute frequency, but rather on a logical understanding of differences in the strategies and their relation to an overall instructional program.

This type of reporting methodology would provide a different type of information to policy makers, curriculum specialists, and teachers. Although it would not necessarily lend support to a particular strategy, it is likely to help these educators discern trends and perhaps measure levels of implementation. Given the cross-sectional nature of the study, this would be an appropriate type of information to generate.

References

Adams, M.J. (1990). *Beginning to read: Thinking and learning about print.* Cambridge, MA: Massachusetts Institute of Technology.

Anderson, R.C., and Pearson, P.D. (1984). A schema-theoretic view of basic processes in reading comprehension. In P.D. Pearson (ed.), *Handbook of reading research.* Vol. 1, 255-291. New York: Longman.

Anderson, R.C., Hiebert, E.H., Scott, J.A., and Wilkinson, I.A.G. (1985). *Becoming a nation of readers: The report of the Commission on Reading.* Champaign, IL: Center for the Study of Reading.

Anderson, R.C., Wilson, P.T., and Fielding, L.G. (1988). Growth in reading and how children spend their time outside of school. *Reading Research Quarterly,* 23, 285-303;389-413.

Applebee, A.N., and Langer, J.A. (1983). Instructional scaffolding: Reading and writing as natural language abilities. *Language Arts,* 60, 168-175.

Balmuth, M. (1982). *The roots of phonics.* New York: Teachers College Press.

Barrett, T.A. (1968). A taxonomy of cognitive and affective dimensions of reading comprehension. Outlined by Clymer, T.C. What is reading? Some current concepts. In H.M. Robinson (ed.), *Innovation and change in reading instruction,* 252-258. NSSE 67th Yearbook, Part II. Chicago: University of Chicago Press.

Beach, R., and Hynds, S. (1991). Research on response to literature. In R. Barr, M.L. Kamil, P. Mosenthal, and P.D. Pearson (eds.), *Handbook of reading research.* Vol. 2, 453-489. New York: Longman.

Binkley, M., and Rust, K. (eds.) (1994). *Reading literacy in the United States. Technical report.* (U.S. Department of Education, National Center for Education Statistics). Washington, DC: U.S. Government Printing Office.

Chall, J.S. (1967). *Learning to read: The great debate.* New York: McGraw-Hill.

Chall, J.S. (1983). *Stages of reading development.* New York: McGraw-Hill.

Clay, M.M. (1985). *The early detection of reading difficulties.* 3rd ed. Portsmouth, NH: Heinemann.

Cline, R.K., and Kretke, G.L. (1980). An evaluation of long-term sustained silent reading in the junior high school. *Journal of Reading,* 23, 503-506.

Collins, C. (1980). Sustained silent reading periods: Effect on teachers' behaviors and students' achievement. *Elementary School Journal,* 81, 108-114.

Cullinan, B.E. (1989). Literature for young children. In D.S. Strickland and L.M. Morrow (eds.), *Emerging literacy: Young children learn to read and write,* 35-51. Newark, DE: International Reading Association.

Davis, F.B. (1944). Fundamental factors of comprehension in reading. *Psychometrika,* 9, 185-197.

Durkin, D. (1988). *Teaching them to read.* 5th ed. Boston: Allyn & Bacon.

Ehri, L.C. (1987). Learning to read and spell words. *Journal of Reading Behavior,* 19, 5-31.

Elley, W., and Mangubhai, F. (1983). The impact of reading on second language learning. *Reading Research Quarterly,* 19, 53-67.

Farr, R., et al. (1991). Writing in response to reading. *Educational Leadership,* 66-69.

Flesch, R. (1955). *Why Johnny can't read.* New York: Harper & Row.

Frye, N. (1957). *Anatomy of criticism.* Princeton, NJ: Princeton University Press.

Goelman, H., Oberg, A., and Smith, F. (eds.) (1984). *Awakening to literacy.* London: Heinemann.

Goodman, K.S. (1970). Behind the eye: What happens in reading. In K.S. Goodman and O.S. Niles (eds.), *Reading, process and program*. Urbana, IL: National Council of Teachers of English.

Goodman, K.S., and Goodman, Y.M. (1979). Learning to read is natural. In L.B. Resnick and P.A. Weaver (eds.), *Theory and practice of early reading*. Vol. 1, 137-154. Hillsdale, NJ: Lawrence Erlbaum Associates.

Goodman, Y.M. (1986). Children coming to know literacy. In W.H. Teale, and E. Sulzby (eds.), *Emergent literacy: Writing and reading*, 1-14. Norwood, NJ: Ablex.

Gough, P.B. (1972). One second of reading. In J.F. Kavanagh and I.G. Mattingly (eds.), *Language by ear and by eye*, 331-368. Cambridge, MA: MIT Press.

Gough, P.B. (1985). One second of reading: Postscript. In H. Singer and R.B. Ruddell (eds.), *Theoretical models and processes of reading*. 3rd ed., 687-688. Newark, DE: International Reading Association.

Gough, P.B., and Hillinger, M.L. (1980). Learning to read: An unnatural act. *Bulletin of the Orlon Society*, 30, 179-196.

Gray, W.S. (1960). *The major aspects of reading. Sequential development of reading abilities*. Supplementary Educational Monographs, No. 90. Chicago: University of Chicago Press.

Greaney, V. (1980). Factors related to amount and type of leisure time reading. *Reading Research Quarterly*, 15, 337-357.

Greaney, V., and Hegarty, M. (1984). Correlates of leisure-time reading. Unpublished manuscript. Dublin: Educational Research Center, St. Patricks College.

Guthrie, J.T., and Greaney, V. (1991). Literacy acts. In R. Barr, M.L. Kamil, P. Mosenthal, and P.D. Pearson (eds.), *Handbook of reading research*. Vol. 2, 68-96. New York: Longman.

Harste, J.C., and Woodward, V.A. (1989). Fostering needed change in early literacy programs. In D.S. Strickland and L.M. Morrow (eds.), *Emerging literacy: Young children learn to read and write*, 147-159. Newark, DE: International Reading Association.

Harste, J.C., Burke, C.L., and Woodward, V.A. (1982). Children's language and world: Initial encounters with print. In J.A. Langer and M.T. Smith-Burke (eds.), *Reader meets the author: Bridging the gap*, 105-131. Newark, DE: International Reading Association.

Heyns, B. (1978). *Summer learning and the effects of schooling*. New York: Academic Press.

Holdaway, D. (1979). *The foundations of literacy*. Portsmouth, NH: Heinemann.

Hunt, R.A. (1990). The parallel socialization of reading research and literary theory. In S.B. Straw and D. Bogdan (eds.), *Beyond communication: Reading comprehension and criticism*. Portsmouth, NH: Boynton/Cook-Heinemann.

Hynds, S. (1990). Reading as a social event: Comprehension and response in the text, classroom, and world. In D. Bogdan and S. Straw (eds.), *Beyond communication, reading comprehension and criticism*. Portsmouth, NH: Boyton/Cook Publishers, Heinemann.

Ingham, J. (1982). *Books and reading development: The Bradford bookflood experiment*. 2nd ed. Exeter, NH: Heinemann.

Juel, C. (1991). Beginning reading. In R. Barr, M.L. Kamil, P. Mosenthal, and P.D. Pearson (eds.), *Handbook of reading research*. Vol. 2, 759-788. New York: Longman.

Just, M.A., and Carpenter, P.A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87, 329-354.

Just, M.A., and Carpenter, P.A. (1985). A theory of reading: From eye fixations to comprehension. In H. Singer and R.B. Ruddell (eds.), *Theoretical models and processes of reading*. 3rd ed., 174-208. Newark, DE: International Reading Association.

Kastler, L.A., Rosen, N.L., and Hoffman, J.V. (1987). Understanding of the forms and functions of written language: Insights from children and parents. In J.E. Readance and R.S. Baldwin (eds.), *Research in literacy: Merging perspectives*, 85-92. Rochester, NY: National Reading Conference.

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

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

Krashen, S.D. (1988). Do we learn to read by reading? The relationship between ·e reading and reading ability. In D. Tannen (ed.), *Linguistics in context: Connecting observation and understanding* (Lectures from the 1985 LSA/TESOL and NEH Institutes), 269-298. Norwood, NJ: Ablex.

LaBerge, D., and Samuels, S.J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, 6, 293-323.

LaBerge, D., and Samuels, S.J. (1985). Toward a theory of automatic information processing in reading: Updated. In H. Singer and R.B. Ruddell (eds.), *Theoretical models and processes of reading*. 3rd ed., 719-721. Newark, DE: International Reading Association.

Lewin, L. (1992). Integrating reading and writing strategies using an alternating teacher-led/student-selected instructional pattern. *The Reading Teacher*, 45 (8), 586-591.

Loban, W. (1963). *The language of elementary school children.* Urbana, IL: National Council of Teachers of English.

Lundberg, I., and Linnakyla, P. (1992). *Teaching reading around the world.* The Hague: International Association for the Evaluation of Educational Achievement.

Manning, G.L., and Manning, M. (1984). What models of recreational reading make a difference? *Reading World*, 23, 375-380.

Mason, J.M. (1980). When do children begin to read: An exploration of four-year-old children's letter and word reading competencies. *Reading Research Quarterly*, 15, 203-227.

McCracken, R.A. (1971). Initiating sustained silent reading. *Journal of Reading*, 14, 521-524, 582-583.

McGinley, W., and Madigan, D. (1990). The research story: A form for integrating reaaing, writing and learning. *Language Arts*, 67, 474-483..

Moffett, J. (1983). *Teaching the universe of discourse*. 2nd ed. Boston: Houghton Mifflin.

Moffet, J., and Wagner, B.J. (1983). *Student-centered language arts and reading K-3.* 3rd ed. Boston: Houghton Mifflin.

Morrow, L.M., and Weinstein, C.S. (1986). Encouraging voluntary reading. The impact of a literature program on children's use of library corners. *Reading Research Quarterly*, 21, 330-346.

Olson, D.R., Torrance, N., and Hildyard, A. (eds.). (1985). *Literacy, language, and learning*. Cambridge, England: Cambridge University Press.

Pearson, P.D. (1985). Changing the face of reading comprehension instruction. *The Reading Teacher*, 38, 724 738.

Pearson, P.D., and Fielding, L. (1991). Comprehension instruction. In R. Barr, M.L. Kamil. P. Mosenthal, and P.D. Pearson (eds.), *Handbook of reading research*. Vol. 2. New York: Longman.

Reid, I. (1990). Reading as framing, writing as reframing. In M. Hayhoe and S. Parker (eds.), *Reading and response*. Philadelphia: Open University Press.

Reinking, D., and Bridwell-Bowles, L. (1991). Computers in reading and writing. In R. Barr, M.L. Kamil, P. Mosenthal, and P.D Pearson (eds.), *Handbook of reading research*. Vol. 2, 310-340. New York: Longman.

Rosenblatt, L.M. (1978). *The reader, the text, the poem: The transactional theory of the literary work.* Carbondale, IL: Southern Illinois University Press.

Ruddell, R.B., and Speaker, R. (1985). The interactive reading process: A model. In H. Singer and R.B. Ruddell (eds.), *Theoretical models and processes of reading*. 3rd ed., 751-793. Newark, DE: International Reading Association.

Rumelhart, D. (1977). Toward an interactive model of reading. In S. Dornic (ed.), *Theoretical models and processes of reading*. Newark, DE: International Reading Association.

Rumelhart, D.E. (1985). Toward an interactive model of reading. In H. Singer and R.B. Ruddell (eds.), *Theoretical models and processes of reading*. 3rd ed., 722-750. Newark, DE: International Reading Association.

Smith, F. (1973). *Psycholinguistics and reading*. New York: Holt, Rinehart, and Winston.

Snow, E., and Perlman (1985). Assessing children's knowledge about book reading. In L. Galda and A. Pelligrini (eds.), *Play, language, and stories*, 165-189. Norwood, NJ: Ablex.

Stahl, S.A. (1992). Saying the "P" word: Nine guidelines for exemplary phonics instruction. *The Reading Teacher*, 34, 618-625.

Stanovich, K.E. (1991). Word recognition: Changing perspectives. In R. Barr, M.L. Kamil, P.B. Mosenthal, and P.D. Pearson (eds.), *Handbook of reading research*. Vol. 2, 419-452. New York: Longman.

Straw, S. (1989). The actualization of reading and writing: Public policy and conceptualizations of literacy. In S.P. Norris and L.M. Phillips (eds.), *Foundations of literacy policy in Canada*, 165-181. Calgary, Alberta: Detselig.

Strickland, D.S. and Morrow, L.M. (eds.). (1989). *Emerging literacy: Young children learn to read and write*. Newark, DE: International Reading Association.

Sulzby, E. (1985). Children's emergent reading of favorite books: A developmental study. *Reading Research Quarterly*, 20, 458-481.

Thurlow, M., Graden, J., Ysseldyke, J.E., and Algozzine, R. (1984). Student reading during reading class: The lost activity in reading instruction. *Journal of Educational Research*, 77 (5), 267-272.

Tierney, R.J., and Shanahan, T. (1991). Research on the reading-writing relationship: Interactions, transactions, and outcomes. In R. Barr, M. L. Kamil, P. Mosenthal, and P.D. Pearson (eds.), *Handbook of reading research*. Vol. 2. New York: Longman.

Tompkins, J.P. (1980). An introduction to reader-response criticism. In J.P. Tompkins (ed.), *Reader-response criticism: From formalism to post-structuralism*. Baltimore: Johns Hopkins University Press.

Vygotsky, L.S. (1978). *Mind in society* (1938). (eds. and trans. M. Cole, V. John-Steiner, S. Scribner, and E. Sonberman). Cambridge, MA: Harvard University Press.

Wells, G. (1986). *The meaning makers*. Portsmouth, NH: Heinemann.

Appendix
Teacher Questionnaire, Selected Items

C. Questions 30 to 53 have to do with your teaching activities.

30. **How often are your <u>students</u> typically <u>involved</u> in the following reading activities?** *(Circle one number per line only.)*

	Reading Activities	Frequency			
		Almost never	About 1 or 2 times a month	About 1 or 2 times a week	Almost every day
a.	Learning letter-sound relationships and/or phonics	1	2	3	4
b.	Word-attack skills (e.g., prediction)	1	2	3	4
c.	Silent reading in class	1	2	3	4
d.	Answering reading comprehension exercises in writing	1	2	3	4
e.	Independent silent reading in a library	1	2	3	4
f.	Listening to students reading aloud to a whole class	1	2	3	4
g.	Listening to students reading aloud to small groups or pairs	1	2	3	4
h.	Listening to teachers reading stories aloud	1	2	3	4
i.	Discussion of books read by students	1	2	3	4
j.	Learning new vocabulary systematically (e.g., from lists)	1	2	3	4
k.	Learning new vocabulary from texts	1	2	3	4
l.	Learning library skills	1	2	3	4
m.	Reading plays or dramas	1	2	3	4
n.	Playing reading games (e.g., forming sentences from jumbled words)	1	2	3	4
o.	Dramatizing stories	1	2	3	4
p.	Drawing in response to reading	1	2	3	4
q.	Orally summarizing their reading	1	2	3	4
r.	Relating experiences to reading	1	2	3	4
s.	Reading other students' writing	1	2	3	4
t.	Making predictions during reading	1	2	3	4
u.	Diagramming story content	1	2	3	4
v.	Looking for the theme or message	1	2	3	4
w.	Making generalizations and inferences	1	2	3	4
x.	Studying the style or structure of a text	1	2	3	4
y.	Comparing pictures and stories	1	2	3	4
z.	Student leading discussion about passage	1	2	3	4
aa.	Reading in other subject areas	1	2	3	4
bb.	Writing in response to reading	1	2	3	4

43. **Below you will find a number of statements about issues in reading instruction. Please state your degree of agreement/disagreement with each statement by circling the appropriate number.** *(Circle one number on each line.)*

		Strongly disagree	Disagree	Uncertain	Agree	Strongly agree
a.	When my students read to me, I expect them to read every word accurately	1	2	3	4	5
b.	Teachers should keep careful records of every student's reading progress	1	2	3	4	5
c.	Students should not be encouraged to read a word they don't know	1	2	3	4	5
d.	All students should enjoy reading	1	2	3	4	5
e.	Most of what a student reads should be assessed	1	2	3	4	5
f.	Every day students should be read to by the teacher from a story book	1	2	3	4	5
g.	Reading aloud by students to a class is a waste of time	1	2	3	4	5
h.	Most students improve their reading best by extensive reading on their own	1	2	3	4	5
i.	Students should always understand why they are reading	1	2	3	4	5
j.	Teachers should always group students according to their reading ability	1	2	3	4	5
k.	9-year-olds should not have access to books they will read in the next year at school	1	2	3	4	5
l.	Class sets of graded reading material should be used as the basis for the reading program	1	2	3	4	5
m.	Students who can't understand what they read haven't been taught proper comprehension skills	1	2	3	4	5
n.	Every mistake a student makes in reading aloud should be corrected at once	1	2	3	4	5
o.	All students' comprehension assignments should be marked carefully to provide them with feedback	1	2	3	4	5
p.	Students should not start a new book until they have finished the last	1	2	3	4	5
q.	Parents should be actively encouraged to help their students with reading	1	2	3	4	5
r.	Students should learn most of their new words from lessons designed to enhance their vocabulary	1	2	3	4	5
s.	Reading learning materials should be carefully sequenced in terms of language structures and vocabulary	1	2	3	4	5
t.	Students should take a book home to read every day	1	2	3	4	5

43. (Continued)

		Strongly disagree	Disagree	Uncertain	Agree	Strongly agree
u.	Students should be encouraged to read texts they have written	1	2	3	4	5
v.	Students should always understand what they are reading	1	2	3	4	5
w.	Students should always choose their own books to read	1	2	3	4	5
x.	A word recognition test is sufficient for assessing students' reading levels	1	2	3	4	5
y.	Teachers should carefully follow the sequence of the textbook	1	2	3	4	5
z.	Students should undertake research projects to improve their reading	1	2	3	4	5

46. **How often do you assess these aspects of reading with all or most of your class?** *(Circle one number per line only.)*

		Never	About once a year	About once a term	About once a month	About once a week or more
a.	Word recognition	1	2	3	4	5
b.	Vocabulary	1	2	3	4	5
c.	Text comprehension	1	2	3	4	5
d.	Literary appreciation	1	2	3	4	5
e.	Use of background knowledge	1	2	3	4	5
f.	Sentence understanding	1	2	3	4	5
g.	Phonic skills	1	2	3	4	5
h.	Reading study skills	1	2	3	4	5
i.	Amount of reading	1	2	3	4	5
j.	Decoding	1	2	3	4	5

53. **How often are the following teaching practices reflected in your class?** *(Circle only one number per line.)*

		Frequency				
			Less than once a week	1 or 2 times a week	3 or 4 times a week	More than 4 times a week
		Never				
a.	Students are assigned specific topics to study	1	2	3	4	5
b.	Students are told how what they know relates to a topic	1	2	3	4	5
c.	Students are informed as to the purposes of lessons	1	2	3	4	5
d.	Students receive feedback from the teacher on their ideas	1	2	3	4	5
e.	Students are directed to proceed based upon set guidelines	1	2	3	4	5
f.	Students deal with issues and topics related to their own experiences	1	2	3	4	5
g.	Students establish their own purposes and goals	1	2	3	4	5
h.	Students have a choice in what they will do	1	2	3	4	5
i.	Students decide how they will approach their texts	1	2	3	4	5
j.	Students share their ideas with each other	1	2	3	4	5
k.	Students are directed to answer a set of the teacher's questions	1	2	3	4	5
l.	Students are given feedback by the teacher on the themes or main ideas of the selections they read	1	2	3	4	5
m.	Students are given the opportunity to discuss various possible themes for the selection	1	2	3	4	5
n.	Spontaneous student responses are discouraged	1	2	3	4	5
o.	Students are encouraged to compare their written texts with other students' written texts	1	2	3	4	5

53. (Continued)

		Frequency				
		Never	Less than once a week	1 or 2 times a week	3 or 4 times a week	More than 4 times a week
p.	Students are encouraged to compare their written texts with the reading selection	1	2	3	4	5
q.	Students are given guided practice with skills	1	2	3	4	5
r.	Students are invited to consider how skills apply to what they have written	1	2	3	4	5
s.	Students are encouraged to work independently on classwork	1	2	3	4	5
t.	Spontaneous student responses are encouraged	1	2	3	4	5
u.	Students are encouraged to use the reading selection as a source for ideas when writing their texts	1	2	3	4	5
v.	Students are told what they have leamed and have yet to learn	1	2	3	4	5
w.	Students are given the opportunity to consider what they think they have learned, as well as their perception of their strengths and weaknesses	1	2	3	4	5
x.	Students are given the opportunity to assess their own progress	1	2	3	4	5
y.	Students are given the opportunity to provide input on how they will be assessed	1	2	3	4	5
z.	Specific skills are taught at certain times	1	2	3	4	5
aa.	Students are given teacher feedback on how they compare with other students	1	2	3	4	5
bb.	Students are expected to follow the activities outlined in the lesson the teacher has planned	1	2	3	4	5
cc.	Student needs necessitate changes to the lesson	1	2	3	4	5
dd.	Students are given the opportunity to work on a variety of different projects	1	2	3	4	5

8 Hierarchical Models: The Case of School Effects on Literacy

Stephen W. Raudenbush

National studies of school effects on pupils' educational achievement commonly involve a two-stage design in which schools are first sampled, and then, within each selected school, students are sampled. The aim in such studies is to link variation in school and teacher characteristics to variation in pupil outcomes, adjusting for exogenous pupil background and school context variables. Most studies of this type have assumed that pupil-level errors are independent and that school or teacher variables have uniform effects on all students. In light of these assumptions, two features of the U.S. portion of the IEA Reading Literacy Study are remarkable. First, the analytic model explicitly represents the nested structure of the data through random effects specification. Second, following Raudenbush and Bryk (1986), school effects are reconceptualized to alter the *distribution* of outcomes rather than just the *mean level* of outcomes. In the presence of many school, teacher, and pupil characteristics, this richer conceptualization confronts the researcher with dilemmas as well as opportunities for learning.

8.1 Background

Beginning with the Coleman report in the United States (Coleman et al. 1966) and the Plowden repurt in Britain (Plowden 1967), many researchers have used survey data to assess the effects of measurable school and teacher characteristics on pupils' educational outcomes. The Coleman and Plowden reports epitomize earlier efforts that focused primarily on relationships between physical resources and pupil outcomes, controlling differences in pupil background and school context.

The more recent wave of such studies, often based on more local samples, has emphasized features of schools as social organizations, including dimensions of climate or ethos (Rutter et al. 1979) and the organization and delivery of instruction in classrooms (Mortimore et al. 1988). Though the choice of critical explanatory variables has shifted over time, these studies have employed analytic models having in common the form

$$Y = f\,(\text{school and teacher characteristics} + \text{school context indicators} + \text{pupil covariates}) + \text{error}$$

where Y is a pupil outcome such as educational achievement. The model has two features that have been targets of persistent criticism regarding methodology.

1. The errors of the model have generally been assumed statistically independent despite the nested character of the design, in which students share membership in classrooms and schools.

2. The analyses have assumed implicitly that schools and teachers have uniform effects on their pupils. That is, the effect of, say, adding a school resource (e.g., a science laboratory) is assumed to be constant for every child in the school, even though in reality some children may never have access to that resource.

Burstein (1980) provided a comprehensive critique of school effects analyses based on these two assumptions and called for an alternative "multilevel" analytic strategy that would address both. Key characteristics of that strategy include specification of random effects to incorporate the shared effects of schools or classrooms on their members and the concept of "systematically varying slopes" or "slopes as outcomes" to include in the models the possibility that schools and classrooms might affect the distribution of outcomes as a function, say, of social class, sex, or ethnicity.

Widespread use of the multilevel strategy, however, was delayed by the limitations of estimation procedures for variance components models in the face of the "messy" data yielded in large-scale field studies. School effects survey designs are, in general, unbalanced, nested designs with random effects of schools and fixed effects of covariates described at each level. Covariates may be discrete or continuous at any level. The idea that schools affect the social distribution of outcomes implies that regression coefficients as well as regression intercepts ought to be specified as random, and, to be realistic, models must allow these random components defined on the same schools (or classrooms) to covary. Hence, to implement the multilevel conceptualization requires a family of *co*variance components models for nested unbalanced data, models that allow flexible incorporation of covariates at each level.

The kind of model that fits these demanding specifications has, in fact, been around since the publication of Lindley and Smith's (1972) classic article that introduced a hierarchical linear model with Bayes estimation theory. However, efficient estimation of the parameters of this model required first algorithmic advances beginning with the estimation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977).

During the mideighties, several investigators, working independently, developed hierarchical linear models of the type envisioned by Lindley and Smith and illustrated their application in the context of educational survey research (Aitkin and Longford 1986; deLeeuw and Kreft 1986; Goldstein 1987; Longford 1987; Raudenbush and Bryk 1986). Dempster, Laird, and Rubin (1977) provided statistical theory for this approach based on EM, which found early application in cross-national demography (Mason, Wong, and Entwistle 1983). Raudenbush (1988) reviews the statistical theory and applications of these approaches. Bryk and Raudenbush (1992) provide a comprehensive discussion of alternative models and data analysis procedures with many detailed examples.

8.2 A Hierarchical Linear Model

Raudenbush and Bryk (1986) used a hierarchical linear model to represent characteristics of the distribution of educational achievement as a set of multiple outcomes. Their level-1 or within-school model conceived of each school as having "its own" regression equation relating student social class to mathematics achievement:

$$Y_{ij} = \beta_{oj} + \beta_{1j} X_{ij} + r_{ij}, \qquad (1)$$
$$r_{ij} \sim N(0, \sigma^2).$$

where Y_{ij} is the math achievement of student i in school j and X_{ij} is a measure of student social class for that student centered around group mean. The model intercept, β_{oj}, represented the mean level of outcomes (educational excellence) while the slope, β_{1j}, represented the strength of association between social class and outcomes (educational equity). The ideal school would produce both excellence (high average achievement) and equity (weak effects of social class on achievement). At level 2 (between schools) the intercept and slope were therefore conceived as outcome variables, i.e,

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \sum_{s=1}^{S_0} \gamma_{0s} W_{sj} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \sum_{s=1}^{S_0} \gamma_{1s} W_{sj} + u_{1j} \end{aligned} \qquad (2)$$

Here the *W*s are school characteristics hypothesized to influence the level of excellence and equity within a school.

In essence, the level-1 model describes the distribution of outcomes within each school in terms of two school-specific parameters, an intercept and a slope. The level-2 model describes the joint distribution of those parameters across the entire population of schools. Student-level covariates can be added at level 1. School-level covariates can be added at level 2. The level-2 model errors are the random effects associated with school *j* and are assumed bivariate normal

$$\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix}\right]. \qquad (3)$$

By constraining the slope to have no variance, equation 3 could be modified to become

$$\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & 0 \\ 0 & 0 \end{pmatrix}\right]. \qquad (4)$$

In this case, the model is a simple variance component or random intercept model. If the intercept were constrained to zero as well, the level 2 would have no random components so that the model would reduce to an ordinary least squares regression model. All parameters are estimated by means of restricted maximum likelihood using the EM algorithm. Likelihood ratio tests are therefore available to test the appropriateness of simplifying the covariance structure of the model.

The explanatory power of the model at each level can be monitored by examining the reduction in estimated variance at each level. The explanatory power of the level-1 model is assessed by estimating equation 1 as well as an "unconditional" level-1 model

$$\begin{aligned} Y_{ij} &= \beta_{oj} + r_{ij}, \\ r_{ij} &\sim N\,(0,\ \sigma^2_{uncond}). \end{aligned} \tag{5}$$

The explanatory power of the level-1 model is then estimated to be

$$R^2_{level\ 1} = \frac{\hat{\sigma}^2 - \hat{\sigma}^2_{uncond}}{\hat{\sigma}^2_{uncond}}. \tag{6}$$

To evaluate the explanatory power of the level-2 model, the level-1 model must be held constant. Suppose the level-1 model were that specified in equation 1. Then an unconditional level-2 model could be estimated

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \end{aligned} \tag{7}$$

where

$$\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00uncond} & \tau_{01uncond} \\ \tau_{10uncond} & T_{11uncond} \end{pmatrix} \right]. \tag{8}$$

The explanatory power of the level-2 model could then be estimated by comparing the unconditional and conditional variances. For example, the power of the model in accounting for variance in the slope would be

$$R^2_{level\ 2} = \frac{\hat{\tau}_{11} - \hat{\tau}_{11uncond}}{\hat{\tau}_{11uncond}}. \tag{9}$$

However, one must use the same *Ws* in the equation for the intercept as in the equation for the slope in order to justify this interpretation.

8.3 Advantages of HLM

The advantages of using hierarchical linear models rather than ordinary least squares regression are that the hierarchical models often give better answers to old questions and may make it possible to get answers to new questions.

Better Answers to Old Questions

The standard applications of regression in school effects research are focused on main effects of variables measured at the pupil and school (or class) level. However, the level-1 effects may be biased if the nested structure of the data is ignored, and estimated standard errors of the level-2 effects will be negatively biased (see Bryk and Raudenbush 1992, chapter 5). Standard regression models have also been widely tailored to discover especially effective or ineffective schools. The approach involves the computation of school-level residuals—discrepancies between a school's mean achievement and the achievement expected on the basis of its student intake. However, the search for effective schools based on this residual analysis can become highly distorted. The larger the number of schools and the smaller the sample of students per school (or classroom), the more likely the researcher is simply to discover artifacts of chance. However, by facilitating estimation of variance components, hierarchical linear models enable the researcher to compute improved empirical Bayes estimates of model residuals (the *u*'s). The empirical Bayes approach shrinks least squares residuals toward 0; the degree of shrinkage is proportional to the unreliability of the least squares residual. Advantages of such estimators are discussed in Morris (1983) and Raudenbush (1988).

Answers to New Questions

Bryk and Raudenbush (1992) describe three kinds of effects that can now be readily addressed—effects that were inaccessible or difficult to assess with standard met ods. One is now able

1. To estimate effects of schools or classrooms on the distribution of outcomes (as in model 2 above);

2. To estimate components of variation and covariation at each level; and

3. To compute empirical Bayes estimates of coefficients that cannot be estimated because of deficient rank data (Braun et al. 1983).

The first two effects, which we shall consider next, are of central interest in the IEA Reading Literacy Study. By estimating these effects, the researcher is encouraged to reconceptualize school and classroom effects research. Rather than conceiving the outcome as school mean achievement, the researcher can conceptualize the social distribution of achievement as the outcome. That distribution is affected by schoolwide differences in resources, processes, and contexts, differences that presumably lead to between-school differences in outcomes; but it is also affected by within-school differences in how children are grouped and taught. These within-school differences modify each school's social distribution of achievement (Lee and Bryk 1989). The level-1 model characterizes each school's distribution of outcomes. The distribution of the level-ı parameters across schools captures the between-school differences.

8.4 Utilization of the Model in the IEA Reading Literacy Study

The General Modeling Strategy

The analytic team of the IEA Reading Literacy Study has adopted a two-phase procedure using the framework of the hierarchical linear model (Williams 1994). The first phase involves deciding on an appropriate level-1 model, estimating the parameters of that model, and developing interpretations of the parameter estimates. The products of this phase of the work are (a) a set of statements about the characteristics of pupils and their families that predict reading literacy, and (b) an associated set of statements about how schools vary in the importance of these pupil characteristics as determinants of achievement. For example, the wealth of a student's family may be (a) a moderately important predictor of reading achievement overall, but (b) the importance of wealth on achievement may be much greater in some schools than in others. Statements of type "a" (statements about the average effect of a pupil variable) depend on estimation of fixed effects in the hierarchical model, while statements of type "b" (statements about the varying effect of a predictor) depend on estimation of the covariance components.

Phase 2 assesses the effects of school and class variables on the level-1 parameters. Two types of statements emanate from this phase: (a) statements about the effects of school- or class-level variables on average achievement within schools (controlling student covariates); and (b) statements about the effects of school- or class-level variables on distributional equity (i.e., statements about the links between school- or class-level variables and level-1 slopes).

In its general structure, this modeling strategy of the IEA Reading Literacy Study is not different from that described by Raudenbush and Bryk (1986). What makes implementation quite different, however, is the large number of candidate variables in the study. The IEA Reading Literacy Study involves 19 level-1 variables and a potentially even larger number of level-2 variables.

The Problem of Many Variables

The IEA Reading Literacy Study involves a reasonably large sample. Restricting our attention to the data for fourth graders, there are 6,428 students nested within 303 classes and 167 schools. However, a model with 19 random slopes and a random intercept would produce up to 210 variance components and $20(S+1)$ fixed effects estimates where S is the number of school or class variables. The sample is not large enough to support estimation of so many parameters. What is to be done? Three solutions might be advocated.

First, the researcher might impose a priori theory. The model could be prespecified to include only those predictors suggested strongly in the literature to be important, or only those predictors of interest to the researcher. A variant on this approach is to prespecify alternative models representing competing theories. However, the IEA Reading Literacy Study is a broad-based, publicly supported survey conducted in conjunction with parallel surveys in a number of countries. The research team would have had considerable difficulty in justifying the imposition of a single favorite theory. Nor is the literature on school or class effects sufficiently cry allized to identify a small set of theories as relevant. So while it is possible to identify a broad set of variables as reasonably comprehensive (excluding, therefore, an infinite number of other variables as irrelevant), it is hard to go much further in theoretical specification without damaging the mission of the study.

Secondly, the investigators might go to the other extreme and propose a purely data-based approach to model simplification—a multilevel version of step-wise regression. However, we know from experience in standard regression that when there are many candidate variables, the resulting fitted model will not be robust under cross-validation; and the absolute values of the coefficient estimates will be positively biased while the standard error estimates will be negatively biased.

A third possible strategy is that chosen by the researchers. Variables at level 1 were first classified in conceptually related blocks. For example, level-1 variables were divided into *status* variables (e.g., gender, ethnicity, parental education, and wealth) and *process* (or *intervening*) variables (e.g., books in the home, hours spent on television, and parental help with homework). For phase one, that is, specification of the level-1 model, the analysis was conducted separately for each block. Within each block, empirical procedures were used to decide (a) which level-1 variables should be specified as having random effects, (b) which should have fixed effects but no random effects, and (c) which should be discarded from the model. The survivors of the separate analyses for each block were then combined into a pooled analysis. This permitted estimation of total, direct, and indirect effects of the status variables found important and of direct effects of the intervening variables found important. Once the level-1 model was decided upon, the phase-two analysis proceeded in a similar manner. Level-2 variables were blocked and the blocks ordered causally. Analyses within blocks relied on empirical indicators (*t*-values) to discard some variables. Total, direct, and indirect effects of the survivors were then estimated.

An Example of Empirically Based Model Selection

Within phase one, that is, specification of the level-1 model, the first task was to study the effects of the status variables, of which there were nine. However, even this part of the analysis posed complications, because the data were insufficient to support estimation of 10 random coefficients (9 slopes plus the intercept). That is, the joint distribution of 10 random coefficients involves 55 variance-covariance parameters; yet there are, on average, only about 22 students per classroom. Again the dilemma facing the researchers contrasts theoretically driven versus empirically driven variable-selection procedures. The analysts' decision again used some of both. The nine variables were broken down into three subblocks: personal characteristics (ethnicity, sex, age); socioeconomic indicators (father's education, mother's education, family wealth); and family structure/culture (family composition, nuclear versus extended family, language). Within subblocks, decision making was empirically informed. For example, within the socioeconomic indicators subblock, the following level-1 model was estimated:

$$\begin{aligned} Y_{ij} &= B_{0j} + B_{1j}\,(\mathit{Dad.Ed.})_{ij} + B_{2j}\,(\mathit{Mom.Ed.})_{ij} \\ &\quad + B_{3j}\,(\mathit{Wealth})_{ij} + B_{4j}\,(\mathit{Age})_{ij} + B_{5j}\,(\mathit{Sex})_{ij} \\ &\quad + B_{6j}\,(\mathit{Ethnic})_{ij} + B_{7j}\,(\mathit{Ext.Fam.})_{ij} \\ &\quad + B_{8j}(\mathit{Fam.Comp.})_{ij} + \mathrm{B}_{9j}(\mathit{Language})_{ij} + r_{ij}, \\ &\quad r_{ij} \sim N\,(0,\sigma^2) \end{aligned} \tag{10}$$

where *Y* was a measure of reading comprehension. At level 2 the model was

$$\begin{aligned} B_{0j} &= \gamma_{00} + u_{0j} \\ B_{1j} &= \gamma_{10} + u_{1j} \\ B_{2j} &= \gamma_{20} + u_{2j} \\ B_{3j} &= \gamma_{30} + u_{3j} \\ B_{pj} &= \gamma_{po},\ p > 3. \end{aligned} \tag{11}$$

However, even within this level-2 model, three models were estimated. Each represented a more or less simple covariance structure. The first model included 11 covariance components, namely

$$\begin{bmatrix} \tau_{00} & \tau_{01} & \tau_{02} & \tau_{03} \\ & \tau_{11} & \tau_{12} & \tau_{13} \\ & & \tau_{22} & \tau_{23} \\ & & & \tau_{33} \end{bmatrix}, \ \sigma^2. \tag{12}$$

Given the estimates of the first model, it was decided that the second model should constrain the mother education slope to have zero variance, so that seven covariance components were to be estimated, namely

$$\begin{bmatrix} \tau_{00} & \tau_{01} & 0 & \tau_{03} \\ & \tau_{11} & 0 & \tau_{13} \\ & & 0 & 0 \\ & & & \tau_{33} \end{bmatrix}, \ \sigma^2. \tag{13}$$

Next, a model was estimated with only the wealth slope random, so that four parameters were estimated:

$$\begin{bmatrix} \tau_{00} & 0 & 0 & \tau_{03} \\ & 0 & 0 & 0 \\ & & 0 & 0 \\ & & & \tau_{33} \end{bmatrix}, \ \sigma^2. \tag{14}$$

Finally, a model with a random intercept was estimated, leaving two parameters to estimate:

$$\begin{bmatrix} \tau_{00} & 0 & 0 & 0 \\ & 0 & 0 & 0 \\ & & 0 & 0 \\ & & & 0 \end{bmatrix}, \ \sigma^2. \tag{15}$$

Results of the estimation of the four models are given in Table 8-1.

Table 8-1. Comparison of covariance components models

Model	Free slopes	Number of parameters	Deviance
1	Father's education Mother's education Wealth	11	72981.6
2	Father's education Wealth	7	72987.5
3	Wealth	4	72994.9
4	None	2	73002.9

SOURCE: IEA Reading Literacy Study, U.S. National Study data, National Center for Education Statistics, 1991.

In comparing models, the difference between deviances has an asymptotic chi-square distribution with degrees of freedom equal to the difference between the numbers of parameters for the two models. Large values of chi-square imply that the simplified model is not justifiable. The results suggest that the simplest model cannot be justified when compared to any of the others. However, the model with just wealth random (model 3) can be justified against any of the others. Hence, the investigators settled on a model with the wealth slope random and the others fixed.

Within the subblocks of variables representing pupil and family status, only the ethnicity and wealth slopes showed signs of significant randomness. When both wealth and ethnicity were included, results suggested that a model with the ethnicity effect random and the wealth effect fixed was adequate. As a result, the phase-one analysis of student status suggested that subsequent models should represent the social distribution of achievement in terms of equity-based ethnic minority status, adjusted for the fixed effect of wealth. Variation in this slope and in the intercept therefore constituted the primary targets of interest in subsequent models employing level-2 variables. More details about the selection of student status and process variables are included in Chapter 14 of the technical report for the study (Williams 1994).

8.5 Conclusions

The IEA Reading Literacy Study has adopted an analytic strategy of studying school effects on educational achievement using a hierarchical linear model with random coefficients. The model explicitly represents the clustered character of the data, thereby enabling the researchers to avoid some of the pitfalls of past studies of school effects, namely, biased effects of pupil background and negatively biased standard error estimates for the level-2 coefficients. More importantly, the model facilitates a richer conceptualization of school effects. Instead of assuming that schools or teachers have uniform effects on every student under their supervision, the model enables the researcher to formulate the social distribution of achievement as the outcome. This reformulation is more realistic in allowing that organizational processes differentially affect organization members; and it opens for policy consideration issues of equity based on social class, sex, and gender.

Because hierarchical analysis allows a richer set of models, however, data analytic decision making in the face of many explanatory variables, always a concern in standard regression analysis, becomes even more complex. Not only must the researcher decide, as in conventional analysis, which main effects to incorporate and which to discard, he or she must also decide which slopes should be random and which slopes explicitly modeled as functions of level-2 variables. This multiplicity of possibilities calls for an emphasis on theoretically driven models as opposed to the use of empirical techniques such as hypothesis testing to inform decisions about model simplification. However, in public policy research in general, and cross-national comparative studies like the IEA Reading Literacy Study in particular, strict a priori model specification may not be feasible.

The IEA Reading Literacy Study approach represents an attempt to find a compromise between model specification based purely on theory and model specification based purely on empirical results. The result in this case is promising: both empirical and theoretical grounds support conceptualizing equity differences between schools in terms of equity with respect to ethnic minority status and family wealth. Of course, there is no guarantee that others using this approach will arrive at such conceptually sensible conclusions. The problem of model specification in the presence of many variables is not solved in the case of standard regression models and so cannot be solved for the richer class of models considered here.

The future holds promise for meaningful extension of these analyses. First, the empirical Bayes approach to estimating effects of individual schools may prove valuable given the IEA's interest in identifying and understanding especially effective schools. Second, the hierarchical model provides a promising approach to cross-national comparisons, as discussed in Chapter 9 of this volume. The country becomes the third level in the model, facilitating study of country-level variation in mean outcomes as well as in the equity with which achievement is distributed. Third, Raudenbush (1988) describes how the model can readily be expanded to incorporate multiple waves of data, including either panel data on students or panel data on countries (repeated cross-sections within each country).

References

Aitkin, M., and Longford, N. (1986). Statistical modelıng issues in school effectiveness studies. *Journal of the Royal Statistical Society*, Series A, 149 (1), 1-43.

Braun, H.I., Jones, D.H., Rubin, D.B., and Thayer, D.T. (1983). Empirical Bayes estimation of coefficients in the general linear model with data of deficient rank. *Psychometrika*, 48 (2), 171-181.

Bryk, A.S., and Raudenbush, S.W. (1992). *Hierarchical linear models: Applications and data analysis techniques*. Newbury Park, CA: Sage.

Burstein, L. (1980). The analysis of multi-level data in educational research and evaluation. *Review of Research in Education*, 8, 158-233.

Coleman, J.S., Campbell, E.Q., Hobson, C.F., McPartland, J., Mood, A.M., Weinfeld, F.D., and York, R.L. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Government Printing Office.

DeLeeuw, J., and Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics*, 11, (1), 57-85.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Series B, 39, 1-8.

Goldstein, H.I. (1987). *Multilevel models in educationa! and social research*. London: Oxford University Press.

Lee, V., and Bryk, A.S. (1989). A multilevel model of the social distribution of high school achievement. *Sociology of Education*, 62, 172-192.

Lindley, D.V., and Smith, A.F.M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society*, Series B, 34, 1-14.

Mason, W.M., Wong, G.M., and Entwistle, B. (1983). Contextual analysis through the multilevel linear model. In S. Leinhardt (ed.), *Sociological methodology*, 72-103. San Francisco: Jossey-Bass.

Longford, N. (1987). A fast-scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, 74, (4), 817-827.

Morris, C. (1983). Parametric empirical Bayes inference: Theory and applications. *Jc⁻rnal of the American Statistical Association*, 78, (381), 47-65.

Mortimore, P., Sammons, P., Stoll, L., Lewis, D., and Ecob, R. (1988). *School matters: The junior years*. Wells, Somerset: Open Books.

Plowden, E. (1967) *Children and their primary schools*. London: Central Advisory Council for Education (United Kingdom).

Raudenbush, S.W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics*, 13, (2), 85-116.

Raudenbush, S.W., and Bryk, A.S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, 59, 1-17.

Rutter, M., Maughan, B., Mortimore, P., Ousten, J., and Smith, M. (1979). *Fifteen thousand hours: secondary schools and their effects on children*.

Williams, T. (1994). Modeling the reading literacy ot fourth and ninth graders. *Reading Literacy in the United States: Technical Report*, Chapter 14. (NCES 94-259). Washington, DC: U.S. Government Printing Office.

9 Synthesizing Cross-National Classroom Effects Data: Alternative Models and Methods

Steven W. Raudenbush, Yuk Fai Cheong, and Randall P. Fotiu

9.1 Introduction

As concern for improving educational effectiveness has intensified during recent years, policymakers in many countries have demanded better information about the outcomes of schooling (Willms 1992). Governments in a number of countries have recently developed performance indicators for the purpose of monitoring levels of achievement, often with the aim of holding national, provincial, or local officials accountable for the productivity of schools under their supervision (Fitz-Gibbon 1991; Bosker and Guldemond 1991; Wheeler, Raudenbush, and Pasigna 1992). The policy of using better information to guide educational improvement poses, within each country, a series of methodological issues and dilemmas discussed in detail by Willms (1992).

The policy environments impelling governments to monitor learning within each society also inspire a need to synthesize information about educational outcomes across societies. It is widely held within many countries that improved educational attainment is a key to improved global competitiveness. Thus it is natural for policymakers within a country to inquire about the standing of their country relative to other countries with respect to educational attainment, and a finding that one's country is faring poorly in the educational competition becomes the occasion for intensified efforts to improve schooling within that country.

Moreover, the availability of cross-national data on learning enables policy-relevant research that cannot be conducted with data on particular society. For example, it might be argued that a nation's level of educational attainment could be enhanced by lengthening the school year. Such a policy option might be evaluated by asking whether schools requiring more days of attendance have higher mean achievement than similar schools requiring fewer days of attendance. However, the length of the school year varies little or not at all within some countries. For those countries, a researcher proposing to study the relationship between the length of the school year and educational achievement requires cross-national data.

Given the current policy climate in many countries, it is hardly surprising to find growing interest in international surveys of educational outcomes. However, the valid synthesis of such cross-national survey data poses a host of methodological challenges in addition to the challenges that arise within a country that attempts to study its own level of educational attainment. The primary purpose of this paper is to consider issues of statistical modeling that arise in synthesizing cross-national achievement data. In pursuing this purpose, we shall assume that comparable measures of achievement and predictors of achievement are available across countries. Thus, we are side-stepping one of the most difficult challenges in cross-national research—developing such comparable measures—in the interest of focusing on statistical issues.

This chapter is organized as follows. In Section 9.2 we consider three uses of cross-national data: comparing countries' literacy levels; testing cross-national hypotheses about predictors of mean literacy; and studying cross-national differences in the equity in the distribution of literacy. Associated with each use of cross-national data is a set of characteristic problems of statistical inference, and we consider these in each case. Section 9.3 proposes a general statistical framework for model formulation and estimation, indicating how this framework addresses the issues of statistical inference that arise in the three uses of cross-national data. Section 9.4 briefly describes the illustrative data, drawn from the IEA Reading Literacy Study. Section 9.5 illustrates application of the statistical methodology to each of the three uses of cross-national data. The final section considers application of the methodology to other cross-national research problems.

9.2 Three Uses of Cross-National Data

Having data from multiple countries makes it possible to ask a variety of questions that cannot be studied when data are available on only one country. We shall consider three broad kinds of questions that become accessible. First and most basically, cross-national data allow one to assess the level of mean literacy in a country relative to the level of literacy in other countries. Second, cross-national data allow one to test a hypothesis about how differences between countries are related to mean literacy. Third, cross-national data allow one to study differences in the equity of the distribution of literacy within countries. A classic example of the third use is Heyneman and Loxley's (1983) hypothesis that effects of student socioeconomic status on student achievement are more pronounced in highly developed nations than in less developed nations. Associated with each type of question is a set of inferential challenges. Our understanding of these challenges motivates our choice of statistical model and estimation procedure.

9.2.1 Comparing Country Means

Perhaps the simplest use of cross-national data is to compare countries with respect to their mean achievement. Such comparisons might involve contrasting a particular pair of countries or locating a particular country within a distribution of country means. The seeming simplicity of this task should not obscure substantial obstacles, both statistical and conceptual, to valid inference.

Taking Uncertainty into Account. Comparing point estimates of country means will be misleading to the extent that significant uncertainty is associated with those point estimates. Thus, confidence intervals and significance levels are needed. These must be computed with care because, within each country, data will typically be collected via a multistage cluster sample. For example, in most countries participating in the IEA Reading Literacy Study, which provides data for the illustrative examples below, schools were first selected, and then, within schools, students were selected.[1] Thus, standard errors for country-level mean estimates must take into account the extra component of variability associated with schools. Methods that ignore the clustering within schools will underestimate the standard error of the country mean. Two common approaches are used to incorporate the clustering effect: resampling approaches, such as the bootstrap or jackknife, and model-based approaches.

[1] In most cases, one classroom was selected at random within each school.

In addition to the clustering effect, researchers will need to consider the possible effects of stratification. Clusters, or students within clusters, may be sampled with unequal probability to achieve adequate representation of rare cluster types or student types. Appropriate weighting of observations is required to achieve unbiased estimates of country means and unbiased standard error estimates.

To cope with effects of clustering and stratification, we have opted to use a hierarchical linear model (Raudenbush and Bryk 1986) within each country. Effects of clusters are represented via random effects, the variance of which is incorporated into standard error estimates for means. The HLM3.0 program (Bryk et al. 1988) allows unequal weighting at either the cluster or the student level to account for stratification. Although the resampling approaches might be viewed as more robust than the model-based approach, the model-based approach extends better to the more complex estimation tasks described below.

Assessing Between-Country Heterogeneity. The substantive significance of a mean difference between two countries can be assessed by comparing that mean difference to the variation within countries (e.g., via a standardized effect size). In addition, a summary measure of between-country heterogeneity may also be important. If the between-country component of variability were trivially small, for example, the search for country differences would take on far less urgency than if between-country variability were large. However, estimating the extent of between-country variability is nontrivial statistically because each country's mean is estimated with different precision. Thus, an iterative computational procedure is needed to estimate the between-country variance. More important, given a modest number of countries (n=22 in the analyses below), a point estimate of the between-country variance will be imprecise. A confidence interval is needed, but large-sample confidence intervals based on the asymptotic normality of maximum likelihood estimators often will be inappropriate in this small-sample setting.

A more profound conceptual problem is in interpreting a measure of between-country variance. In what sense are countries random? For example, the IEA Reading Literacy Study countries volunteered for the study, and so they cannot constitute a random sample.

To address the problems of estimating and interpreting between-country heterogeneity, we adopt a Bayesian approach with estimation via Gibbs sampling (Gelfand and Smith 1990; Seltzer 1993). In the Bayesian framework, the between-country variance represents the investigator's uncertainty about the degree to which countries vary in their means. Thus, we need not assume countries to have been sampled randomly. We postulate a relatively noninformative prior distribution for the between-country variance component. Then the posterior distribution of this variance gives us a range of plausible values of the extent of between-country heterogeneity and, for each value, a degree of plausibility (technically the posterior density). These posterior distributions can readily be displayed and explained to nontechnical audiences.

A final and important advantage of the Bayesian approach to studying between-country variation is that it extends well to the more complex modeling tasks described below. Thus, when we estimate the relationship between gross national product (GNP) and country means, the standard error of the estimated regression coefficient will take fully into account the uncertainty with which the degree of between-country heterogeneity is estimated. Rubin (1981) lucidly describes the difficulties that arise with maximum likelihood estimation in this case and advocates the Bayesian method in this type of setting where the number of countries is small.

Comparability of Countries. It might be argued that countries should be compared with respect to mean literacy only if those countries are comparable in other ways. For example, one might wish to compare the means of countries that have similar resources as indicated by GNP. A poor country with low mean literacy might, nonetheless, be viewed as having a relatively efficient educational system if it scores higher than other countries of similar GNP. Thus, rather than comparing simple means, one might estimate the effect of GNP on literacy and compute GNP-adjusted "country effects." We consider the problem of estimating relationships between country-level characteristics (such as GNP) and achievement means in Section 9.2.2.

Differential Demographic Effects. Comparing means tells us nothing about the equity of distribution of the outcome. Thus, a country with high average literacy might nonetheless be comparatively ineffective for some children. For example, such a country might produce large literacy gaps between males and females or between rich and poor students, and the high average literacy of that country would provide little consolation to those who are so disadvantaged. It is well known in experimental research that interpreting main effects of treatments in the presence of statistical interactions between treatments and subject background can be highly misleading. Similarly, to compare country means is to assess main effects of countries, and these main effects will be misleading if country interacts with student demography. We consider models for country variation in the equity of distribution of literacy in Section 9.2.3.

9.2.2 Testing Hypotheses About the Relationship Between Country Characteristics and Mean Literacy

Though comparing country means is a plausible use of cross-national literacy data, many researchers would seek to account for the variability among country means. Do countries with longer school years score higher than similar countries with shorter school years? How do countries with national examinations compare to similar countries without such examinations? What relations with mean literacy are associated with having a national curriculum? Answers to such questions may be of interest to national policymakers considering options for increasing literacy. Again, however, challenges to valid statistical inference arise.

Controlling Confounding Variables. The policy-relevant questions defined above require that we compare countries with different policies that are similar in other regards such as student background, school resources, and gross national product. The implication is that covariates at each level should be measured and incorporated into the analysis.

Efficient Estimation and Valid Assessment of Uncertainty. Country means, whether adjusted or unadjusted for student and school characteristics, will be estimated with unequal precision because of the varying sample sizes across countries and because of the varying explanatory power of the student-level and school-level predictors within countries. Efficient estimation requires that the varying precision of the outcome from each country be taken into account via weighted least squares (Seber 1978). However, the precision of the country mean (the inverse of its variance) depends not only on the data within each country but also on the variance between countries. Let b_k denote the estimated mean outcome for country k and let β_k denote the "true" mean. We may write

$$\begin{aligned} b_k &= \beta_k + e_k, \\ e_k &\sim N(0, v_k), \end{aligned} \qquad (1)$$

that is, e_k is the error by which b_k estimates β_k and ν_k is thus the sampling variance of b_k. However, the true means β_k, $k = 1,\ldots,K$ are themselves viewed as randomly varying about their predicted values. For example, using GNP as a predictor, we have

$$\begin{gathered}\beta_k=\gamma_0+\gamma_1(GNP)_k+u_k,\\ u_k \sim N(0,\tau),\end{gathered} \tag{2}$$

where u_k is the unique effect associated with country k assumed normally distributed and τ is the between-country variance. Combining equations (1) and (2), we have

$$\begin{gathered}b_k=\gamma_0+\gamma_1(GNP)_k+ u_k+ e_k,\\ u_k+ e_k \sim N(0,\tau+\nu_k).\end{gathered} \tag{3}$$

Under the model of equation (3) and with τ and ν_k known, the maximum likelihood estimator of the GNP coefficient and its variance are given by weighted least squares with weights $\omega_k = 1(\tau + \nu_k)$ according to the formulas

$$\hat{\gamma}_1=\frac{\sum \omega_k(GNP_k-G\bar{N}P)(b_k-\bar{b})}{\sum \omega_k(GNP_k-G\bar{N}P)^2} \tag{4}$$

and

$$Var(\hat{\gamma}_1)=\frac{1}{\sum \omega_k(GNP_k-G\bar{N}P)^2} \tag{5}$$

where

$$G\bar{N}P=\frac{\sum \omega_k GNP_k}{\sum \omega_k} \quad \text{and} \quad \bar{b}=\frac{\sum \omega_k b_k}{\sum \omega_k}.$$

When the data are balanced, the weights ω_k are equal for every country and equation (4) reduces to ordinary least squares, eliminating dependence of the coefficient estimate on τ. Moreover, equation (5) simplifies and an exact t test becomes available, eliminating dependence of hypothesis testing on τ. (See Raudenbush (1992) for detailed applications in the balanced case.)

However, when the data are unbalanced, which will generally be the case in international studies, equations (4) and (5) will depend upon τ via the dependence of the weights ω_k on τ, and τ will not be known.[2] When τ is not known, the maximum likelihood estimates (MLEs) of γ_1 and its standard error are equations (4) and (5) with the MLE of τ substituted in the construction of ω_k. These MLEs will be sensible when τ is estimated with reasonable precision. However, this precision depends heavily on the number of countries, which will tend to be limited (K=22 in our case). When the precision is poor, equation (5) will underestimate the uncertainty associated with the MLE of γ_1.

Our strategy for coping with the small number of countries and the consequent limited precision of the MLE of τ is to employ the Bayesian strategy described in Section 9.2.1. Using this approach, the posterior distribution of γ_1 gives a range of plausible values for that parameter, and, associated with each

[2] The sampling variance ν_k can be precisely estimated and assumed known, given the large amount of data typically gathered within countries in international educational surveys.

value, its degree of plausibility (posterior density). This posterior density fully incorporates the uncertainty about τ. An important byproduct of this analysis is a good approximation to the posterior density of τ itself, which indicates the range of plausible degrees of heterogeneity in country means that remains after controlling for the effects of GNP.

Interactions with Demographic Background. We mentioned earlier that comparisons between country means are misleading when countries vary in the equity of distribution of achievement. Similarly, statements about the effects of a country-level predictor on country-level mean outcomes will be misleading if the effect of that predictor differs for different demographic groups. We turn our attention to modeling such interactions in the next section.

9.2.3 Studying Cross-National Differences in the Equity of the Literacy Distribution

In a widely cited article, Heyneman,and Loxley (1983) reported an analysis of data from 29 countries indicating that student social status was comparatively less important for predicting educational achievement in developing nations than in developed nations. Within a multilevel modeling framework, we might therefore hypothesize that the greater the level of economic development of a country, as measured by GNP, the larger the magnitude of the regression coefficient relating student social status to academic achievement. This hypothesis exemplifies a broader class of important cross-national hypotheses concerning relationships between country characteristics and the distribution of outcomes within countries. Mason, Wong, and Entwistle (1983-84) have employed multilevel models to investigate similar cross-national hypotheses in their work on the world fertility survey. Once again, inferential challenges arise in this setting, and our modeling and estimation strategies are designed to cope with these.

Individual and Contextual Effects of Demographic Background. A multilevel perspective on social status shows that the link between student social status and an educational outcome has two components: a student-level component and a contextual component (Burstein 1980; Raudenbush and Bryk 1986). The student-level component indicates the extent to which students attending the same school but varying in social status vary on the outcome of interest. The contextual component indicates the extent to which attendance at schools having varied social status compositions has consequences for students who are similar in personal social status. A common finding in school effects research is that the social status composition of the school predicts achievement even after controlling for the social status of persons (see Willms 1986 for a review). Thus, the varying association between social status and educational achievement across societies, as reported by Heyneman and Loxley (1983), reflects a blend of two effects. The student-level effect can apparently be reduced only by equalizing resources available to students within a school who vary in social status.[3] However, to reduce the contextual effect requires a potentially different set of policy options, including reducing between-school segregation by social status and equalizing resources available to schools. To render the association between social status and achievement interpretable requires disentangling the individual and contextual effects.

Efficient Estimation and Valid Assessment of Uncertainty. In formulating models for the relationship between country characteristics and regression coefficients characterizing demographic effects, the same statistical concerns discussed earlier with respect to means as outcomes arise. Specifically, the precision of an estimated regression coefficient for a given country will depend both on the information in that country's data and on the between-country variance in the coefficient of interest (see equations (1) to (3) above). The IEA Reading Literacy Study data provide sufficient information to

[3] An alternative policy option is to reduce or eliminate social status differences between students, but we consider this option beyond the purview of educational policy.

estimate the within-country parameters with good precision. However, the small number of countries available results in estimates of between-country parameters that are comparatively imprecise. One goal of the analysis is to accurately reflect these multiple sources of uncertainty in our inferences. The next section considers our strategy to accomplish this goal.

9.3 Statistical Methodology

9.3.1 Overview

We have selected a statistical approach that is designed to facilitate the kinds of uses described in the previous section: comparing country means, testing hypotheses about relationships between country characteristics and country means, and studying country differences in the equity of distribution of outcomes within countries. The approach is tailored to take into account the multilevel design within each country and to represent between-country heterogeneity via random effects defined on countries.

Formally, the data collected by the IEA Reading Literacy Study have a three-level hierarchical structure: students are nested within classrooms, which, in turn, are nested within countries.[4] Such a structure suggests analysis by means of a three-level hierarchical linear model as described, for example, by Goldstein (1987) or Bryk and Raudenbush (1992). However, these data have special characteristics that contradict the assumptions of standard three-level applications. Our approach is adapted to these special characteristics, as described below.

Sparse Data at Level 3. As described in Section 9.2, inferences about fixed effects (regression coefficients) based on maximum likelihood are conditional on point estimates of variance and covariance components when the data are unbalanced. The dependence of key inferences on point estimates of these variance-covariance components poses no serious problem when adequate data are available at the highest level of the hierarchy. However, international studies of achievement will typically include a modest number of countries. Thus we have rejected the maximum likelihood approach in favor of a Bayesian methodology.

Randomness of Between-Country Variability. Classical random effects models require the assumption that available countries are randomly sampled from a population of such countries. This assumption is unrealistic since, in fact, countries have volunteered for the survey. Under the Bayesian approach, the conception of countries as random represents the investigator's uncertain state of knowledge about the sources of variation between countries. Following DeFinetti (1964) and Lindley and Smith (1972), we view the country-level effects as exchangeable. Once hypothesized predictors of country differences (e.g., GNP) have been specified in the regression model of the form of equation (2), the investigator's knowledge about country differences has been exhausted and the residuals u_k, $k=1,\ldots,K$, are exchangeable: one has no a priori reason to expect that u_k will be larger or smaller than $u_{k'}$ or that their variance will differ or that knowledge of one predicts the other. This assumption of exchangeability is functionally equivalent to assuming that the u_k form an *iid* (independent, identical, distribution) random sample, but does not require the existence of a sampling mechanism.

Varying Covariance Structures. Standard applications of hierarchical models would require the assumption that the covariance structure within countries (between students and schools) is homogeneous—or at least can be predicted on the basis of country characteristics such as size or GNP. In reality, each country is likely to have a unique covariance structure. Fortunately, enough data are

[4] In most IEA study countries one classroom per school was selected so that classroom and school variance are confounded.

available within each country participating in the IEA Reading Literacy Study (150-200 schools and 2,000-3,000 students in most countries) to provide a stable estimate of these variances and covariances based on that country's data. No pooling of information is required across countries to achieve reasonably precise estimation. Thus, our approach will employ maximum likelihood to estimate the variances and covariances separately within each country. The results will then be synthesized in a between-country analysis based on the Bayesian approach described above.

Computational Considerations. Computations associated with Bayes estimation via Gibbs sampling are known to be intensive. However, using our two-stage approach, the method of maximum likelihood is employed to summarize the data within each country and the Bayesian analysis is computed on a data set having only 22 cases (the number of countries). Thus, where the data are dense (i.e., within countries), the computational method—maximum likelihood—is highly efficient. However, where the data are sparse—at the country level—the more sophisticated and computationally intensive Bayesian approach is used. The resulting Bayesian computations, based on only 22 data points, are relatively inexpensive.

Summary. The approach we have adopted has, therefore, the following elements:

- A two-level hierarchical model is first estimated separately for each country's data separately. Estimation is via maximum likelihood. The output for each country is a set of regression coefficient estimates and their variance-covariance matrix. These separate analyses are highly efficient because data are summarized within each classroom so that, for each country, the effective sample size for the computations is the number of classrooms rather than the number of students.

- A Bayes regression model is formulated to describe variation between countries. The input data are the maximum likelihood parameter estimates from the separate countries along with their standard errors. The output is constituted by estimates of the posterior densities of all quantities of interest.

- Bayesian computations are achieved via Gibbs sampling as described in detail in the appendix to this chapter. This approach avoids the need for difficult numerical integrations and produces an empirical representation of the relevant posterior distributions.

The structure of the analytic model and assumptions are described in more detail in the next section.

9.3.2 The Model

The choice of variables for the model at each of its levels was made after extensive exploratory analysis of the data country by country. Many potentially relevant predictors were rejected because they were clearly not measured on comparable metrics across countries, because of missing data, or because of anomalous features of their distributions. As a result, the specification of the model is quite thin. For example, our sole indicator for the social status of the students is the availability of books in the home. While related to social status, this indicator better reflects the literacy environment of the home. Our sense is that this variable is a better indicator of the social status composition of a classroom or country at the aggregate level than of the child.

Because the model is underspecified, substantive conclusions are made with extreme caution. However, we believe the analysis and preliminary results give a sense of the kinds of questions that become accessible using our approach. We leave to the future two kinds of activity needed to develop a more credible base for substantive conclusions: a) an investigation into the sources of noncomparability across countries in the metrics of key IEA Reading Literacy Study variables, such as years of parental or teacher education leading to construction of equated measures and a reanalysis of these data; and b) the collection of future international data with greater care to comparability across countries.

In principle, the estimation of a single model within each country can produce evidence relevant to each of the three types of research questions we have identified: comparing means, testing hypotheses about the relationships between country characteristics and country means, and examining country differences in the equity of the literacy distribution. Using the fourth grade study data, we illustrate this idea by formulating a within-country model having two levels. At level 1—the student level—overall reading literacy is predicted by our indicator of social status and gender. This model defines, for each classroom, three quantities of interest: a) the adjusted overall reading literacy mean for the class; b) a regression coefficient indicating, for that class, the strength of association between the social status indicator and the literacy outcome; and c) a regression coefficient, indicating, for that class, the gap in overall reading literacy between males and females. These three quantities in essence define the distribution of literacy within each class, in terms of the average level of literacy and the equity of distribution of literacy with respect to social status and gender. At level 2—between classrooms within each country—these three quantities become the outcome variables. We use the school mean of the social status indicator, the school size, the class size, and the urban versus rural location of the school to predict the classroom means. This level-2 model defines a vector of regression coefficients for each country that become outcome variables at the country level. Key country-level outcomes of interest are

- The country's mean overall reading literacy;
- The gender gap in overall reading literacy;
- The effect of student social status on overall reading literacy; and
- The contextual effect of social status.

Each of these is adjusted for the other variables in the model, including urban versus rural location, school size, class size, student social status, school mean social status, and gender. Variation in these outcomes across countries is then studied by means of a multivariate Bayes regression model. We now turn to specification of this model in detail.

Level-1 or Student-Level Model. Within each classroom *j* of country *k*, we formulate a model to predict the overall reading literacy of fourth grade student *i*:

$$Y_{ijk}=\pi_{0jk}+\pi_{1jk}(Books)_{ijk}+\pi_{2jk}(Gender)_{ijk}+e_{ijk} \tag{6}$$

where

Y_{ijk} is the combined reading literacy outcome for child *i* in classroom *j* of country *k*;[5]

π_{ojk} is the mean outcome for class *j*, country *k* (assuming books and gender are scaled as deviations about their country means);

[5] The outcome Y_{ijk} is the simple average of the narrative, document, and expository reading subtest scores for student *ijk*.

(Books)$_{ijk}$ is a measure of the availability of books in a the home of student *ijk*; so that

π_{1jk} is the expected increase in literacy per unit increase in books for students within classroom *j* of country *k*;

(Gender)$_{ijk}$ is an indicator for males (1 = male; 0 = female) that has then been centered about its country mean; so that

π_{2jk} is the mean difference between males and females within classroom *jk,* adjusted for the effect of books; and

e_{ijk} is a within-classroom random error assumed normally distributed with mean 0 and a country-specific within-classroom variance, that is, $e_{ijk} \sim N(0, \sigma_k^2)$.

Level-2 or Classroom-Level Model. The level-1 model defines three quantities (the π's) as characterizing the distribution of overall reading literacy within each classroom. These now become the outcomes in the level-2 model

$$\begin{aligned}\pi_{0jk}&=\beta_{00k}+\beta_{01k}(mean\ books)_{jk}+\beta_{02k}(class\ size)_{jk}\\&\quad+\beta_{03k}(schoolsize)_{jk}+\beta_{04k}(urban)_{jk}+u_{0jk}\\\pi_{1ik}&=\beta_{10k}+u_{1jk},\\\pi_{2jk}&=\beta_{20k}+u_{2jk}\end{aligned}\tag{7}$$

where

β_{00k} is the mean outcome for country *k* (all class-level predictors are expressed as deviations from their country means);

(mean books)$_{jk}$ is the mean availability of books in the homes of members of class *jk*;
so that

β_{01k} is the compositional (or contextual) effect of books in the home within country *k*;

(class size)$_{jk}$, (school size)$_{jk}$, and (urban)$_{jk}$ are, respectively, the enrollment of the class, the enrollment of the school, and an indicator for urban location, each deviated around their country means; so that

β_{02k}, β_{03k}, and β_{04k} are the associated regression coefficients within country *k*;

β_{10k} and β_{20k} are the average within-classroom effects of books and gender, respectively; and

u_{0jk}, u_{1jk}, and u_{2jk} are random effects defined on classrooms within country *k* and are assumed trivariate normal in distribution, that is

$$\begin{pmatrix}u_{0jk}\\u_{1jk}\\u_{2jk}\end{pmatrix}\sim N\left[\begin{pmatrix}0\\0\\0\end{pmatrix},\begin{pmatrix}\tau_{\pi 00k}&\tau_{\pi 01k}&\tau_{\pi 02k}\\\tau_{\pi 10k}&\tau_{\pi 11k}&\tau_{\pi 12k}\\\tau_{\pi 20k}&\tau_{\pi 21k}&\tau_{\pi 22k}\end{pmatrix}\right]\tag{8}$$

Restricted Maximum Likelihood Estimation Within Countries. With between 67 and 232 schools and between 1,331 and 7,277 students per country, sufficient data were available to permit estimation of all country-level parameters in separate, within-country analyses. Specifically, the computer package HLM3.0 (Bryk et al. 1988) was used to produce restricted maximum likelihood (REML) estimates of the variance-covariance components (σ_k^2, $\tau_{\pi k}$), where $\tau_{\pi k}$ is the 3-by-3 covariance matrix described in equation (8). As described in Raudenbush (1988), inferences about the regression coefficients (the β_k's) are then based on their posterior means and variances given the REML variance-covariance components.[6]

We define b_k as the country-specific vector of estimates of the regression coefficients (the β_k's) and V_k as its covariance matrix. These summarize the results of estimation in country k and provide input into the third level of the model, the between-country level.

Four β's are of particular interest in the between-country analysis: β_{00k} (mean literacy), β_{01k} (the contextual effects of social status), β_{10k} (the effect of student social status), and β_{20k} (the gender gap). These are the latent outcomes to be synthesized in the Bayesian between-country analysis.

9.3.3 A Bayesian Synthesis of Results Across Countries

The two-level analyses based on each country's data produce the input for the between- country synthesis. A new computing algorithm was needed to compute the posterior distributions using Gibbs sampling, and this algorithm is described in detail in the appendix to this chapter. For comparative purposes and to help check the results from the new algorithm, the between-country synthesis was computed using restricted maximum likelihood as well, by means of the "v-known" subroutine of HLM3.0 as described by Bryk and Raudenbush (1992, chapter 7). Both the restricted maximum likelihood and Bayes approaches are outlined below.

The Likelihood. Conditional on the true value of the regression coefficients, the estimates b_k are assumed normal, i.e.,

$$b_k|\beta_k \sim N(\beta_k, V_k) \tag{9}$$

where b_k is a vector of estimates from country k, β_k is the corresponding vector of parameters, and V_k is the variance-covariance matrix of the estimates b_k. The dimensions of b_k and V_k vary according to the analytic task at hand.

An Exchangeable Prior for β_k. Conditional on a set of known country-level predictors contained in the matrix W_k, the parameters β_k are assumed exchangeable. That is,

$$\begin{gathered}\beta_k = W_k\gamma + u_k,\\ u_k \sim N(0,T).\end{gathered} \tag{10}$$

Estimation via REML. It is possible to estimate T in equation (10) via REML and then, conditioning on this point estimate, to base inferences about γ on its posterior mean vector and covariance matrix. In fact, we did compute such analyses. As mentioned earlier, the difficulty with this approach is that T will be estimated imprecisely based on only 22 countries. Inferences about γ may be sensitive

[6] A vague prior is specified for the regression coefficients so that their posterior means are equivalent to generalized least squares estimates given the REML variance-covariance estimates.

to this imprecision. The Bayesian approach via Gibbs sampling, designed to overcome this problem, uses the REML estimates as starting values. We checked the new Bayes results against the REML results, and they behaved as expected in comparison.

Estimation via Bayes. We now formulate noninformative priors for β, γ, and T (Fotiu 1989) as described in detail in the appendix to this chapter. Then the joint posterior density of the parameters is

$$p(\beta,\gamma,T\_b,V) = const. * L(b|\beta,V,\gamma,T)f(b|\gamma,T)p_1(\gamma)p_2(T) \tag{11}$$

where $L(b|\beta,V,\gamma,T)$ is the likelihood of equation (9), $f(\beta|\gamma,T)$ is the exchangeable prior of equation (10), and p_1 and p_2 are noninformative priors described in the appendix. Inferences about the country-level regression coefficients, β, the between-country regression coefficients, γ, and the between-country variance-covariance matrix, T, are then based on their marginal posteriors:

$$\begin{aligned} g_1(\beta|b,V) &= \int\int p(\beta,\gamma,T|b,V)\partial\gamma\partial T \\ g_2(\gamma|b,V) &= \int\int p(\beta,\gamma,T|b,V)\partial\beta\partial T \\ g_3(T|b,V) &= \int\int p(\beta,\gamma,T|b,V)\partial\beta\partial\gamma. \end{aligned} \tag{12}$$

Gibbs Sampling. Unfortunately, the integrals in equation (12) are difficult to evaluate numerically, as is the integral required to find the normalizing constant of equation (11). Recently, Gibbs sampling (Gelfand and Smith 1990) has become a popular approach to approximate such integrals. We refer the interested reader to Fotiu (1989) for details; also see Seltzer (1993) in the univariate case. We used the final 2,000 realizations from the Gibbs sampling process to approximate the marginal posteriors of the parameters in equation (12).

9.4 Data

The 27 countries participating in the IEA Reading Literacy Study at Population A are predominately high-income countries, the majority of which are in Europe. Not all of these countries were used in this analysis. Two low-income countries were excluded because their income levels were substantially different from those of the other countries. If there had been a larger number of low-income countries, these would have been retained. Two other countries were excluded because of apparent irregularities in test administration. One other country was excluded because of insufficient data on key predictor variables. The analytic sample included the 22 countries listed in Tables 9-1 to 9-7, which give descriptive statistics for each country on each variable.

Within each country, schools were selected at random. Within most countries, one Population A class was selected at random, although some schools had only one class at that level. In a few countries, two classrooms were selected. We view the design as a two-stage cluster sample within each country having students clustered within classrooms/schools and classrooms/schools clustered within countries. As mentioned above, we view the countries as exchangeable (conditional on the model at hand), so that we conceive of three levels of random variability in the data.

Table 9-1. **Descriptive statistics: Urban versus rural location**

Country	Mean	Standard deviation	Number of classrooms
ENTIRE SAMPLE	.64	.48	2,908
Belgium (French)	.55	.50	113
Canada (B.C.)	.88	.33	123
Finland	.55	.50	67
France	.35	.48	108
Germany, East	.61	.49	82
Germany, West	.60	.49	89
Greece	.76	.43	141
Hong Kong	.94	.23	124
Hungary	.63	.48	135
Iceland	.59	.49	153
Ireland	.54	.50	114
Italy	.52	.50	105
Netherlands	.32	.47	77
New Zealand	.81	.40	176
Norway	.51	.50	158
Portugal	.25	.44	124
Singapore	1.00	.00	206
Spain	.80	.40	232
Sweden	.52	.50	118
Switzerland	.35	.48	173
USA	.80	.40	152
Slovenia	.64	.48	138

NOTE: 0 = rural; 1 = urban.

SOURCE: IEA Reading Literacy Study, International Association for the Evaluation of Educational Achievement, 1991.

Table 9-2. Descriptive statistics: School enrollment

Country	Mean	Standard Deviation	Number of Classrooms
ENTIRE SAMPLE	416.47	394.78	2,809
Belgium (French)	241.31	137.37	113
Canada (B.C.)	327.46	147.09	123
Finland	271.58	145.99	67
France	111.99	85.42	108
Germany, East	382.55	153.09	82
Germany, West	293.12	146.69	89
Greece	249.27	163.94	141
Hong Kong	696.73	352.84	124
Hungary	560.13	249.63	135
Iceland	215.71	233.29	153
Ireland	285.93	215.89	114
Italy	470.38	277.16	105
Netherlands	178.13	78.97	77
New Zealand	268.56	135.93	176
Norway	156.55	129.08	158
Portugal	168.02	167.53	124
Singapore	1,261.22	489.50	206
Spain	597.33	409.39	232
Sweden	233.75	161.37	118
Switzerland	196.34	276.06	173
USA	505.93	306.60	152
Slovenia	701.03	365.86	138

SOURCE: IEA Reading Literacy Study, International Association for the Evaluation of Educational Achievement, 1991.

Table 9-3. Descriptive statistics: Class size

County	Mean	Standard deviation	Number of classrooms
ENTIRE SAMPLE	24.04	8.34	2,908
Belgium (French)	20.20	4.47	113
Canada (B.C.)	23.33	3.19	123
Finland	24.48	5.08	67
France	21.34	6.03	108
Germany, East	20.41	3.45	82
Germany, West	22.20	4.31	89
Greece	23.59	5.40	141
Hong Kong	36.38	6.47	124
Hungary	23.39	4.68	135
Iceland	14.84	6.46	153
Ireland	29.71	8.86	114
Italy	16.36	5.17	105
Netherlands	24.29	5.75	77
New Zealand	29.75	7.35	176
Norway	15.52	6.46	158
Portugal	20.90	5.62	124
Singapore	36.83	5.21	206
Spain	27.91	7.12	232
Sweden	19.97	4.07	118
Switzerland	18.40	4.25	173
USA	23.94	5.55	152
Slovenia	24.64	3.91	138

SOURCE: IEA Reading Literacy Study, International Association for the Evaluation of Educational Achievement, 1991.

Table 9-4. Descriptive statistics: Books at home

Country	Mean	Standard deviation	Number of classrooms
ENTIRE SAMPLE	2.14	.48	2,908
Belgium (French)	2.31	.41	113
Canada (B.C.)	2.42	.34	123
Finland	2.33	.26	67
France	2.08	.44	108
Germany, East	2.00	.40	82
Germany, West	1.98	.31	89
Greece	1.71	.40	141
Hong Kong	1.38	.25	124
Hungary	2.26	.36	135
Iceland	2.54	.32	153
Ireland	2.09	.44	114
Italy	1.80	.38	105
Netherlands	2.42	.34	77
New Zealand	2.37	.41	176
Norway	2.44	.41	158
Portugal	1.60	.51	124
Singapore	1.89	.33	206
Spain	2.14	.41	232
Sweden	2.61	.26	118
Switzerland	2.28	.45	173
USA	2.30	.37	152
Slovenia	2.14	.30	138

NOTE: School means, 1 = 0-50; 2 = 51-100; 3 = more than 100.

SOURCE: IEA Reading Literacy Study, International Association for the Evaluation of Educational Achievement, 1991.

Table 9-5. Descriptive statistics: Overall reading literacy

Country	Mean	Standard deviation	Number of students
ENTIRE SAMPLE	516.80	78.47	55,651
Belgium (French)	510.69	72.83	1,916
Canada (B.C.)	505.44	74.64	2,013
Finland	569.20	69.32	1,376
France	533.50	69.53	1,458
Germany, East	501.30	81.05	1,437
Germany, West	512.20	81.07	1.605
Greece	513.10	73.29	2,821
Hong Kong	524.74	67.86	2,395
Hungary	503.85	75.29	2,690
Iceland	517.73	85.51	1,738
Ireland	509.35	76.68	2,388
Italy	538.69	77.04	1,474
Netherlands	487.10	72.06	1,331
New Zealand	532.58	83.37	2,906
Norway	529.02	86.65	2,011
Portugal	484.23	70.10	2,121
Singapore	513.47	72.01	7,277
Spain	511.06	76.97	5,794
Sweden	538.87	91.34	2,084
Switzerland	509.55	80.51	2,417
USA	546.78	74.45	3,232
Slovenia	500.44	77.80	3,167

SOURCE: IEA Reading Literacy Study, International Association for the Evaluation of Educational Achievement, 1991.

Table 9-6. Descriptive statistics: Gender

Country	Mean	Standard deviation	Number of students
ENTIRE SAMPLE	.51	.50	55,651
Belgium (French)	.49	.50	1,916
Canada (B.C.)	.52	.50	2,013
Finland	.52	.50	1,376
France	.49	.50	1,458
Germany, East	.49	.50	1,437
Germany, West	.52	.50	1,605
Greece	.50	.50	2,821
Hong Kong	.54	.50	2,395
Hungary	.50	.50	2,690
Iceland	.51	.50	1,738
Ireland	.49	.50	2,388
Italy	.52	.50	1,474
Netherlands	.48	.50	1,331
New Zealand	.52	.50	2,906
Norway	.49	.50	2,011
Portugal	.51	.50	2,121
Singapore	.52	.50	7,277
Spain	.49	.50	5,794
Sweden	.51	.50	2,084
Switzerland	.52	.50	2,417
USA	.50	.50	3,232
Slovenia	.51	.50	3,167

NOTE: 0 = female; 1 = male.

SOURCE: IEA Reading Literacy Study, International Association for the Evaluation of Educational Achievement, 1991.

Table 9-7. Descriptive statistics: Books at home

Country	Mean	Standard deviation	Number of students
ENTIRE SAMPLE	2.13	.88	55,651
Belgium (French)	2.32	.83	1,916
Canada (B.C.)	2.44	.78	2,013
Finland	2.34	.78	1,376
France	2.09	.87	1,458
Germany, East	1.99	.87	1,437
Germany, West	1.98	.85	1,605
Greece	1.78	.85	2,821
Hong Kong	1.38	.71	2,395
Hungary	2.28	.81	2,690
Iceland	2.53	.71	1,738
Ireland	2.13	.88	2,388
Italy	1.82	.86	1,474
Netherlands	2.44	.80	1,331
New Zealand	2.40	.81	2,906
Norway	2.49	.74	2,011
Portugal	1.70	.86	2,121
Singapore	1.90	.88	7,277
Spain	2.16	.86	5,794
Sweden	2.61	.68	2,084
Switzerland	2.33	.82	2,417
USA	2.30	.84	3,232
Slovenia	2.16	.84	3,167

NOTE: Student means, 1 = 0-50; 2 = 51-100; 3 = more than 100.

SOURCE: IEA Reading Literacy Study, International Association for the Evaluation of Educational Achievement, 1991.

As indicated in Tables 9-1 to 9-7, the analytic sample includes 2,908 classrooms and 55,651 students, an average of 132 classrooms and 2,530 students per country and 19 students per classroom. School and classroom variables include urban versus nonurban location, school size, class size, and classroom mean availability of books in the home. Somewhat more than half of the schools are urban (Table 9-1), though all schools in Singapore and 94 percent of the schools in Hong Kong are urban. In contrast, fewer than half the schools in France, the Netherlands, Portugal, and Switzerland are classified as urban. Average school enrollment (Table 9-2) is 416, with Singapore having exceptionally large schools (mean enrollment of 1,261) and France having the smallest schools (mean of 112). Class sizes (Table 9-3) average 24 students overall, with the largest classes found in Singapore (mean of 37) and the smallest classes in Iceland (mean of just under 15). The school-aggregate of the student-level variable books at home is described in Table 9-4.

Student-level variables include overall reading literacy, gender, and books in the home (Tables 9-5 to 9-7). Overall reading literacy (Table 9-5) is the average of three test scores, each of which indicates proficiency in reading a different type of text (narrative, expository, and documents). Country sample means range from 484.23 in Portugal to 569.20 in Finland, with standard deviations ranging from 67.86 in Hong Kong to 91.34 in Sweden. As one might expect, every country shows near-equal proportions of males and females (Table 9-6). Availability of books at home (Table 9-7) is measured ordinally (low = 0-50 books; medium = 51-100 books; high = more than 100 books). However, exploratory analyses indicated that it was reasonable to treat this as a linear contrast. The lowest sample mean on this contrast was found in Hong Kong and the highest in Iceland.

A number of variables we had hoped to use in the analysis were found unusable. These included a home possession score and a student possession score. These were measured on different metrics in different countries with different types of possessions and commodities listed in different countries and no attempt to equate the scales. Similarly, years of parental education was available in the Population B sample but not the Population A sample. One result of excluding these variables was underspecification of social status.

Histograms of all candidate variables and scatter plots between pairs of variables were examined for each country's data. These analyses led to the exclusion of some variables as mentioned above and informed choice of metric for those variables that remained. Small numbers of anomalous cases (at the student level) were removed within a number of countries. These included students who achieved the minimum on all three tests, likely indicating that they had not tried to respond to the test. Two countries having large numbers of such cases and, as a result, displaying unexpectedly low mean overall literacy, were also excluded.

9.5 Results

As explained earlier, our interest focused on three potential uses of the data: comparing country means, relating country characteristics to mean levels of literacy, and studying country differences in the equity of distribution of literacy with respect to gender and social status as indicated by books available in the home. However, as mentioned in Section 9.3, all of the necessary information for these purposes was obtained by estimating within each country the two-level model described by equations (6) and (7). The key output from each country's analysis is a vector of four estimates of the parameters β_{00k} (mean literacy), β_{01k} (the contextual effect of books at home), β_{10k} (the student-level effect of books at home), and β_{20k} (the gender gap) along with their variance-covariance matrix. Table 9-8 summarizes the marginal posterior distributions of these four parameters, as defined in equation (12) for each country.

Table 9-9. Posterior estimates, by country

Country	GNP[1]	β_{00k} Mean Literacy		β_{01k} School Books[2]		β_{10k} Student Books[2]		β_{20k} Gender[2]	
		Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation	Mean	Standard deviation
Hungary	2.46	502.65	2.54	35.08	8.12	16.78	1.67	-9.13	2.54
Portugal	3.65	480.80	3.05	28.71	7.26	10.45	1.84	-7.04	2.64
Greece	4.80	509.85	3.46	20.89	9.21	10.00	1.45	-1.69	2.21
Slovenia	6.50	499.85	2.26	20.24	7.78	16.77	1.63	14.12	2.52
Spain	7.74	507.50	1.97	33.42	5.13	12.76	1.18	-5.57	1.92
Ireland	7.75	507.16	2.74	14.69	6.54	17.91	1.72	14.23	3.08
Singapore	9.07	511.11	1.98	37.26	5.87	11.60	0.90	10.37	1.60
Hong Kong	9.22	523.40	2.94	22.53	10.03	5.92	1.72	-8.46	2.41
New Zealand . . .	10.00	529.87	2.38	38.39	6.07	20.06	2.01	18.95	2.91
Germany, East . .	11.30	501.60	3.40	20.93	8.10	13.43	2.18	13.67	3.57
Italy	13.33	536.75	4.59	-4.04	11.32	13.41	1.90	-9.00	2.94
Belgium (French) .	14.49	511.25	2.68	38.80	6.68	12.49	1.81	12.51	2.69
Netherlands	14.52	486.71	3.32	26.69	9.50	14.71	2.25	-8.82	3.10
France	16.09	534.02	3.08	-4.44	7.69	14.31	2.06	-6.40	3.13
Iceland	16.59	515.27	2.92	21.50	9.02	12.31	2.48	19.15	3.60
Canada	16.96	503.43	2.59	20.55	7.71	12.19	1.98	13.08	2.88
Germany, West . .	18.48	510.81	3.42	33.77	9.72	15.98	2.03	-8.86	3.23
Finland	18.59	568.36	2.54	-4.63	9.74	11.48	2.29	11.89	3.32
Sweden	19.30	538.25	2.83	27.00	10.31	16.21	2.57	11.97	3.37
USA	19.84	547.29	2.48	58.30	7.00	11.15	1.40	-8.60	2.13
Norway	19.99	528.68	2.56	4.34	6.96	16.55	2.42	14.28	3.16
Switzerland	27.50	506.73	2.53	-4.38	6.21	18.03	2.02	-5.45	2.74
Column Mean . .	**13.10**	**516.43**	**2.83**	**22.07**	**8.00**	**13.84**	**1.89**	**10.60**	**2.80**

[1] GNP per capita in thousands of dollars.

[2] The variables school books, student books, and gender were centered around the grand mean for the country.

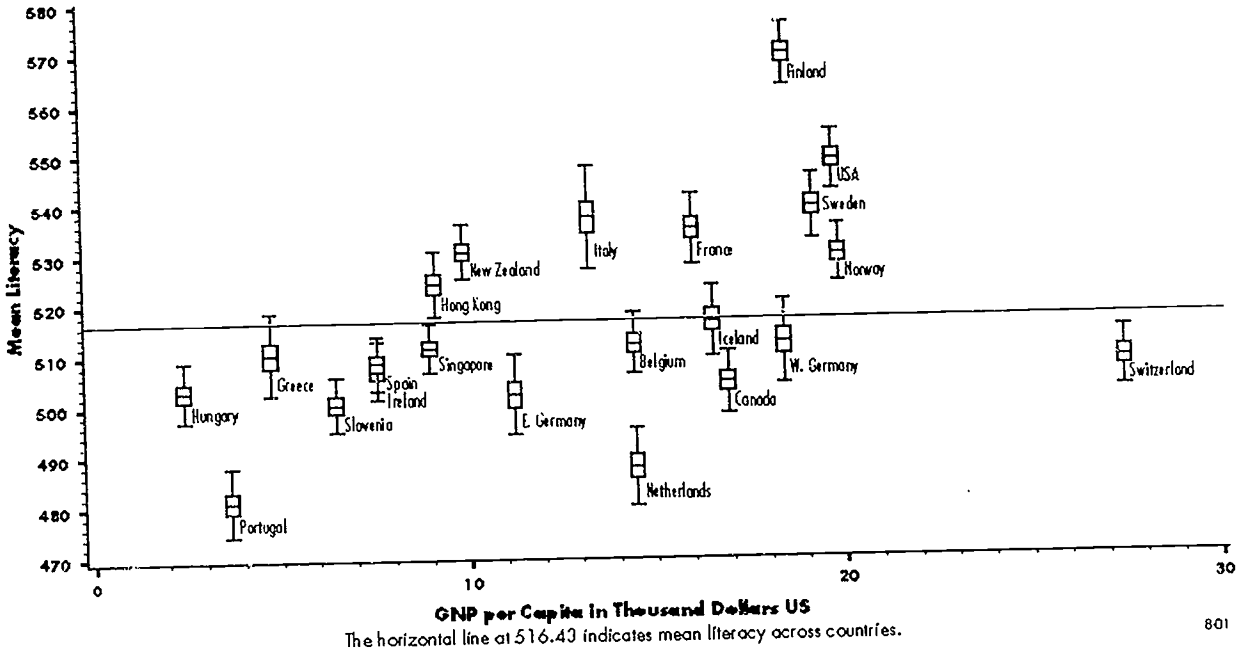
SOURCE: IEA Reading Literacy Study, International Association for the Evaluation of Educational Achievement, 1991.

9.5.1 Comparing Country Means

The first column of Table 9-8 gives the name of the country, and, for reference, the second gives that country's GNP. For convenience, the countries are listed in ascending rank order by GNP. Column 3 summarizes the posterior distribution of β_{00k}, that is, mean literacy, for each country, by listing the posterior mean and standard deviation of the country mean. A moderate tendency for these posterior means to increase with GNP is manifest.

Assessing Uncertainty. Figure 9-1 displays box plots describing the posterior distribution of mean literacy for each county. Each box plot summarizes 2,000 sampled values based on the Gibbs method after convergence. The horizontal line within each box marks the median point of the distribution. These posterior distributions have a normal shape, ana the median and mean are approximately equivalent. Inside the box are the middle 75 percent of the sampled values, while the ends of the whiskers delimit the top and bottom 1 percent of the sampled values. Thus, the plots give 75 percent and 98 percent credibility intervals for the mean literacy parameter for each country. Within the Bayesian framework, one is justified in saying that the posterior probability is .75 or .98 and that the true mean lies in that interval. We note that these credibility intervals are virtually identical to the confidence intervals based on the two-level analyses using maximum likelihood. This equivalence is expected given the substantial amount of data within each country and the asymptotic equivalence of classical and Bayes estimators. Many of these intervals overlap, though the interval for Finland stands alone as by far the most positive interval.

Figure 9-1. Posterior distribution of mean literacy 98 percent credibility intervals, by country



The horizontal line at 516.43 indicates mean literacy across countries.

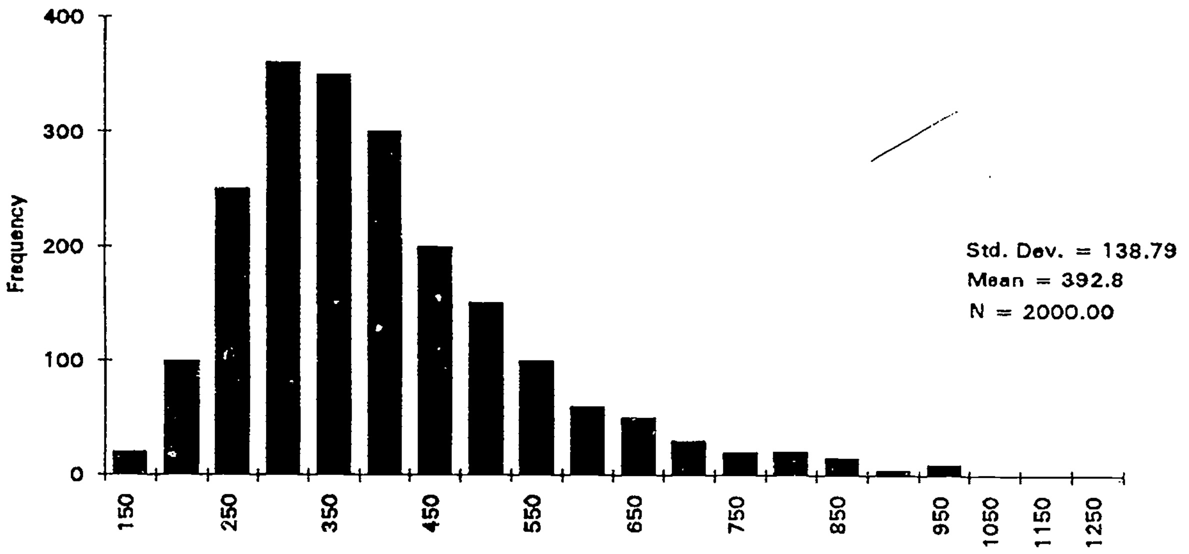
SOURCE: IEA Reading Literacy Study, International Association for the Evaluation of Educational Achievement, 1991.

The analysis takes into account the sampling design within each country in that the sampling variance matrix V_k was computed by the two-level hierarchical analysis within country k and therefore incorporates the clustering of students within classrooms.

Assessing Heterogeneity. How much do countries vary in their means relative to the variation within countries? For our data, the between-country variance in their means will tend to be estimated with uncertainty despite the fact that each country has a substantial amount of data. The precision of that variance estimate depends quite heavily on the number of countries providing data. Figure 9-2 is a histogram that approximates the posterior distribution of the variance of the means, that is, the posterior distribution of $Var(\beta_{00k})=(\tau_{\beta00})$.

This histogram is based on 2,000 sampled values of $\tau_{\beta00.}$ As the figure indicates, all plausible values of this variance are positive, implying clearly that the country means are heterogeneous. The posterior mean of this between-country variance is 392.8. Recall that the overall standard deviation of the outcome across all countries is 78.47 (Table 9-5). Thus, it appears that the proportion of variance in the outcome that lies between countries is about $392.8/(392.8+78.5^2)$ = .060, so that about 6.0 percent of the variance is between countries. However, as Figure 9-2 implies, values of $\tau_{.00}$ as small as 150 and as large as 850 are plausible, implying that the percentage of variance lying between countries could be as small as 2.4 percent or as large as 12.3 percent, giving some sense of the degree of uncertainty about the extent to which literacy means vary across countries. Although no more than a fraction of the variability in literacy lies between countries by any estimate, this does not imply that country differences are trivial. As Table 9-8 indicates, it is common to find pairs of countries with posterior means differing by more than half the overall standard deviation, a quite substantial effect size.

Figure 9-2. Posterior distribution of the variance of the country mean coefficient



SOURCE: IEA Reading Literacy Study, International Association for the Evaluation of Educational Achievement, 1991.

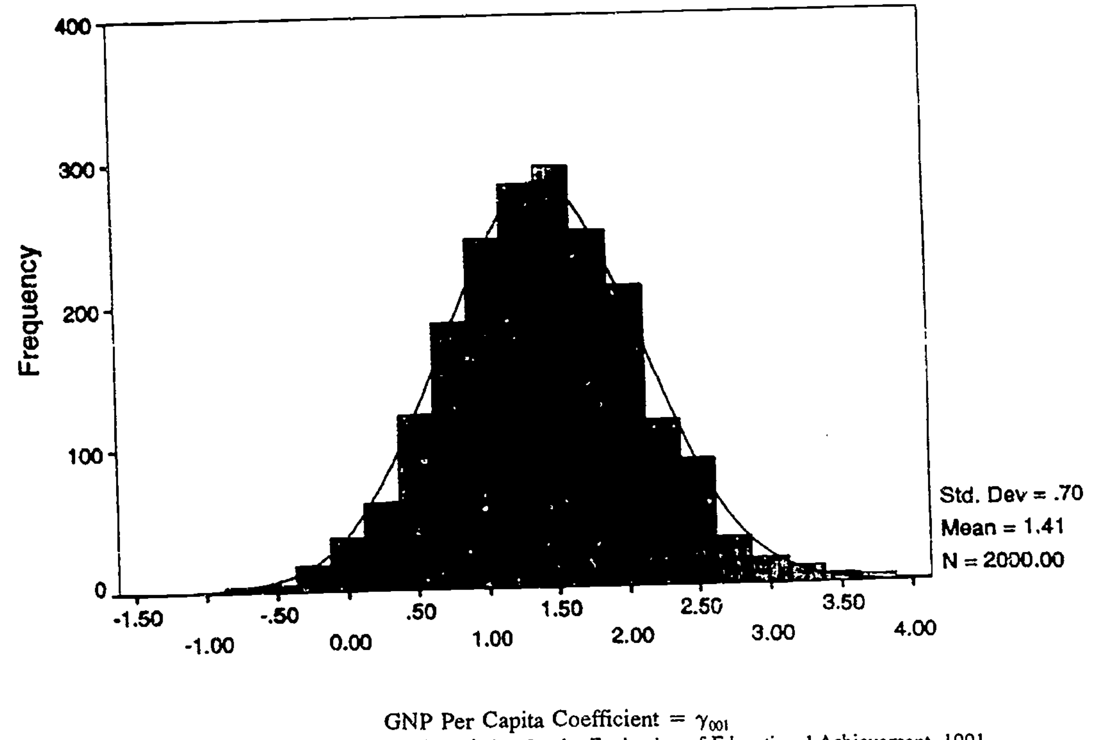
9.5.2 Modeling Country Means

As described in Section 9.2, there are compelling reasons to formulate models to predict variability between countries in mean literacy. One might want to investigate whether policy manipulable variables such as the length of the school day or the use of a national examination are related to greater school achievement. One may also wish to control for GNP in assessing mean differences between countries to avoid unfair judgments about poor countries' educational systems. Our analysis of country means in the preceding section clearly signals the existence of heterogeneity in country means, encouraging a search for explanatory variables.

Figure 9-3 displays the posterior distribution of the regression coefficient γ_{001} relating GNP to mean literacy. Our belief about the magnitude of this relationship does depend upon our opinion about the variance between country means (see equations (4) and (5) and the associated discussion). The posterior distribution displayed in Figure 9-3 fully takes into account the uncertainty about this variance. Thus, the posterior standard deviation of 0.70 is larger than would be found using maximum likelihood estimation, which conditions upon a given value of this variance (effectively the posterior mode).

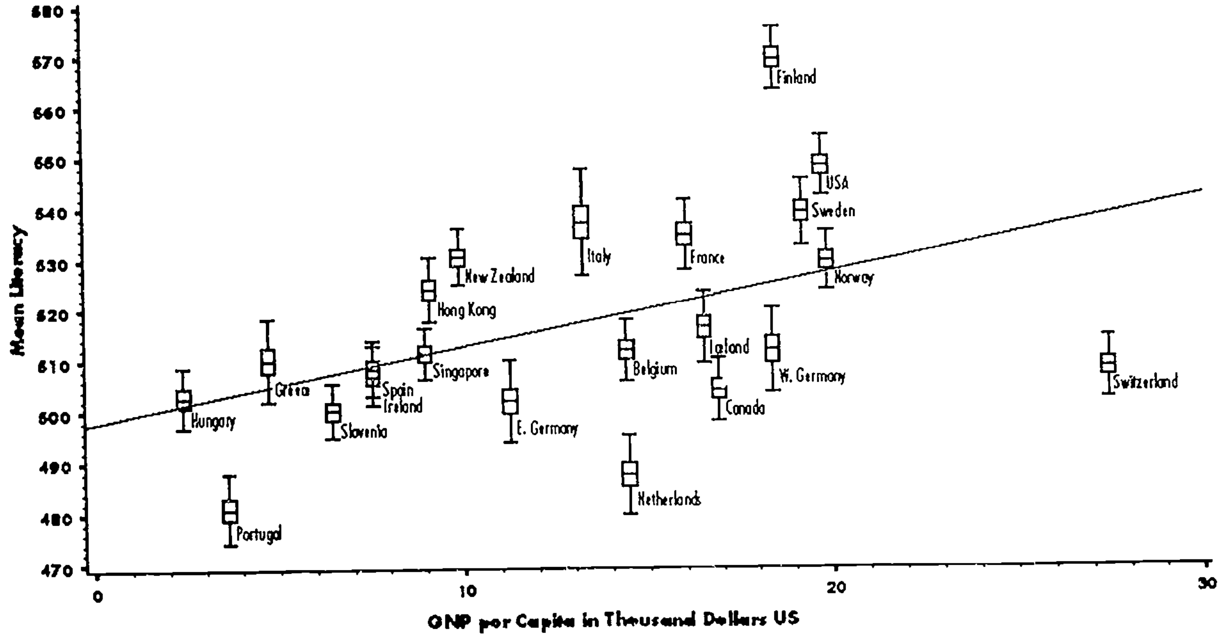
As Figure 9-3 indicates, the posterior probability is concentrated on values greater than 0, implying the existence of a positive relationship between GNP and mean literacy. The posterior mean is 1.41. Given the standard deviation of GNP of 6.38 we see that the posterior mean of γ_{001} = 1.41 is equivalent to a standardized regression coefficient of 1.41*6.38/78.57 = 0.11. However, values of γ_{001} quite near 0 are plausible, and the posterior mean is twice the posterior standard deviation. Values as large as 3.0 are also plausible, implying that the standardized regression coefficient could be as small as 0 or as large as 0.25.

Figure 9-3. Posterior distribution of the GNP per capita coefficient



SOURCE: IEA Reading Literacy Study, International Association for the Evaluation of Educational Achievement, 1991.

Figure 9-4. Posterior distribution of mean literacy 98 percent credibility intervals, by country



The regression line is 516.43 + 1.41(GNP - 13.10).

8-04

SOURCE: IEA Reading Literacy Study, International Association for the Evaluation of Educational Achievement, 1991.

Taking GNP into account may revise our opinion about the relative efficiency of countries' educational systems. Figure 9-4 displays the posterior distributions of mean literacy as a function of GNP. Certain countries that had appeared quite different from each other (e.g., Hungary and Norway) are achieving about as expected given their GNP. However, the low performance of Portugal and the high performance of Finland do not appear completely attributable to their substantial GNP differences.

9.5.3 Modeling Equity Differences Between Countries

Gender Equity. A study of mean differences between countries will be misleading if equity in outcomes varies from country to country. Consider the relationship between gender and reading literacy. The fourth column of Table 9-8 lists the posterior means and standard deviations of the gender gap in reading literacy for 22 countries. Note that every posterior mean is negative, implying that females tend to outperform males. Boxplots displaying the posterior 75 percent and 98 percent credibility intervals are displayed in Figure 9-5.

Under the model

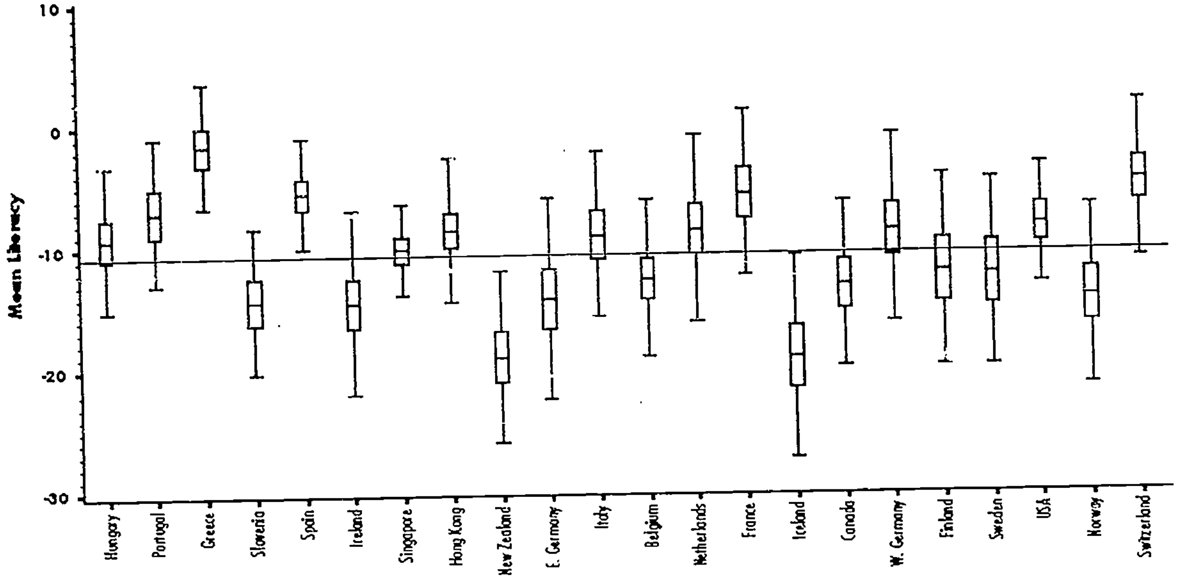
$$\begin{aligned}\beta_{20k} &= \gamma_{200} + u_{20k}, \\ u_{20k} &\sim N(o, \tau_{\beta 20})\end{aligned} \tag{13}$$

γ_{200} represents the average gender gap across the 22 countries and $\tau_{\beta 20}$ represents the variance in the gender gaps. The posterior distribution of γ_{200}, displayed in Figure 9-6, shows a posterior mean of -10.59, indicating that males score about 10.59/78.57 = 0.13 standard deviations on average behind females at the fourth grade across these countries. Note that this estimate is quite precise (the posterior standard deviation is only 1.36 = 0.02 standard deviation units). However, the posterior distribution of the variance of the gender gaps ($\tau_{\beta 20}$, see Figure 9-7) indicates that this variance is unmistakably greater than 0. Thus, gender gaps differ significantly from country to country. Unfortunately, these data contain little information to help us understand the sources of variation of the gender gap.

Figure 9-8 plots posterior expected male and female means for the 22 countries. Points near the diagonal line most closely approximate gender equity. There is little evidence that countries scoring high on average are more or less equitable than countries scoring low. Apparently excellence in terms of a high mean does not guarantee gender equity, nor does gender equity preclude excellence.

Equity with Respect to Social Status. Heyneman and Loxley (1983), synthesizing data from 29 countries, found the relationship between student social status and literacy to be stronger in developed than in developing countries, with a correspondingly greater effect of school resources in developing than in developed countries. This hypothesis exemplifies an important class of hypotheses regarding the relationship between country characteristics and the distribution of outcomes within countries. Our modeling framework is well suited to examine such hypotheses. The data at hand are not well suited to test Heyneman and Loxley's hypothesis because the available countries are uniformly quite highly developed. However, the essential methodological principles involved in such a test become clear in our illustrative analysis.

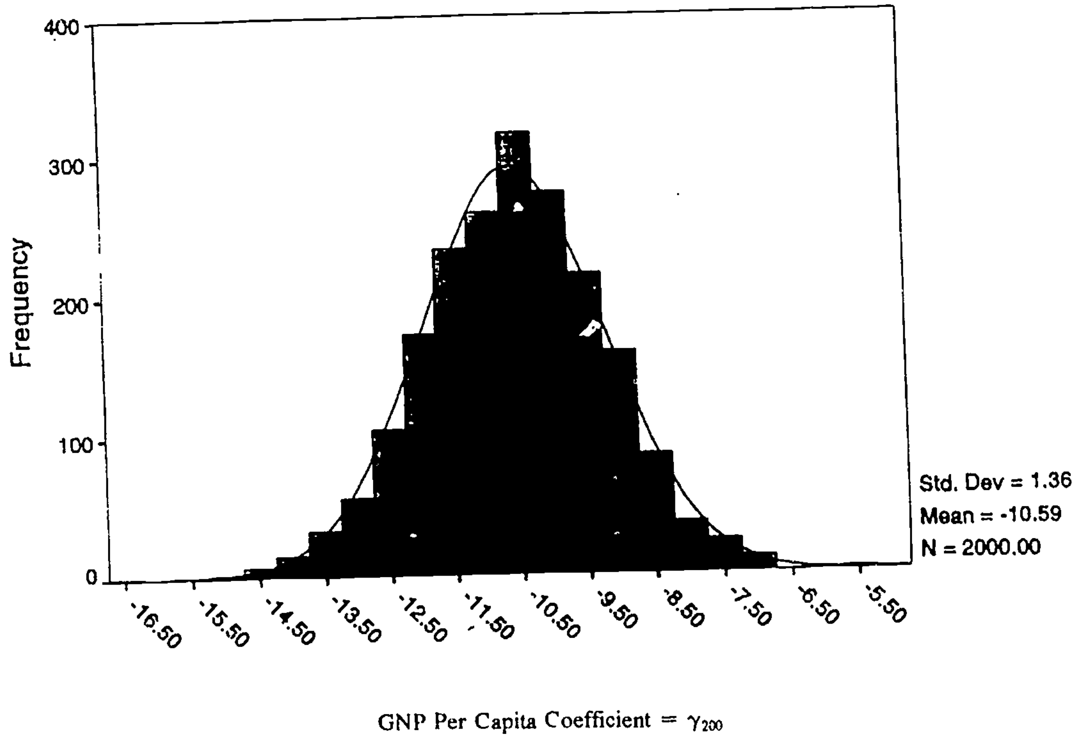
Figure 9-5. Posterior distribution of gender effect 98 percent credibility intervals, by country



The horizontal line at -10.60 indicates the mean gender effect across countries.

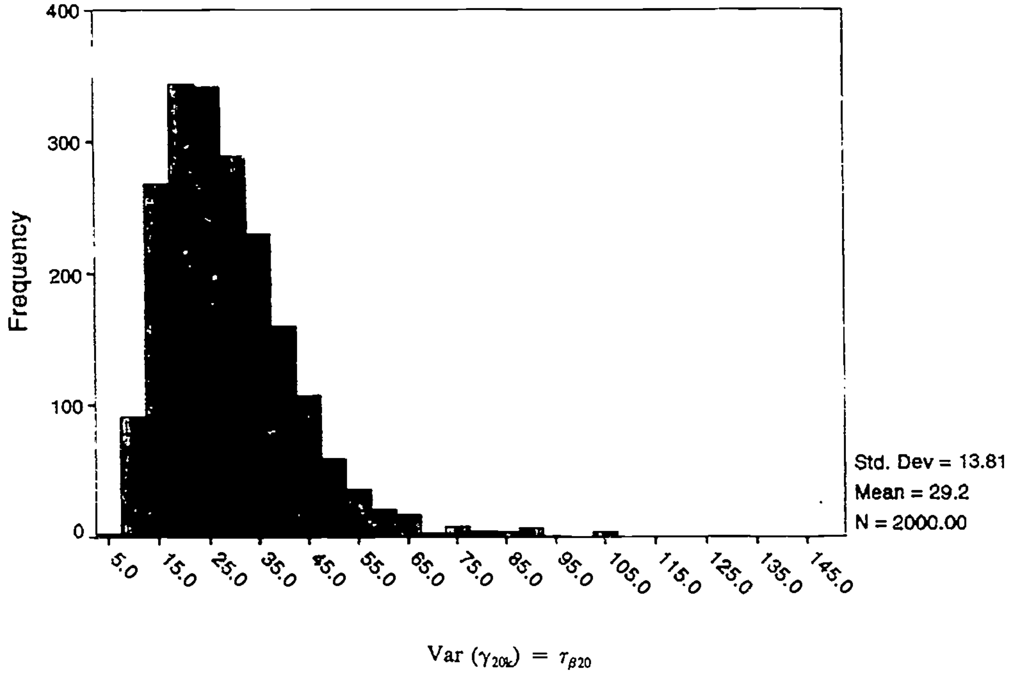
SOURCE: IEA Reading Literacy Study, International Association for the Evaluation of Educational Achievement, 1991.

Figure 9-6. Posterior distribution of the mean country gender gap coefficient



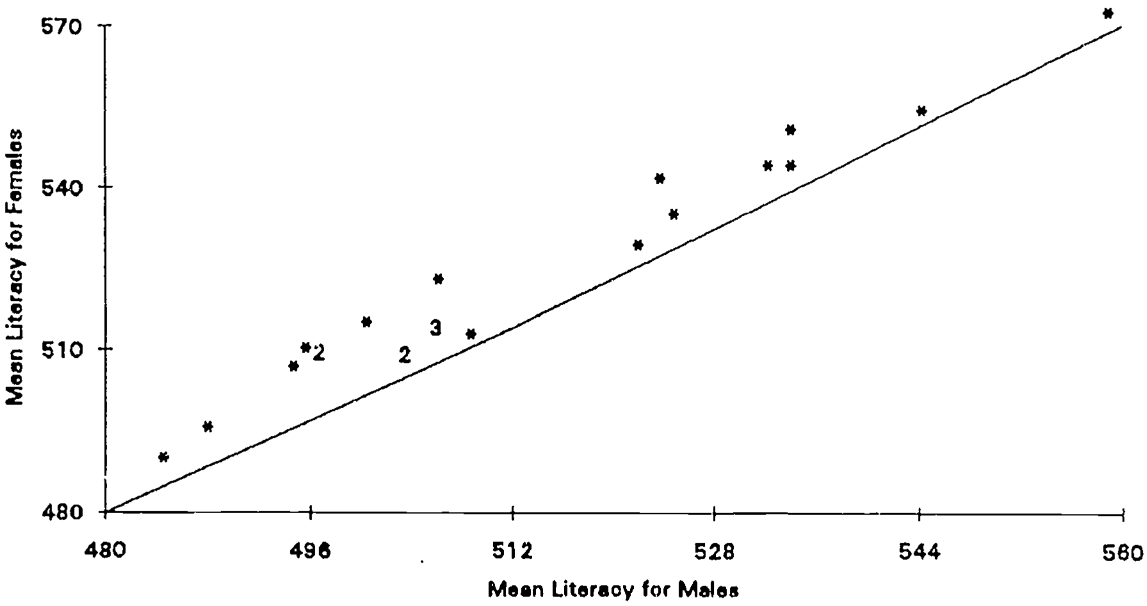
SOURCE: IEA Reading Literacy Study, International Association for the Evaluation of Educational Achievement, 1991.

Figure 9-7. Posterior distribution of the variance of the country gender gap coefficient



SOURCE: IEA Reading Literacy Study, International Association for the Evaluation of Educational Achievement, 1991.

Figure 9-8. Posterior mean literacy for males and females in 22 countries



NOTE: An asterisk indicates that one country lies at this point. An integer indicates that the specified number of countries lie at that point. Diagonal line denotes equal mean for males and females.

SOURCE: IEA Reading Literacy Study, International Association for the Evaluation of Educational Achievement, 1991.

The essential difficulty in examining hypotheses of this type is that the relationship between a student variable such as social status and the outcome has two sources (as discussed in Section 9.2.3): the relationship between social status and literacy for students within the same school, and the relationship between the social status composition of the school and student literacy for students of the same SES. Thus, for any country, we can characterize the literacy gap between two students, one of high and one of low social status, as

$$\beta^*_k=\beta_{10k}(X_{High}-X_{Low})+\beta_{01}(\overline{X}_{High\ k}-\overline{X}_{Low\ k}) \tag{14}$$

where

X_{High} is the value of social status indicator for a student of high social status;

X_{Low} is the value of the social status for a student of low social status;

$X_{High\ k}$ is the mean social status of the typical school attended by high social status students in country *k*;

$X_{Low\ k}$ is the mean social status of the typical school attended by low social status students in country *k*;

β_{10k} is the student-level regression coefficient for social status in country *k*; and

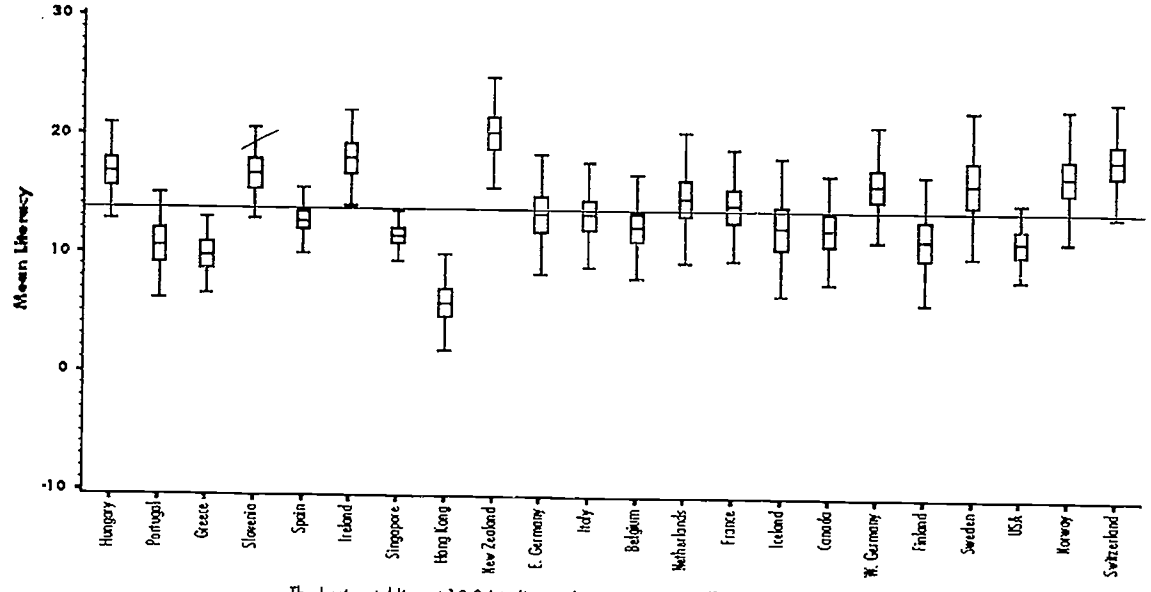
β_{01k} is the contextual social status coefficient in country *k*.

With this formulation in mind, equation 14 shows that the gap between high and low social status students within a country depends on the within-school coefficient β_{10k}, the contextual coefficient β_{01k}, and the degree of social status segregation of the schooling system as characterized by the discrepancy between the mean social status of schools attended by high and low social status students within that country. Thus, a country interested in closing the social status achievement gap might consider strategies that a) close the within-school gap, b) reduce the contextual effect, or c) reduce social status segregation of schools. The summary measure β_k^* is by itself uninterpretable because it is composed of these three contributors. Thus, it appears essential to decompose β_k^* into its constituent parts.

To approximate these quantities, we have used the availability of books in the home as the sole indicator at hand for social status. Figures 9-9 and 9-10 display the posterior distributions of the student-level and school-level effects and contextual effect of this social status indicator in 22 countries. The figures indicate that countries having large individual effects may have small contextual effects and vice versa. For example, the United States exhibits a very large posterior mean for the contextual effect and a comparatively modest individual effect.

Clearly, both of these effects, on average, are significantly positive as indicated by their posterior distributions, with the student-level effect having a posterior mean of 13.85 (posterior s.d. = 0.98) and the contextual effect having a mean of 22.1 (s.d. = 4.60) (Figures 9-11 and 9-12). Moreover, these effects do vary significantly from country to country, as indicated by the posterior distributions of their variance components (Figures 9-13 and 9-14). The contextual effects are particularly highly variable.

Figure 9-9. Posterior distribution of student-level booklet effect 98 percent credibility intervals by country

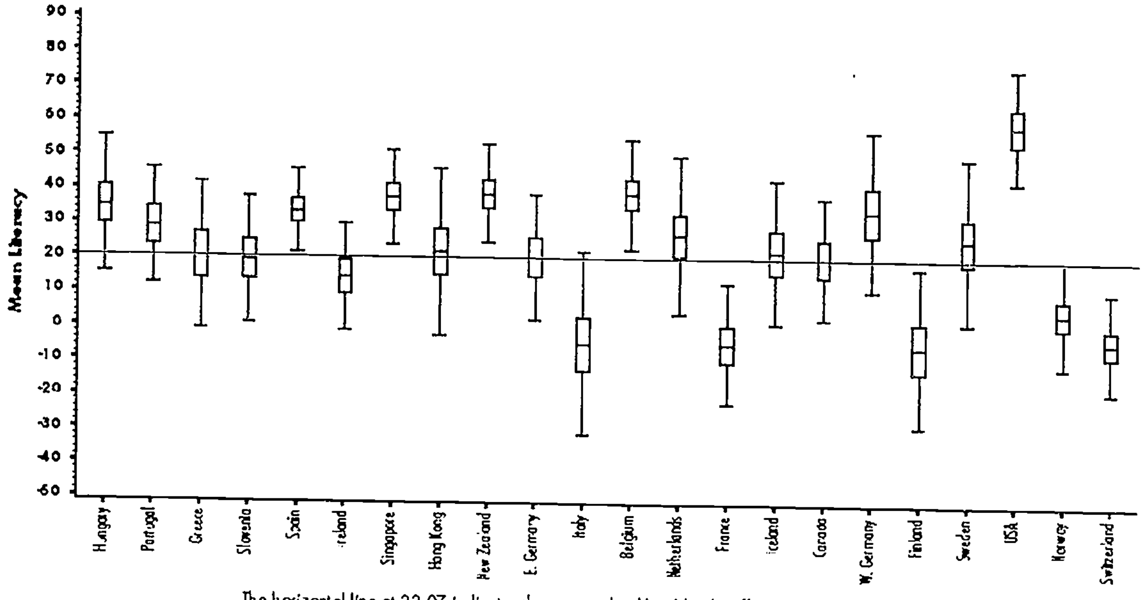


The horizontal line at 13.84 indicates the mean student effect across countries.

6-09

SOURCE: IEA Reading Literacy Study, International Association for the Evaluation of Educational Achievement, 1991.

Figure 9-10. Posterior distribution of school-level books effect 98 percent credibility intervals, by country

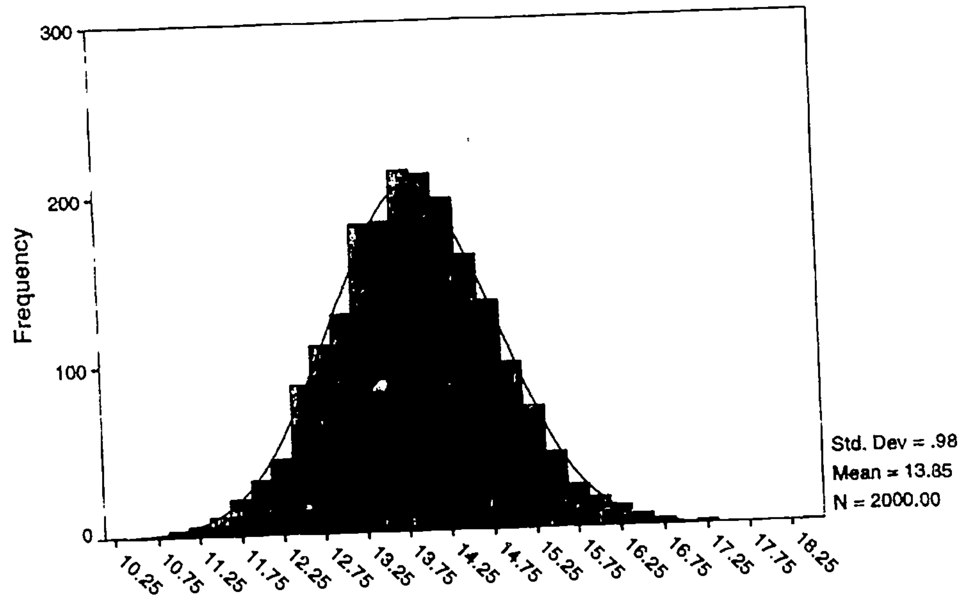


The horizontal line at 22.07 indicates the mean school-level books effect across countries.

6-10

SOURCE: IEA Reading Literacy Study, International Association for the Evaluation of Educational Achievement, 1991.

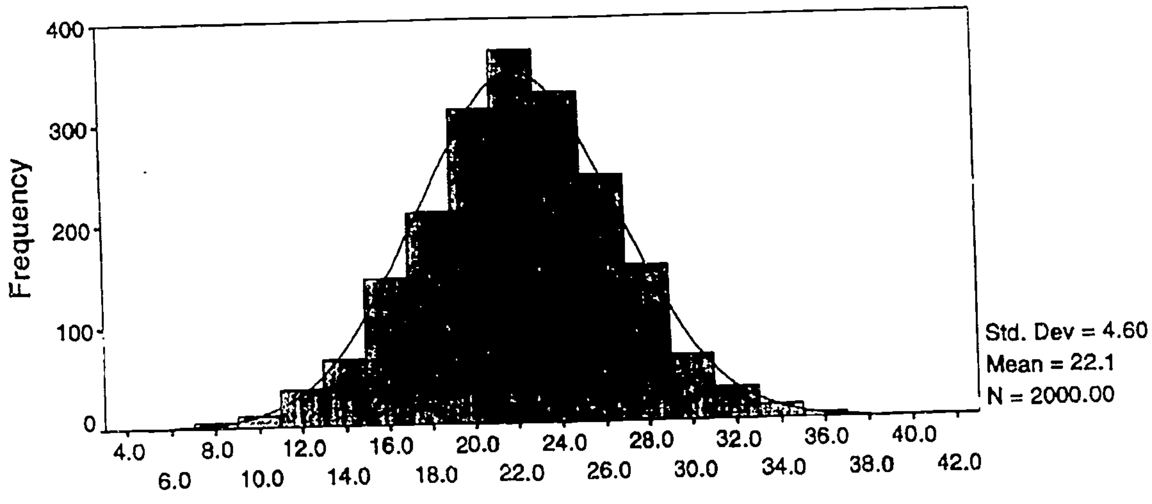
Figure 9-11. Posterior distribution of the mean student-level books coefficient



Mean Student-Level Books Coefficient = γ_{001}

SOURCE: IEA Reading Literacy Study, International Association for the Evaluation of Educational Achievement, 1991.

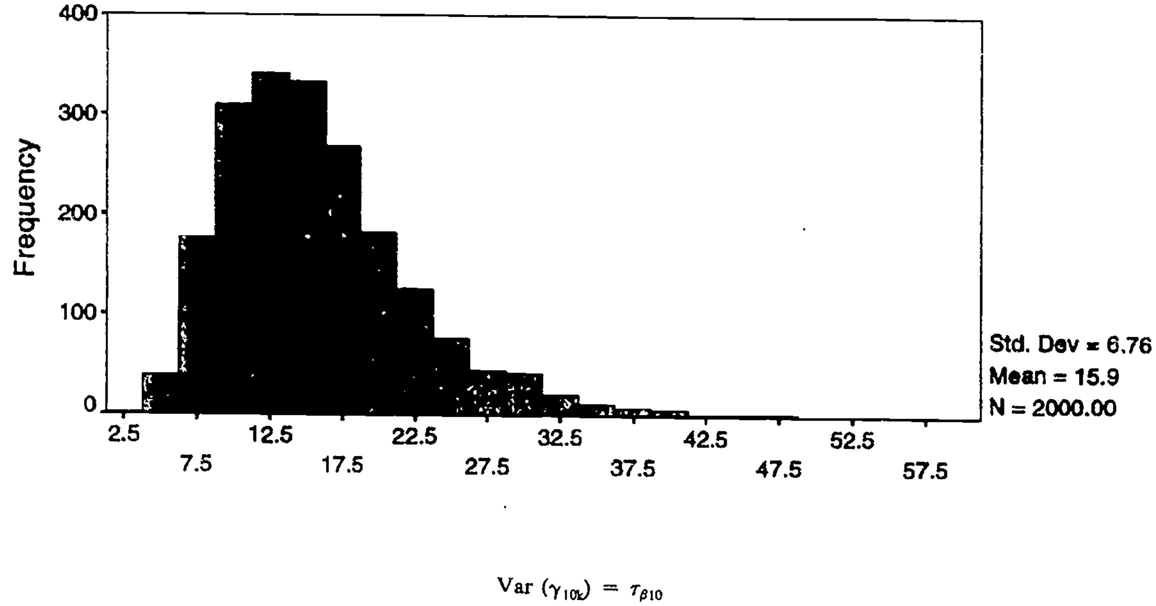
Figure 9-12. Posterior distribution of the mean school-level books coefficient



Mean Student-Level Books Coefficient = γ_{010}

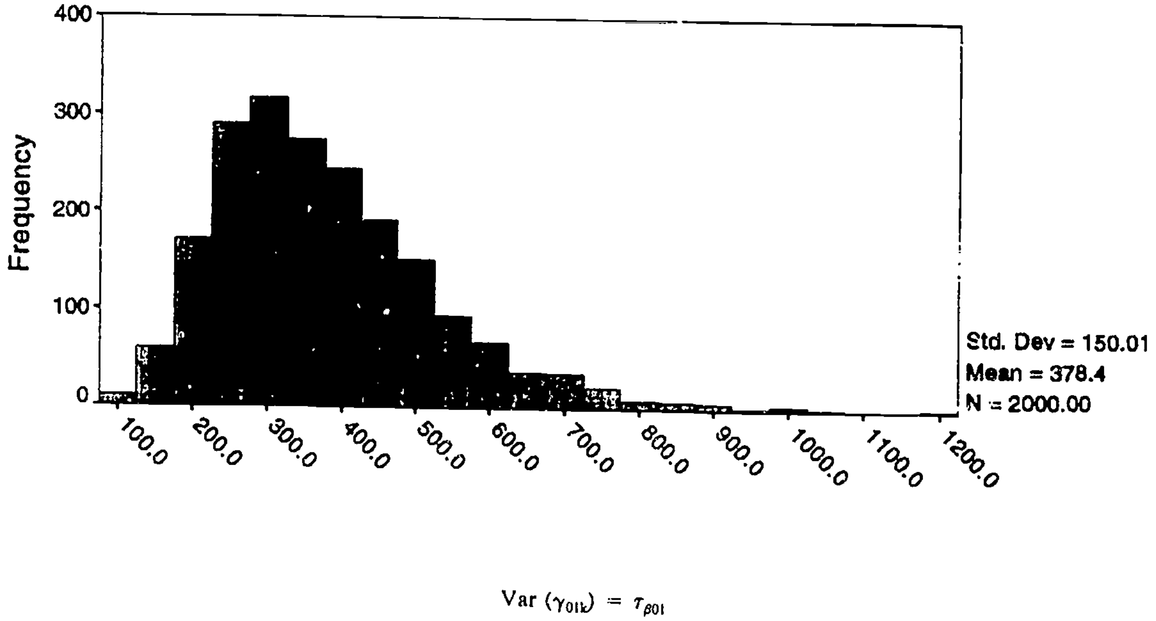
SOURCE: IEA Reading Literacy Study. International Association for the Evaluation of Educational Achievement, 1991.

Figure 9-13. Posterior distribution of the variance of the student-level books coefficient



SOURCE: IEA Reading Literacy Study, International Association for the Evaluation of Educational Achievement, 1991.

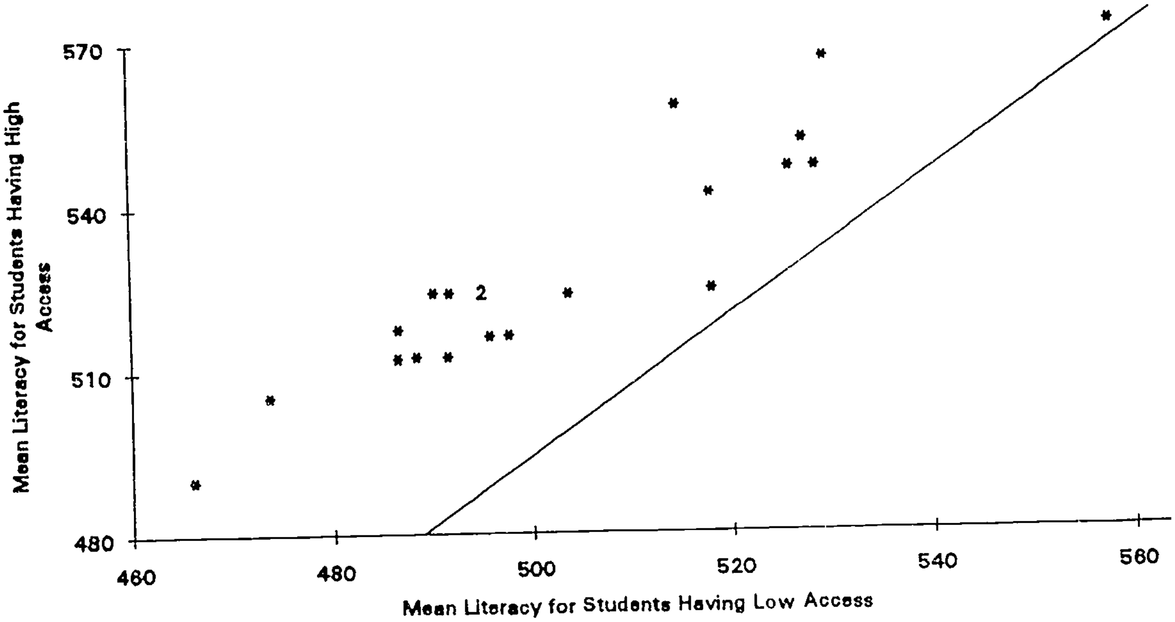
Figure 9-14. Posterior distribution of the variance of the mean school-level books coefficient



SOURCE: IEA Reading Literacy Study, International Association for the Evaluation of Educational Achievement, 1991.

Finally, Figure 9-15 plots the expected literacy mean for students with X_{High} = more than 100 books at home and X_{Low} = less than 50 books at home in the 22 countries. These points are, in fact, the β_k^* of equation (15). Countries with points close to the diagonal line have comparatively equitable distributions of outcomes. We see that the highest achieving country (Finland) has a comparatively equitable distribution of literacy. There is some indication that countries with low means tend to exhibit less equitable distributions, although the country with the least equitable distribution (Norway) is about average on mean literacy. No relationship was manifest between GNP and β_k^*, a result that cannot be viewed as evidence against Heyne..ian and Loxley (1983) given the absence of low-income countries in this sample.

Figure 9-15. Posterior mean literacy for students having high and low access to books at home in 22 countries



NOTE: An asterisk indicates that one country lies at this point. An integer indicates that the specified number of countries lies at that point. Diagonal line denotes equal mean for students having high and low access.

SOURCE: IEA Reading Literacy Study, International Association for the Evaluation of Educational Achievement, 1991.

9.6 Conclusions

In this chapter we have identified three common uses of cross-national classroom data on student literacy: comparing country means, testing hypotheses about the relationship between country characteristics and country means, and studying the equity with which literacy outcomes are distributed. Associated with each use is a set of challenges to valid statistical inference. We have proposed a two-stage statistical methodology for coping with these challenges. The first stage involved estimation of a hierarchical linear model separately for each country. This model copes with the multistage cluster sampling designs commonly employed in such surveys and incorpo ates weights inversely proportional to the probability of selection of units within strata. Typically, as in the IEA Reading Literacy Study, each country will have a reasonably large sample, and maximum likelihood estimates of model parameters based on each country's data will be quite precise. At the second stage, we synthesized results from the first stage. The number of countries involved will typically be small, leading to imprecision in estimates of between-country variance of mean literacy and certain regression coefficients that characterize the equity of distribution of literacy. This between-country variance is important in its own right, and its estimation is critical for inference about the other parameters in the between-country distribution. Because maximum likelihood estimates of relevant means and regression coefficients are conditioned on point estimates of such variances, and because such point estimates will tend to be imprecise, we have avoided use of maximum likelihood at the country level, opting instead for a Bayesian approach.

The Bayesian approach bases all inferences on the posterior distributions of the parameters of interest. The posterior distribution of the between-country variances is interesting in itself and useful in interpreting other results. For example, we found that the posterior distribution of the variance of the country literacy means was quite dispersed, conveying a realistic degree of uncertainty about the proportion of variation in literacy outcomes that lies between countries. When we examined the relationship between GNP and mean literacy, our inference about that relationship fully took into account the uncertainty about this variance. As a result, the posterior distribution of the regression coefficient relating GNP to mean literacy is quite dispersed. In contrast, other posterior distributions, such as that of the gap between males and females in literacy, were very concentrated about their means reflecting the substantial amount of information in the data about that parameter. Nevertheless, gaps between males and females were found to vary significantly from country to country, as were gaps between students of high and low social status.

Heyneman and Loxley (1983) had hypothesized that effects of social status on literacy outcomes are greater in more developed than in less developed societies. Applying a multilevel perspective to this problem, we found that such social status gaps are more complicated than might be expected. They can arise because of gaps between high and low status students attending the same school, because schools are segregated (to some degree) with respect to social status, and because the expected literacy level for students of the same social status depends on the mean social status of the school attended. The modeling framework we have used enables one, given adequate data, to separate these effects.

Our substantive inferences, especially those regarding social status, must be viewed with caution in light of data limitations. We were disappointed to find that potentially important indicators of social status were not measured on comparable metrics across countries and therefore could not be used for our purpose. Thus, social status was underspecified. There are two possible antidotes to this problem. First, it may be possible to discover the sources of incommensurability of some of the measures in the study data. If so, a more adequate specification of the social status construct may be the basis for firmer inferences about cross-national differences in social status effects as they operate within and between schools and countries. Second, this experience may encourage designers of future cross-national surveys of literacy to take pains to construct equitable measures of these key constructs.

There are potentially important uses of cross-national data other than those considered here to which our modeling framework could be applied. Three come immediately to mind.

Controlling for Student- and Class-Level Variables in Estimating Country Means. Our analyses of country means was based on unadjusted means or means adjusted for GNP. One might wish instead to study country means adjusted for social status or other student, classroom, or school variables. Our approach can easily be adapted to this task. One can simply scale the covariates used for the adjustments as deviations from a common international mean. However, interpretations must be made with care. Student and school social status are probably best viewed as endogenous to GNP, so the effect of GNP on literacy adjusting for social status would underestimate the total effect of GNP.

Assessing the Variable Effect of Educational Reforms Implemented in a Number of Countries. The analytic approach we have outlined is well suited to compare relationships between policy-relevant variables and outcomes in several countries. Consider, for example, a reform requiring postgraduate education for teachers. One could examine the relationship between postgraduate education and classroom literacy in each country, assess the degree to which these relationships are heterogeneous, estimate the average relationship between postgraduate education and literacy across countries, and test hypotheses about student, school, or country characteristics that moderate that relationship.

Studying the Varying Structure of Variation at Several Levels of Social Organization Across Countries. The proportion of variation in outcomes that lies within and between schools (or classrooms) may vary from country to country for a number of reasons. Some countries may have greater segregation of classrooms or schools with respect to demographic predictors, and some may use selective admission of students to elite schools, inflating the between-school variance. Decentralized governance of schools may lead to more variability in instructional inputs and outcomes, leading again to greater between-school variance. Some countries' teachers may be more variable than others in prior education, creating variability between classroom means. Identifying and accounting for discrepancies between societies in the proportion of variance at each level is a potentially interesting task that can be approached via the modeling framework we have described.

There are undoubtedly other cross-national research issues that will require other analytic strategies. Our broader agenda is to stimulate thinking about how such data can be productively exploited, how statistical methods can be tailored to such uses, and about how new cross-national surveys can be designed to facilitate such analyses.

References

Bartlett, M.S. (1993). On the theory of statistical regression. *Proceedings of the Royal Statistical Society of Edinburgh*, 53, 260-283.

Bayes, T.R. (1763). An essay towards solving a problem in the doctrine of changes. *Philosophical Transactions of the Royal Society*, 53, 370 (reprinted in *Biometrika* (1958), 45, 293-315).

Bosker, R.J. and Guldemond, H. (1991). Interdependency of performance indicators: An empirical study of a categorical school system. In S.W. Raudenbush and J.D. Willms (eds.), *Schools, pupils and classrooms: International studies of schooling from a multilevel perspective*. San Diego: Academic Press.

Bryk, A.S., and Raudenbush, S.W. (1992). *Hierarchical linear models in social and behavioral research: Appiications and data analysis methods*. Beverly Hills, CA: Sage Publications.

Bryk, A.S., Raudenbush, S.W., Congdon, R.T., and Seltzer, M. (1988). *An introduction to HLM: Computer program and user's guide*. Chicago: Scientific Software, Inc.

Burstein, L. (1980). The analysis of multi-level data in educational research and evaluation. *Review of Research in Education*, 8, 158-233.

De Finetti, B. (1964). Foresight: its logical laws, its subjective sources. In H.E. Kyburg, Jr., and H.E. Smokler (eds.), *Studies in subjective probability*, 93-158. New York: Wiley.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, Series B, 39, 1-38.

Dempster, A.P., Rubin, D.B., and Tsutakawa, R.K. (1991). Estimation of covariance components models. *Journal of the American Statistical Association*, 76, 341-353.

Fitz-Gibbon, C. (1991). Multilevel modeling in an indicator system. In S.W. Raudenbush and J.D. Willms (eds.), *Schools, pupils and classrooms: International studies of schooling from a multilevel perspective*. San Diego: Academic Press.

Fotiu, R.P. (1989). *A comparison of the EM and data augmentation algorithms on simulated small sample hierarchical data from research on education*. Michigan State University, Ph.D. diss., East Lansing.

Gelfand, A.E., and Smith, A.F.M. (1990). Sampling-based approaches to calcuiating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

Gelfand, A.E., Hills, S.E., Racine-Poon, A., and Smith, A.F.M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, 85, (412), 972-985.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

Goldstein, H. (1987). *Multilevel models in educational and social research*. London: Oxford University Press.

Heyneman, S.P., and Loxley, W.A. (1983). The effect of primary school quality on academic achievement across twenty-nine high and low-income countries. *American Journal of Sociology*, 88, 1162-1194.

Lindley, D.V., and Smith A.F.M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Society*, Series B, 34, 1-41.

Mason, W.M., Wong, G.M., and Entwistle, B. (1983-84). Contextual analysis through the multilevel linear model. In S. Leinhardt (ed.), *Sociological Methodology*, 72-103. San Francisco: Jossy-Bass.

Raudenbush, S.W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics*, 13, (2), 85-116.

Raudenbush, S.W. (1992). Hierarchical linear models and experimental design. In L. Edwards, (ed.), *Applied analysis of variance in behavioral science*. New York: Marcell-Decker.

Raudenbush, S.W., and Bryk, A.S. (1986). A hierarchical model for studying school effects. *Sociology of Education,* 59, 1-17.

Rubin, D.B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics*, 6 (4), 377-400.

Seber, G.A.F. (1978). *Linear progression analysis*. New York: Wiley.

Seltzer, M.H. (1993). Sensitivity analysis for fixed effects in the hierarchical model: A Gibbs sampling approach. *Journal of Educational Statistics*, 18 (3), 207-235.

Smith A.F.M. (1973). A general Bayesian linear model. *Journal of the Royal Society*, Series B, 35, 61-75.

Tanner, M.A. (1992). *Tools for statistical inference*. New York: Springer-Verlag.

Tanner, M.A., and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, 82, 528-550.

Wheeler, C.W., Raudenbush, S.W., and Pasigna, A. (1992). Policy initiatives to produce teacher productivity in Thailand: An essay on implementation, constraints, and opportunities for educational reform. *International Journal of Educational Research*, 17, 2.

Willms, J.D. (1986). Social class segregation and its relationship to pupils' examination results in Scotland. *American Sociological Review,* 51, 224-241.

Willms, J.D. (1992). *Monitoring school performance: A guide for educators.* Lewes: Falmer.

Appendix
Empirical Bayes and Bayes Estimation Theory for Two-Level Models with Normal Errors

1. Introduction

Historically, the hierarchical linear model (HLM) has been developed and promoted from a Bayesian perspective.[1] More generally, Bayesian approaches to statistical problems have been studied since Thomas R. Bayes's (1763) famous paper,[2] but only recently have practical estimation techniques been available to implement Bayesian statistical methods for many current applications. One difficulty encountered with traditional implementation of Bayesian methods is the required integration over one or more parameter spaces. Many applications of scientific interest have complicated, multidimensional parameter spaces. Some of these integration problems can be solved with sophisticated numerical analytic techniques, while others have been resistant to analytic solution.

Two estimation techniques, the EM algorithm developed by Dempster, Laird, and Rubin[3] and the Gibbs sampler introduced by Geman and Geman[4] have been instrumental in making the Bayesian approach to the HLM a practical alternative. The EM approach to HLM as first described by Dempster, Rubin, and Tsutakawa[5] can be viewed either as a strictly classical procedure or as providing an approximation to the Bayesian posterior distribution. This approximation is known as an empirical Bayes approach because the parameters of certain prior distributions are estimated from the data rather than specified a priori. The empirical Bayes strategy we have adopted for within-country analysis is described briefly in the next section. In Section 3, we discuss the Gibbs sampler as an improved Bayes solution in the context HLM. The Gibbs sampler is a sampling-based algorithm for calculating finite approximations to posterior distributions enabling one to incorporate more information into the calculation of a posterior distribution and provide a better account of the uncertainty associated with parameter estimation than is possible using EM. We refer the reader to Fotiu,[6] Gelfand and Smith,[7] and Seltzer[8] for more detailed treatments.

We note that the stage-1 analysis employs the empirical Bayes approach within each country. The stage-2 analysis employs the Gibbs sampler to synthesize results from the several countries.

[1] D.V. Lindley, and A.F.M. Smith. Bayes Estimates for the Linear Model (with discussion). *Journal of the Royal Society*, Series B, 34, 1-41, 1972.

[2] T.R. Bayes. An Essay Towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society*, 53, 370, 1763 (reprinted in *Biometrika*, 45, 293-315, 1958).

[3] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data Via the EM Algorithm (with discussion). *Journal of the Royal Statistical Society*, Series B, 39, 1-38, 1977.

[4] S. Geman, and D. Geman. Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741, 1984.

[5] A.P. Dempster, D.B. Rubin, and R.K. Tsutakawa. Estimation in Covariance Components Models. *Journal of the American Statistical Association*, 76, 341-353, 1981.

[6] R.P. Fotiu. *A Comparison of the EM and Data Augmentation Algorithms on Simulated Small Sample Hierarchical Data from Research on Education*. Unpublished doctoral dissertation, East Lansing, MI: Michigan State University, 1989.

[7] A.E. Gelfand, and A.F.M. Smitn. Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85, 398-409, 1990.

[8] M.H. Seltzer. Sensitivity Analysis for Fixed Effects in the Hierarchical Model: A Gibbs Sampling Approach. *Journal of Educational Statistics*, 18(3), 207-235, 1993.

2. Empirical Bayes Estimation with the EM Algorithm

2.1 The Model

We now consider the two-level HLM and its assumptions for the empirical Bayes estimation approach.[9] The model is formulated in submodels: a level-1 model that describes variation within clusters and a level-2 model that describes variation between clusters.

Level-1 Model. Within clusters such as classrooms, the outcome Y is viewed as depending on characteristics of level-1 units according to the model

$$Y = X\beta + r,\ r \sim N(0, \Sigma) \tag{1}$$

where Y is a vector of outcomes, X is a matrix of known predictors, β is a vector of unknown level-1 regression coefficients describing the relationship between X and Y within the clusters, r is a vector of level-1 random effects, and Σ is a positive-definite level-1 covariance matrix. Assuming X to be of full rank and β known, one might estimate β via generalized least squares, i.e.,

$$\hat{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y \tag{2}$$

$$V = Var(\hat{\beta}) = (X^T \Sigma^{-1} X)^{-1} \tag{3}$$

Typically, it is assumed that $\Sigma = \sigma^2 I$ in which case equation (2) reduces to ordinary least squares with $V = \sigma^2 (X^T X)^{-1}$.

Level-2 Model. Between clusters, the coefficients β are viewed depending upon cluster characteristics and random error according to the model

$$\beta = W_\gamma + u,\quad u \sim N(0, T) \tag{4}$$

where W is a matrix of known cluster characteristics, γ is a vector of unknown level-2 regression coefficients describing the relationship between W and β between clusters, u is a vector of level-2 random effects, and T is a positive definite level-2 covariance matrix, having block diagonal structure with J identical submatrices τ along the main diagonal, one submatrix for every cluster $j = 1, 2, \ldots, J$, i.e.. T = subdiag (τ).

Combined Model. Substituting equation (4) into equation (1)gives the combined model

$$Y = XW_\gamma + Xu + r. \tag{5}$$

[9] We present the model in its "hierarchical form" as opposed to the more general mixed model form. Raudenbush (S.W. Raudenbush. Educational Applications of Hierarchical Linear Models: A Review. *Journal of Educational Statistics*, 13,2,85-116, 1988) discusses the two forms of the model. This clarifies parallels with our application of Gibbs sampling, although the mixed model form is actually more general and will be employed in the stage-1 analysis.

Premultiplying equation (5) by $VX^T\Sigma^{-1}$ yields the equivalent model

$$\hat{\beta} = W_{\gamma}+u+VX^{T}\,\Sigma^{-1}r \tag{6}$$

showing that the marginal distribution of $\hat{\beta}$ is $N\left(W_{\gamma},\Delta\right)$ with $\Delta = V + T$. Thus, the generalized least squares estimator of γ and its covariance matrix are given by

$$\gamma^{*} = (W^{T}\Delta^{-1}W)^{-1}W^{T}\Delta^{-1}\hat{\beta} \tag{7}$$

and

$$Var(\hat{\gamma}) = D_{\gamma}=(W^{T}\Delta^{-1}W)^{-1}. \tag{8}$$

Empirical Bayes Estimation. Following Dempster, Rubin, and Tsutakawa,[10] we now formulate a noninformative prior distribution for γ such that, a priori,

$$\gamma \sim N\,(0,\Gamma),\ \Gamma^{-1} \rightarrow 0 \tag{9}$$

Equation (9) assumes that the prior precision, Γ^{-1}, of our knowledge about the value of γ approaches 0. As a result, the specific value of the location parameter is inconsequential, and we have chosen 0 for convenience. Then the conditional posterior density of $\gamma\,|\,Y,\Sigma,T$ is $N(\gamma^{*},D_{\gamma})$ and the conditional density of $\beta\,|\,Y,\Sigma,T$ is $N(\beta^{*},D_{\beta})$ where γ^{*} is given by equation (7), D_{γ} is given by equation (8), and we have

$$\beta^{*} = \Lambda\hat{\beta}+(I-\Lambda)W_{\gamma}^{*} \tag{10}$$

and

$$D_{\beta} = L^{-1}+(I-\Lambda)WD_{\gamma}W^{T}(I-\Lambda)^{T}, \tag{11}$$

where

$$L = V^{-1}+T^{-1}, \tag{12}$$

$$\Lambda = L^{-1}V^{-1}.$$

We note that the conditional covariance between β and γ, is

$$Cov(\beta,\gamma\,|\,Y,\Sigma,T) = -L^{-1}X^{T}\,\Sigma^{-1}XWD_{\gamma}. \tag{13}$$

Empirical Bayes inferences about β and γ are typically made by substituting maximum likelihood (ML) estimates of Σ and T in equations (7), (8), (10) and (11). Such inferences do not take into account the uncertainty of the ML estimates.

[10] See footnote 5.

2.2 Covariance Estimation via EM

Suppose that, in addition to the data Y, the level-1 random effects r and the level-2 random effects u were also observed. Then, with T = subdiag (τ) and $\Sigma = \sigma^2 I$, ML estimators of the covariance components τ and σ^2 could be computed simply as

$$\hat{\tau} = \frac{1}{J}\sum u_j u_j^T \tag{14}$$

$$\hat{\sigma}^2 = \frac{1}{N} r^T r$$

where J is the number of clusters, N is the number of level-1 units, u_j is the jth subvector of u, and r_j is the jth subvector of r. Of course, the quantities u and r are not observed. However, given current estimates of the covariance parameters, the sufficient statistics defined by equation (14) (termed "complete-data sufficient statistics")[11] can be *estimated* by their conditional expectations given the data and these current parameter estimates. Thus, based on equations 10 to 13, and denoting current estimates with the superscript "^{p}", we have

$$\begin{aligned} E\left(\sum u_j u_j^T \middle| Y,\tau^p,\sigma^{2p}\right) &= \sum \left(\beta_j^{*p} - W_j \gamma^{*p}\right)\left(\beta_j^{*p} - W_j \gamma^{*p}\right)^T \\ &\quad \sum Var\left(u_j \middle| Y,\tau^p,\sigma^{2p}\right) \\ E\left(r^T r \middle| Y,\tau^p,\sigma^{2p}\right) &= \left(Y - X\beta^{*p}\right)^T \left(Y - X\beta^{*p}\right) \\ &\quad + \; Trace\left(X^T X * D_\beta^p\right) \end{aligned} \tag{15}$$

where

$$Var\left(u_j \middle| Y,\tau,\sigma^2\right) = L_j^{-1} + L_j^{-1} W D_\gamma W^T L_j^{-1}. \tag{16}$$

Given an initial estimate of τ and σ^2, and therefore of the posterior distribution of γ (from equations (7) and (8) and ß (from equations 10 to 13), the EM algorithm iteratively computes the complete-data sufficient statistics (equation 15) and then uses these to compute new complete-data ML estimators using equation (14). Equation (15) is called the "E" or "Expectation" step and Equation (14) is called the "M" or "Maximization" step. Under quite mild conditions, each E-M cycle increases the observed data likelihood

$$L\left(Y|\tau,\sigma^2\right) = \frac{f\left(Y|\beta,\tau,\sigma^2\right) * g\left(\beta|\tau,\sigma^2\right)}{h\left(\beta|Y,\tau,\sigma^2\right)} \tag{17}$$

until convergence to a maximum.

[11] See footnote 5.

At convergence, empirical Bayes estimates are based on

$$p(\beta,\gamma|Y,\sigma^2=\sigma^{2*},\tau=\tau^*)=\text{const.}\times f(Y|\beta,\hat{\sigma}^2)g(\beta|\gamma,\hat{\tau})p(\gamma). \qquad (18)$$

As mentioned, the empirical Bayes approach does not take into consideration the uncertainty of our knowledge of the unknown variance-covariance components σ^2 and τ.

3. Bayesian Estimation with the Gibbs Sampler

Bayes (Via Gibbs) Versus Empirical Bayes (Via EM). The Gibbs sampler is a special case of the data-augmentation algorithm described by Tanner and Wong.[12] These two approaches are compared by Gelfand and Smith.[13] A number of methodologies that offer solutions on a continuum between the Gibbs sampler and the EM algorithm are discussed by Tanner.[14] The essential difference between empirical Bayes estimation via the EM algorithm and Bayesian estimation via the Gibbs sampler applied to the HLM is that the Bayesian approach using the Gibbs sampler computes a posterior distribution for the variance-covariance components in the model, rather than summarizing this information into a point estimate as illustrated in equation 18. Hence, Bayesian inferences about β, γ are based on

$$p(\beta,\gamma|Y)=\text{const.}\times\int\int f(y|\beta,\sigma^2)g(\beta|\gamma,\tau)p(\gamma)p(\tau)p(\sigma^2)\partial\tau\partial\sigma^2. \qquad (19)$$

This Bayesian approach provides more information about the posterior distribution of a model's parameters than is available with empirical Bayes because more elements of uncertainty are accounted for explicitly.

The following assumptions for the Bayesian formulation are the same as those specified earlier for the empirical Bayes approach, except that we now add prior distributions for the parameters σ^2 and τ. The variance parameter, σ^2, is assumed a priori to have an inverse chi-square distribution given by

$$\sigma^2\sim\nu_0\sigma_0^2\chi^{-2}(\nu_0) \qquad (20)$$

with the degrees of freedom parameter ν_0 and noncentrality parameter σ_0^2. This prior distribution for σ^2 is considered noninformative in its contribution to the posterior distribution of σ^2 as ν_0 approaches 0. In addition, the variance-covariance matrix, τ, is assumed a priori to have an inverse Wishart prior distribution given by

$$\tau\sim W^{-1}(\Psi,\nu) \qquad (21)$$

where Ψ is the precision matrix of the inverse Wishart distribution and ν is the degrees of freedom parameter. This prior distribution for ν is assumed to be noninformative in its contribution to the posterior distribution of τ as the degrees of freedom parameter, ν, approaches 0, and Ψ approaches 0.

[12] M.A. Tanner. *Tools for Statistical Inference* (New York: Springer-Verlag, 1992).

[13] See footnote 7.

[14] See footnote 12.

It is the assumptions concerning the model in conjunction with the data that determine the joint distribution of all unknowns given by

$$p(Y,\beta,\sigma^2,\gamma,\tau)=f(Y|\beta,\sigma^2)g(\beta|\gamma,\tau)p(\sigma^2,\gamma,\tau). \tag{22}$$

Two alternative expressions for this joint density[15] are

$$\begin{aligned} p(Y,\beta,\sigma^2,\gamma,\tau)&=q_1(Y)q_2(\beta,\sigma^2|Y)q_3(\gamma,\tau|\beta,\sigma^2) \\ &=r_1(Y)r_2(\beta,\sigma^2|Y,\gamma,\tau)r_3(\gamma,\tau|Y) \end{aligned} \tag{23}$$

with $q_1 = r_1$. Gibbs sampling exploits the fact that, although this joint density is not tractable, both q_3 and r_2 are readily accessible (as shown below) so that it is simple to sample from those. Starting from rough guesses at the values of γ and τ, Gibbs works by sampling from r_2 to obtain new values of β and σ^2. Knowing those values, it is easy to sample from q_3, yielding new values of γ and τ. This process iterates as described in more detail in the next section. The goal is to obtain the joint posterior densities of all unknowns, i.e.,

$$p(\beta,\sigma^2,\gamma,\tau|y)=\frac{p(Y,\beta,\sigma^2,\gamma,\tau)}{q_1(y)} \tag{24}$$

The marginal posteriors are readily derived from this joint posterior.

The Gibbs Sampler. The Gibbs sampler uses the data and distribution assumptions to generate approximate posterior distributions by Monte Carlo sampling. Successive iterations move closer to the true posterior distribution until stochastic convergence is achieved. After convergence, we can collect a sufficiently large set of generated parameters from subsequent iterations as a finite approximation to the true posterior distribution.

There are two basic steps to this algorithm. The first step is to calculate a current approximation of a required posterior distribution. The second step is to sample from this distribution.

Initially, suppose the parameters γ and τ from q_3 in (23) and σ^2 were observed. Then β^* and D_β could be calculated, where the asterisk (*) indicates an estimated posterior mean. Next, given β^* and D_β just calculated, β can be sampled by Monte Carlo methods from the posterior distribution of $\beta|Y,\gamma,\tau,\sigma^2$. With the knowledge of β, an estimate of the central tendency of the conditional distribution of σ^2 (given β) can be calculated and then σ^2 is sampled from its conditional distribution given Y and β (see below). The resulting parameter pair of β and σ^2 approximates a sample from r_2 in (23).

In a similar manner, a sample of γ and τ can be obtained. Given parameters β and σ^2 from r_2 just realized along with τ from the previous iteration, the posterior mean and variance of γ can be calculated from the distribution of γ given β,σ^2, and τ. Next, the location parameter for the conditional density of γ (given β and this new γ) can be calculated and then τ can be sampled from its conditional density given β and γ. This results in the parameter pair of γ and τ approximating a sample from q_3 in (23).

[15] C.N. Morris. Comment on article by Tanner and Wong. *Journal of American Statistician*, 82, 542-543, 1987.

To improve the approximations, the resulting sample from q_3 is considered an intermediate approximation of q_3 and recycled back to calculate estimated conditional distributions and generate a new sample to update r_2. The new sample from r_2 is used in the same manner to calculate estimated conditional distributions and obtain a sample from q_3. This iteration scheme is repeated until convergence. Afterwards, m more iterations are completed and the parameter values from each iteration are collected. If m is large, the mixture of the densities can be considered a finite approximation to the joint posterior distribution given in (24).

One advantage of this algorithm is that not only are point estimates generated, but the results also include finite approximations to the true joint posterior distribution. For example, the sampled values for a parameter of interest can be sorted in order and then the $\alpha/2$ percent tails of the distribution can be easily determined. As a consequence, highest posterior densities can be easily determined for both symmetric and nonsymmetric distributions.

Initial Values. It does not matter at what level parameter estimation begins. In the example detailed next, we shall begin at the first level in the hierarchy to obtain values for β and σ^2. Initial values for γ, τ, and σ^2 are required for the start of this algorithm, in addition to the ordinary least squares (OLS) estimate of β. Initial parameter estimates may be calculated by a variety of techniques. Of course, better initial estimates will result in faster convergence. One strategy is to use the empirical Bayes modal estimates of the posterior distributions as a starting point.

Calculating and Sampling the First-Level Parameters β and σ^2. The sampling of the first-level parameters β requires the knowledge of γ, τ and σ^2. The data Y are summarized in the OLS estimator, $\hat{\beta}$ and its sampling variance, V. The first iteration of the algorithm uses the initial values, while subsequent iterations use the previous iteration's generated values.

The desired posterior density r_2 can be rewritten as

$$r_2(\beta,\sigma^2,|Y,\gamma,\tau)=r_2'(\beta|Y,\gamma,\sigma^2,\tau)r_2''(\sigma^2). \tag{25}$$

Let $\sigma^{2(i-1)}$, $\gamma^{(i-1)}$, and $\tau^{(i-1)}$ indicate the previous iteration's sample. New values for the $\beta_j^{*(i)}$'s can be calculated as

$$\beta_j^{*(i)} = \Lambda^{(i-1)}\hat{\beta}+(I-\Lambda^{(i-1)})W_j\gamma^{(i-1)}. \tag{26}$$

A new set of βj's can be sampled from the density of β given γ, τ, σ^2, which is $N(\beta_j^{*(i)}$, $D(\beta_j^{i-1}))$, where

$$D(\beta_j^{(i-1)}) = \mathrm{Var}(\beta_j|\sigma^{2(i-1)},\gamma^{(i-1)},\tau^{(i-1)}) = L_j^{(i-1)-1}. \tag{27}$$

First, the matrix $D\left(\beta_j^{(i-1)}\right)$ is factored by the Cholesky method such that $D\left(\beta_j^{i-1}\right) = M_j^{(i)}M_j^{(i)T}$, where $M_j^{(i)}$ is a lower triangular matrix. Next, the $\beta_j^{(i)}$'s are sampled with the following equation:

$$\beta_j^{(i)} = \beta^{*(i)}+M_j^{(i)}x_j^{(i)} \tag{28}$$

where $x_j^{(i)}$ is a vector containing independent and identically distributed elements sampled from $N(0, 1)$.

After generating new $\beta_j^{(i)}$'s from (28), $\sigma^{2*(i)}$ can be calculated as follows:

$$\sigma^{2*(i)} = \sum \left(Y_j-X_j\beta_j^{(i)}\right)^T\left(Y_j-X_j\beta_j^{(i)}\right)/N, \tag{29}$$

where $N = \Sigma n_j$. Alternatively, equation (29) can be expressed as

$$\sigma^{2*(i)} = \left[\sum \left(Y_j-X_j\hat{\beta}_j\right)^T\left(Y_j-X_j\hat{\beta}_j\right)+\sum \left(\hat{\beta}_j-\beta_j^{(i)}\right)^T X_j^T X_j\left(\hat{\beta}_j-\beta^{(i)}\right)\right] /N, \tag{30}$$

to minimize computation and illustrate the partitioning of the sources of variation. The first part of the expression enclosed in brackets computes the sum of the squared deviations of the ordinary least squares prediction from the observed data vector Y. This expression can be computed once because its value is constant across iterations of the algorithm. The second part of the expression adds the variance as a function of the deviation of $\hat{\beta}_j$ from the ith iteration's realization of the parameter $\beta_j^{(i)}$. It has been assumed that σ^2 has an inverse chi-square prior distribution with υ_0 degrees of freedom. The posterior distribution of σ^2 given the data and $\beta^{(i)}$ is

$$\sigma^2 \sim \left(\upsilon_0\sigma_0^2+N\sigma^{2*(i)}\right)|\chi^{-2}\left(\upsilon_0+N\right). \tag{31}$$

For a noninformative prior, we can let υ_0 approach 0 in its contribution to the posterior distribution and (31) becomes

$$\sigma^2 \sim N\sigma^{2*(i)}|\chi^{-2}(N). \tag{32}$$

To sample σ^2, we generate a chi-square variate with N degrees of freedom, invert it, and substitute it in (32) to obtain $\sigma^{2(i)}$.

Now we have a sample from r_2 of the parameter pair $\left(\beta^{(i)},\sigma^{2(i)}\right)$.. These are passed on to calculate and sample the second-level parameters in the HLM.

Calculating and Sampling the Second-Level Parameters γ and τ. Given a pair of$\left(\beta,\sigma^2\right)$ drawn from r_2 $\left(\beta,\sigma^2|Y,\gamma,\tau\right)$ we can calculate the posterior mean, γ^* and with $\tau^{(i-1)}$ from the previous iteration a sample is drawn from the posterior distribution of γ. In a similar manner, we can then use our sampled γ to calculate a conditional distribution for τ and then sample from it. The goal is to achieve a realization of the parameter pair (γ, τ) from q_3. The distribution of q_3 may be expressed in the following form:

$$q_3(\gamma,\tau\,|\,\beta,\sigma^2) = q_3'(\gamma\,|\,\tau,\beta,\sigma^2)q_3''(\tau). \tag{33}$$

We note in passing that the calculation and sampling of γ and τ does not directly depend on σ^2.

A posterior sample of $\gamma^{(i)}$ given β and τ is drawn from the normal distribution

$$\gamma^{i} \sim N\!\left(\gamma^{*(i)}, \left[\sum W_j^T \tau^{-1^{(i-1)}} W_j\right]^{-1}\right). \tag{34}$$

The posterior mean value for γ^* can be calculated as:

$$\gamma^{*(i)} = \left(\sum W_j^T W_j\right)^{-1} \sum W_j^T \beta_j^{(i)}. \tag{35}$$

A sample is drawn from equations 34 and 35 given $\beta^{(i)}$ and $\tau^{(i-1)}$ as follows. Let

$$A^{(i)} = \left(\sum W_j^T \tau^{(i-1)} W_j\right)^{-1}. \tag{36}$$

The matrix $A^{(i)}$ is then factored such that

$$A^{(i)} = B^{(i)}B^{(i)T}, \tag{37}$$

where $B^{(i)}$ is a lower triangular Cholesky factor of $A^{(i)}$. The matrix equation used to generate a new $\gamma^{(i)}$ is

$$\gamma^{(i)} = \gamma^{*(i)} + B^{(i)}x^{(i)}. \tag{38}$$

The column vector $x^{(i)}$ contains elements that are independently and identically distributed $N(0, 1)$.

The new $\gamma^{(i)}$ vector is used to update the posterior distribution of τ. Let

$$C^{(i)} = J^{-1}\Sigma\left(\beta_j^{(i)} - W_j\gamma^{(i)}\right)\left(\beta_j^{(i)} - W_j\gamma^{(i)}\right)^T. \tag{39}$$

Based on the noninformative inverse Wishart prior distribution with parameters Ψ and υ, the posterior distribution of τ given β and γ is given by

$$\tau \sim W^{1}(C^{(1)} + \Psi,\ J + \nu) \tag{40}$$

If we assume that the prior precision matrix Ψ approaches 0 and that the prior degrees of freedom parameter υ approaches 0, we can sample τ from

$$\tau \sim W^{-1}(C^{(i)},\ J). \tag{41}$$

To sample τ from equation (41), we find the Cholesky factor of the $C^{(i)}$ such that $C^{(i)} = D^{(i)}D^{(i)}T$, where $D^{(i)}$ is a lower triangular matrix. For $E^{(i)}$, a lower triangular matrix, define

$$F^{(i)} = J^{-1}D^{(i)}E^{(i)}E^{(i)T}D^{(i)T} = J^{-1}(D^{(i)}E^{(i)})(D^{(i)}E^{(i)})^{T}. \tag{42}$$

If e_{ij} is an element of E where each e^{2}_{jj} element on the main diagonal is an independent chi-square variable with (J) degrees of freedom and the elements below the diagonal are independently distributed $N(0, 1)$, then $F^{(i)}$ will have an inverted Wishart distribution with J degrees of freedom.[16]

This concludes one complete iteration of the Gibbs sampler. To improve the approximation to the posterior distribution of interest the new values for $\gamma^{(i)}$ and $\tau^{(i)}$ are passed to the next iteration to generate new updated values for β and σ^2. This process of calculation and sampling is continued until convergence. After the algorithm has converged, a sequence of m further iterations are performed. The parameter samples resulting from each iteration are collected. This sample of size m of the model's parameters is considered a finite approximation to the true joint posterior distribution.

3.1 The *V*-Known Modification

There are some situations when the dispersion matrix V can be estimated with enough precision to be considered known. This may be a reasonable assumption when the sample size used to compute V is sufficiently large, as in the case of the IEA Reading Literacy Study data from each country. An advantage resulting when the V-known simplifying assumption is tenable is a reduction in the algorithm's computational burden. In this situation, computing an estimate of V and sampling from its posterior distribution every iteration is not required. Another advantage occurs frequently in meta-analysis situations. Typically, access to the raw data is impossible and one is forced to work with summary statistics. For the case where V is considered known, the only modification required of the Gibbs sampler developed above is to skip the estimation and sampling of elements of V such as σ^2. Otherwise the algorithm is the same.

[16] M.S. Bartlett. On the Theory of Statistical Regression. *Proceedings of the Royal Statistical Society of Edinburgh*, 53, 260-283, 1933.

Biographical Sketches of the Authors

Marilyn R. Binkley is a Senior Research Associate at the National Center for Education Statistics, U.S. Department of Education. Her research interests are in literacy, cognition, instruction, assessment, and the impact of school policy. She has taught at the elementary, secondary, and undergraduate levels. She has published articles and booklets dealing with text, text features, learning to read, and reading instruction. Most recently she served as the U.S. National Research Coordinator for the IEA Reading Literacy Study, and she is currently leading the U.S. portion of the International Adult Literacy Survey. She received her doctorate in teacher education with specialties in reading and international education from the George Washington University.

Edward C. Bryant, Chairman Emeritus and Founder, Westat, Inc., is a statistical consultant. Dr. Bryant was formerly Head of the Department of Statistics, University of Wyoming. His principal interests are in sample design of surveys and quality improvement. Dr. Bryant received his Ph.D. in Statistics from Iowa State University.

Yuk Fai Cheong is a doctoral candidate at the College of Education, Michigan State University. His main fields of interest are research on processes and effects in schools and classrooms. He has been awarded a dissertation grant by the American Educational Research Association to develop and test new statistical methodologies used to study middle grades curriculum. Mr. Cheong received his M.A. in Educational Systems from Michigan State University.

Graham Kalton is a Senior Statistician and Vice President at Westat, Inc. Dr. Kalton was formerly Research Scientist at the Survey Research Center, Professor of Biostatistics, and Professor of Statistics at the University of Michigan. He received his Ph.D. in Survey Statistics from the University of Southampton. He is a Fellow of the American Statistical Association and of the American Association for the Advancement of Science, and a member of the International Statistical Institute. Dr. Kalton serves as Associate Editor of *Survey Methodology* and *Statistics in Transition*. His research interests lie in survey sampling and survey methodology.

Barbara A. Kapinus is Program Director for the Council of Chief State School Officers. She previously served as Specialist in Reading and Communication Skills for the Maryland State Department of Education and as Project Director of the National Assessment of Educational Progress Consensus Project. Dr. Kapinus received her Ph.D. in Education at the University of Maryland. She was President of the State of Maryland International Reading Association and was Co-editor of the assessment column in *The Reading Teacher* during 1993-94. Her major areas of research interest include reading assessment, vocabulary development, and metacognition.

Daniel Kasprzyk serves as Chief, Special Survey and Analysis Branch, Elementary and Secondary Education Statistics Division, National Center for Education Statistics, U.S. Department of Education. He formerly held various positions associated with the management of the Survey of Income and Program Participation, U.S. Bureau of the Census. Dr. Kasprzyk received his Ph.D. in Mathematical Statistics from the George Washington University. He serves as Associate Editor for the *Journal of Official Statistics*, and he is a Fellow of the American Statistical Association. Dr. Kasprzyk's principal research interests include panel surveys, compensating for missing survey data, and improving the knowledge and quality of federal data sets.

Irwin S. Kirsch is Executive Director of the Literacy Learning and Assessment Group at Educational Testing Service. He earned his Ph.D. in Educational Measurement; Reading/Literacy from the University of Delaware. Since joining ETS in 1984, he has directed a number of large-scale assessments in the area of literacy and is currently the ETS project director for the first International Adult Literacy Survey. In 1987, he received the ETS Research Scientist Award for his work in this area. Dr. Kirsch has served on a number of panels concerned with literacy issues including *Literacy and Technology* for the Office of Technology Assessment; *Developing a Research Agenda* for the U.S. Department of Education; *What Works* for the Division of Adult and Vocational Education; and the Hudson Institute panel convened to link Workforce 2000 data to the adult literacy scales. He also serves as a reviewer for *Reading Research Quarterly, American Educational Research Journal, Review of Educational Research*, and *Adult Basic Education*. In addition, he co-authored a monthly column on reading/literacy for the *Journal of Reading* from 1988 to 1992. His research interests include the psychology of literacy, issues of comparability and interpretability in large-scale assessments, and using technology to link learning and assessment.

Peter B. Mosenthal is a Professor of Education at Syracuse University where he chairs the Reading and Language Arts Center. He received his master's degree in Linguistics and Ph.D. in Educational Psychology from Ohio State University. For the past several years, he has served as a consultant to Educational Testing Service in the area of Adult Literacy Learning and Assessment, and he has helped to identify and validate the constructs underlying three national adult literacy assessments. His areas of research interest include literacy instruction, policy, and research; literacy in the workplace; literacy assessment; and computer adaptive testing/instruction. He has published extensively in the *Reading Research Quarterly, Journal of Education Psychology*, and the *Elementary School Journal* and serves on the editorial boards for the *Reading Research Quarterly* and *Discourse Processes*. He has co-edited the two volumes of the *Handbook of Reading Research* and has co-authored a column in the *Reading Teacher* addressing issues of learning from text and understanding document literacy. He is currently completing a textbook on teaching and assessing literacy proficiency in the middle grades.

Stephen P. Norris is a Professor of Educational Research and Philosophy of Education at Memorial University of Newfoundland. He received his Ph.D. in Philosophy of Education with a specialty in the Theory of Educational Measurement from the University of Illinois at Urbana-Champaign. Dr. Norris is the author or editor of four books, and has published numerous articles and book chapters dealing with the application of philosophy of science to science education, the philosophy of reading, critical thinking assessment, and educational measurement theory and methods.

Linda M. Phillips is a Professor of Reading Research at Memorial University of Newfoundland. She has published numerous articles, chapters, and technical reports, and has co-authored and co-edited several books and tests. She received her Ph.D. in Reading and Language from the University of Alberta. Her research interests are in cognition and instruction, including inferential processing and text interpretation, self-regulatory thinking strategies, and the epistemological, normative, and pragmatic bases of interpretation.

Stephen W. Raudenbush is Professor of Education in the Measurement and Quantitative Methods Program at Michigan State University. He is Chair of the Management Committee of the *Journal of Educational Statistics*, serves on the Editorial Board of *Psychological Bulletin*, and is a member of the Human Development and Aging Study Section of the National Institutes of Health. Dr. Raudenbush has won the Raymond Cattell Research Award of the American Educational Research Association and the George Z.F. Bereday Outstanding Scholarship Award of the Comparative and International Education Society. His research focuses on developing, testing, and refining statistical models that account for effects of social contexts on individual development.

Keith Rust is an Associate Director of the Statistical Group at Westat, Inc., and was formerly with the Australian Bureau of Statistics. His expertise is in the design and analysis of complex sample surveys. He is the Director, Sample Design and Statistical Operations, for the National Assessment of Educational Progress, and is the International Sampling Coordinator for the Third International Mathematics and Science Study, conducted by IEA. He was a member of the U.S. Steering Committee for the IEA Reading Literacy Study. Dr. Rust is a member of the Committee on National Statistics of the National Academy of Sciences, an Associate Editor of the *Journal of Official Statistics*, and an elected member of the International Statistical Institute. He received his Ph.D. in biostatistics from the University of Michigan.

Marianne Winglee is a Senior Statistician at Westat with experience in research, survey design, survey analyses, and quality control. She has participated in a broad range of research projects in the areas of education policy research, health care, and assessments of children with behavioral problems. Ms. Winglee has provided statistical analysis support to education policy research dealing with financial aid for postsecondary students and special education programs for disadvantaged students. In addition, Ms. Winglee has provided statistical design support, including sample design, weight estimation, and imputation of missing data on projects including the Fast Response Survey System, National Postsecondary Financial Aid Study (NPSAS), Postsecondary Education Quick Information System (PEQIS), and the IEA International Reading Literacy Study. She also co-authored a book, *Who Reads Literature? The Future of the United States as a Nation of Readers*. Ms. Winglee received an M.S. degree in Applied Statistics and an M.A. degree in Psychology from the George Washington University.

☆ U.S. GOVERNMENT PRINTING OFFICE: 1995 - 390-210 - 814/20501

United States
Department of Education
Washington, DC 20208–5650

Official Business
Penalty for Private Use, $300

Postage and Fees Paid
U.S. Department of Education
Permit No. G–17

Fourth Class Special
Special Handling

NCES 95–469