VOLUME 4

Analyzing Data from a National Assessment of Educational Achievement

Gerry Shiel
Fernando Cartwright
Analyzing Data from a National Assessment of Educational Achievement
VOLUME 4

Analyzing Data from a National Assessment of Educational Achievement

Gerry Shiel
Fernando Cartwright

Vincent Greaney and
Thomas Kellaghan, Series Editors
CONTENTS

PREFACE xvii
ABOUT THE AUTHORS AND EDITORS xix
ACKNOWLEDGMENTS xxi
ABBREVIATIONS xxiii
INTRODUCTION 1
  Note 5

Part I
An Introduction to the Statistical Analysis of National Assessment Data
Gerry Shiel

1. THE DATABASE FOR ANALYSES 9
  Saving the CD Files to Your Hard Drive or Server 11
  Survey Instruments 12
  Sampling Weights 14
  SPSS 16
  WesVar 19
  Notes 19
I.A.  NAEA DATA ANALYSIS: FILE DIRECTORY STRUCTURE 109

I.B.  NAEA DATA ANALYSIS: SUBFOLDERS AND FILES 111

I.C.  OPENING AN SPSS FILE IN WESVAR 115

Notes 122

---

Part II

Item and Test Analysis

Fernando Cartwright

8.  INTRODUCTION TO IATA 125

Installing IATA 125
Assessment Data 126
Data Produced by IATA 137
Interpreting IATA Results 139
Sample Data 140
IATA Analysis Workflows and Interfaces 141
Navigating IATA Workflows 144
Notes 145

9.  ANALYZING DATA FROM A PILOT TEST ADMINISTRATION 147

Step 1: Loading Response Data 148
Step 2: Loading the Answer Key 151
Step 3: Analysis Specifications 151
Step 4: Item Analysis 155
Step 5: Test Dimensionality 164
Step 6: Differential Item Functioning 170
Step 7: Scale Review 176
Step 8: Selecting Test Items 180
Step 9: Performance Standards 185
Step 10: Viewing and Saving Results 185
Notes 187

10.  PERFORMING A FULL ANALYSIS OF A FINAL TEST DATA ADMINISTRATION 189

Step 1: Setting up the Analysis 190
Step 2: Basic Analysis Results 192
Step 3: Analysis of Differential Item Functioning 192
Step 4: Scaling 194
Step 5: Selecting Test Items 199
Step 6: Setting Performance Standards 200
Step 7: Saving Results 209
Note 209

11. ANALYSIS OF ROTATED BOOKLETS 211
   Step 1: Loading Data 211
   Step 2: Analysis Specifications 213
   Step 3: Item Analysis Results 214

12. ANALYSIS OF PARTIAL CREDIT ITEMS 217
   Step 1: Loading Data 217
   Step 2: Analysis Specifications 219
   Step 3: Item Analysis Results 220

13. COMPARING ASSESSMENTS 225
   Step 1: Analysis Setup 227
   Step 2: Common Item Linking 230
   Step 3: Rescaling Linked Results 235
   Step 4: Assigning Performance Standards 236
   Notes 239

14. SPECIALIZED METHODS IN IATA 241
   Linking Item Data 242
   Selecting Optimal Test Items 245
   Developing and Assigning Performance Standards 247
   Response Data Analysis with Anchored Item Parameters 250
   Note 256

15. SUMMARY OF IATA WALK-THROUGHS 257

II.A. ITEM RESPONSE THEORY 259
   Note 266

REFERENCES 267

BOX
6.1 Variables in Standard Regression 81
EXERCISES

1.1 Running Descriptive Statistics in SPSS and Saving the Files 17

2.1 Running Explore in SPSS, Single Dependent Variable
(One Level) 26

2.2 Running Explore in SPSS, Single Dependent Variable
(More Than One Level) 31

3.1 Generating Descriptive Statistics in WesVar 36

3.2 Computing a Mean Score and Its Standard Error in WesVar 41

3.3 Computing Mean Scores and Standard Errors in WesVar,
Four Regions 43

4.1 Evaluating the Difference between Two Mean Scores 49

4.2 Evaluating Differences among Three or More Mean Scores 53

5.1 Computing National Percentile Scores 60

5.2 Computing Percentile Scores by Region 63

5.3 Recoding a Variable into Percentile Categories Using WesVar 65

5.4 Computing Percentages of Students Scoring below
Key National Percentile Benchmarks and Standard
Errors by Region 67

6.1 Drawing a Scatterplot in SPSS 73

6.2 Computing a Correlation Coefficient, National Level 76

6.3 Running a Regression in WesVar, One Independent Variable
(Continuous) 82

6.4 Running Regression in WesVar, One Independent Variable
(Categorical) 86

6.5 Estimating Correlation Coefficients 88

6.6 Running Regression in WesVar, More Than One
Independent Variable 89

7.1 Drawing a Column Chart to Show Performance by
Proficiency Level, National Data 98

7.2 Drawing a Bar Chart to Show Percentage at Each
Proficiency Level by Region 100

7.3 Drawing 95 Percent Confidence Intervals for a Series
of Mean Scores 104

7.4 Showing Trend Data with a Line Graph 107

EXERCISE FIGURES

1.1.A Weight Cases Dialog Box 17

1.1.B SPSS Descriptives Dialog Box 18

2.1.A Stem-and-Leaf Plot for Mathematics Scale Scores 29
2.1.B Box Plot for Mathematics Scale Scores
2.2.A Box Plots for Mathematics Scale Scores by Region
3.1.A New WesVar Workbook
3.1.B Specifying Variables for Analysis in WesVar Descriptives
3.1.C Output from WesVar Descriptives
3.1.D Exporting a WesVar File
3.2.A Specifying a Computed Statistic in a WesVar Table
3.2.B Output for WesVar Tables: Computing Mean Score
3.3.A WesVar Workbook before Computing Mean Scores by Region
3.3.B WesVar Output for Computing Mean Scores by Region
4.1.A WesVar Workbook before Assessing the Difference between Two Mean Scores
4.1.B WesVar Output: Mean Mathematics Scores of Students with and without Electricity at Home
4.1.C WesVar Output: Mean Score Difference in Mathematics between Students with and without Electricity at Home
4.2.A WesVar Workbook Showing Adjustment to Alpha Level
4.2.B Completing Cell Definitions in WesVar
4.2.C WesVar Workbook Showing Cell Functions
4.2.D WesVar Output: Mean Mathematics Scores by Region
4.2.E WesVar Output: Mean Mathematics Score Differences by Region
5.1.A WesVar Workbook: Computing Percentile Scores
5.1.B WesVar Output: Computing Percentile Scores
5.2.A WesVar Workbook before Computing Percentile Scores by Region
5.2.B Partial WesVar Output: Computing 10th Percentile Scores by Region
5.3.A WesVar Workbook: Recoding Mathss to Discrete Variable
5.3.B Labeling Percentile Categories in WesVar
5.4.A WesVar Workbook Screenshot before Computing Percentages Scoring below Key National Benchmarks by Region
5.4.B Partial Output: Percentages of Students Scoring below Key National Benchmarks by Region
6.1.A Partial SPSS Dialog Box before Drawing Scatterplot
6.1.B Scatterplot of Relationship between Implementation of Procedures and Problem Solving in Mathematics
CONTENTS
9.5 Item Analysis Results for PILOT1 Data, MATHC1019 155
9.6 Item Analysis Results for PILOT1 Data, MATHC1027 161
9.7 Item Analysis Results for PILOT1 Data, MATHC1075 162
9.8 Item Analysis Results for PILOT1 Data, after Removal of MATHC1075 163
9.9 Test and Item Dimensionality for PILOT1 Data, MATHC1019 166
9.10 Item Dimensionality Results for PILOT1 Data, MATHC1035 167
9.11 Item Dimensionality Results for PILOT1 Data, MATHC1002 170
9.12 DIF Analysis Results for PILOT1 Data by Sex, MATHC1046 171
9.13 DIF Analysis Results for PILOT1 Data by Sex, MATHC1035 173
9.14 DIF Analysis Results for PILOT1 Data by Sex, MATHC1042 174
9.15 DIF Analysis Results for PILOT1 Data by Language of Students' Home, MATHC1006 175
9.16 The Scale Review and Scale Setting Interface 177
9.17 Item Selection Results for PILOT1 Data, 50 Items 181
9.18 Item Selection Results for PILOT1 Data, 79 Items 184
9.19 Viewing Results from the Analysis of PILOT1 Data 186
10.1 Analysis Specifications for CYCLE1 Data 191
10.2 DIF Analysis Results for CYCLE1 Data by Location, MATHC1043 193
10.3 Distribution of Proficiency (IRT Score) and Test Information, CYCLE1 Data 196
10.4 A Comparison of Ideal Test Information and the Normal Distribution 197
10.5 Distribution and Summary Statistics for New Scale Score (NAMscore), CYCLE1 Data 198
10.6 Selecting Items, CYCLE1 Data 200
10.7 Default Performance Standards Interface, CYCLE1 Data 202
10.8 Performance Standards Interface, RP = 50 percent, CYCLE1 Data 205
10.9 Bookmark Data, RP = 50 percent, CYCLE1 Data 205
10.10 Performance Standards Interface with Manually Set Thresholds, CYCLE1 Data 208
11.1 Student Responses, PILOT2 Data 212
11.2 Analysis Specifications, Rotated Booklets, PILOT2 Data 213
11.3 Item Analysis Results, PILOT2 Data, MATHC2003 214
12.1 Item Answer Keys and Metadata, PILOT2 Data 218
12.2 Analysis Specifications, Rotated Booklets with Partial Credit Items, PILOT2 Data 219
12.3 Item Analysis Results, PILOT2 Data, MATHC2003 220
12.4 Partial Credit Item Response Function, CYCLE2 Data, MATHSA001, Score = 2 221
13.1 Response Data Analysis with Linking Workflow 227
13.2 Reference Item Data from CYCLE1 to Link CYCLE2 Data 228
13.3 Item Analysis Results for CYCLE2 Data, MATHSA005, Score = 1 230
13.4 Common Item Linking Results, CYCLE2 to CYCLE1 231
13.5 Common Item Linking Results, CYCLE2 to CYCLE1, MATHC1052 233
13.6 CYCLE2 Test Scores Expressed on the CYCLE1 (NAMscore) Scale 236
13.7 Assigning Performance Standards, CYCLE2 Data 239
14.1 Selecting Optimal Test Items, CYCLE1 Data 247
14.2 Item Data for CYCLE3 with Anchored Item Parameters 253
14.3 Item Analysis Results with Anchored Item Parameters, CYCLE3 Data, MATHC2047 254
II.A.1 Distributions of Proficiency for Correct and Incorrect Respondents to a Single Test Item 262
II.A.2 Distributions of Proficiency for Correct and Incorrect Respondents to a Single Test Item 263
II.A.3 Distributions of Proficiency for Correct and Incorrect Respondents to a Single Test Item and Conditional Probability of Correctly Responding 264

TABLES
1.1 Mathematics Test: Distribution of Items by Content Area and Process 13
1.2 Abbreviated Questionnaire Descriptions 14
5.1 Percentages of Students Scoring below the National 25th Percentile Benchmark by Region 70
5.2 Percentage of Students Scoring at or above the National 75th Percentile Benchmark by Region 70
8.1 Variables Produced or Used by IATA to Describe Student Proficiency and Test Performance 132
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.2</td>
<td>Variables in an Item Data File</td>
<td>133</td>
</tr>
<tr>
<td>8.3</td>
<td>Sample Section of an Item Data File</td>
<td>134</td>
</tr>
<tr>
<td>8.4</td>
<td>Sample Section of an Item Data File for a Partial Credit Item</td>
<td>136</td>
</tr>
<tr>
<td>8.5</td>
<td>Data Tables Produced by IATA</td>
<td>138</td>
</tr>
<tr>
<td>8.6</td>
<td>Traffic Symbols in IATA and Their Meaning</td>
<td>139</td>
</tr>
<tr>
<td>8.7</td>
<td>Tasks in IATA and the Workflows in Which They Are Used</td>
<td>144</td>
</tr>
<tr>
<td>9.1</td>
<td>Distractor Analysis for MATHC1019, PILOT1 Data</td>
<td>159</td>
</tr>
</tbody>
</table>
Measuring student learning outcomes is essential for monitoring a school system’s success and for improving education quality. Student achievement information can be used to inform a wide variety of education policies and decisions, including ones related to the design and implementation of programs to improve teaching and learning in classrooms and the provision of appropriate support and training where it is most needed.

The *National Assessments of Educational Achievement* series of publications, of which this is the fourth volume, focuses on state-of-the-art procedures that need to be followed to ensure that the data (such as test scores and background information) produced by a national large-scale assessment exercise are of high technical quality and address the concerns of policy makers, decision makers, and other stakeholders in the education system.

Volume 1 in the series describes the key purposes and features of national assessments of educational achievement and is mainly aimed at policy makers and decision makers. Volume 2 addresses the design of two types of data collection instruments for national assessment exercises: student achievement tests and background questionnaires. Volume 3 focuses on the practical tasks involved in implementing a large-scale assessment exercise, including detailed step-by-step instructions on logistics, sampling, and data cleaning and management.
This fourth volume, *Analyzing Data from a National Assessment of Educational Achievement*, deals with how to generate information on test items and test scores and how to relate the test scores to educational and social factors. Like volumes 2 and 3, the volume is intended primarily for teams in developing and emerging economies with responsibility for conducting national assessments.

Finally, volume 5 describes how to write reports that are based on the national assessment findings and how to use the results to improve the quality of educational policy and decision making. It is of particular relevance to those with responsibility for preparing assessment reports and for communicating and using the findings.

As readers work through this fourth volume, the complexities and potential of the analysis of data generated by a national large-scale assessment will become evident. To fully explore what these data say about quality, equity, and other aspects of achievement in an education system, the analyst must use a variety of techniques described in part I of this volume. Part II describes a key analytical technique, Item Response Theory (IRT). The volume comes with specially designed user-friendly IRT software called Item and Test Analysis (IATA). Assessment teams everywhere, whether they are learning about IRT or are familiar with it, should find IATA to be a very useful addition to their collection of data analytical tools.

Marguerite Clarke
Senior Education Specialist/Learning Assessment Coordinator
August 2014
AUTHORS

Fernando Cartwright is a psychometrician, social science researcher, and software developer/architect. He has worked on numerous national and international skill and learning assessments, including the Programme for International Student Assessment (PISA) and the International Adult Literacy and Lifeskills Survey. He is the architect of several social measurement projects, including the Composite Learning Index, the European Lifelong Learning Index, and the Third Billion Index. He has produced software related to educational measurement, including Item and Test Analysis (IATA) and Web-based applications for test development, item banking, test delivery, and data analysis and evaluation. He resides in Ottawa.

Gerry Shiel is a research fellow at the Educational Research Centre at St. Patrick’s College, Dublin. He has led the development of a range of standardized tests of achievement in the areas of reading literacy (both English and Irish), mathematics, and science. He has worked extensively on national assessments at the primary level and has led the national implementation in Ireland of the Organisation for Economic Co-operation and Development Teaching and Learning International Survey (TALIS) and PISA. He has worked on assessment issues in Africa, South and East Asia, and Eastern Europe.
EDITORS

Vincent Greaney was lead education specialist at the World Bank. A former teacher, research fellow at the Educational Research Centre at St. Patrick’s College, Dublin, and visiting Fulbright Professor at Western Michigan University, he is a member of the International Reading Association’s Reading Hall of Fame. Areas of interest include national assessment, public examinations, teacher education, reading, and promotion of social cohesion through textbook reform. He has worked on educational projects mainly in Africa, South and East Asia, Eastern Europe, and the Middle East.

Thomas Kellaghan was director of the Educational Research Centre at St. Patrick’s College, Dublin, and is a fellow of the International Academy of Education. He has worked at the University of Ibadan in Nigeria and at Queen’s University, Belfast. Areas of research interest include national assessment, public examinations, educational disadvantage, teacher education, and home-school relationships. He served as president of the International Association for Educational Assessment from 1997 to 2001. He has worked on assessment issues in Africa, Eastern Europe, South and East Asia, Latin America, and the Middle East.
ACKNOWLEDGMENTS

A team led by Vincent Greaney (consultant, Education Global Practice, World Bank) and Thomas Kellaghan (Educational Research Centre, St. Patrick’s College, Dublin) prepared the series of books titled National Assessments of Educational Achievement, of which this is the fourth volume. Other contributors to the series are Sylvia Acana (Uganda National Examinations Board); Prue Anderson (Australian Council for Educational Research); Fernando Cartwright (Polymetrika, Canada); Jean Dumais (Statistics Canada); Chris Freeman (Australian Council for Educational Research); J. Heward Gough (Statistics Canada); Sara J. Howie (University of Pretoria); George Morgan (Australian Council for Educational Research); T. Scott Murray (Data Angel, Canada); Kate O’Malley (Australian Council for Educational Research); and Gerry Shiel (Educational Research Centre, St. Patrick’s College, Dublin).

The work was conducted under the general direction of Ruth Kagia, director of education; her successors, Elizabeth King and Amit Dar; and Robin Horn and Harry Patrinos, managers, all at the World Bank. Robert Prouty initiated the project and managed it up to August 2007. Marguerite Clarke has managed it since then through review and publication.

We are very grateful for the contributions of the review panel: Eugenio Gonzalez (Educational Testing Service), Pei-tseng Jenny Hsieh (University of Oxford), and Laura Jane Lewis (World Bank).
Diana Manevskaya (World Bank) facilitated the preparation of this volume. At the Educational Research Centre, St. Patrick’s College, Dublin, support was received from Hilary Walshe, as well as from Peter Archer, John Coyle, and Mary Rohan. Copy editing was carried out by Laura Glassman, Mary-Ann Moalli, and Linda Stringer of Publications Professionals LLC. Book design, editing, and production were coordinated by Janice Tuten and Paola Scalabrin of the World Bank’s Publishing and Knowledge Division; printing was coordinated by Andrés Meneses.

The Australian Council for Educational Research, the Bank Netherlands Partnership Program, the Educational Research Centre, the Irish Educational Trust Fund, Statistics Canada, and the Russia Education Aid for Development (READ) Trust Fund have generously supported preparation and publication of the series.
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CI</td>
<td>confidence interval</td>
</tr>
<tr>
<td>CTT</td>
<td>classical test theory</td>
</tr>
<tr>
<td>DIF</td>
<td>differential item functioning</td>
</tr>
<tr>
<td>HLM</td>
<td>hierarchical linear modeling</td>
</tr>
<tr>
<td>IATA</td>
<td>Item and Test Analysis</td>
</tr>
<tr>
<td>ICCC</td>
<td>item category characteristic curve</td>
</tr>
<tr>
<td>ID</td>
<td>identification</td>
</tr>
<tr>
<td>IQR</td>
<td>interquartile range</td>
</tr>
<tr>
<td>IRF</td>
<td>item response function</td>
</tr>
<tr>
<td>IRT</td>
<td>item response theory</td>
</tr>
<tr>
<td>JK</td>
<td>jackknife</td>
</tr>
<tr>
<td>NAEP</td>
<td>National Assessment of Educational Progress</td>
</tr>
<tr>
<td>PIRLS</td>
<td>Progress in International Reading Literacy Study</td>
</tr>
<tr>
<td>PISA</td>
<td>Programme for International Student Assessment</td>
</tr>
<tr>
<td>PSU</td>
<td>primary sampling unit</td>
</tr>
<tr>
<td>RP</td>
<td>response probability</td>
</tr>
<tr>
<td>SE</td>
<td>standard error</td>
</tr>
<tr>
<td>SPSS</td>
<td>Statistical Package for Social Sciences</td>
</tr>
<tr>
<td>TCC</td>
<td>test characteristic curve</td>
</tr>
<tr>
<td>TIMSS</td>
<td>Trends in International Mathematics and Science Study</td>
</tr>
</tbody>
</table>
Today’s global knowledge economy requires governments, education systems, and schools to closely monitor a variety of educational outcomes, including student achievements. A national assessment of student achievement in key curriculum areas contributes to this effort by addressing issues relating to

- **Quality**—providing information on student learning with reference to implementation of the curriculum, achievement of expected educational standards, or preparation for future learning
- **Equity**—determining if the education system is underserving particular groups of students as evidenced in differences in achievement related to gender, location, ethnic or language group membership, socioeconomic group, or school governance (public-private)
- **Provision**—identifying factors related to student learning (for example, school resources; curriculum implementation; teachers’ level of training, qualifications, and experience; and students’ home circumstances)
- **Change**—referring to change in educational outcomes over time (Greaney and Kellaghan 2008; Kellaghan and Greaney 2001; Kellaghan, Greaney, and Murray 2009).
Earlier volumes in this series, *National Assessments of Educational Achievement*, describe the components of a sample-based national assessment. These components include specification of the content of tests and questionnaires, definition of a population of interest and selection of a probability sample to represent the population, administration of assessment and other instruments to students and other respondents, scoring of student responses, and cleaning and management of data. The final set of data generated by these activities, in which test items have been created and assembled in a test booklet and response data have been collected, provides the source for the analyses described in this volume.

Part I of the volume is designed to assist national assessment teams in performing analyses of data normally carried out in a national assessment. Chapter 1 provides an overview of the data sets used in the worked examples on the CD accompanying the volume. It is followed in chapter 2 by an exploratory analysis of the data using SPSS. Concepts such as mean, median, mode, and standard deviation are defined, and a number of illustrative analyses are run. Chapter 3 introduces the concept of standard error of the estimate and describes procedures to estimate the extent to which data from a sample may be expected to differ from population data. How WesVar computes standard errors for a complex sample, an important feature of a well-designed national assessment, is described. Chapter 4 describes ways to address issues related to equity by analyzing differences between the mean scores of categories of students to determine if an obtained difference is statistically significant. In chapter 5, the focus shifts to ways in which the performance of high and low achievers can be described. Chapter 6 deals with associations between variables (for example, relationships between school resources and student learning), as evidenced in correlation, and provides an introduction to regression analysis. Chapter 7 contains examples of how data can be presented using charts and graphs.

Part II of the volume focuses on the development of scales to describe student learning. Two popular statistical frameworks (within which several models have been formulated) are used to address this issue. The first framework, classical test theory (CTT) (see Crocker and Algina 2006; Haladyna 2004; Lord and Novick 1968), has been in
use for most of the 20th century and was used to describe test development in volume 2 of the present series (Anderson and Morgan 2008). The second framework, which is described in part II, is item response theory (IRT) (see De Ayala 2009; De Mars 2010; Hambleton, Swaminathan, and Rogers 1991; Lord and Novick 1968). It had its origins in the middle of the 20th century and is widely used today in national and international assessments of student achievement.

The Item and Test Analysis (IATA) software described in this volume uses IRT to analyze test data. It was designed to offer a user-friendly way to address two major statistical considerations related to national assessments: (a) to increase the usability and interpretability of test scores and (b) to establish meaningful and consistent scales on which to report scores. The latter requires reduction in the error of measurement and the provision of information that can be generalized beyond the sample from which data were obtained. The sequence of analyses in part II is designed to mimic the phases of development and implementation of a national assessment program, from pilot testing to full-scale testing and follow-up testing in subsequent assessment cycles. Chapter 8 provides a description of the main IATA menu, its interactive elements, and the results it produces. Chapter 9 describes the steps in analyzing data from a pilot test administration, following which, in chapter 10, the steps in an analysis of a final test data administration are described. Analyses of rotated test booklets (chapter 11) and of partial credit items (chapter 12) are described. Comparing assessments through linking and specialized methods in IATA are covered in chapters 13 and 14, respectively. The volume concludes with an annex on IRT. Note that IATA runs only on Windows.

The main advantages of IRT are that, in contrast to CTT, it produces item statistics that are independent of the ability distribution of a set of examinees and parameters that characterize an examinee that are independent of the particular set of test items from which they are calibrated. Its advantages are considered to be particularly appropriate in situations that require test equating, identification of item bias, and design of computerized adaptive testing.

A disadvantage of IRT is that it requires advanced analytic skills and complex computing procedures, which may not be available to
a national assessment team. Many national assessments in developing countries continue to base their test development on the item facility and discrimination indices of CTT. It should be acknowledged that these forms of data provide test developers with useful information regardless of the measurement model that is applied at later stages of the test development process. Furthermore, CTT and IRT produce very similar results in terms of the comparability of item and person statistics as well as of the degree of invariance of item statistics across examinee samples (Fan 1998).

Whether one’s choice for test development is CTT or IRT, two issues related to current practice in national and international assessments, which follows practice in the development of tests designed to assess the achievements of individual students, merit attention: (a) the assumption that a single dimension underlies the trait or ability being assessed and (b) the focus on maximizing differences between the achievements of examinees. Both have implications for test validity.

The assumption of unidimensionality underlying test development has important implications in a national or international assessment, not just for the content validity of tests, but also for determining item bias and test linking. However, the assumption is challenged by evidence that students vary at the rate in which they acquire competence in different areas of achievement (illustrated, for example, when mathematics achievement is described in terms of number, measurement, shape, and data). This variation is due, most likely, to differences in students’ educational and broader cultural experiences (Goldstein and Wood 1989). Rejecting items for inclusion in a test because the statistical data do not support the assumption of unidimensionality may have the effect of excluding important content, thereby resulting in inadequate representation of a construct, which, of course, would affect the content validity of a test—an aspect of validity that is generally regarded as more important than inferences based on statistical data.¹ The assumption of unidimensionality must be of particular concern in international assessments, in which students’ experiences, in school and out of school, are known to vary widely.

The goal of maximizing differences between examinees, another feature of procedures designed to develop tests to assess the
achievements of individual students, is of concern in a national (or international) assessment because the purpose of such an assessment is to describe the achievements of the education system, not to differentiate between the achievements of individual students. The implication of this situation is that factors other than discrimination and facility have to be considered when deciding whether to include items in a test. For example, items that all students answered correctly or items that no students answered correctly might not normally be included in a test designed for individual students because they would not contribute to differentiating between the students. However, in the case of a national assessment, it might be important to know that all or no students had mastered certain areas of achievement. Hence, items representing those areas would be included in the assessment. To ensure that tests used in a national assessment adequately represent the construct being assessed and provide comprehensive information on the range of achievements acquired by students in the education system, it is imperative that test developers liaise with curriculum specialists and teachers in an iterative fashion throughout the process of test development.

The general introduction to statistical analysis precedes the section on IRT in this volume because it introduces the reader to many of the analytic procedures used in IRT. However, in the real-life situation of a national assessment, the scaling of data to describe student achievement, as described in part II, would have to be completed before carrying out the analyses in part I.

Users of this volume are assumed to have a basic knowledge of using folders and files, Excel, and SPSS and to have the ability to navigate, without difficulty, between components of SPSS.

**NOTE**

1. Cronbach (1970, 457) has pointed out that, even in the case of tests developed to assess individual students, “nothing in the logic of content validation requires that the universe or the test be homogenous in content.”
PART I
AN INTRODUCTION TO THE STATISTICAL ANALYSIS OF NATIONAL ASSESSMENT DATA

Gerry Shiel
National assessment data always contain a measure of student achievement, which can be represented in a variety of ways, such as number of items on a test that the student has answered correctly (though this measure is not always very meaningful); percentage of items correctly answered; and scaled scores in which a distribution of scores with an obtained mean and standard deviation is transformed into a distribution with a different mean and standard deviation. Most national assessments also collect additional data. These data may relate to schools (such as type, size); teachers (such as qualifications, experience); students (such as age, time spent doing homework); and parents and home environment (such as educational level of parents, number of books in the home).

The data collected will contain a range of variable types. Some variables will be *categorical* and involve placing individuals into clearly defined categories or groups, such as educational level or gender. Other variables, described as *discrete*, consist of numerical measurements or counts, such as the number of children in a family. They are obtained by counting and have values for which there are no in-between values. *Continuous* variables, in contrast, describe numerical measurements that can be any value between two specified values,
such as the distance from a student’s home to school. The type of data imposes constraints on the kind of statistical analysis that can be carried out, as well as on the way in which data can be represented graphically.

Analyses will usually begin with an exploration of simple numerical data, presented in summary statistics, in graphs or plots, or both ways. The focus at this stage, as elaborated later in this chapter, is on description, though what is learned may generate hypotheses to be tested at a later stage. The exploratory stage of the data analysis also provides the opportunity to inspect the quality of the data by checking for missing values, outliers, gaps, and erroneous values, though these should have been identified at the data-cleaning stage (see Freeman and O’Malley 2012). It also reveals the nature of the data, indicating if the distribution is symmetric, skewed, or clustered. At this early stage, a graphical display, in the form of a bar graph, histogram, or box and whiskers plot, can be very informative in identifying patterns in the data.

When more than one observation is available on individuals, one can investigate relationships between variables, such as the relationship between students’ literacy and numeracy achievements or between mathematics achievement and home background factors. An association between a pair of variables is termed bivariate. Because many of the variables in a national assessment will be interrelated, it is necessary to carry out multivariate analyses that involve procedures to predict performance on one variable (for example, reading achievements) from the values of a set of other variables (for example, student gender, home background factors). A first step in multivariate analysis is to display and examine pair-wise correlations between variables in a correlation matrix. This volume contains an introduction to multivariate analysis (regression analysis; see chapter 6). However, it does not deal with more complex forms of analysis, such as multilevel modeling, in which analyses are designed to reflect the structure found in education systems (students grouped in classes, classes in schools, schools in regions).

Readers can develop their analytical skills by working through a set of exercises using the database on the CD that accompanies this book. The database, which is similar to that used in the section
on sampling in *Implementing a National Assessment of Educational Achievement* (Dumais and Gough 2012a), contains achievement test and other data that were modified from data collected in an actual assessment of mathematics achievement carried out in grade 4 in a small country and are presented in this series as coming from a country with the pseudonym Sentz.

The following chapters describe a series of analytical tasks that are typically carried out with data obtained in a national assessment. By carrying out these analyses, readers should become familiar with a set of statistical techniques that they can apply to their own data. Most of the analyses use WesVar software. Unlike many other software packages, WesVar takes into account the complexity of the design of the national assessment when conducting statistical analyses, such as estimating variance and sampling error. Parts II and IV of volume 3 of this series, *Implementing a National Assessment of Educational Achievement*, describe complex sampling in detail (Dumais and Gough 2012a, 2012b).

**SAVING THE CD FILES TO YOUR HARD DRIVE OR SERVER**

Files can be saved from the CD to a hard drive or server. Copy or create a folder called *NAEA DATA ANALYSIS* from the CD to your desktop. You should have seven subfolders within the *NAEA DATA ANALYSIS* folder: *SPSS DATA*, *EXERCISE SOLUTIONS*, *WESVAR UNLABELED DATA*, *MY WESVAR FILES*, *WESVAR DATA & WORKBOOKS*, *MY SPSS DATA*, and *MY SOLUTIONS*. To copy the folder *NAEA DATA ANALYSIS* from the CD to your desktop, locate the folder on the CD, and right-click Copy. Then open Desktop, and right-click Paste. Check that the folder *NAEA DATA ANALYSIS* has been copied successfully. Details of the seven subfolders follow.

- **SPSS DATA.** The SPSS data files (*NATASSESS.SAV* and *NATASSESS4.SAV*) used to complete the exercises in chapter 2 of this volume, as well a file on schools (*SCHOOLS.SAV*), can be found in this folder.
• **EXERCISE SOLUTIONS.** Here you can find solutions, mainly in the form of text files, for the exercises in chapters 2 to 7 of this volume. When you complete the exercises, you can check the solutions you obtain against those in this subfolder.

• **WESVAR UNLABELED DATA.** Use this source for the WesVar exercises in chapter 3. This data file (*NATASSESS4.VAR*) should be the same as the data file obtained when you create your own WesVar data file using the steps outlined in annex I.C. The data file in this directory can serve as a backup.

• **MY WESVAR FILES.** Use this subfolder to save WesVar data files and workbooks you create when completing the exercises in chapters 3 to 6. When you open this folder for the first time, you will find that it is empty. This is because you have not yet saved any files to the folder. It is strongly recommended that you create your own WesVar data files and workbooks using the procedures outlined later and in annex I.C. Note that all the WesVar data files and workbooks that you create should be saved in **MY WESVAR FILES**.

• **WESVAR DATA & WORKBOOKS.** This subfolder contains the data file *NATASSESS4.VAR*, its associated log file *NATASSESS4.LOG*, and four completed workbooks, *CHAPTER3 WORKBOOK.WVB*, *CHAPTER4 WORKBOOK.WVB*, *CHAPTER5 WORKBOOK.WVB*, and *CHAPTER6 WORKBOOK.WVB*. You can refer to these sources to check the accuracy of your work in WesVar.

• **MY SPSS DATA.** Use this folder to save new or modified SPSS data files, such as those you create before transferring an SPSS data file to WesVar (see annex I.C).

• **MY SOLUTIONS.** Save your solutions to the exercises in chapters 2 to 7 in this subfolder. As in the cases of **MY WESVAR FILES** and **MY SPSS DATA**, it will be empty when opened for the first time.

Annex I.B contains details of the contents of each folder and file.

**SURVEY INSTRUMENTS**

This section describes the main instruments used to gather the data used in the database.
Test of Mathematics Achievement

The test consisted of 125 items based on the national curriculum framework for grade 4. Table 1.1 shows the distribution of items across the main mathematics content areas and cognitive processes (or intellectual behaviors). Most items assessed the content areas of number and measures, reflecting the weights assigned to these areas in the national curriculum and in textbooks. Over half the items assessed two cognitive processes: “implement procedures” (28 percent) and “apply and problem solve” (32 percent).

Items were clustered into five blocks (A, B, C, D, E), each consisting of 25 items. Each test booklet contained 75 items from the total pool of 125. Each block (except common block B) appeared once in the initial position and once in the final position in a test booklet.

Each item used either a multiple-choice or short-answer format. Multiple-choice items had four possible responses (A, B, C, D). Students were expected to mark the response they believed to be correct. For short-answer items, students were expected to write answers to the questions or make drawings (for example, draw the lines of symmetry through a two-dimensional shape such as a rectangle). Each multiple-choice question had one correct response option. Each short-answer item was scored either right or wrong according to a scoring rubric provided to item scorers.

<table>
<thead>
<tr>
<th>Table 1.1</th>
<th>Mathematics Test: Distribution of Items by Content Area and Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content areas</td>
<td>Number of items</td>
</tr>
<tr>
<td>Number</td>
<td>46</td>
</tr>
<tr>
<td>Algebra</td>
<td>6</td>
</tr>
<tr>
<td>Shape and space</td>
<td>18</td>
</tr>
<tr>
<td>Measures</td>
<td>44</td>
</tr>
<tr>
<td>Data and chance</td>
<td>11</td>
</tr>
<tr>
<td>Total</td>
<td>125</td>
</tr>
</tbody>
</table>
Context Questionnaires

The national assessment included separate principal, teacher, student, and parental questionnaires (table 1.2). Teachers also completed a rating form for each student in the assessment.

**SAMPLING WEIGHTS**

Sampling weights were computed and included in the file. The weights reflect the probability of selection for each student. How these weights are computed and how they are used are described in volume 3 of this series (Dumais and Gough 2012b). For each student, a design weight that included the following components was computed:

- **School selection component.** Schools were selected with probability proportional to size. For school $i$, in stratum $h$, this was the reciprocal of the product of the number of schools selected

<table>
<thead>
<tr>
<th>TABLE 1.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abbreviated Questionnaire Descriptions</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Questionnaire</th>
<th>Completed by</th>
<th>Topics covered included</th>
</tr>
</thead>
<tbody>
<tr>
<td>School questionnaire</td>
<td>Principals of schools</td>
<td>School size, school resources, teaching staff, school development planning, qualifications of school principal</td>
</tr>
<tr>
<td>Teacher questionnaire</td>
<td>Teachers of participating students in grade 4</td>
<td>Teacher qualifications, years of teaching experience, distance traveled to school every day, class size, time spent teaching mathematics, frequency of assessing students’ progress, availability and use of educational resources in the classroom</td>
</tr>
<tr>
<td>Student questionnaire</td>
<td>Students</td>
<td>Age, frequency of doing homework, interest in mathematics</td>
</tr>
<tr>
<td>Parent questionnaire</td>
<td>Parents of participating students</td>
<td>Educational attainment (own and that of spouse or partner), number of books in the home, size of holding (land), availability of electric light at home, parental support and encouragement</td>
</tr>
<tr>
<td>Student rating form</td>
<td>Teachers of participating students in grade 4 with respect to each student</td>
<td>Student attendance at school, student proficiency in the language of instruction, teacher rating of student achievement, behavior, parental support</td>
</tr>
</tbody>
</table>
multiplied by the number of students in the target grade level in the school (measure of size), divided by the number of students in the stratum in the population. For example, if 5,000 students were in the stratum, and 10 schools in the stratum were selected, with 50 students in school \(i\), the school selection component for school \(i\) (Schwgt) would be \(5,000/(10*50) = 10\).

- **School nonresponse correction component.** Because all selected schools participated, the school nonresponse adjustment factor was set at 1.0 (Schnrfac).\(^1\)

- **Student selection component.** Because all students in grade 4 in a school were selected, the probability of a student in a selected school being tested was 1.0, and its reciprocal was also 1.0 (Studfac).\(^2\)

- **Student nonresponse within-school adjustment component.** An adjustment correction was made for student nonresponse within schools. This was the inverse of the number of valid test booklets returned for students in the school over the number of students in grade 4 in the school minus the exempted students (Stunrfac). For example, if 90 students were enrolled in grade 4 at the time of the study, none of whom was entitled to an exemption, and 80 took part, the adjustment factor would be 90/80.

The weight for each student was obtained by calculating the product of these four components (Schwgt × Schnrfac × Studfac × Stunrfac). Using the preceding examples, for student \(x\) in school \(i\), the weight would have been \(10 \times 1 \times 1 \times 90/80\). This product yields the design weight (Wgtpop in the data file). When data in the data file are weighted using the design (population) weight, the estimated population size is 51,713 (the projected number of students in grade 4 in the population). Each student in the sample represented on average \(51,713/4,747 = 10.89\) students.

In performing analyses on the SPSS national assessment data file accompanying this workbook, it is recommended that you apply the population weight (Wgtpop). This weights the data to ensure proportionate representation of each stratum.

The computation of survey weights is described in chapter 14 of volume 3 in the series, *Implementing a National Assessment of Educational Achievement* (Dumais and Gough 2012b). The steps
outlined in the chapter will automatically generate the weights needed to analyze national assessment data. In the analyses in this volume, from chapter 2 onward, \texttt{Wgtpop} is used to weight the data.

\textbf{SPSS}

Some of the data files on the CD that accompanies this book (such as \texttt{NATASSESS.SAV}) are in SPSS format. The particular version of SPSS used to analyze the data in this file was SPSS, Version 18; the data files have also been analyzed using more recent versions of SPSS. For the purpose of the exercises presented in this volume, all of the data files (the assessment data and the files based on each of the questionnaires) were merged into a single SPSS file consisting of student achievement and other data on 4,747 cases.

For the purpose of analysis, school- and teacher-level variables were disaggregated to the student level. In other words, each student was assigned values for these variables corresponding to values assigned to his or her school and teacher. For example, one of the variables in the teacher questionnaire was the number of minutes allocated to mathematics teaching each week. When this variable was disaggregated, each student in a class was assigned the same number of minutes of instruction per week provided by his or her teacher. Data for several hundred variables were collected in the course of the national assessment. However, the data file \texttt{NATASSESS.SAV} is limited to a subset of these variables to keep the size and structure of the file at a manageable level.

\textbf{Opening an SPSS Data File}

There are two ways to open a data file. One is to go into (My) Computer on the Windows (Start) menu or on your desktop and to click on the drive and directory in which your SPSS data file is saved: for example, \texttt{NAEA DATA ANALYSIS – SPSS DATA – NATASSESS.SAV}.

You can also open SPSS by clicking on Start, All Programs – (IBM) SPSS Statistics. Click the specific version of SPSS that appears on your screen. Once launched, locate the required SPSS data
file by selecting File – Open – Data, and then finding **NAEA DATA ANALYSIS – SPSS DATA – NATASSESS.SAV**. Double-click on **NATASSESS.SAV** to open it.

**Using the Toolbar to Conduct Preliminary Analyses**

You can do analyses in SPSS in two main ways, using syntax files or using the toolbar. Here, the toolbar is used. This can be found at the top of the open SPSS data file. You simply click on the procedures you want to run, as shown in exercise 1.1.

---

### EXERCISE 1.1

**Running Descriptive Statistics in SPSS and Saving the Files**

1. Open the SPSS data file **NATASSESS.SAV**, which can be found in **NAEA DATA ANALYSIS\SPSS DATA**.

2. Check that the weights are turned on: Data – Weight Cases – Weight Cases by – Wgtpop (exercise figure 1.1.A) and click OK. You use Wgtpop to ensure that the computed statistics represent the population. You should see the message Weight On in the bottom right of the screen.

---

**EXERCISE FIGURE 1.1.A  Weight Cases Dialog Box**

---

(continued)
3. Select **Analyze – Descriptive Statistics – Descriptives**.

4. In the Descriptives dialog box, highlight the required variable in the left panel (in this case, **Mathss**, the scale scores on the mathematics achievement test). Click on the arrow to move it to **Variable(s)** (see exercise figure 1.1.B). Click **OK**. Your output should show a table with a weighted mean of 249.99 (which rounds to 250) and a standard deviation of 49.99780 (which rounds to 50).

5. Use **File – Save As** to assign a suitable name to your SPSS output file (for example, **EXERCISE 1.1.SPV**), and save to **NAEA DATA ANALYSIS\MY SOLUTIONS**. Then select **File – Close**.

6. To save your SPSS data file, which should be in the data editor mode, select **File – Save As .... Save to NAEA DATA ANALYSIS\MY SPSS DATA** using the filename **NATASSESS.SAV**. Then select **File – Exit**.

a. If you see text rather than data on the screen, change from the viewer mode to the data editor mode by clicking on **Window** and on (IBM) **SPSS Statistics Data Editor**.

b. When you open a dialog box, you may find the list of variable labels (labels assigned to each variable name) rather than variable names. Similarly, you may find that the variables are in alphabetical order rather than in the order in which they appear in the data file. To adjust these settings, close the dialog box, and click **Edit – Options – General**. Then select the options you want in the variables list box.

c. To reduce the number of decimal places to one, double-click and highlight the digits in the relevant cell (such as 249.99) in the table. Right-click **Cell Properties – Format Value – Number – Decimals – 1**.
WESVAR

WesVar is a statistical package that is often used in conjunction with SPSS to analyze national assessment data. In addition to presenting some preliminary exercises using SPSS, chapter 2 describes the rationale for using WesVar, while chapters 3 to 6 describe a variety of analyses using WesVar. The WesVar software (which includes an extensive help menu) may be downloaded from the Westat Web site.5

NOTES

1. If 20 schools were in a stratum, and 18 participated, the appropriate correction factor would have been 20/18 or 1.11.

2. If there had been five classes of grade 4 students in the school, and three had been selected to take part, the student selection component would have been 5/3. Alternatively, if there had been 100 grade 4 students, and 35 had been selected at random to participate, the component would have been 100/35.

3. The suffix .SAV is used when you save an SPSS data file, whereas .SPV is used when you save an SPSS output (results) file.

4. Chapter 12 in volume 3 of this series, Implementing a National Assessment of Educational Achievement, contains details on how to merge files using Access (Freeman and O’Malley 2012). Files can also be merged in SPSS, using Data and Merge Files on the toolbar (see annex I.C).

This chapter explores a national assessment data set using SPSS. Exercises are designed to enable the analyst to understand and compute data such as the overall mean score, the mean scores of constituent groups (such as regions), and the variability of group test scores. The analyses described in the chapter are based on weighted data.

The notion of a distribution of scores is a central concept in the chapter. A distribution is a group of scores from a sample on a single variable, such as scores on a test of achievement. For example, if a test of mathematics with a maximum score of 10 points is administered to a sample of 20 students, the following distribution of scores may result: 0, 2, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 5, 5, 6, 6, 6, 7, 10. In a national assessment, where hundreds or even thousands of students may take the same test, the number of scores will of course be much larger. This chapter describes measures of (a) central tendency in scores, (b) spread in scores, (c) the position of scores, and (d) the shape of distributions. Examples are based on a weighted distribution of mathematics test scores for the 4,747 cases for which national assessment test data are available.
MEASURES OF CENTRAL TENDENCY

The most common summary measures representing the typical or central value of a set of test scores are the mean, the median, and the mode.

To calculate the (unweighted) mean of a data set, such as the mathematics achievement test scores of students, sum each of the values. Then divide the sum by the number of data points that contributed to the sum (the number of students who took the test).

The median is the midpoint in a set of numbers arranged in order of magnitude.

The mode is the most frequently occurring value in a data set. A distribution with two modes is called bimodal.

The following is a set of nine test scores for students who took a history test: 45, 52, 55, 55, 59, 60, 70, 71, and 73. The mean score is 60, the median score is 59, and the modal score is 55.

MEASURES OF SPREAD

Dispersion is a central concept in statistics. The most commonly used statistical measures of dispersion are variance, standard deviation, and range.

The variance is a measure of how much test scores vary or are dispersed. To compute the variance of a set of scores, the distance (called a deviation) between each score and the mean score is calculated. The deviations are squared and summed and then divided by the number of cases less one. Thus, the variance is the average squared difference between each of the points in the distribution and the mean.

A related statistic, the standard deviation, is the square root of the variance.

Other less widely used measures of the spread of scores are available. The range of scores in a distribution is the difference between the highest and lowest scores. If the lowest score is 30 and the highest is 70, the range is 40. The interquartile range (IQR) is the difference between scores at the 25th (quartile 1) and 75th (quartile 3)
percentiles in a distribution. (Percentile is described in the following section.) The IQR is useful as a benchmark in identifying outliers (such as values that are more than 1.5 IQRs below the value at quartile 1 or above the value at quartile 3).

**MEASURES OF POSITION**

The relative position of a specific member of a set, such as a student’s score compared to the scores of others who took a test, can be identified in a number of ways. One is the percentile ranking of a specific score or value. This is the percentage of scores or values that fall below a particular score. For example, a score with a percentile rank of 62 in a national assessment means that 62 percent of students had a lower score. To calculate a percentile rank, sort test scores from the lowest to the highest, following which the percentage of scores that are lower than a specified score is calculated. Some national and international assessments report test scores along with their standard errors (see chapter 3) for specific percentiles such as 10th, 25th, 50th, 75th, and 90th. The percentile rank is easy to understand, but meaningful statistical analysis is limited because the interval property of the measurement system is destroyed in the transformation of scores into percentiles.

A score or value can be denoted in terms of the number of standard deviations by which it deviates from the mean. In a normal distribution, approximately 68 percent of scores fall within one standard deviation of the mean, 95 percent within two standard deviations, and almost 100 percent within three standard deviations. Figure 2.1 illustrates this graphically.

Consider, for example, a normal distribution of scores with a mean of 250 and a standard deviation of 50. Because the scores are normally distributed, roughly 34 percent of the students scored between 250 and 300, and a further 34 percent scored between 200 and 250. A score of 325 would be 1.5 standard deviations (75 points) above the mean, while a score of 125 would be 2.5 standard deviations (125 points) below the mean.
When examining a distribution of test scores, one must consider the shape of the data, that is, whether the distribution is clustered (bunched together) in one direction or the other, because a significant departure from normality may violate assumptions for some statistical techniques. In a positively skewed distribution, most of the scores are clustered at the lower end, with a few scores spread out toward the upper end (figure 2.2). This may occur when a test is particularly difficult and most students achieve low scores. Some national assessments encounter this problem if the test is too difficult for the population. In a positively skewed distribution, the mean is typically greater than the median.
In a negatively skewed distribution, most of the scores are clustered at the upper end, with fewer scores spread out toward the lower end (figure 2.2). This may occur when a test is particularly easy and many students achieve high scores. In a negatively skewed distribution, the mean is typically smaller than the median.

In a symmetric distribution (figure 2.2), the mean, the median, and the mode are close to one another, near the center.

Kurtosis is a measure of the “peakedness” of scores around the mean. It indicates whether the graph of the distribution of scores is more peaked or flat than in a normal distribution. A data set with high kurtosis (leptokurtic) tends to have a pronounced peak at the mean while a data set with low kurtosis (platykurtic) tends to have a relatively flat peak over the mean compared to the normal distribution. In a normal distribution, the value of the kurtosis statistic is at, or close to, zero.¹
EXPLORING A DATA SET USING SPSS

The SPSS command Explore provides a range of statistics and associated graphs that are very useful for exploring a distribution of scores. In addition to descriptive statistics such as mean, median, mode, and standard deviation, Explore provides measures of skewness and kurtosis, stem-and-leaf plots, histograms, box plots, and normal plots. Use Explore to analyze all the cases in a distribution of national assessment results or to focus on subgroups, such as male and female students or students attending schools in different regions.

In this next exercise, Explore is run using the same data set used in Implementing a National Assessment of Educational Achievement (Greaney and Kellaghan 2012) (NATASSESS4.SAV). The focus is on the results for one variable, Mathss (mathematics scale score). Exercise 2.1 presents a number of alternative approaches to obtaining descriptive statistics. Before starting the exercise, go to the SPSS toolbar. Click Edit – Options – General, and check that Display Names has been selected under Variable Lists. Click OK.

One can also use Explore to carry out an initial analysis to compare levels of single variables such as Gender or Region. Exercise 2.2 describes how to view summary statistics for the four regions for which data were obtained in the national assessment.

EXERCISE 2.1

Running Explore in SPSS, Single Dependent Variable (One Level)

1. Open the data file NAEA DATA ANALYSIS\MY SPSS DATA\NATASSESS.SAV. (Note you saved the data from exercise 1.1 into this subfolder.)

2. Check that the weights have been turned on: Click Data – Weight Cases – Weight Cases by – Wgtpop – OK.


4. Move Mathss (mathematics scale score) into the Dependent List. Move Studid into the Label Cases by box.¹

5. Confirm that Both has been checked under Display (this will ensure that both plots and statistics are displayed in your output). Click Statistics (top right corner).
EXERCISE 2.1 (continued)

Make sure that a check mark appears before **Descriptives**. Click **Continue – Plots**. Ensure that **Stem-and-Leaf** is checked. Click **Continue – OK**.

6. Click **Window** on the **toolbar**, and click **Output1**. Save the output to **NAEA DATA ANALYSIS\MY SOLUTIONS\EXERCISE 2.1.SPV**.

Exercise table 2.1.A provides a case-processing summary. Because the data were weighted to the population size, the 4,747 cases in the database represent a population of 51,713. There are no missing cases (Valid percent: 100).

**EXERCISE TABLE 2.1.A**  Case-Processing Summary

<table>
<thead>
<tr>
<th>Variable</th>
<th>Cases</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Valid</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>Percent</td>
<td></td>
</tr>
<tr>
<td>Mathss</td>
<td>51,713</td>
<td>100.0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>Percent</td>
<td></td>
</tr>
<tr>
<td>Mathss</td>
<td>51,713</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

The following explanations describe how to interpret the statistics in exercise table 2.1.B:

- **Mean score** (250.0) is the arithmetic weighted average. The standard error of the mean is 0.22. (See chapter 4 for a description of standard errors.)

- **95% confidence interval for the mean** is the estimated approximate range of values that has a 95 percent probability of including the unknown mean of the national

**EXERCISE TABLE 2.1.B**  Descriptive Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Statistic</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mathss</td>
<td>Mean</td>
<td>250.0</td>
<td>0.22</td>
</tr>
<tr>
<td></td>
<td>95% confidence interval for mean</td>
<td>Lower bound</td>
<td>249.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Upper bound</td>
<td>250.4</td>
</tr>
<tr>
<td></td>
<td>5% trimmed mean</td>
<td>251.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>256.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>2,499.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Standard deviation</td>
<td>50.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>88.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>400.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>311.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Interquartile range</td>
<td>67.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td>−0.380</td>
<td>0.011</td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td>−0.101</td>
<td>0.022</td>
</tr>
</tbody>
</table>

(continued)
• **5% trimmed mean** is the arithmetic mean calculated when the top 5 percent and bottom 5 percent of scores (cases) have been dropped. It provides a better measure of central tendency if the data are nonsymmetrical. The 5 percent trimmed mean is 251.1.

• **Median** is the value below which 50 percent of cases fall. It is also the 50th percentile. The median has been calculated as 256.3.

• **Variance** is a measure of the extent to which the distribution of test scores is dispersed or spread out. The variance is 2,499.8.

• **Standard deviation** (equal to the square root of the variance) is 50.0.

• **Minimum and maximum** are the smallest (88.4) and largest (400.0) values in the distribution.

• **Range** of scores is the difference (311.6) between the largest and smallest score values in the distribution.

• **Interquartile range** is the distance between the third quartile (75th percentile) and first quartile (25th percentile) values and provides a measure of the spread of the data. The IQR for the scores is 67.1.

• **Skewness** provides a measure of the asymmetry of a distribution. A normal distribution is symmetric and has a skewness value of around zero. The skewness is slightly negative (−0.38). A skewness value between −1 and +1 is considered very good for most psychometric uses, but a value between −2 and +2 is usually acceptable.

• **Kurtosis** is a measure of the extent to which observations cluster around a central point (the “peakedness” of the probability distribution). For a normal distribution, the value of the kurtosis statistic is at, or close to, zero. Excessive positive kurtosis indicates that the observations (scores) cluster more and have flatter tails (a leptokurtic distribution) than those in a normal distribution, whereas excess negative kurtosis indicates that observations cluster less and have higher tails (a platykurtic distribution). As with skewness, a kurtosis value between −1 and +1 is considered very good, but a value between −2 and +2 is also usually acceptable. Our obtained value of −0.101 is well within both limits.

The analysis produces a stem-and-leaf plot (exercise figure 2.1.A), which displays the relative density and shape of the data. It is a method of presenting score frequencies. Each observed value (sorted in ascending order) is divided into two components: leading digits (stem) and trailing digits (leaf). The stem represents the tens digits (or higher) of a score, and the leaf contains the last digits. Stem 15, for example, shows that 821 (weighted) students achieved scores between 150 and 159 (inclusive). The data also
EXPLORING NATIONAL ASSESSMENT DATA USING SPSS

indicate that values less than or equal to 117, and equal to or greater than 386, are deemed to be “extreme” for reasons explained later.

Close the SPSS data file NATASSESS.SAV by selecting File – Exit on the toolbar. There is a file for this exercise in the Exercise Solutions folder.

The Explore command also produces a box plot (or box-and-whiskers diagram) (exercise figure 2.1.B). This is a graphic representation of the distribution of test scores that includes the median (50th percentile) and the 25th and 75th percentiles. The distance between the top and the bottom of the box (between the 25th and 75th percentiles) is the IQR, or the distance between the highest and lowest scores in the middle 50 percent of scores in the distribution.

The box plot also shows outliers and extreme values. The whiskers (the lines that extend from the top and bottom of the box) represent the highest and lowest values that are not outliers or extreme values. Outliers (values that are between 1.5 and 3 times the IQR) and extreme values (values that are more than 3 times the IQR) are represented by small circles beyond the whiskers. The listed numbers (such as 423410) are the student identification (Studid) numbers of students who recorded outlier or extreme scores.

The box plot (exercise figure 2.1.B) can provide useful information in a visual format. It portrays the following data characteristics:

- The median (the line through the center of the box) is the middle point in the distribution and, like the mean score, is a measure of central tendency.
• The height of the box (the IQR) shows the extent to which the test score values in the distribution vary.

• A median located in the lower half of the box suggests positive skewness, whereas one located in the upper half of the box would suggest negative skewness. In exercise figure 2.1.B, the median is toward the middle of the box, indicating relatively minor skewness.

Outlier scores, defined as scores between 1.5 and 3 box-lengths from the upper (75th percentile) and lower (25th percentile) box values, and extreme scores, defined as scores that are more than 3 box-lengths from these points, should be examined to ascertain if they are incorrect or are in fact valid scores.

a. You may also wish to run some optional statistics or generate some optional charts. By clicking Statistics after step 3, you can select students with extreme values by selecting Outliers (which lists the five highest and five lowest scores in the distribution). In a similar manner, by selecting Percentiles you can get scores at the 5th, 10th, 25th, 50th, 75th, 90th, and 95th percentiles, in addition to the default Descriptives. By clicking Plots, you can select Histogram in addition to the default Stem-and-Leaf plot. Both can be copied into a Word document.

b. Note that a copy of the output for exercise 2.1 is also available at NAEA Data Analysis\Exercise Solutions\Exercise2.1.SPV.

c. Note that all but the final estimates in table 2.1.B have been rounded—to one decimal place in the case of estimates and two places in the case of standard errors. This was done by highlighting the values in the table (except those for skewness and kurtosis), right clicking, selecting Cell – Properties – Format – Value – Number, and setting the number of decimal places to 1 (or 2).
Running Explore in SPSS, Single Dependent Variable (More Than One Level)

1. Open the SPSS data file NAEA DATA ANALYSIS\MY SPSS DATA\NATASSESS.SAV.


   Move Mathss to the Dependent List. Move Region to the Factor List. Move Studid into the Label Cases by box. Make sure Both has been checked under Display (this will ensure that both plots and statistics are displayed in your output). Click Statistics (top right corner). Make sure there is a check mark before Descriptives. Click Continue – Plots. Ensure that Stem-and-Leaf is checked. Click Continue – OK.

In your output file, scroll down to see descriptive statistics for each of the four regions. Note, for example, that the mean score for the Northwest region is 233.3. The corresponding data for the Metro Area, Eastern Highlands, and Southwest Coast are 265.7, 249.1, and 251.2, respectively.* Scroll down to the end to see the box plots for each of the four regions (exercise figure 2.2.A).

3. The box plot (exercise figure 2.2.A) shows the median scores for the four regions (Northwest, Metro Area, Eastern Highlands, Southwest Coast). The analyst will note

EXERCISE 2.2

EXERCISE FIGURE 2.2.A Box Plots for Mathematics Scale Scores by Region

(continued)
the relatively large number of “extreme” scores in the Metro Area, which is a function of the relatively high score at the 25th percentile in that region, compared to, for example, the Northeast.

4. Save the output to NAEA DATA ANALYSIS\MY SOLUTIONS\EXERCISE 2.2.SPV.
   Save the SPSS data file and exit SPSS: File – Save and File – Exit.
   a. The appropriate standard errors are computed in exercise 3.3.

NOTES

1. Skewness and kurtosis are not usually reported in national assessments. However, they can be of diagnostic value, identifying distribution shapes that may be problematic.

2. Unlike NATASSESS4.SAV, NATASSESS.SAV does not contain jackknife zones and jackknife replicates (indicators). See annex 1.C.
This chapter describes procedures for calculating population estimates such as mean scores and percentile ranks and their accompanying standard errors, using weighted national assessment data. The analyses in this and subsequent chapters are performed using the WesVar statistical package, which takes into account the complexity of a sample that is selected in multiple steps.

**SETTING UP A DATA FILE IN WESVAR**

Ensure that WesVar is installed on your computer. You can download it from the Westat Web site. You must modify the SPSS data file *NATASSESS.SAV* before saving it as a WesVar data file. To create a data file in WesVar, your SPSS data file should include variables such as the following (though the actual variable names are arbitrary):

- **Studid**: student identification number, a unique number assigned to each student
- **Wgtpop**: the population weight that weights the data to provide a good estimate of a population characteristic
- **Jkzone**: jackknife zone; each pair of schools is allocated to a different jackknife zone
• Jkindic: jackknife indicator; within each jackknife zone, one school is randomly assigned a value of 0 and the other a value of 1.

The SPSS file, NATASSESS4.SAV, which you can access at NAEA DATA ANALYSIS\SPSS DATA, contains the data on each of these key variables. To bring the SPSS data file containing test and questionnaire data into WesVar and create replicate weights, follow the directions in annex I.C of this volume and save your new WesVar data file, NATASSESS4.VAR, to NAEA DATA ANALYSIS\MY WESVAR FILES. Alternatively, for the purpose of carrying out the following exercises, you can use the ready-made but unlabeled WesVar data file NATASSESS4.VAR in the WESVAR UNLABELED DATA subfolder.

ADDING VARIABLE LABELS

Follow these directions below:

1. Launch WesVar – Open WesVar Data File – NAEA DATA ANALYSIS\WESVAR UNLABELED DATA and open NATASSESS4.VAR.
2. On the toolbar, select Format – Label.
3. Select Region in the Source Variables box (left side).
4. In the cell after 1, type Northwest in label column; in the cell after 2, type Metro_Area; in the cell after 3, type Eastern_Highlands; and in the cell after 4, type Southwest_Coast (figure 3.1).
5. Click OK. You will get this message: This operation will create a new VAR file. Click OK. Save as NATASSESS4.VAR in NAEA DATA ANALYSIS\MY WESVAR FILES. This will overwrite any existing data file of the same name.

COMPUTING DESCRIPTIVE STATISTICS IN WESVAR

Descriptive statistics can be generated in a number of ways in WesVar. Here, Descriptive Stats menu command is used to generate some descriptive statistics for one variable, Mathss (mathematics scale score).
You are presented with four separate paths when you open WesVar:

1. **New WesVar Data File**: Use this path to create a WesVar data file from another file format such as SPSS. (Annex I.C describes the process for creating a new WesVar data file.)

2. **Open WesVar Data File**: Use this path to open an existing WesVar data file for modification purposes, such as when you wish to label or recode variables. In the following exercises, you will use the created WesVar data file `NATASSESS4.VAR`.

3. **New WesVar Workbook**: WesVar requires all analyses to be run in a workbook (see exercise 3.1). The workbook must be linked to a WesVar data file. You will create a new workbook for exercise 3.1.

4. **Open WesVar Workbook**: This path involves opening an already saved workbook to run new analyses or modify existing ones. You can open the WesVar workbook you created for exercise 3.1 and use it to carry out additional analyses within the same chapter (such as exercises 3.2 and 3.3).

Follow the steps in exercise 3.1.
EXERCISE 3.1

Generating Descriptive Statistics in WesVar

1. Launch WesVar, and click New WesVar Workbook. The workbook will allow you to tell WesVar which analyses you wish to run. You will receive the following warning: Before creating a new Workbook, you will be asked to specify a data file that will be used as the default data file for new Workbook requests. Click OK.

2. A window titled Open WesVar Data File for Workbook will appear. Locate the data file NAEA DATA ANALYSIS\MY WESVAR FILES\NATASSESS4.VAR.

3. Select Open – NATASSESS4.VAR. Click Descriptive Stats (bottom right of screen). Highlight Workbook Title 1 on the left panel. Change this title to a more specific title by highlighting the text in the Title box on the right panel and then entering the words Chapter 3 Exercises. In the same manner, change the Request Name from Descriptive Request One by typing in a new name, Exercise 3.1 (see exercise figure 3.1.A). Save your workbook by selecting File – Save As and locating the directory NAEA DATA ANALYSIS\MY WESVAR FILES. Save your file as CHAPTER 3 WORKBOOK.

EXERCISE FIGURE 3.1.A New WesVar Workbook

![New WesVar Workbook](image-url)
4. Select **Options – Output Control** in the left panel. This allows you to control the number of decimal points in your output. Many national assessments use one decimal place for estimates (such as mean scores or percentile ranks) and two for standard errors. Type these figures into the two last boxes in the right panel. (You can make these specifications permanent for the current workbook by going to **File – Preferences – General**. Specify the relevant numbers of decimal places and click **Save**.)

5. Select **Analysis Variables** in the left panel. Move **Mathss** from **Source Variables** to **Selected** (exercise figure 3.1.B). Additional variables can be brought into the **Selected** column if desired.

**EXERCISE FIGURE 3.1.B  Specifying Variables for Analysis in WesVar Descriptives**

6. Click the first **Green Arrow** icon on the toolbar (or select **Requests – Run Workbook Requests** in the menu bar) to run the analysis. Allow time to compute the analysis. Click the **Open Book** icon in the toolbar.

7. To view the output, expand the display by clicking the + signs in front of **Exercise 3.1**, **Variables**, and **Mathss**. Select **Statistics** to see the results (see exercise figure 3.1.C).

**Descriptives** in WesVar generates the following statistics:

- **N** is the number of cases. The column titled “Weighted” denotes that, when **Wgtpop** is applied, the population estimate is 51,713 cases. The unweighted sample size is 4,747.
• **Minimum and Maximum.** The obtained values are 88.3 and 400.0, respectively.

• **Percentile ranks.** The unweighted and weighted values for the 1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th, and 99th quantiles (percentile ranks) are given. The weighted values are normally reported. For example, the weighted score at the 25th percentile is 217.3. This is the estimated value below which 25 percent of the grade 4 population scores. The standard error (2.73) corresponding to this estimate can be found in the final column, “SE Weighted.” The next section in this chapter contains additional information on the meaning of standard error and how to compute it.

• **Mean.** The unweighted mean is 250.2. The weighted mean is 250.0, and the corresponding standard error is 2.20. We should not be surprised to obtain a weighted mean of 250 because the mean was set at this value during scaling. If you wish, you can establish a 95 percent confidence interval around the mean, using the standard error of 2.20 (see below).

• **GeoMean (the geometric mean)** is calculated by taking the product of the values in a distribution and obtaining its nth root, where n is the count of the numbers in the distribution. In general, the geometric mean is not reported in national assessments.
• Sum is the sum of values in a distribution. It is not reported in national assessments.

• Variance. The standard deviation is the square root of the variance. The weighted variance is 2,499.73 (almost 2,500). The square root of this value is 50, which is the standard deviation of Mathss set during scaling.

• CV (coefficient of variation) is computed by dividing the square root of the variance by the mean score. The unweighted and weighted values for the CV are the same (0.20).

• Skewness. The Mathss distribution is slightly negatively skewed (−0.38). Note that a skewness value in the range ±1 is considered satisfactory (see chapter 2).

• Kurtosis. The weighted value for kurtosis is −0.10. Again, this value is well within the recommended ±1 range, indicating no serious concerns with kurtosis.

8. Save the WesVar output as a text file by clicking File and Export on the menu bar on your output file. Select Single File and One Selection, and click Export (exercise figure 3.1.D). Save to NAEA DATA ANALYSIS\MY SOLUTIONS using a suitable file name (such as EXERCISE 3.1.TXT).
CALCULATING A MEAN SCORE AND ITS STANDARD ERROR

The standard error of a statistic is the estimate of the standard deviation of that statistic were one to draw infinite samples from the population such as the one at hand (for example, all students in grade 4). The standard error of the mean is an important statistic because it is used in testing for statistical significance. It should always be reported with national assessment results. Standard errors can also be computed for other statistics, such as percentile ranks. Exercise 3.2 describes how to compute a mean score, its standard error, and its confidence interval in WesVar.
Computing a Mean Score and Its Standard Error in WesVar

1. Launch WesVar, and click Open WesVar Workbook. Locate the workbook used in exercise 3.1 (NAEA DATA ANALYSIS\MY WESVAR FILES\CHAPTER 3 WORKBOOK).

2. Select Chapter 3 Exercises (left panel). Then click Table (right panel). Highlight Table Request One (left panel). Click Add Table Set Single (right panel). Click Table Request One (left panel). Change the Request Name (right panel) to Exercise 3.2.

3. Select Options – Generated Statistics (left panel), and ensure that Estimate, Standard Error, and Confidence Interval (Standard) are checked. Uncheck the other boxes. Under Exercise 3.2 (left panel), click Computed Statistics. Highlight Mathss in the Source Variables menu (right panel), and click Block Mean (also right panel) (exercise figure 3.2.A). Click the Green Arrow icon in the toolbar (Run Workbook Request). Click the Open Book icon in the toolbar to view the output.

4. In Output, select Exercise 3.2, Table Set #1, and then Overall. Click the + (plus) icon to expand nodes as needed. The output is shown in exercise figure 3.2.B.

   The WesVar output shown in exercise figure 3.2.B gives the following data:
   - $M_{Maths}$ is the weighted mean score of 250.0.
   - StdError (Standard Error) is 2.20.
   - Lower 95% and Upper 95% form the 95 percent confidence interval around the mean score of 250. It ranges from 245.6 (lower limit) to 254.4 (upper limit). One can expect to find the true mean score of the population in this interval 95 percent of the time. This interval could also have been computed with the data in figure 3.1.C where the mean and standard error were given as 250.0 and 2.20, respectively.

5. Save your output by clicking on File and Export. Select the Single File and One Selection options. Export your output as a text file to NAEA DATA ANALYSIS\MY SOLUTIONS, using a suitable file name (such as Exercise 3.2.Txt).

6. Exit the output through the Exit Door icon in the toolbar, and save your WesVar workbook by selecting File – Save and then File – Close. This should save your workbook to NAEA DATA ANALYSIS\MY WESVAR FILES.

(continued)
a. Note that WesVar multiplies the standard error by 2.00 rather than the more conventional 1.96 when computing 95 percent confidence intervals. This results in a slightly larger confidence interval.
AN INTRODUCTION TO WESVAR

COMPUTING MEAN SCORES AND STANDARD ERRORS FOR SUBGROUPS IN THE POPULATION

Policy makers, researchers, or others may wish to look at descriptive statistics for different levels of a variable. For example, they may be interested in the national-level mean scores on mathematics achievement for male and female students or in the mean scores for students attending school in different regions in a country.

A simple addition to the WesVar workbook allows us to compute the mean score and standard error for each of the four regions in the Sentz national assessment (Northwest, Metro Area, Eastern Highlands, Southwest Coast) (exercise 3.3).

EXERCISE 3.3
Computing Mean Scores and Standard Errors in WesVar, Four Regions

1. Launch WesVar. Select Open WesVar Workbook. Open the WesVar workbook you saved when you completed exercise 3.2. This should be NAEA DATA ANALYSIS\MY WESVAR FILES\CHAPTER 3 WORKBOOK.

2. Highlight the node Exercise 3.2 on the left panel of your workbook screen, and select Clone (by right-clicking). This makes a copy of the Compute Mean Score table request. Click Table Request Two and label it on the right panel as Exercise 3.3. Expand Exercise 3.3 in the left panel. In Options – Output Control, set Estimate to one decimal place and Std. Error to two. Ensure that Variable Label and Value Label are checked. Uncheck the other options.

3. Click Table Set #1. Then, on the right panel, move Region from Source Variables into the box labeled Selected. Click Add as New Entry.

4. Apply labels for Region if they have not already been applied to the WesVar data file (Natassess4.var). In the left panel, click Region – Cells. Under Cell Definition (right panel), click on 1 (under Region), and type Northwest in the Label panel. Press Return on your keyboard, or click Add as New Entry. Continue this process by assigning the label Metro_Area to 2, Eastern_Highlands to 3, and Southwest_Coast to 4. (Because WesVar does not allow you to leave blank spaces between words, you must use an underscore.)

5. Select Region (left panel) and check Percent under Sum of Weights in the right panel; uncheck the other boxes under Sum of Weights. The Percent option will give the percentage of students in each region, while the mean mathematics score will be generated separately for each region because Mathss has already been specified in Computed Statistics (see exercise figure 3.3.A).

(continued)
6. Click the Green Arrow icon in the toolbar to run the analysis, and then click the Open Book icon to view the output. The output is shown in exercise figure 3.3.B. You may need to click the + (plus) icon to expand Exercise 3.3, Table Set #1, and then Region to view the output.

7. Save your output by clicking File and then Export. Select the Single File and One Selection options. Export your output as a text file to NAEA DATA ANALYSIS\MY SOLUTIONS using a suitable file name (such as EXERCISE 3.3.TXT).

8. Exit the output via the Exit Door icon, and save your WesVar workbook by selecting File – Save and then File – Close. Alternatively, use File – Save As, and overwrite the existing Chapter 3 Workbook in NAEA DATA ANALYSIS\MY WESVAR FILES\CHAPTER 3 WORKBOOK.

The output shown in exercise figure 3.3.B comprises the mean scores, standard errors, and 95 percent confidence intervals for each region. For example, the mean score for the Northwest region is 233.3, and its standard error is 3.28. The corresponding 95 percent confidence interval is 226.8 to 239.9. The figure also shows the percentage of the total grade 4 population in each region (for example, 26.1 percent are in the Metro Area).
NOTES


2. This variable may also be called Jkpair or “cluster pairing code for variance estimation.” Jackknife zones (JKpairs) must be established to create the appropriate weights (replicates).

3. This variable may also be called Jkrep or “within cluster pair replicate code.”

4. The directions for labeling ask you to open a data file in WESVAR UNLABELED DATA. However, if you have already created a data file by following the instructions in annex 1.C, you can use the data file saved to NAEA DATA ANALYSIS\MY WESVAR FILES instead of the data file in WESVAR UNLABELED DATA.

5. If you copy NATASSESS4.VAR from WESVAR UNLABELED DATA into MY WESVAR FILES using Copy and Paste, the file will not run unless you also copy the associated log file NATASSESS4.LOG into the same subfolder.
Policy makers may ask if the achievement levels of subpopulations of students who participated in a national assessment (for example, boys and girls; students attending schools in different regions) differ significantly from each other. This chapter describes procedures that allow such questions to be answered.

**EXAMINING THE DIFFERENCE BETWEEN TWO MEAN SCORES**

To assess whether the difference in achievement between two groups is statistically significant, one needs a categorical variable with two or more levels and a continuous or interval-scaled variable. The following are examples of questions of interest involving categorical and continuous variables:

- **Gender** (female/male) (categorical variable) and **mathematics achievement** (continuous variable). Do male students differ significantly from female students in their mathematics achievements?
- **Availability of an electric light for study** (yes/no) (categorical variable) and **reading achievement** (continuous variable). Do students who
have an electric light at home for study differ significantly in their reading achievements from students who do not have an electric light at home?

• *Access to help with homework at home* (yes/no) (categorical variable) and *science achievement* (continuous variable). Do students who have access to help with homework at home differ significantly in their science achievements from students who do not have access to help?

• *Use of language of instruction at home* (yes/no) (categorical variable) and *civic knowledge* (continuous variable). Do students who speak the language of instruction of the school at home differ significantly in their performance on a test of civic knowledge from students who speak a different language at home?

• *Access to own reading textbook at school* (yes/no) (categorical variable) and *reading achievement* (continuous variable). Do students who have their own reading textbooks at school achieve a mean reading score that differs significantly from the mean score of students who are required to share a textbook?

A test comparing the mean scores of two groups answers the following question: Do two populations represented by the two samples differ significantly in their average score on some variable? Results are reported in terms of significance levels, known as *p*-values. The term “statistically significant” is used to indicate that an observed difference is unlikely to have been caused by chance. If, for example, the results show the mean difference in favor of girls is significant at the 0.05 level (*p*-value), this means that there is less than a 5 percent chance that the difference was caused by chance. A *p*-value of 0.01 means that there is less than one in a hundred chances that the observed result could have happened if the groups did not differ on the variable of interest.

The standard error of the difference (between means) is an important concept in looking at the statistical significance of mean score differences. If the mean score difference is large enough to fall outside a confidence interval around the difference, which is based on the standard error of the difference, one can conclude that the difference is
statistically significant. If the difference falls inside the confidence interval, one can conclude that the difference is not statistically significant.

In the following example (exercise 4.1), the aim is to determine if the mean difference in mathematics achievement between students who have an electric light at home and those who do not is statistically significant.

Before doing exercise 4.1, label the variable Electric in your data file. To do this, launch WesVar, and select Open WesVar Data File. Open the data file NAEA DATA ANALYSIS\MY WESVAR FILES\NATASSESS4.VAR. Then follow the steps for labeling a variable that were outlined in chapter 3 (pages 34–35). Label 1 as Yes to indicate the availability of electricity at home. Label 2 as No to indicate a lack of electricity. Save your WesVar data file to NAEA DATA ANALYSIS\MY WESVAR FILES, overwriting the existing NATASSESS4.VAR if necessary.

**EXERCISE 4.1**

**Evaluating the Difference between Two Mean Scores**

1. Launch WesVar, and click New WesVar Workbook. You will receive the following warning: Before creating a new Workbook, you will be asked to specify a Data file that will be used as the default Data file for new Workbook requests. Click OK.

2. A window called Open WesVar Data File for Workbook will open. Select the data file NATASSESS4.VAR in NAEA DATA ANALYSIS\MY WESVAR FILES. Click Open.

3. Save your new workbook as NAEA DATA ANALYSIS\MY WESVAR FILES\CHAPTER 4 WORKBOOK.WVB.

4. Highlight Workbook Title 1 (left panel) and type Chapter 4 Exercises in the Title box (right panel). In the box New Request (right panel, bottom half), click Table. Highlight Table Request One (left panel) and change Request Name (right panel) to Exercise 4.1. Click Add Table Set Single (right panel).

5. Go to Options – Generated Statistics in the left panel. Ensure that Estimate, Standard Error, and Confidence Interval (Standard) are checked. Uncheck the other boxes.

(continued)
6. Go to Options – Output Control in the left panel. Ensure that Estimates has been set to one decimal place and Std. Error to two decimal places. Ensure that variable and value labels have been checked.

7. Select Computed Statistics, highlight Mathss in Source Variables (right panel), and click Block Mean.

8. Click Table Set #1 (left panel). Move Electric from the Source Variables box into the box marked Selected in the right panel. Ensure that Percent has been checked under Sum of Weights. If selected, uncheck Value, Row Percent, and Column Percent. Click the Add as New Entry button.

9. Highlight Electric (left panel), and click Cells below it. Click on 1 (right panel), and type Yes in the label box and press Return (or click Add as New Entry). Click on 2 and type No in the label box, and press Return.

10. Now select Cell Functions (left panel). Type the following into the box labeled Function Statistic: MeanDiff = Yes − No. Click Add As New Entry.

11. Under MeanDiff = Yes − No (left panel), click For. Ensure that M_Mathss is in the box labeled Selected (exercise figure 4.1.A). If necessary, move Sum_Wts to the box labeled Source Variables.

EXERCISE FIGURE 4.1.A WesVar Workbook before Assessing the Difference between Two Mean Scores
12. Run the analysis by clicking the Green Arrow in the toolbar. Allow time for the program to run. View the output by clicking the Open Book icon in the toolbar.

13. To view mean scores associated with having or not having an electric light at home, expand Exercise 4.1 – Table Set #1 and click Electric. The data for students in households with and without electricity can be viewed in exercise figure 4.1.B. Note that students in households with electricity recorded a mean Mathss score of 254.3 and a standard error of 2.30.

**EXERCISE FIGURE 4.1.B** WesVar Output: Mean Mathematics Scores of Students with and without Electricity at Home

14. To save the output, select File – Export – Single File – One Selection, and click Export. Save to NAEA DATA ANALYSIS\MY SOLUTIONS using a suitable file name (such as EXERCISE 4.1A.TXT).

15. To view the estimate of the difference in mathematics scores between students in households with and without electricity (26.5), click Functions (below Electric) (exercise figure 4.1.C).

**EXERCISE FIGURE 4.1.C** WesVar Output: Mean Score Difference in Mathematics between Students with and without Electricity at Home
Exercise figure 4.1.B gives the mean mathematics scores for students with electricity at home (254.3) and for those without electricity (227.8). The corresponding standard errors are 2.30 and 4.95. Exercise figure 4.1.C gives the mean score difference (26.5). This is the difference in mean mathematics scores between those with electricity at home and those without (Yes − No). The standard error of the difference is 5.64. The 95 percent confidence interval (around the mean score difference) extends from 15.2 (lower limit) to 37.8 (upper limit). The confidence interval can help to quickly determine if a significant difference exists between means. If the confidence interval includes the value zero (such as from −4.5 to +7.9), one can assert that the mean difference is not significantly different from zero at the 0.05 level. In the case of the present data, because the 95 percent confidence interval (15.2 to 37.8) does not include zero, one concludes that the mean difference of 26.5 is significantly different from zero ($p < 0.05$). The difference in mean mathematics achievement between students with and without electricity at home is statistically significant.

The information obtained in this analysis can be summarized in a table (exercise table 4.1.A). A table such as this might be included in a report on a national assessment.

**EXERCISE TABLE 4.1.A  Comparison of Mean Mathematics Scores of Students with and without Electricity at Home**

<table>
<thead>
<tr>
<th>Status</th>
<th>Percentage of students (SE)</th>
<th>Mean score (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electricity at home</td>
<td>83.9 (3.41)</td>
<td>254.3 (2.30)</td>
</tr>
<tr>
<td>No electricity at home</td>
<td>16.1 (3.41)</td>
<td>227.8 (4.95)</td>
</tr>
<tr>
<td>Comparison Electricity – No electricity at home</td>
<td>Difference (SED)</td>
<td>CI (95%)</td>
</tr>
<tr>
<td></td>
<td>26.5 (5.64)</td>
<td>15.2 to 37.8</td>
</tr>
</tbody>
</table>

Note: CI (95%) = 95% confidence interval; SE = standard error of the estimate; SED = standard error of the difference. Confidence intervals associated with statistically significant differences are in boldface type.

16. Return to Functions, and save the comparison of mean scores as **EXERCISE 4.1B.TXT** in NAEA DATA ANALYSIS/MY SOLUTIONS.

17. Return to the workbook Chapter 4 Exercises by clicking the Open Door icon. Click Save on the toolbar to save the changes. Select File – Close in the menu bar to close the workbook, or continue to exercise 4.2.
EXAMINING DIFFERENCES BETWEEN THREE OR MORE MEAN SCORES

In this section, we examine differences between three or more categories of students (for example, students attending school in different regions of a country). We compare the achievements of students in one region of Sentz (such as the Metro Area) with that of students in each of the country’s three other regions. In the analysis, Metro Area is termed the reference group, and performance in each of the remaining regions is compared to it.

When making multiple comparisons (for example, comparing the mean score of students in one region with the mean scores of students in three other regions), one must adjust the alpha or significance level. If it is not adjusted, one runs the risk of reporting that a difference is statistically significant when in fact it is not. The standard value of alpha (0.05) that is used when comparing a pair of mean scores (using a 95 percent confidence interval around the mean score difference) needs to be adjusted downward (that is, divided by the number of comparisons to be made) if one makes more than one comparison. For example, if three comparisons are made, the alpha level value of 0.05 should be divided by 3, giving an adjusted value of 0.0167 (0.05/3).

In the example in exercise 4.2, the mean score of students in the Metro Area (the reference group) is compared to the mean score of students in each of the other regions. Hence, three comparisons are made:

- Metro Area – Northwest
- Metro Area – Eastern Highlands
- Metro Area – Southwest Coast

EXERCISE 4.2

Evaluating Differences among Three or More Mean Scores

1. Launch WesVar. Select Open WesVar Workbook. Open the WesVar workbook you saved when you completed exercise 4.1. This is NAEA DATA ANALYSIS\MY WESVAR FILES\CHAPTER 4 WORKBOOK.
2. Highlight the node Exercise 4.1. Right-click and select Clone. This makes a copy of the Exercise 4.1 (Differences between Two Means request table). Highlight Table Request Two on the right panel, and label it Exercise 4.2.

3. Expand Exercise 4.2 (left panel). Select Options (left panel), and in the right panel, change the Alpha level to 0.0167 (because three comparisons will be made) (exercise figure 4.2.A).

4. Expand Options in the left panel. Under Generated Statistics, ensure that Estimate, Standard Error, and Confidence Interval (Standard) are checked. Under Options – Output Control, ensure that Variable Label and Value Label are checked. Uncheck the other boxes. Expand Table Set #1 in the left panel, and highlight Electric, deselect it in the right panel, and select Region. Click Replace Current Entry. If you receive a message: Table structure has changed… Do you want to make this change?, select Yes. Ensure that only Percent is checked. Expand Region (left panel). Select Cells, and define the cells as follows: 1 = Northwest; 2 = Metro_Area; 3 = Eastern_Highlands; 4 = Southwest_Coast. (Because WesVar does not allow you to leave blank spaces between words, you must use an underscore.) After entering each label, click Add as New Entry or press Return on your keyboard (exercise figure 4.2.B).
5. Select **Cell Functions** (immediately below **Cells** in the left panel). Enter the following, into the **Function Statistic** box, and click **Add as New Entry** after each one.

- **MeanDiffMetro_NW** = Metro_Area − Northwest (exercise figure 4.2.C).

Click **For** below each function (left panel), and ensure that **Mathss** appears under **Selected** each time. You may have to move **Sum_Wts** to **Source Variables**.

Repeat the process for

- **MeanDiffMetro_EHighlands** = Metro_Area − Eastern_Highlands
- **MeanDiffMetro_SW** = Metro_Area − Southwest_Coast

6. Click the **Green Arrow** in the toolbar to run the analyses.

7. View the output by clicking the **Open Book** icon in the toolbar. To access mean scores for each region, select **Exercise 4.2 – Table Set #1** and **Region** (exercise figure 4.2.D). To access the data on differences between mean scores, select **Functions** (the node immediately below **Region**) (exercise figure 4.2.E).
8. Use **File** and **Export (Single File, One Selection)** to save your output as text files. First save the mean scores (Regions), and then the mean score differences (Functions) in **NAEA DATA ANALYSIS MY SOLUTIONS** using the file names **EXERCISE 4.2A.TXT** and **EXERCISE 4.2B.TXT**, respectively.

9. Return to **CHAPTER 4 WORKBOOK** by clicking the **Open Door** icon in the toolbar. Select **Save – Close** on the menu bar (or select the **Save** icon on the toolbar).

Exercise figure 4.2.E provides the three comparisons requested. The difference in mean scores between the Metro Area and the Northwest is 32.4 scale score points, and
its standard error is 5.74. The confidence interval around the difference is 18.2 to 46.5. Because the interval does not include zero, the mean score of the Metro Area is statistically significantly different from the mean score of the Northwest region. Similarly, the difference in mean scores between the Metro Area and the Eastern Highlands (16.6 points) is statistically significant because the interval around the difference (3.3 to 30.0) does not include zero. Finally, the difference between the Metro Area and the Southwest Coast (14.5 points) is not statistically significant, because the interval around the difference (−0.0 to 29.0) includes zero.

Exercise table 4.2.A presents the output of this analysis as it might be presented in a national assessment report. The table shows each regional mean score and the national mean score and associated standard errors. The bottom half of the table indicates, in boldface type, statistically significant regional differences.

<table>
<thead>
<tr>
<th>Region</th>
<th>Mean score (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metro Area</td>
<td>233.3 (3.28)</td>
</tr>
<tr>
<td>Northwest</td>
<td>265.7 (4.46)</td>
</tr>
<tr>
<td>Eastern Highlands</td>
<td>249.1 (3.59)</td>
</tr>
<tr>
<td>Southwest Coast</td>
<td>251.2 (3.35)</td>
</tr>
<tr>
<td>National</td>
<td>250.0 (2.20)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Difference (SED)</th>
<th>BCI (95%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metro Area – Northwest</td>
<td>32.4 (5.74)</td>
<td>18.2 to 46.5</td>
</tr>
<tr>
<td>Metro Area – Eastern Highlands</td>
<td>16.6 (5.42)</td>
<td>3.3 to 30.3</td>
</tr>
<tr>
<td>Metro Area – Southwest Coast</td>
<td>14.5 (5.90)</td>
<td>−0.0 to 29.0</td>
</tr>
</tbody>
</table>

Note: BCI (95%) = adjusted 95% confidence interval; SE = standard error of the estimate; SED = standard error of the difference. Confidence intervals associated with statistically significant differences are in boldface type.
Other comparisons that might interest policy makers (depending on the variables in the database) include the following:

- **Ethnic group** and **mathematics achievement**: Are there significant differences between ethnic groups in their mean mathematics achievements? Which group has the highest mean score?
- **Parental educational level** and **reading achievement**: Are there significant mean reading achievement differences between students of parents who have completed college and students in each of the other groups (such as parents who have had no formal education; parents whose highest level of education was in grades 1–3, grades 4–6, grades 7–9, or grades 10–12)?
- **Access to communications** and **language achievement**: Are there significant differences in mean language achievement between students who live in homes that have (a) both radio and television, compared to those in homes that have (b) a radio only, (c) a television only, or (d) no radio and no television?
IDENTIFYING HIGH AND LOW ACHIEVERS

In addition to considering differences between the mean achievement levels of subpopulations of students in a national assessment (chapter 4), policy makers and other users of results may be interested in identifying factors related to the distribution of achievement. Questions such as the following may arise:

- Is there a greater proportion of at-risk students (lower achievers) in one region (province) in the country than in another?
- Are boys and girls equally represented among high achievers in mathematics?
- In which school type (such as rural or urban, public or private) is the largest proportion of low achievers?

In this chapter, questions such as these are answered by identifying the proportions of students above or below key points, or benchmarks, such as the 10th or 90th percentile.
ESTIMATING SCORES CORRESPONDING TO KEY NATIONAL PERCENTILE RANKS

This section describes an approach to estimating the scores of students at key percentile ranks in a national assessment. In chapter 3, the Descriptive Stats routine in WesVar was used to find the unweighted and weighted scores at different percentile ranks, along with their standard errors (see exercise figure 3.1.C). In exercise 5.1, an alternative approach—WesVar Tables—is used to estimate the scores that correspond to the 10th, 25th, 50th, 75th, and 90th percentile ranks and their standard errors for each region in Sentz. This approach can also be used to identify scores corresponding to other percentile ranks.

EXERCISE 5.1

Computing National Percentile Scores

1. Launch WesVar, and click New WesVar Workbook. You may receive the following warning: Before creating a new Workbook, you will be asked to specify a Data file that will be used as the default Data file for new Workbook requests. If so, click OK.

2. A window titled Open WesVar Data File for Workbook will appear. Select the data file NAEA DATA ANALYSIS\MY WESVAR FILES\NATASSESS4.VAR.

3. Save your new workbook in NAEA DATA ANALYSIS\MY WESVAR FILES, giving it the name CHAPTER 5 WORKBOOK.WVB.

4. Change Workbook Title One (right panel) to Chapter 5 Exercises. In New Request, highlight Table. Click Table Request One (left panel). Change the Request Name to Exercise 5.1 (right panel). Click Add Table Set Single (right panel).

5. Select Options – Generated Statistics and ensure that Estimate, Standard Error, and Confidence Interval (Standard) are checked. Uncheck the other boxes. In Output Control, set Estimate to one decimal place and Std. Error to two. Ensure that Variable Label and Value Label are checked.
6. In the left panel, select Computed Statistics. In the right panel, type $Pcile10 = \text{Quantile(Mathss, 0.1)}$ and click Add as New Entry (exercise figure 5.1.A). This tells WesVar to compute the score at the 10th percentile.

**EXERCISE FIGURE 5.1.A** WesVar Workbook: Computing Percentile Scores

7. Follow the same procedure for the 25th, 50th, 75th, and 90th percentiles, clicking on Computed Statistics first each time. For example, the formula for the 25th percentile is $PCile25 = \text{Quantile(Mathss, 0.25)}$. Remember to click Add as New Entry after you enter each formula.³

8. Run the analysis by clicking the Green Arrow in the toolbar.

9. View the output (exercise figure 5.1.B) by clicking the Open Book icon in the toolbar. Expand Exercise 5.1 – Table Set #1 in the left panel. Click Overall to view percentile scores.

(continued)
Once the percentile rank scores at the national level have been obtained, the corresponding scores for each region in Sentz can be computed (exercise 5.2).
EXERCISE 5.2

Computing Percentile Scores by Region

1. Launch WesVar. Select Open WesVar Workbook. Open the WesVar workbook you saved when completing exercise 5.1. This is NAEA DATA ANALYSIS\My WesVar FILES\Chapter 5 Workbook.

2. Minimize (if necessary) Exercise 5.1 (left panel), by clicking on the – (minus) sign.

3. Right-click Exercise 5.1 in the left panel, and select Clone. Rename Table Request Two in the right panel as Exercise 5.2. All subsequent parts of exercise 5.2 will be completed in this node.

4. Expand Exercise 5.2 (if necessary). Select Table Set #1. Move Region from Source Variables to Selected (right panel).

5. Click Add as New Entry (right panel).

6. Select Region in the left panel, and check the Percent box under Sum of Weights in the right panel. Uncheck Value, Row Percent, and Column Percent if necessary (see exercise figure 5.2.A).

EXERCISE FIGURE 5.2.A  WesVar Workbook before Computing Percentile Scores by Region

(continued)
7. Click the **Green Arrow** in the toolbar to run the analyses.
8. Click the **Open Book** icon in the toolbar. Expand **Exercise 5.2** (if necessary). Select **Table Set #1 – REGION** to see the percentile estimate for each region.

Exercise figure 5.2.B gives the scores at the 10th percentile for each region, together with associated standard errors and 95 percent confidence intervals. Data for other percentile scores can be accessed by scrolling down the output file. Exercise table 5.2.A presents the data in a format that could be used in a national assessment report.

**EXERCISE 5.2 (continued)**

Exercise table 5.2.A shows that the scores corresponding to percentile ranks vary by region. For example, scores at the 10th percentile range from 162.2 in the Northwest to 205.1 in the Metro Area. An inspection of regional scores suggests that lower-achieving students (performing at the 10th percentile) in the Northwest, the Eastern Highlands, and the Southwest Coast perform less well than lower-achieving students in the Metro Area. Scores of students at the 90th percentile in the Northwest are lower than those of students in other regions.
ESTIMATING PERCENTAGES OF STUDENTS IN SUBGROUPS, USING NATIONAL PERCENTILE RANKS

Information on the percentage of students in subgroups in the population (such as region, ethnicity, or gender) who score below a selected benchmark, such as the national 10th percentile score, can be obtained by first identifying the scores that correspond to national percentile ranks on the national assessment database in WesVar (see exercise table 5.1.A). Exercise 5.3 shows how to create new variables corresponding to the 25th, 50th, and 75th percentile ranks in the WesVar data file. Create a new variable corresponding to each percentile benchmark, and allocate a “1” to students who score below the benchmark and a “2” to those who score at or above it.

EXERCISE 5.3

Recoding a Variable into Percentile Categories Using WesVar

1. Launch WesVar. Select Open WesVar Data File, and select NAEA DATA ANALYSIS\MY WESVAR FILES\NATASSESS4.VAR.

2. Click Format in the menu bar, and select Recode. In Pending Records, click New Continuous (to Discrete).

3. Assign a New Variable Name in the box provided. For this exercise, assign a group variable Pcie25, which will divide students into two categories: those scoring below the national 25th percentile and those scoring at or above it.

4. In the column labeled Range of Original Variables, type Mathss < 217.3 (national 25th percentile value; see table 5.2). On the same row, in the column labeled Pcie25,
5. Click **New Continuous to Discrete** again, and follow the same procedure to create the following variables:

   - **Pcile50**: Mathss < 256.3 → 1 Mathss ≥ 256.3 → 2 (OK)
   - **Pcile75**: Mathss < 284.4 → 1 Mathss ≥ 284.4 → 2 (OK)

6. After entering the last variable (**Pcile75**), click **OK**. Then click **OK** in the Pending Recodes screen. You will see the message *This operation will create a new VAR file*. You will be asked to specify a file name. Save as NAEA DATA ANALYSIS\MY WESVAR FILES\NATASSESS4.VAR, overwriting the previously saved version.

7. Select **Format** and **Label** in the menu bar. Under **Source Variables**, click **Pcile25**. For **Values**, in the column titled **Label**, replace the number 1 with **Bottom 25%** and replace 2 with **Top 75%** (exercise figure 5.3.B). Then click **OK**. You will see the message *This operation will create a new VAR file*. You will be asked to specify a file name. Click **OK**. Save as NAEA DATA ANALYSIS\MY WESVAR FILES\NATASSESS4.VAR to overwrite the previously saved file version.
8. Repeat the process for Pcile50 and Pcile75.

9. Select File and Exit. 

   a. Note this exercise has no output file to save. Instead, you saved the modified WesVar data file (NATASSESS4.VAR), which is used in the next exercise.

Next, the percentages of students scoring below key national percentile benchmarks in each region are computed (exercise 5.4).

**EXERCISE 5.4**

**Computing Percentages of Students Scoring below Key National Percentile Benchmarks and Standard Errors by Region**

1. Launch WesVar. Select Open WesVar Workbook. Open NAEA DATA ANALYSIS /MY WESVAR FILES/CHAPTER 5 WORKBOOK.

2. Select Chapter 5 Exercises (left panel). Select New Request – Table (right panel). Click Table Request Three (left panel), and rename it Exercise 5.4. Select Exercise 5.4 (left panel) and then Add Table Set – Single (right panel).
EXERCISE 5.4 (continued)

3. Select **Table Set #1** (left panel), click **Pcile25** in the **Source Variables** box (right panel), and move it to the **Selected** box. Then move **Region** from **Source Variables** to **Selected**. Under **Sum of Weights**, ensure that the box for **Column Percent** (right panel) is checked and that the other options (**Value**, **Percent**, and **Row Percent**) are unchecked. Click **Add as New Entry** (right panel). Return **Pcile25** and **Region** to **Source Variables**.

4. Repeat step 3 for **Pcile50** and **Region** and for **Pcile75** and **Region**, ensuring that you enter **Pcile** before **Region** each time. Click **Add as New Entry** after each modification (exercise figure 5.4.A).

EXERCISE FIGURE 5.4.A  WesVar Workbook Screenshot before Computing Percentages

Scoring below Key National Benchmarks by Region

5. Select **Output Control** in the left panel, and ensure that **Variable Label** and **Value Label** are checked.

6. Select **Generated Statistics**, and ensure that only **Estimate**, **Standard Error**, and **Confidence Interval (Standard)** are checked.
7. Run the analysis by clicking the **Green Arrow** in the toolbar. View the output by selecting the **Open Book** icon in the toolbar. Expand the left side by expanding **Exercise 5.4, Table Set #1**, and **Pcile25*Region** rank to view the first block of output (exercise figure 5.4.B).

**EXERCISE FIGURE 5.4.B**  Partial Output: Percentages of Students Scoring below Key National Benchmarks by Region

<table>
<thead>
<tr>
<th>Pcile25</th>
<th>Region</th>
<th>STATISTIC</th>
<th>EST_TYPE</th>
<th>ESTIMATE</th>
<th>STDERROR</th>
<th>LOWER_95%</th>
<th>UPPER_95%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottom 25%</td>
<td>Northwest</td>
<td>SIM_WTS</td>
<td>COLCPT</td>
<td>36.5</td>
<td>2.74</td>
<td>31.0</td>
<td>42.0</td>
</tr>
<tr>
<td>Bottom 25%</td>
<td>Metro Area</td>
<td>SIM_WTS</td>
<td>COLCPT</td>
<td>14.6</td>
<td>2.62</td>
<td>9.7</td>
<td>20.0</td>
</tr>
<tr>
<td>Bottom 25%</td>
<td>Eastern Highland</td>
<td>SIM_WTS</td>
<td>COLCPT</td>
<td>25.1</td>
<td>2.83</td>
<td>19.5</td>
<td>30.8</td>
</tr>
<tr>
<td>Bottom 25%</td>
<td>Southwest Coast</td>
<td>SIM_WTS</td>
<td>COLCPT</td>
<td>24.4</td>
<td>2.30</td>
<td>19.8</td>
<td>29.0</td>
</tr>
<tr>
<td>Bottom 25%</td>
<td>MARGINAL</td>
<td>SIM_WTS</td>
<td>COLCPT</td>
<td>25.0</td>
<td>1.66</td>
<td>22.1</td>
<td>27.9</td>
</tr>
<tr>
<td>Top 75%</td>
<td>Northwest</td>
<td>SIM_WTS</td>
<td>COLCPT</td>
<td>63.5</td>
<td>2.74</td>
<td>58.0</td>
<td>69.0</td>
</tr>
<tr>
<td>Top 75%</td>
<td>Metro Area</td>
<td>SIM_WTS</td>
<td>COLCPT</td>
<td>34.6</td>
<td>2.63</td>
<td>29.0</td>
<td>41.3</td>
</tr>
<tr>
<td>Top 75%</td>
<td>Eastern Highland</td>
<td>SIM_WTS</td>
<td>COLCPT</td>
<td>74.9</td>
<td>2.63</td>
<td>69.2</td>
<td>80.6</td>
</tr>
<tr>
<td>Top 75%</td>
<td>Southwest Coast</td>
<td>SIM_WTS</td>
<td>COLCPT</td>
<td>75.6</td>
<td>2.30</td>
<td>71.0</td>
<td>80.2</td>
</tr>
<tr>
<td>Top 75%</td>
<td>MARGINAL</td>
<td>SIM_WTS</td>
<td>COLCPT</td>
<td>75.0</td>
<td>1.46</td>
<td>72.1</td>
<td>77.9</td>
</tr>
<tr>
<td>MARGINAL</td>
<td>Northwest</td>
<td>SIM_WTS</td>
<td>COLCPT</td>
<td>100.0</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MARGINAL</td>
<td>Metro Area</td>
<td>SIM_WTS</td>
<td>COLCPT</td>
<td>100.0</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MARGINAL</td>
<td>Eastern Highland</td>
<td>SIM_WTS</td>
<td>COLCPT</td>
<td>100.0</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MARGINAL</td>
<td>Southwest Coast</td>
<td>SIM_WTS</td>
<td>COLCPT</td>
<td>100.0</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MARGINAL</td>
<td>MARGINAL</td>
<td>SIM_WTS</td>
<td>COLCPT</td>
<td>100.0</td>
<td>0.00</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

8. Save the output by selecting **File – Export – Single File – One Selection** from the menu bar. Click **Export**. Save as a text file to **NAEA DATA ANALYSIS\MY SOLUTIONS\EXERCISE 5.4 – 25th** (for output relating to the 25th percentile), **EXERCISE 5.4 – 50th** (for output relating to the 50th percentile), and so on. Exit the output file through the **Open Door** icon in the toolbar.

9. Save the WesVar workbook by clicking the **Save** icon in the toolbar (or by selecting **File – Save** on the menu bar). Click **File – Close**.

The output can be used to ascertain the percentage of students in each region scoring below the national 25th percentile (exercise figure 5.4.B). It is estimated that, in the Northwest region, 36.5 percent of students score below this percentile benchmark. The corresponding standard error of the estimate is 2.74, and the 95 percent confidence interval is 31.0 percent to 42.0 percent. Table 5.1 presents the data in tabular form. It can be seen that 36.5 percent of students in
the Northwest, compared with 14.4 percent in the Metro Area, score below the national 25th percentile benchmark. In the remaining two regions, Eastern Highlands and Southwest Coast, the percentages scoring at this level are similar to the national percentage (25 percent).

In a similar manner, you can tabulate the percentages of students scoring below (or at or above) other national benchmarks, such as the 50th percentile (Pcile50) and the 75th percentile (Pcile75). Table 5.2 shows the percentages of higher-achieving students (defined as those scoring at or above the national 75th percentile) in each region. The data in this table are based on the output of exercise 5.4 and show that 15.1 percent of students in the Northwest score at or above the national 75th percentile benchmark, compared to 35.3 percent of students in the Metro Area. In the remaining two regions, the percentages achieving at or above the national 75th percentile benchmark (24.7 percent in both cases) are similar to the national percentage (25 percent).

TABLE 5.1
Percentages of Students Scoring below the National 25th Percentile Benchmark by Region

<table>
<thead>
<tr>
<th>Region</th>
<th>Students scoring below national 25th percentile benchmark</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Percent</td>
</tr>
<tr>
<td>Northwest</td>
<td>36.5</td>
</tr>
<tr>
<td>Metro Area</td>
<td>14.4</td>
</tr>
<tr>
<td>Eastern Highlands</td>
<td>25.1</td>
</tr>
<tr>
<td>Southwest Coast</td>
<td>24.4</td>
</tr>
<tr>
<td>National</td>
<td>25.0</td>
</tr>
</tbody>
</table>

TABLE 5.2
Percentage of Students Scoring at or above the National 75th Percentile Benchmark by Region

<table>
<thead>
<tr>
<th>Region</th>
<th>Students scoring at or above national 75th percentile benchmark</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Percent</td>
</tr>
<tr>
<td>Northwest</td>
<td>15.1</td>
</tr>
<tr>
<td>Metro Area</td>
<td>35.3</td>
</tr>
<tr>
<td>Eastern Highlands</td>
<td>24.7</td>
</tr>
<tr>
<td>Southwest Coast</td>
<td>24.7</td>
</tr>
<tr>
<td>National</td>
<td>25.0</td>
</tr>
</tbody>
</table>
ASSOCIATION BETWEEN VARIABLES: CORRELATION AND REGRESSION

CORRELATION

Policy makers may be interested in identifying the extent to which student learning is related to a range of factors. The correlation coefficient ($r$), which is a measure of the linear association between two variables, provides this information. The following are examples of questions that correlations are designed to answer:

- Is there an association between frequency of attendance at school and students’ mathematics achievement?
- Is level of parents’ educational attainment related to students’ reading achievement?
- Is there an association between the distance a student travels to school and performance in mathematics?
- Is teachers’ experience (number of years teaching) related to students’ achievements in key curriculum areas?

The correlation coefficient can tell us the direction of the relationship between two variables and the strength or magnitude of the relationship between them.
Direction of the Relationship

Correlation coefficients can be positive, negative, or zero. A positive correlation indicates that the values of two variables tend to move in the same direction; as scores on one variable increase, scores on the other, on average, also increase. A positive correlation (for example, 0.60) between performance in reading and writing would indicate that, on average, as performance in reading increases, performance in writing also increases, and vice versa. Conversely, a negative correlation would indicate that as the value of one variable increases, the value of the other variable tends to decrease. For example, a negative correlation (−0.28) between anxiety about mathematics and performance on a test of mathematics would indicate that, in general, as student anxiety increases, performance decreases (and vice versa). A correlation of zero indicates that no evidence exists of a relationship between two variables (such as student height and achievement). Figure 6.1 illustrates graphically positive and negative relationships between variables. In the first diagram, the relationship is positive; in the second it is negative.

Strength or Magnitude of the Relationship

Correlation coefficients that are close to −1.0 or +1.0 indicate a strong relationship. Values between these extremes indicate somewhat weaker relationships.
In general, correlations between measures of socioeconomic status and achievement tend to range between 0.20 and 0.30. Correlations between students’ scores on tests of reading and science are often in the range of 0.80 to 0.90 (see, for example, OECD 2007), suggesting that performance on one test (for example, reading) draws on the same skills as performance on another (science). In national assessments or in educational research in general, one rarely finds correlations that are perfect or almost perfect.

**Drawing a Scatterplot**

Before computing a correlation coefficient and evaluating its statistical significance, drawing a scatterplot that illustrates the relationship between two variables in the form of a graph may be useful. If the relationship is linear, points will tend to fall around a straight line running through the data. The closer the points are to a straight line, the stronger the relationship between the variables. In exercise 6.1, SPSS is used to describe the strength of the relationship between two mathematics subscale scores in the *NATASSESS* database, Impl_pc (percent correct score on items dealing with implementation of mathematics procedures) and Solve_pc (percent correct score on analyzing and solving mathematical problems).1

---

**EXERCISE 6.1**

**Drawing a Scatterplot in SPSS**

1. Open the data file *NAEA DATA ANALYSIS*\*SPSS DATA\*NATASSESS4.SAV.
2. Select Data – Weight Cases – Weight Cases by ..., move Wgtpop into the box labeled Frequency Variable, and click OK.
4. Assign Impl_pc to the Y Axis and Solve_pc to the X Axis (exercise figure 6.1.A). Click OK. Allow time for processing. The output is shown in exercise figure 6.1.B.

(continued)
EXERCISE 6.1 (continued)

EXERCISE FIGURE 6.1.A  Partial SPSS Dialog Box before Drawing Scatterplot

EXERCISE FIGURE 6.1.B  Scatterplot of Relationship between Implementation of Procedures and Problem Solving in Mathematics
We can add a line of best fit to the scatterplot. This is a straight line that best summarizes the data in the plot. To draw the line of best fit on an SPSS scatterplot, double-click on the scatterplot to enter the Chart Editor and, from the menu bar, select Elements – Fit Line at Total – Linear – Confidence Intervals – None. The output is shown in exercise figure 6.1.C.

In a scatterplot, one or more individuals are represented by a point that is the intersection of their scores on two variables. For example, student N (exercise figure 6.1.B) scored 100 percent correct on mathematics implementation (y-axis) and 83 percent correct on mathematics problem solving (x-axis). Note that points cluster
Computing a Correlation Coefficient and Evaluating Its Statistical Significance

This section shows how to compute a correlation coefficient using WesVar (exercise 6.2). The aim is to determine the magnitude of the relationship between performance on implementation of mathematical procedures (Impl_pc) and performance on mathematical problem solving (Prob_pc). A measure of the error around the obtained correlation coefficient is also needed to allow one to test whether the correlation coefficient differs significantly from zero.

**EXERCISE 6.1 (continued)**

in a band, which goes from lower left to upper right, an indication of a positive correlation between the two variables.

The output for the scatterplot (exercise figures 6.1.B and 6.1.C) shows that performance on mathematical problem solving tends to increase as performance on implementing mathematical procedures increases (and vice versa).

5. In SPSS output, click outside the chart area. Save the output by selecting File – Save As: NAEA DATA ANALYSIS\MY SOLUTIONS, and name the file EXERCISE 6.1.SPV.

---

**EXERCISE 6.2**

Computing a Correlation Coefficient, National Level

1. Launch WesVar, and click New WesVar Workbook. You may receive the following warning: Before creating a new Workbook, you will be asked to specify a Data file that will be used as the default Data file for new Workbook requests. If so, click OK.

2. A window titled Open WesVar Data File for Workbook will appear. Select the data file NAEA DATA ANALYSIS\MY WESVAR FILES\NATASSESS4.VAR.

3. Save your new workbook in NAEA DATA ANALYSIS\MY WESVAR FILES\CHAPTER 6 WORKBOOK.WVB. Click Workbook Title 1 (left panel) and type in Chapter 6 Exercises (right panel).

4. Click Descriptive Stats. Highlight Descriptive Request One (left panel) and type in Exercise 6.2 (right panel). Select Options – Output Control in the left panel. Make sure that the number of decimal places for Estimates and Std. Error is set at three. Also, make sure that Variable Name and Variable Label have been checked.
5. Select **Correlations** in the left panel. Activate **List 1** (right panel), and move **Solve_pc** from **Source Variables** into **List 1**. Activate **List 2**, and move **Impl_pc** from **Source Variables** into **List 2** (exercise figure 6.2.A). You can add additional variables to each list if desired. Each variable in **List 1** will be correlated with each one in **List 2**.

6. Run the analysis by clicking the **Green Arrow** in the toolbar.

7. View the output by clicking the **Open Book** icon in the toolbar. Click the + (plus) sign to expand **Correlations**. Select **Overall** (exercise figure 6.2.B).

**EXERCISE FIGURE 6.2.A** WesVar Workbook before Running Correlation Analysis

**EXERCISE FIGURE 6.2.B** WesVar Output: Correlation between Problem Solving and Implementing Mathematics Procedures

(continued)
Regression differs from correlation in a number of ways. First, the correlation model does not specify the nature of the relationship between variables. Regression, in contrast, models the dependence of one variable on one or more other variables. On the basis of theory or research, scores on one variable (the dependent variable, usually represented on the vertical [y] axis, in a graph) are considered to be dependent on, or contingent on, scores on another variable (the independent variable, usually represented on the horizontal [x] axis). For example, scores on a reading test could be expected to depend on the amount of time a student spends reading for leisure.2

Second, in regression, the functional relationship between independent and dependent variables can be formally stated as an equation with associated values that describe how well the equation fits the data. Information about the performance of a group of individuals is used to specify an equation (known as a regression equation).
(equation) on the assumption that the relationship is linear (that is, a change in value in one variable will be similar for all values of the variable).

Analysts may need to go beyond this form of regression when analyzing data from a national assessment. More sophisticated approaches, such as hierarchical linear modeling (HLM), are generally better suited to take into account the hierarchical or multilevel structure of the data obtained in these studies (see Raudenbush and Bryk 2002; Snijders and Bosker 1999). HLM can separate the effects of school- and student-level variables. For example, if you have data on school- and student-level socioeconomic status, both can be included in the model and the effects of each on student achievement ascertained. A two-level model also allows one to identify the proportion of variance between schools and the proportion within schools that the variables in the model explain. Similarly, a three-level model (schools, classes, students) provides an estimate of the proportion of variance explained by variables at the school level (such as location and size), class level (such as teacher characteristics and available classroom resources), and student level (such as age and anxiety about mathematics). Multilevel models are especially useful when between-school variance in the dependent variable is large (for example, when it exceeds 5 percent of total variance).

Multilevel modeling is beyond the scope of the current volume. However, the form of regression analysis that is described provides an introduction to some of the concepts underlying multilevel modeling and may be used when multilevel modeling is not appropriate.

The regression equation, in cases with one dependent variable, \( y \), and one independent variable, \( x \), is as follows:

\[
\hat{y} = \alpha + bX + \varepsilon,
\]

where

- \( \alpha \) = intercept (the point on the \( y \) axis when \( x \) is zero)
- \( b \) = gradient or slope of the regression line (the regression coefficient)
- \( X \) = score on the independent variable
- \( \varepsilon \) = error term (reflected in the residuals, or differences between expected and observed values).\(^3\)
Figure 6.2 shows the regression equation and regression line on a scatterplot for two variables: x (independent) and y (dependent).

The regression equation (a) indicates if a tendency exists for y (predicted) scores to increase or decrease as values of x (predictor) change, (b) can be used to estimate or predict values of y from known values of x, and (c) estimates the value of y when the value of x is zero (see exercise 6.3).

Regression describes the relationship between two or more variables in the form of an equation. This makes it possible to predict, for example, a student’s score based on an achievement test from what is known about the student’s home background or other variables.

In a typical national assessment, many variables are likely to correlate significantly with test scores in mathematics, language, or science. In this situation, one can use multiple regression to quantify the association between multiple independent variables and a dependent variable. Examples of dependent and independent variables that feature frequently in national assessments are provided in box 6.1.
Implementing Regression with One Dependent Variable and One Independent Variable

The sections that follow provide examples of how to conduct analysis using regression in WesVar and how to interpret the output. WesVar is used because it takes the complex nature of the sample in the national assessment into account (see chapter 3) when assessing significance levels (for example, differences arising from the clustering of students in schools and classes). The first example involves the simplest type of regression. It looks at the association between one dependent variable, mathematics achievement (Maths), which ranges from 88 to 400, and one independent variable, number of books in the home (Books), which ranges from no books (0) to 120.

The WesVar linear regression program asks you to select independent variables from two lists: Class Variables and Source Variables. The Class Variables list comprises categorical variables that have 255 or fewer response categories, excluding missing values (a feature of WesVar). The Source Variables are continuous variables. Some variables appear on both the Class and Source Variables lists. An example is Books. Here, however, we treat Books as a Source variable. See the example in exercise 6.3.

Regression can be used with categorical variables as well as with continuous variables. In the example in exercise 6.4, the independent variable is Region. Recall that in chapter 4, it was established that students in the Metro Area performed significantly better than students in the other three regions in Sentz (exercise 4.2). Note that, when a categorical variable such as Region is selected as the independent variable in a regression analysis, a series of variables

---

**BOX 6.1**

**Variables in Standard Regression**

*Single dependent (outcome) variable*, such as reading or mathematics achievement

*One or more independent (explanatory) variables*, such as class size, geographic region, teacher qualifications, parent education, student gender
Running a Regression in WesVar, One Independent Variable (Continuous)

1. Launch WesVar, and open the workbook used in exercise 6.2, NAEA DATA ANALYSIS\MY WESVAR FILES\CHAPTER 6 WORKBOOK.

2. Select Chapter 6 Exercises in the left panel, and click Regression in the right panel. Select Regression Request One (left panel). Type Exercise 6.3 (right panel).

3. Select Options in the left panel, and ensure that Linear has been checked. Within Options, select Generated Statistics – Confidence Interval. Within Output Control, set Estimates and Std. Error at three decimal places.

4. Select Models in the left panel. Drag Mathss from the Source Variables list to the box labeled Dependent. This is the dependent variable for your regression analysis. (If you can see only one or two variables in the boxes of the right panel, move your cursor to the bottom border and drag down to show more variables.)

5. Go to the Source Variables box and drag Books to the first space labeled Independent. Ensure that the Intercept box has been checked. Click Add as New Entry (exercise figure 6.3.A).
6. Run the regression analysis (click the Green Arrow in the toolbar), and open the output (click the Open Book icon in the toolbar). Expand Exercise 6.3, Models, and Mathss = Books. Click Sum of Squares (exercise figure 6.3.B). The R-square value (R_Square Value) is 0.099. This indicates that Books explained almost 10 percent of the variance in mathematics achievement test scores. The R-square value is obtained by dividing the model (explained) sum of squares by the total sum of squares.

7. Save this output as a text file in MY SOLUTIONS using File – Export – Single File – One Selection. Use the file name EXERCISE 6.3 SUM OF SQUARES.

8. Select Estimated Coefficients in the output file (exercise figure 6.3.C). This shows the parameter estimate or expected score change associated with the number of books in a student’s home. Apply the formula described earlier, \( \hat{y} = \alpha + bX \), to calculate the relationship or association between number of books and achievement in mathematics. Do not use \( \epsilon \) in these calculations.* Note that \( X \) represents the number of books and \( b \) is the slope of the regression line. The value of the estimate for books (0.515) indicates that an increase of one book is associated with a 0.515-point increase in performance in mathematics. The expected mathematics score of a student with no books is the value of the intercept\( b \) 225.196 or, using the formula, 225.196 + 0 * 0.515. The expected score for a student with 10 books is 230.346 (225.196 + 10 * 0.515). Hence, having 10 books at home is associated with a five-point increase in mathematics achievement. The mean score for Books is 48.2. Hence, the expected mathematics score for a student with an average number of books at home is 225.196 + 0.515 * 48.2, or 250.019, which is very close to the overall recorded mean mathematics score of 250.0.

(continued)
The \( t \) values in exercise figure 6.3.C are computed by dividing each estimate by its standard error. The \( t \) value is a measure of statistical significance and tests the likelihood that the actual value of the parameter is not zero. For Books, the \( t \) value is 11.449 and the probability (\( p \)) value (Prob>|T|) is at or close to zero (0.000). This indicates a very small chance exists that the actual value of the Books parameter is zero. The 95 percent confidence interval around a parameter is estimated approximately by adding twice its standard error to the parameter and subtracting twice its standard error. Thus, after rounding, the 95 percent confidence interval for Books is 0.425 to 0.605 (0.515 \( \pm \) 0.045). There is 95 percent certainty that the estimate for the value Books in the population is between 0.425 and 0.59. (These values are almost identical to those presented in exercise figure 6.3.C.)


10. Next, the “goodness of fit” of the statistical model, which estimates the value of Mathss on the basis of data from one independent variable, Books, is of interest. Select Tests (exercise figure 6.3.D). Note that the overall fit of the regression model is
must be created to indicate the region in which a student’s school is located. WesVar creates a series of dummy variables, each one corresponding to a single region, which are coded 1 or 0, depending on whether a student belongs to the region or not. For example, when WesVar creates a dummy variable **Northwest**, all students attending school in that region are coded 1, and students in each of the other three regions are coded 0. Similarly, students in schools in the Metro Area would be coded 1 on the dummy variable **Metro**, and students in each of the other three regions would be coded 0. The same applies to students in Eastern Highlands. The final region or category, which is known as the **reference category** and is not coded separately, is included in the analysis. In the example in exercise 6.4, where **Region** is the categorical (class) variable, the performance of students in each of the first three regions (Northwest, Metro Area, Eastern Highlands) is compared to that of students in the fourth region (Southwest Coast).
Running Regression in WesVar, One Independent Variable (Categorical)

1. Launch WesVar, and open the workbook used in exercise 6.3, NAEA DATA ANALYSIS\MY WESVAR FILES\CHAPTER 6 WORKBOOK.

2. Select Chapter 6 Exercises (left panel), and click Regression (right panel). Select Regression Request Two* (left panel). Type Exercise 6.4 (right panel). Under Options – Generated Statistics, check Confidence Interval. Under Options – Output Control, set the number of decimal points for estimate and standard error to three decimal places.

3. Select Models in the left panel.

4. Drag Mathss from the Source Variables list to the box labeled Dependent in the right panel. This is the dependent variable for your regression analysis.

5. Go to the Class Variables box and drag down Region[4] to the row for the first independent variable. Ensure that the Intercept box has been checked. Click Add as New Entry. Region[4] is taken from the Class Variables list here because it is a categorical variable (because each student was assigned to one of four regions).

6. Run the regression analysis (click the Green Arrow in the toolbar), and open the output (click the Open Book icon in the toolbar). Expand Exercise 6.4, Models, and Mathss = region[4]. Select Sum of Squares in the left panel. The R-square value is 0.054. This indicates that Region[4] explained or accounted for 5 percent of the variance in mathematics achievement test scores.

7. Click Estimated Coefficients in the left panel.

   The parameter estimates for three of the four regions are given (see exercise figure 6.4.A). The parameter estimate for the reference region is the intercept (251.248). This corresponds to the mean score for the Southwest Coast in exercise 4.2. The parameter estimate for Region.1 (Northwest) is −17.898 (that is, 17.898 points lower than in the reference category). Hence, the expected score of a student with average performance in the Northwest is 233.350 (251.248 − 17.898). This corresponds to the estimated mean score for the Northwest in exercise 4.2.

   The estimated score of a student performing at the average level in the Metro Area (Region.2) is 265.735 (that is, 14.487 points higher than that of a student in the reference region). Finally, the estimated score for a student performing at the average level in the Eastern Highlands (Region.3) is 249.108, which is 2.140 points lower than in the reference region. The nonsignificant t value associated with the parameter estimate for the Eastern Highlands (Prob|T| = 0.667) indicates that −2.14 is not significantly different from zero and, therefore, the performance of an average student in this region is not statistically significantly different from that of an average student in the reference region (Southwest Coast).
Implementing Multiple Regression with One Dependent Variable and Two or More Independent Variables

In this section, the effect of increasing the number of independent variables on explaining or accounting for the mathematics test results is described. Three independent variables are considered:

- **Books**, the number of books in a student’s home, a discrete variable
- **Distance**, the distance in kilometers between a student’s home and school, a continuous variable
- **Parented [6]**, the highest level of education achieved by either parent, a categorical variable with six categories: 1 = no formal education; 2 = grades 1–3; 3 = grades 4–6; 4 = grades 7–9; 5 = upper secondary; and 6 = degree

The intercorrelations among the independent variables should be scrutinized before running a regression. Of particular interest should be **multicollinearity**, which arises when two or more independent variables are strongly correlated. When this occurs, the standard errors of the regressions increase, making it more difficult to assess the unique role of each independent variable in explaining performance. Exercise 6.5 shows how correlation coefficients may be estimated in WesVar.
**EXERCISE 6.5**

**Estimating Correlation Coefficients**

1. Launch WesVar, and open the workbook used in exercise 6.4, NAEA DATA ANALYSIS MY WESVAR FILES CHAPTER 6 WORKBOOK.

2. Select Chapter 6 Exercises (left panel), and click Descriptive Stats (right panel). Select Descriptive Request Two in the left panel, and type Exercise 6.5 in the space in the right panel.

3. Select Options – Output Control, and set the number of decimal places for Estimates and Std. Error to three decimal places.


5. Run correlations (click the Green Arrow in the toolbar). Open the output (click the Open Book icon in the toolbar). Select and expand Exercise 6.5 in the left panel. Select Correlations – Overall.

The output data in the Weighted column (exercise figure 6.5.A) show that there is no evidence of multicollinearity, since none of the correlations approaches 0.80 (see Hutcheson and Sofroniou 1999). The negative correlation between Books and Distance (−0.077) indicates that as the distance between home and school tends to increase, the number of books in a student’s home tends to decrease. The correlation between Books and Parented (treated here as a continuous variable) is 0.331. This indicates that higher parental levels of education are associated with a larger number of books in the home.

**EXERCISE FIGURE 6.5.A Output for Correlations among Independent Variables**

Because the three variables are not too highly intercorrelated, you can now run a regression with one dependent variable (Mathss) and three independent variables (Books, Distance, Parented [6]) (exercise 6.6).

In summary, the data show that a model with three independent variables (Books, Distance, Parented [6]) explains 24 percent of the variation in mathematics achievement. The model suggests a positive association between number of books in a student’s home and mathematics performance, even after taking the other two variables into account. It also suggests that level of parent education is associated with student performance in mathematics; students whose parents have a higher level of education tend to have higher expected scores than students whose parents have a lower level, taking into account the other two factors (number of books and distance). Finally, the model indicates a negative association between distance from school and mathematics achievement, taking number of books and parent education into account; students living farther from their school tend to do less well than students who live near their school.
three-variable model accounts for 24 percent of the variance in mathematics achievement \( (R^2 = 0.242) \). This is an improvement on the earlier model in which Books, as a single independent variable, accounted for just under 10 percent of the variance in mathematics achievement (see exercise figure 6.3.B).

**EXERCISE FIGURE 6.6.A** WesVar Screen before Running Regression with More Than One Independent Variable

9. Select **Estimated Coefficients** in the output file (left panel, expanding as needed). The output gives parameter estimates for **Intercept**, **Books**, **Distance**, and five of the six levels of **Parented**. Note that all the parameters in the model are statistically significant; the value of Prob>|T|) is at or close to zero. This indicates a very low probability that any of the parameters is zero.

10. The parameter estimate for the **Intercept** is 278.909. This corresponds to the expected score of a student with no **Books** at home, who lives no **Distance** from the school (zero kilometers), and who has at least one parent at the highest level of parent education (**Parented.6**, the reference category for **Parented [6])**. The parameter estimate for **Books** is 0.309. This is the increment in achievement associated with one additional book in the home. Thus, a student with the average number of **Books** in the home (48.191), who lives less than one kilometer from the school, and who has at least one parent with a degree would have an estimated score of 278.909 + (0.309 * 48.191) + (−5.620 * 0), or 293.800. (Note that **Parented [6]** has a zero weight in this calculation because it is the reference category.)

11. The parameter estimate for **Distance** is −5.620. The negative sign means that the students’ expected mathematics scores decrease the farther they live from the school. The average distance from school is 4.257 kilometers.b Thus, the expected mathematics score of a student who lives 4.257 kilometers from the school, who has zero books in the home, and who has at least one parent at the reference category of **Parented [6]** is 278.909 + (−5.620 * 4.257) + (0.309 * 0), or 254.985.

12. Parameter estimates are provided for five levels of **Parented [6]**. As noted previously, **Parented.6** (at least one parent possesses a degree), the reference category, is the highest level of parental education. Negative parameter estimates are associated with lower levels of parental education. For example, the parameter estimate for **Parented.2** (grades 1–3 completed) is −30.156 (see exercise figure 6.6.C). Thus, the expected score of a student who has no books at home, who lives beside or very close to the school, and who has at least one parent who has completed grade 3 (**Parented.2**) is 278.909 − 30.156, or 248.753.

**EXERCISE FIGURE 6.6.C** Output for Regression in WesVar, More Than One Independent Variable: Estimated Coefficients
13. In a similar manner, you can generate expected scores for students having varying numbers of books in the home, living varying distances from the school, and having parents with varying levels of education. Thus, the expected score for a student who has 50 books in the home, who lives a distance of 5 kilometers from the school, and whose highest level of parental education is grades 7–9 (Parented.4) is equal to

\[ 278.909 + (0.309 \times 50) + (-5.620 \times 5) + (-15.615) \], or 250.644.

14. Save the output in **MY SOLUTIONS** using File – Export – Single File – One Selection. Use the file name **EXERCISE 6.6 ESTIMATES.TXT**.

15. Select **Tests** in the output. Here you can see (first row of exercise figure 6.6.D) that the overall fit of the regression model is statistically significant, with the probability of obtaining an F value of 134.636 approaching zero (0.000). This means that at least one of the variables **Books**, **Distance**, or **Parented [6]** is statistically significant. The p-values for **Books**, **Distance**, and **Parented [6]** show that for each of these variables, the regression coefficient differs significantly from zero, after controlling for the others. For example, **Distance** is statistically significant even after taking into account the effects of **Books** and **Parented [6]**.

**EXERCISE FIGURE 6.6.D** Output for Regression in WesVar, More Than One Independent Variable: Test of Model Fit

16. Save this output as a text file in **My Solutions** using File – Export – Single File – One Selection. Use the file name **EXERCISE 6.6 TESTS.TXT**.

17. Return to your workbook (via the Exit Door icon in the toolbar). Select File – Save and then File – Close. Your workbook should save to **NAEA DATA ANALYSIS\MY WESVAR FILES\CHAPTER 6 WORKBOOK**.

a. Note that **Parented** also appears in the **Source Variables** list. In this instance, however, you must select **Parented [6]** from the **Class Variables** list because it is treated as a categorical variable in a regression.

b. To calculate this, select **Descriptives** in WesVar, and move **Distance** from **Source Variables** to **Selected**, according to the procedure outlined in exercise 3.1.
CORRELATION AND CAUSATION

In most national assessments, information is obtained about a range of personal and situational variables in questionnaires completed by students, teachers, and (sometimes) parents. Variables are usually selected in the belief that they are related to student achievement (whether supported by research or not). Policy makers may also include variables to guide the choice of interventions designed to improve student achievement.

The association between a background variable and student achievement can be represented in a correlation. The finding that variables are correlated (even strongly) in a national assessment, however, does not mean that one variable is the cause of another. There are a number of reasons for saying this. First, most national assessments are cross-sectional in nature. Data relating to background factors and achievement are collected at the same time. Thus, the temporal sequence between events, in which the cause precedes the effect, which is normally required to support an inference of causality, will not be present. This problem may be circumvented when some background variables describe past events (for example, the amount of instructional time spent teaching a curriculum area). The issue is also addressed in the rare cases in which data on students are collected at more than one point in time. Taking account of student characteristics at an earlier stage in their educational careers (including earlier achievements and background data) will strengthen the inferences that can be made relating to the effects of their school experiences (Kellaghan, Greaney, and Murray 2009).

A second problem in identifying causal relationships in a national assessment is that the factors that affect student achievement are complex and interrelated. A vast literature has identified a wide range of factors associated with achievement, including personal characteristics of students, family and community factors, and instructional practices of individual teachers. Statistical analyses in which a single variable is related to achievement not only will fail to recognize this complexity but may also result in a misleading conclusion. A simple example will serve to illustrate this point. A finding that students in private schools have higher levels of achievement than students in
public schools might be interpreted to mean that the education provided in private schools is superior to that provided in public schools. However, such a conclusion would need to be modified if measures of other factors, such as students’ level of achievement when they enrolled in school or their home circumstances (for example, parents’ socioeconomic level), were included in analyses.

The complexity of the factors that affect student achievement is addressed in multiple regression analysis when the “effects” of a variable are isolated by the systematic removal or adjustment of the “effects” of other variables. An inference relating to causality is strengthened if significant relationships are found to exist following adjustment.

A variety of more elaborate regression methods (a description of which is beyond the scope of the present work), in which patterns of correlations between predictor and outcome variables are examined to identify causal pathways, serve to further strengthen inferences regarding causality. These methods are designed to identify variables that can be considered moderators, which affect the direction and strength of relationship between criterion and independent variables (for example, gender, age, socioeconomic status), and mediators, which speak to how or why effects occur (Bullock, Green, and Ha 2010).

Even the most sophisticated methods of analysis, however, fail to address a further problem that arises in inferring causation on the basis of data obtained in a cross-sectional study (such as a national assessment). For instance, data may not be available on important, often uncontrollable and obscure, information that is likely to affect student achievement. This issue can be addressed only in randomized, controlled research designs in which the direct influence of one condition on another is studied when all other possible causes of variation are eliminated.

Although making causal inferences based on national assessment data is problematic, the information relating to correlates of achievement revealed in an assessment can still have pragmatic value. Data collected in an assessment relating to, for example, gender, location, or ethnic group membership can act as important signals, identifying problem areas in the education system that merit attention in policy formation, possibly pointing to intervention or remedial action. Determining the nature of such intervention or remedial action,
however, requires a consideration of local conditions and resources, the involvement of stakeholders, and a consideration of relevant research findings (see Kellaghan, Greaney, and Murray 2009).

NOTES

1. If a variable label (Implement – Percent Correct) appears in the left column of Simple Scatterplot instead of the variable name (Impl_pc), select Cancel and then select Edit – Options – General – Variable Lists – Display Names – Apply – OK (to message) – OK.

2. Swapping the two would yield different best-fit lines. That is, the line that best predicts y from x is not the same as the line that best predicts x from y.

3. The error term or “noise” reflects the fact that factors in addition to those included in the regression model influence the dependent variable.

4. In preparing to run a regression, the analyst should also look for outliers or extreme scores on a variable (see chapter 2) because they can distort results. It is also important to determine if any of the variables in the regression equation are highly skewed. Outliers or extreme scores can contribute to distortions of parameters and statistical estimates.
“A picture is worth a thousand words.” Many readers of national assessment reports are likely to more readily understand data when presented in the form of a chart or graph than when presented in table format. Earlier chapters contained scatterplots, histograms, and box plots produced by computer programs during the course of various analyses. (For example, exercise 6.1 demonstrated how SPSS constructs a scatterplot.) This chapter provides examples of a number of graphical approaches to preparing and presenting national assessment results. The focus is on charts and line graphs that can be produced in Excel and can be constructed and pasted into a national assessment report. As you become familiar with these techniques, you will come across additional software that can be used to present data in pictorial form. Charts and graphs are not intended to replace tables. They display the data in a different way. If included in a report, the corresponding tables—the data tables on which the charts and graphs are based—should also be included, either in the body of the text or in an appendix (for example, an electronic appendix). The charts in the following section are two dimensional (2-D).
CHARTS

A column chart has vertical rectangular bars of lengths usually proportional to the frequencies they represent. For example, you can use a column chart to compare the populations in four regions or the average scores of students in different types of school in a specific curriculum area.

In exercise 7.1, data on students’ proficiency levels in mathematics are presented in the form of a column chart, where the height of the bar indicates the percentage of students at each level. The data are based on a national assessment that reported results using five proficiency levels (levels 1–4 and below level 1). The procedure for creating a column chart in Excel is also described in exercise 7.1.1.

You can also report the percentage of students at each proficiency level in each region in a country. This time, the regional data on proficiency levels will be represented by a bar chart (see exercise 7.2), which is very similar to a column chart, except that it represents data in a horizontal manner.

**EXERCISE 7.1**

**Drawing a Column Chart to Show Performance by Proficiency Level, National Data**

1. Arrange the data on an Excel sheet (exercise figure 7.1.A). Copy your data from an existing table, or simply type the data into an Excel sheet.
2. Highlight the data (exercise figure 7.1.A). On the menu bar, select Insert and Column (exercise figure 7.1.B).a

![EXERCISE FIGURE 7.1.A](image)

<table>
<thead>
<tr>
<th>Level</th>
<th>Percentage of students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below level 1</td>
<td>19.8</td>
</tr>
<tr>
<td>Level 1</td>
<td>20.4</td>
</tr>
<tr>
<td>Level 2</td>
<td>23.6</td>
</tr>
<tr>
<td>Level 3</td>
<td>19.2</td>
</tr>
<tr>
<td>Level 4</td>
<td>17.1</td>
</tr>
</tbody>
</table>
3. Select **2-D Column** and **Clustered Column**—the first column chart. To label the vertical (y) axis, place the cursor on the left where you see the values (0 to 25) and right-click. On the toolbar, select **Chart Tools** (top of screen) – **Layout** – **Axis Titles** – **Primary Vertical Axis Title** – **Rotated Title**. Click on the box next to the vertical axis, highlight the text in the box, and type *Percentage of students* (exercise figure 7.1.C).

4. To label the horizontal (x) axis, place the cursor at the bottom of the chart and right-click. (The different proficiency levels will be highlighted.) On the toolbar, select **Chart Tools** (top of screen) – **Layout** – **Axis Titles** – **Primary Horizontal** – **Axis Title** – **Title Below Axis**. Highlight the text (Axis Title) in the box, and type *Proficiency levels* (exercise figure 7.1.C).

**EXERCISE 7.1 (continued)**

**EXERCISE FIGURE 7.1.B**  Insert Chart Options in Excel

**EXERCISE FIGURE 7.1.C**  Percentage of Students at Each Mathematics Proficiency Level

(continued)
EXERCISE 7.1 (continued)

5. If Excel has assigned a chart title, you may wish to change it by highlighting the title and changing the wording. To assign a title if none has been provided, place the cursor on the chart and click to highlight the chart. On the toolbar, select Chart Tools (top of screen) – Layout – Chart Title – Above Chart. Highlight the text, and type *Percentage of Students at Each Mathematics Proficiency Level*. The title has been placed above the chart in exercise figure 7.1.C.

6. You can alter various features of the chart (such as font size and style) by clicking in the appropriate area on the graph and using the editing tool (right-click) or by using Design and Layout (under Chart Tools). If you want to show the values (percentages) on each column, right-click on one of the columns (this should highlight all of them). Then select Chart Tools – Layout – Data Labels – Outside End. This shows the percentage of students at each proficiency level.

7. Highlight the chart area in Excel by clicking on the perimeter and select Copy. Paste the chart into your report.

8. Save your Excel worksheet in MY SOLUTIONS, using the file name EXERCISE 7.1.XLS.

a. Earlier versions of Excel may take you to the Chart Wizard where you can also select a column chart.

---

EXERCISE 7.2

**Drawing a Bar Chart to Show Percentage at Each Proficiency Level by Region**

1. Copy the data in exercise figure 7.2.A to an Excel file.

2. Highlight the data, as shown in exercise figure 7.2.A. Click Insert (or select Insert – Charts) – Bar – 2-D Bar – 100% Stacked Bar (third option) (exercise figure 7.2.B). This chart allows you to compare the percentages of students across proficiency levels and regions. You may modify various features of this chart to make it more presentable.

- If your output (see exercise figure 7.2.C) presents the regions in the incorrect order (for example, Southwest Coast on top instead of Northwest), click on the chart and

---

**EXERCISE FIGURE 7.2.A** Percentage of Students at Each Mathematics Proficiency Level by Region

<table>
<thead>
<tr>
<th>Region</th>
<th>Below level 1</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northwest</td>
<td>27.8</td>
<td>23.8</td>
<td>22.9</td>
<td>16.8</td>
<td>8.7</td>
</tr>
<tr>
<td>Metro area</td>
<td>13.5</td>
<td>13.9</td>
<td>16.9</td>
<td>23.8</td>
<td>31.9</td>
</tr>
<tr>
<td>Eastern highlands</td>
<td>17.3</td>
<td>21.6</td>
<td>23.1</td>
<td>25.9</td>
<td>12.1</td>
</tr>
<tr>
<td>Southwest coast</td>
<td>18</td>
<td>16.9</td>
<td>24.7</td>
<td>30.5</td>
<td>9.9</td>
</tr>
</tbody>
</table>

- If the levels appear on the y-axis, and the regions on the x-axis (that is, the reverse of what is shown in exercise figure 7.2.C), return to Excel. Right-click on the chart area. Select Data. Click Switch Row/Column (exercise figure 7.2.D).
You can, if you wish, show the percentages of students at each proficiency level in each region. Click on the Below Level 1 bar in the first region. This will highlight the Below Level 1 bars for all regions. Right-click and select \textbf{Add Data Labels} (or select \\textit{Layout – Labels – Data Labels – Center}). Repeat the process for the other levels (exercise figure 7.2.E).

- If the horizontal axis (with the percentage figures) appears at the top of the chart rather than the bottom, select the horizontal axis (the axis with the percentages) by right-clicking on any label. Then right-click \textit{Axis Labels Hi} or \textit{Lo} (depending on the current position of the percentage figures) and select \textbf{Format Axis – Horizontal Axis Crosses – Automatic}.

- If the regions are in reverse order, select the vertical axis by left-clicking on any label. Then right-click \textbf{Format Axis} and check \textbf{Categories in Reverse Order}.

- If you wish to add a title to the x- or y-axes, highlight the area of the chart in which the percentages appear. Then select \textit{Insert – Layout – Axis Titles – Primary Horizontal Axis Title – Title Below Axis}. Highlight \textit{Axis Title}, and type the new title (for example, \textit{Percentage of students}) in the formula bar or directly into the newly created text box. Follow the same procedure to add a title to the other axis. To add a title to the chart, follow the instructions in step 5 in exercise 7.1

3. Highlight the chart area in Excel by clicking once on the perimeter; select \textbf{Home – Copy}. Then paste your chart into your report, giving it a suitable title.
LINE GRAPHS WITH CONFIDENCE INTERVALS

A graph that shows a series of mean scores, together with their 95 percent confidence intervals, allows the nontechnical reader to learn which regions performed well, compared to others, and also whether a mean score and its confidence interval in one region overlaps with a mean score and its confidence interval in another region.
1. Launch Excel. Enter the data manually into the Excel sheet, arranging it as shown in exercise figure 7.3.A. Alternatively, you can transfer these data into the Excel file from NAEA DATA ANALYSIS\MY SOLUTIONS\EXERCISE 3.3.TXT. This will require you to delete some unnecessary data (such as standard errors and the rows that give percentage data) from the Excel sheet.

2. Highlight the data, as indicated in exercise figure 7.3.A. Select Insert – Other Charts (or Insert Chart) – Stock – High-Low-Close (first option).

3. Highlight the numbers on the vertical axis. Right-click, select Format – Axis. Then, under Axis Options, check Fixed for Minimum (or Scale) and insert 220. Likewise set Maximum at 280 (exercise figure 7.3.B). Click Close (or OK). Note these values are just below and above the highest values in exercise figure 7.3.A.

## EXERCISE 7.3

### Drawing 95 Percent Confidence Intervals for a Series of Mean Scores

1. Launch Excel. Enter the data manually into the Excel sheet, arranging it as shown in exercise figure 7.3.A. Alternatively, you can transfer these data into the Excel file from NAEA DATA ANALYSIS\MY SOLUTIONS\EXERCISE 3.3.TXT. This will require you to delete some unnecessary data (such as standard errors and the rows that give percentage data) from the Excel sheet.

2. Highlight the data, as indicated in exercise figure 7.3.A. Select Insert – Other Charts (or Insert Chart) – Stock – High-Low-Close (first option).

3. Highlight the numbers on the vertical axis. Right-click, select Format – Axis. Then, under Axis Options, check Fixed for Minimum (or Scale) and insert 220. Likewise set Maximum at 280 (exercise figure 7.3.B). Click Close (or OK). Note these values are just below and above the highest values in exercise figure 7.3.A.

## EXERCISE FIGURE 7.3.A

Mathematics Mean Scores and Scores at Upper and Lower Confidence Intervals by Region

<table>
<thead>
<tr>
<th>Region</th>
<th>Upper 95% CI</th>
<th>Mean</th>
<th>Lowest 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northwest</td>
<td>239.9</td>
<td>233.3</td>
<td>226.8</td>
</tr>
<tr>
<td>Metro area</td>
<td>274.7</td>
<td>265.7</td>
<td>256.8</td>
</tr>
<tr>
<td>Eastern highlands</td>
<td>256.3</td>
<td>249.1</td>
<td>241.9</td>
</tr>
<tr>
<td>Southwest coast</td>
<td>257.9</td>
<td>251.2</td>
<td>244.6</td>
</tr>
</tbody>
</table>

Note: CI = confidence interval.
4. On the Excel chart, left-click on the lower limit on the line for the first (Northwest) region. This should highlight the lower limit for all four regions. On the menu bar, select **Format Selection – Marker Options** (or **Marker Style**) – **Built In – Type**. Select a suitable option (such as “–”) from the selection under **Type** (exercise figure 7.3.C) (or use the **Style** dropdown window). Increase **Size** to 10 or higher, ensuring that the size is sufficiently large to stand out in the graph. Click **Close**. Repeat the process for the upper limits.

5. Highlight the centerpoint of each line (the mean score) by clicking on the line for the first region and follow the same procedure as in step 4 to select a suitable marker. Use size 20 to distinguish the mean score marker from the other markers. Note that the legend on the left side of the chart is produced automatically, based on your initial spreadsheet. You have to label the axes and the series boxes on the right side in addition to deciding on a chart title.

Copy the chart (exercise figure 7.3.D) into your report, using **Copy** and **Paste**.

6. Save the Excel sheet in **NAEA DATA ANALYSIS\MY SOLUTIONS** using a suitable name (such as **EXERCISE 7.3.XLS**).

The graph in exercise figure 7.3.D describes differences in mean performance between regions. Where no overlap occurs between confidence intervals, one can say that mean
scores are significantly different. Thus, no overlap occurs between the 95 percent confidence interval for the Northwest and any of the other regions. Average mathematics scores in the Eastern Highlands and the Southwest Coast do not differ significantly from each another (because substantial overlap occurs between their respective lines).

a. As noted in chapter 3, WesVar uses 2.0 rather than 1.96 in computing 95 percent confidence intervals.

**LINE GRAPHS TO REPRESENT TREND DATA**

Exercise 7.4 illustrates how a line graph can be used to present trend data. The graph summarizes the performance of boys and girls for each of the four years in which a national assessment was administered (2004, 2007, 2010, and 2013).
EXERCISE 7.4

Showing Trend Data with a Line Graph

1. Launch Excel. Arrange the data in an Excel worksheet (exercise figure 7.4.A).*

2. Highlight the data and labels to be included in the graph (exercise figure 7.4.A). On the toolbar, select **Insert – Line – Line with Markers** (first option, second row) to create the line graph shown in exercise figure 7.4.B. You can modify the style of the lines by clicking on one of them. Select **Format – Data Series** (or under **Current Selection – Format Selection**) – **Line Style**. Use the **Dash Type** down arrow to select the line style you want. Click **Close**.^

EXERCISE FIGURE 7.4.A  Excel Worksheet with Mean Mathematics Scores by Gender, 2004–13

<table>
<thead>
<tr>
<th></th>
<th>Boys</th>
<th>Girls</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>225</td>
<td>245</td>
</tr>
<tr>
<td>2007</td>
<td>240</td>
<td>245</td>
</tr>
<tr>
<td>2010</td>
<td>245</td>
<td>250</td>
</tr>
<tr>
<td>2013</td>
<td>250</td>
<td>250</td>
</tr>
</tbody>
</table>

EXERCISE FIGURE 7.4.B  Mean Mathematics Scores by Gender, 2004–13

(continued)
3. You can add titles for the x- and y-axes, and for the chart as a whole, following the procedure in exercise 7.1. To add a title to the y-axis, click on the adjacent scale (210 to 255). Select Chart Tools – Layout (or Chart Layout) – Axis Titles (or Labels – Axis Titles) – Primary Vertical Axis Title – Rotated Title. Click on the space for axis title and type *Mean Mathematics Score*.

4. In a similar manner, label the horizontal x-axis *Year*.

5. Use the Chart Title icon to create a label title (such as *Mean Mathematics Scores by Gender, 2004–13*).

6. Copy the chart (exercise figure 7.4.B) into your report. Then save the Excel sheet using a suitable name (*NAEA DATA ANALYSIS\MY SOLUTIONS\EXERCISE 7.4.XLS*).


   a. Note that the data in exercise figure 7.4.A are not drawn from the *NATASSESS4.VAR* data file.
   b. This is *Marked Line* in some versions of Excel.
   c. Some versions of Excel may require: Format – Data Series – Line – Weights & Arrows – Dashed. Use the drop-down arrow to select the chosen line style. Click OK.

**NOTE**

1. Instructions may vary somewhat depending on which version of Excel is used. Excel for Mac and Excel for Windows have some slight differences in capability, depending on the platform and version used.
ANNEX I.A

NAEA DATA ANALYSIS: FILE DIRECTORY STRUCTURE

<table>
<thead>
<tr>
<th>Folder</th>
<th>Subfolder</th>
<th>File</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAEA DATA ANALYSIS</td>
<td>SPSS DATA</td>
<td>NATASSESS NATASSESS4 SCHOOLS</td>
</tr>
<tr>
<td></td>
<td>EXERCISE SOLUTIONS</td>
<td>LISTED IN ANNEX I.B</td>
</tr>
<tr>
<td></td>
<td>WESVAR UNLABELED DATA</td>
<td>DATA FILE CREATED (ANNEX I.C)</td>
</tr>
<tr>
<td></td>
<td>MY WESVAR FILES</td>
<td>EMPTY INITIALLY</td>
</tr>
<tr>
<td></td>
<td>WESVAR DATA &amp; WORKBOOKS</td>
<td>LISTED IN ANNEX I.B</td>
</tr>
<tr>
<td></td>
<td>MY SPSS DATA</td>
<td>EMPTY INITIALLY</td>
</tr>
<tr>
<td></td>
<td>MY SOLUTIONS</td>
<td>EMPTY INITIALLY</td>
</tr>
</tbody>
</table>
## NAEA DATA ANALYSIS: SUBFOLDERS AND FILES

<table>
<thead>
<tr>
<th>Subfolders</th>
<th>Files</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SPSS DATA</strong></td>
<td>NATASSESS.SAV</td>
<td>File includes identifiers, background variables, scores, participation status, and design weight for each selected student, plus class, school, and regional identifiers.</td>
</tr>
<tr>
<td></td>
<td>NATASSESS4.SAV</td>
<td>File includes identifiers, background variables, scores, participation status, and design weight for each selected student, plus class, school, and regional identifiers. Also included are jackknife (JK) strata (zones) and JK replicates (indicators).</td>
</tr>
<tr>
<td></td>
<td>SCHOOLS.SAV</td>
<td>File includes school identification numbers, JK strata (zones), and JK replicates (indicators). Merge with NATASSESS.SAV to create NATASSESS4.SAV.</td>
</tr>
<tr>
<td><strong>EXERCISE SOLUTIONS</strong></td>
<td>EXERCISE 1.1.SPV</td>
<td>SPSS descriptive statistics for Mathss (mean, standard deviation) using Descriptives.</td>
</tr>
<tr>
<td></td>
<td>EXERCISE 2.1.SPV</td>
<td>SPSS descriptive statistics for Mathss (various) using Explore (national).</td>
</tr>
<tr>
<td></td>
<td>EXERCISE 2.2.SPV</td>
<td>SPSS descriptive statistics for Mathss (various) using Explore (by region).</td>
</tr>
<tr>
<td></td>
<td>EXERCISE 3.1.TXT</td>
<td>Mathss WesVar descriptive statistics.</td>
</tr>
<tr>
<td></td>
<td>EXERCISE 3.2.TXT</td>
<td>Mathss mean score, standard error, and confidence interval (national).</td>
</tr>
</tbody>
</table>

(continued)
<table>
<thead>
<tr>
<th>Subfolders</th>
<th>Files</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXERCISE 3.3.TXT</td>
<td>Mathss mean scores, standard errors, and confidence intervals (by region).</td>
<td></td>
</tr>
<tr>
<td>EXERCISE 4.1A.TXT</td>
<td>Mathss mean scores of students with and without electricity at home.</td>
<td></td>
</tr>
<tr>
<td>EXERCISE 4.1B.TXT</td>
<td>Mathss mean scores and confidence intervals of students with and without electricity at home (two-group comparison).</td>
<td></td>
</tr>
<tr>
<td>EXERCISE 4.2A.TXT</td>
<td>Mathss mean scores (by region).</td>
<td></td>
</tr>
<tr>
<td>EXERCISE 4.2B.TXT</td>
<td>Statistical significance of Mathss regional mean score differences (more than two-group comparison).</td>
<td></td>
</tr>
<tr>
<td>EXERCISE 5.1.TXT</td>
<td>Mathss percentile scores (national).</td>
<td></td>
</tr>
<tr>
<td>EXERCISE 5.2.TXT</td>
<td>Mathss percentile scores (by region).</td>
<td></td>
</tr>
<tr>
<td>EXERCISE 5.4-25TH.TXT</td>
<td>Mathss percentage of pupils below 25th percentile (standard errors [SEs] and confidence intervals [CIs]) (by region).</td>
<td></td>
</tr>
<tr>
<td>EXERCISE 5.4-50TH.TXT</td>
<td>Mathss percentage of pupils below 50th percentile (SEs and CIs) (by region).</td>
<td></td>
</tr>
<tr>
<td>EXERCISE 5.4-75TH.TXT</td>
<td>Mathss percentage of pupils below 75th percentile (SEs and CIs) (by region).</td>
<td></td>
</tr>
<tr>
<td>EXERCISE 6.1.SPV</td>
<td>Scatterplot of mathematics Implementation and Analysis/Problem Solving scores.</td>
<td></td>
</tr>
<tr>
<td>EXERCISE 6.2.TXT</td>
<td>Correlation between mathematics Implementation and Analysis/Problem Solving.</td>
<td></td>
</tr>
<tr>
<td>EXERCISE 6.3 SUM OF SQUARES.TXT</td>
<td>Regression sum of squares and R-square value, one independent variable (Books).</td>
<td></td>
</tr>
<tr>
<td>EXERCISE 6.3 ESTIMATES.TXT</td>
<td>Regression estimated coefficient, one independent variable (Books).</td>
<td></td>
</tr>
<tr>
<td>EXERCISE 6.3 TESTS.TXT</td>
<td>Regression, significance tests, one independent variable (Books).</td>
<td></td>
</tr>
<tr>
<td>EXERCISE 6.4 ESTIMATES.TXT</td>
<td>Estimated regression coefficients, one independent variable (Books).</td>
<td></td>
</tr>
<tr>
<td>EXERCISE 6.5 CORRELATIONS.TXT</td>
<td>Intercorrelations among three variables (Books, Parented, Distance).</td>
<td></td>
</tr>
<tr>
<td>EXERCISE 6.6 SUM OF SQUARES .TXT</td>
<td>Regression sum of squares, and R-square value (for three independent variables).</td>
<td></td>
</tr>
<tr>
<td>EXERCISE 6.6 ESTIMATES.TXT</td>
<td>Regression estimate coefficients (for three independent variables).</td>
<td></td>
</tr>
<tr>
<td>EXERCISE 6.6 TESTS.TXT</td>
<td>Regression significance tests (for three independent variables).</td>
<td></td>
</tr>
<tr>
<td>EXERCISE 7.1.XLS</td>
<td>Chart: Percentage of students at each mathematics proficiency level (national).</td>
<td></td>
</tr>
<tr>
<td>Subfolders</td>
<td>Files</td>
<td>Description</td>
</tr>
<tr>
<td>-----------------</td>
<td>--------------------------------</td>
<td>---------------------------------------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td><strong>EXERCISE 7.2.XLS</strong></td>
<td>Chart: Percentage of students at each mathematics proficiency level (by region).</td>
</tr>
<tr>
<td></td>
<td><strong>EXERCISE 7.3.XLS</strong></td>
<td>Line graph: <strong>Maths</strong> Mean Scores and 95% Confidence Intervals (by region).</td>
</tr>
<tr>
<td></td>
<td><strong>EXERCISE 7.4.XLS</strong></td>
<td>Line graph: <strong>Maths</strong> Mean Scores (by gender).</td>
</tr>
<tr>
<td><strong>WESVAR UNLABELED DATA</strong></td>
<td><strong>NATASSES4.VAR</strong> <em>(NATASSES4.LOG)</em></td>
<td>Data file contains identifiers, background variables, scores, participation status, and design weight for each selected student, with class, school, and region identifiers; JK strata (zones); and JK replicates (indicators). Variable labels and created variables are not included.</td>
</tr>
<tr>
<td><strong>MY WESVAR FILES</strong></td>
<td>Subfolder is initially empty. User saves WesVar data files and workbooks created when completing exercises in chapters 3–6 in this subfolder.</td>
<td></td>
</tr>
<tr>
<td><strong>WESVAR DATA &amp; WORKBOOKS</strong></td>
<td><strong>NATASSES4.VAR</strong> <em>(NATASSES4.LOG)</em></td>
<td>Data file contains identifiers, background variables, scores, participation status, and design weight for each selected student, with class, school, and region identifiers; JK strata (zones); and JK replicates (indicators). Also included are some variable labels and created variables.</td>
</tr>
<tr>
<td></td>
<td><strong>CHAPTER 3 WORKBOOK.WVB</strong></td>
<td>Files are completed workbooks for running the exercises in chapters 3–6.</td>
</tr>
<tr>
<td></td>
<td><strong>CHAPTER 4 WORKBOOK.WVB</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>CHAPTER 5 WORKBOOK.WVB</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>CHAPTER 6 WORKBOOK.WVB</strong></td>
<td></td>
</tr>
<tr>
<td><strong>MY SPSS DATA</strong></td>
<td>Subfolder is initially empty. User saves new or modified SPSS data files in this subfolder.</td>
<td></td>
</tr>
<tr>
<td><strong>MY SOLUTIONS</strong></td>
<td>Subfolder is initially empty. User saves all exercise solutions in this subfolder.</td>
<td></td>
</tr>
</tbody>
</table>
Once you have prepared a national assessment data set in SPSS containing test item and questionnaire data and computed survey weights, you will have to make a few data modifications before transferring your file to WesVar.

The following set of instructions guides you through the creation of jackknife weights for a two-stage survey design. You can find more detailed information on jackknifing in volume 3 of this series (Dumais and Gough 2012b, annexes IV.C and IV.D).

SPSS is used to create the important sampling information that WesVar requires to create the replicate weights for the analysis of the national assessment data. Because the replicate weights are created for schools, you must create one record for each participating school. This record must include the actual Schoolid for each school. If you do not have a file containing one ID per school, you should use the following procedure to create one.

If you have not already transferred NATASSESS.SAV to MY SPSS DATA, you should copy it from SPSS DATA into MY SPSS DATA. (Note that this is NATASSESS.SAV not NATASSESS4.SAV.)

1. Go to NAEA DATA ANALYSIS\MY SPSS DATA and open the file NATASSESS.SAV. This can be done by double-clicking on the
file name. If this file is not in MY SPSS DATA, open it in NAEA DATA ANALYSIS\SPSS DATA and save it to NAEA DATA ANALYSIS\MYSPSS DATA using File – Save As.

2. On the SPSS toolbar, select Data – Aggregate. Move Schoolid into Break Variable(s). Move MATHRS into the Aggregated Variables box (figure I.C.1). Under Save, select Write a new data file using only the aggregated variables. Select File (just below this). Select NAEA DATA ANALYSIS\MY SPSS DATA as the directory to which to save the new file, using the file name SCHOOLS.

3. Under Options for Very Large Datasets, check Sort file before aggregating. Click OK.

4. Go to NAEA DATA ANALYSIS\MY SPSS DATA and open Schools. Open Variable View at the bottom, and delete the row Mathrs_Mean using Edit – Clear. Select Data – Sort Cases, and move Schoolid into the Sort by box; select Ascending, and click OK. (If you have to move from the output screen in SPSS, select Window on the toolbar and then the relevant data set.)

5. In this step, the values 1 or 2 are assigned to each participating school.1 Select Transform – Compute Variable, and type Jkzone under Target Variable. Then type \texttt{RND($Casenum/2)} under Numeric Expression, and click OK. Next, select Transform – Compute Variable. Select Reset. Then type Randompick under Target Variable and \texttt{RV.Uniform(0,1)} under Numeric Expression. Click OK. At this point, in Data View, you should see 120 schools in 60 pairs numbered from 1 to 60, each school also bearing a random number between 0 and 1. If the random numbers are displayed as zeros and ones, increase the number of decimal places from the Variable View tab. Now you are ready to create the JK replicates.

6. Steps 5 and 6 allocate one of two values to each school within its pair (zone). Select Data – Sort Cases, then move Jkzone and Randompick to Sort by. Click Ascending and OK. Now select Data – Identify Duplicate Cases, and move Jkzone to Define matching cases by. In Variables to Create, check Indicator of
Primary Cases, and check Last case in each group is primary (PrimaryLast). Click OK.

7. Because WesVar expects replicates to be numbered from 1, not from 0, replicate codes need to be modified using Transform – Recode into Different Variables. Move PrimaryLast to Input
Variable, and type \texttt{Jkindic} in \texttt{Output Variable Name}.\textsuperscript{2} If you wish, type in a label. Click \texttt{Change} and then \texttt{Old and New Values}. In \texttt{Old Value}, click \texttt{Value} and type \texttt{0} under \texttt{Value}. In \texttt{New Value} (on the right side), type \texttt{1} and click \texttt{Add}. Under \texttt{Old Value}, click \texttt{All other values}. Now, in \texttt{New Value}, type \texttt{2}. Click \texttt{Add}, \texttt{Continue}, and \texttt{OK}. All \texttt{PrimaryLast} \texttt{0} values have been transformed to \texttt{Jkindic} values of \texttt{1}, and all values of \texttt{1} have become \texttt{2}. (Note that whether or not the actual values of 0 and 1 appear on your screen for the variable \texttt{PrimaryLast} depends on your SPSS data settings.) Choose \texttt{Data – Sort Cases} from the menu. Select \texttt{Reset}. Move \texttt{Schoolid} into the \texttt{Sort by} box; click \texttt{Ascending} and \texttt{OK}.

8. Open \texttt{Variable View}, and delete the variables \texttt{RANDOM PICK} and \texttt{PrimaryLast}. Select \texttt{File – Save} to resave the data file \texttt{SCHOOLS}. Do not close the file. At this point, the JK zone and JK indicator (replicate) numbers have been assigned to the participating schools. This information now has to be attached to \texttt{NATASSESS.SPV}, the file where the survey data and weights have been saved.

9. Open the file \texttt{NAEA DATA ANALYSIS\MY SPSS DATA\NATASSESS.SAV}. Select \texttt{Data, Sort Cases}, and move \texttt{Schoolid} into the \texttt{Sort by} box. Select \texttt{Ascending} and \texttt{OK}.

10. Next, the variables \texttt{Jkzone} and \texttt{Jkindic} (currently in the \texttt{SCHOOLS.SAV} file) are merged into the \texttt{NATASSESS.SAV} data file. Still in \texttt{NATASSESS.SAV}, select \texttt{Data – Merge files – Add variables}. Choose \texttt{SCHOOLS.SAV} from \texttt{Open Dataset}, and click \texttt{Continue}. Check the box beside \texttt{Match cases on key variables in sorted files}. Highlight \texttt{SCHOOLID}, and move it to \texttt{Key Variables} (figure I.C.2). Select \texttt{Non-active dataset is keyed table}.

11. Click \texttt{OK}. You will receive a warning, \texttt{Keyed Match will fail if data are not sorted in ascending order of Key Variables}. If both files (\texttt{NATASSESS} and \texttt{SCHOOLS}) have been sorted by \texttt{Schoolid}, click \texttt{OK}. In \texttt{NATASSESS.SAV}, open the \texttt{Variable View}. The variables \texttt{Jkzone} and \texttt{Jkindic} should be visible toward the end of the data set. Select \texttt{Save As} and save \texttt{NATASSESS.SAV} as \texttt{NATASSESS4.SAV} in \texttt{NAEA DATA ANALYSIS\MY SPSS DATA}.
12. Close the open files `SCHOOLS.SAV` and `NATASSESS4.SAV`.

13. Now `NATASSESS4.SAV` is brought into WesVar and transformed into a WesVar data file. Launch WesVar. Click `New WesVar Data File`. Under `Input Database`, select `NAEA DATA ANALYSIS\MY SPSS DATA\NATASSESS4.SAV`. Click `Open`. You will see a message: `Create extra formatted variables`. Click `Done`.

14. In `Source Variables`, highlight `Wgtpop` (the variable that weights the data according to their representation in the population). Select `Full Sample`. Move `Wgtpop` to `Full Sample` using the `>` arrow. Then highlight `Studid` in `Source Variables`. Click `ID` (on right side) and move `Studid` to `ID`, again using the `>` arrow. Now
highlight the remaining variables in Source Variables. Select Variables, and click the >> arrows to move the remaining variables into the Variables box (figure I.C.3).

15. Save your data file (File – Save As) to NAEA DATA ANALYSIS\MY WESVAR FILES using the file name NATASSESS4.VAR. (You can use the same name because the format and extension are unique to WesVar; they will not be confused with the SPSS originals.)

16. Before any table can be computed, WesVar must create replication weights to compute the sampling error. Still on the same screen, click the scale button or select Data – Create weights. From the Source Variables, move Jkzone to VarStrat, and move Jkindic to VarUnit. Click JK2 under Method (figure I.C.4). Click OK. WesVar provides a message, This operation will create a
FIGURE I.C.4

Creating Weights in WesVar

FIGURE I.C.5

Replication Weights Created by WesVar
new VAR file. You will be asked to specify a file name. Click OK. Save the file to \texttt{NAEA DATA ANALYSIS\MY WESVAR FILES} using the filename \texttt{NATASSESS4.VAR}. You may be asked to \texttt{Confirm Save As}. Click \texttt{Yes}. WesVar has added replication weights for the estimation of sampling error, and the file now looks like figure I.C.5.

17. To close the WesVar data file, select \texttt{File – Close}. You are now ready to run the analyses in chapter 3, using the WesVar data file, \texttt{NATASSESS4.VAR}. Note that this file is the same as the ready-made file \texttt{NATASSESS4.VAR} in \texttt{NAEA DATA ANALYSIS\WESVAR DATA & WORKBOOKS}.

\begin{notes}

1. In the current data set, there is an even number of primary sampling units (PSUs) (schools in this case). If there is an uneven number of schools, the final jackknife (JK) zone will consist of 3 schools (1, 2, and 3). Where this occurs, it will be necessary to select JK as the analysis method (step 15) because JK2 requires an even number of PSUs.

2. \texttt{Jkindic} is the variable name for Jackknife Indicator.
\end{notes}
ITEM AND TEST ANALYSIS

Fernando Cartwright
This chapter provides a description of data requirements for Item and Test Analysis (IATA) and an overview of the types of information that IATA provides. An introduction is provided to the IATA interface, including task interfaces, main menu, and workflow navigation.

INSTALLING IATA

To run IATA, a computer must have the minimum recommended system requirements of Windows XP (SP3) or a newer Windows-based operating system, such as Windows Vista, Windows 7, or Windows 8. If a previous version of IATA is installed on the computer, it should be uninstalled and the IATA sample data folder deleted from the desktop before beginning the new installation. During the installation process, the system may need to be updated with the latest .NET framework from Microsoft (.NET 4 or later). The .NET framework is a regular part of the Windows operating system and is normally updated automatically by Windows. However, if Internet access is infrequent or the automatic updates feature of
Windows has been turned off, the operating system may be out of date. If IATA is being installed from a CD or USB, the installation file for .NetFx40_Full_x86_x64.exe can be found on the CD in the same directory as the IATASETUP.exe file. The update may also be downloaded and run from the following URL: http://www.microsoft.com/en-us/download/details.aspx?id=17718.

To install IATA, you must be logged in as a system administrator or as a user with rights to install software. Open the CD, open the IATA folder, then double-click the IATASETUP.exe file. Copy this file, and paste it to the Windows desktop, which can be accessed by clicking the desktop icon at the bottom right corner of the windows taskbar or by opening Explorer and double-clicking the desktop icon. You may be required to grant permission for IATASETUP to access or modify your computer. You must click Allow or Yes to continue. Next, you will see the welcome screen that confirms that you are installing IATA. At any point during the installation, you may click Cancel to quit the installation; no changes will be made to your computer. Click Next>> to continue. Please read through the agreement and select I accept the agreement to confirm that you will abide by its terms. Continue through the rest of the installation process by clicking Next>>. You may leave the default installation settings unaltered or change them if you have different directories for your files or menus. Before installation is performed, a summary of the installation specifications will appear. Click Install. Once the installation is complete, you will be presented with the confirmation screen. If you wish to run IATA immediately, mark the checkbox labeled Launch IATA. Click Finish to exit the setup wizard.

ASSESSMENT DATA

Two main types of data are produced by and used in the analysis of assessments: response data and item data. Individual learners produce response data as they answer questions on a test (a specific collection of questions that evaluate a common domain of proficiency
or knowledge). Individual questions on a test are referred to as *items*, which can be multiple-choice, short-answer, open-ended, or performance tasks. Item data are produced by analyzing or reviewing items and recording their statistical or cognitive properties. Each row in a response data file describes the characteristics of a test taker, whereas each row in an item data file describes the characteristics of a test item.

IATA can read and write a variety of common data table formats (for example, Access, Excel, SPSS [Statistical Package for Social Sciences], delimited text files) if they are formatted correctly. If the data are not formatted with the correct structure, IATA will not be able to carry out the analyses. Database-compatible formats, such as Access or SPSS, take care of most data-formatting issues. However, if the data are stored in a less restrictive format, such as Excel or a text file, the following conventions should be followed:

- The names of variables should appear in the cell at the top of each column (known as the *header*). Each column with data must have a column header. The name of each variable must be distinct from the names of other variables in a data file. A variable’s name must begin with a letter and should not contain any spaces.
- The *data range* is the rectangle of cells that contain data, beginning with the variable name of the first variable to appear in the data file and ending with the value of the last variable in the bottom-most row. The rectangle of cells forming the data range must not contain empty rows or columns.
- The data range must begin at the first cell in the spreadsheet or file. In Excel, this cell is labeled A1. In text files, this is the top-left cursor position in the text file.

The two examples in figure 8.1 illustrate incorrect and correct formats. In the incorrect format (a), a blank row exists above the data rectangle and a blank column to the left of it. Blank rows and columns are also found within the rectangle, and a column containing data has no header. In the correct format (b), all of the data are gathered into a single rectangle at the top left of the spreadsheet with no blank rows or columns.
### FIGURE 8.1

**Incorrect and Correct Data-Formatting Examples**

**a. Incorrect format: Empty rows and columns within and around the data**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>Var_A</td>
<td>Var_B</td>
<td>Var_C</td>
<td>Var_E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**b. Correct format: Data range in top corner without empty rows or columns**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Var_A</td>
<td>Var_B</td>
<td>Var_C</td>
<td>Var_D</td>
<td>Var_E</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Response Data**

Response data include the response of each student to each test item. The test results imported in the response data file must allow for automated scoring, that is, the item response data should include the codes representing how students actually responded to items. For example, if the response data are from a multiple-choice test, the data should record codes representing the options endorsed by each student (for example, A, B, C, D). IATA will transform the coded responses into scores using the answer key entered manually or provided as an answer key file.
Other information may be stored in a response data file that may be useful in analyzing test results. Examples include demographic information on variables such as age, grade, gender, school, and region. Other useful information may be collected in questionnaires (such as student and teacher questionnaires) or obtained from administrative records. If a stratified sample of students was used, the sample weighting for each student should be included on this file.

A unique identifier should be provided for each student; IATA will automatically produce unique identifiers based on the record order if a unique identifier is not specified. However, if the results are to be linked to other data sources, such as follow-up surveys or administrative records, using a previously defined identifier, such as student name or number, is a good idea to facilitate future links between data sets.

All responses should be assigned codes. For multiple-choice items, this procedure is straightforward, because each response option is already coded as correct or incorrect. For open-ended items, a scoring rubric is required to score item responses using a common coding framework. Open-ended items may be scored as either correct/incorrect or with partial credit given to different responses. A partial credit test item has more than one score that is greater than zero. Answers to open-ended questions must have been previously coded during the preparation of the response data. Volumes 2 and 3 of this series describe coding procedures for test items (Anderson and Morgan 2008; Greaney and Kellaghan 2012). For scoring of the response data, for most analyses, an answer key must be loaded into IATA. An answer key is a list of response codes that indicates the correct answer or answers for each test item. The answer key may be imported as a data file or entered manually. If the analysis is using anchored item parameters (for example, to link different versions of a test), the parameters must be included in the answer key file; they may not be entered manually (see “Item Data,” below).

**Treatment of missing and omitted data**

Missing data occur when a student does not provide a response to a test item. When this happens, rather than leaving the data field blank, a missing-value code is used to record why the response is missing. There are two types of missing responses: missing and omitted.
Missing codes are assigned to variables when students could have responded to an item but did not, leaving the answer blank. Such missing data are scored as incorrect. In contrast, omitted data codes are used when students were not administered an item, as when a rotated booklet design is used in a national assessment.

Depending on the circumstances of test administration or data processing, one must decide whether codes that apply to student responses that were unreadable or answered inappropriately, such as selection of two multiple-choice options, will be processed as missing or omitted data. Generally, if these data errors are the result of student error, the codes should be treated as missing and scored as incorrect. However, if the errors are the result of limitations in the data processing, such as imprecision in scorecard scanning that was not verified, the codes should be treated as omitted.

Another use of omitted data codes occurs when a balanced rotated booklet design is used. Balanced rotated booklet designs involve giving randomly equivalent samples of items to different students so that not all students answer the same test items (see Anderson and Morgan 2008). These designs permit extensive subject matter coverage while limiting the amount of student test-taking time. In a rotated booklet design, omitted codes must be assigned to all items for a student except those contained in the test booklet presented to the student. Omitted codes are not normally assigned to items in situations where all students are required to answer all of the items.

Common conventions use specific values for the different types of nonresponse data (see Freeman and O’Malley 2012 for information on response codes). The following common values are used:

- 9 for a missing response, where students have not responded to an item
- 8 for an unscorable response, which typically occurs in a multiple-choice test when students provide multiple responses and in open-ended items when responses are illegible
- 7 for items omitted or not presented, which might be used in a rotated booklet design
Regardless of the specific codes used, one must specify how IATA is to treat each nonresponse code, as either missing or omitted.

**Item naming**

Assigning a unique name to each item is important in a national assessment program (see Anderson and Morgan 2008; Freeman and O’Malley 2012). All statistical analyses performed on a test item should be linked clearly to the name or label of an item. If an item is repeated in several cycles of a national assessment, it should retain the same name in the data files for each cycle. For example, a mathematics item first used in 2009 might have the name M003, to indicate that it was the third item to appear in the 2009 test. If this same item is used in a 2010 test, it should still receive the name M003, regardless of where it appears on the test. Naming items by position in a test may cause confusion when items are reused. For this reason, assigning permanent names to test items when they are first developed is more useful than assigning names when they are first used in an assessment.

Using consistent names also facilitates linking the results of different tests. When IATA estimates statistical links between tests, it matches items in the linking procedure using item names. If an item name refers to different items in the two tests being linked, the results of the linkage will not be accurate. Although items can be renamed to facilitate the linking process, maintaining unique item names from the start is simpler and less likely to introduce errors.

**Variables reserved by IATA**

During the analysis of response data, IATA calculates a variety of working or output variables. The names of these variables are restricted and should not be used as names of test items or questionnaire variables. These variables, which IATA adds to the scored test data file, are listed in table 8.1.

In addition to avoiding these specific names, avoid using names that contain the @ symbol, which is reserved for processing *partial credit items*, which are test items that have more than one possible score value greater than zero.
<table>
<thead>
<tr>
<th>Score name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>XWeight</td>
<td>This is the design weight of the case that is used during analysis (if not specified, the value is equal to 1 for all students).</td>
</tr>
<tr>
<td>Missing</td>
<td>This variable describes the number of items that are omitted for a student.</td>
</tr>
<tr>
<td>PercentScore</td>
<td>The percent score is the number of items a student answered correctly expressed as a percentage of the total number of items administered to the student (excluding omitted response data).</td>
</tr>
<tr>
<td>PercentError</td>
<td>This is the error of measurement for the percent score. This estimate is specific to each student; its value depends on the percent score and number of items to which a student responded.</td>
</tr>
<tr>
<td>Percentile</td>
<td>The percentile rank (a number between 0 and 100) describes, for each student, the percentage of other students with lower scores.</td>
</tr>
<tr>
<td>RawZScore</td>
<td>The RawZScore is the percent score, transformed to have a mean of 0 and a standard deviation of 1 within the sample.</td>
</tr>
<tr>
<td>ZScore</td>
<td>This score is the normal distribution equivalent of the percentile score. It is also referred to as the bell curve score. Whereas the distribution of the RawZScore depends on the distribution of percent correct scores, the ZScore distribution tends to be more perfectly bell shaped.</td>
</tr>
<tr>
<td>IRTscore</td>
<td>The IRTscore is the proficiency estimate of the student. This score typically has a mean and standard deviation of 0 and 1, respectively. It facilitates generalization beyond a specific sample of items because its estimation takes account of the statistical properties of the test items.</td>
</tr>
<tr>
<td>IRTerror</td>
<td>The IRTerror is the error of measurement for the IRTscore.</td>
</tr>
<tr>
<td>IRTskew</td>
<td>The skewness of the proficiency estimate indicates if the test is better at measuring the lower or upper bound of a student's proficiency. (For example, an easy test may accurately describe if students have reached a minimum level of proficiency but may not provide an accurate estimate of higher levels of proficiency.)</td>
</tr>
<tr>
<td>IRTkurt</td>
<td>The kurtosis of the proficiency estimate describes how precise the estimate is for a given level of error. (For example, for two scores with the same measurement error, the one with the greater kurtosis is more precise.)</td>
</tr>
<tr>
<td>TrueScore</td>
<td>This score is an estimate of a percent score that is calculated from the IRT score. It is preferable to the raw percent score because it corrects for differences in measurement error between items. This score is calculated as the average of the probability of correct response to each item, given the IRT score of the student and the parameters of the test items.</td>
</tr>
<tr>
<td>Level</td>
<td>This variable is an estimate of the proficiency level for a student that has been assigned on the basis of standard-setting procedures. (If no standard-setting procedures have been performed, the default is for all students to be assigned a value of 1.)</td>
</tr>
</tbody>
</table>

Note: IRT = item response theory.

a. Additional IRT scaling options are available in IATA's advanced functionality; refer to the installation guide on the accompanying CD.
**Item Data**

IATA produces and uses item data files with a specific format. An item data file contains all the information required to perform statistical analysis of items and may contain the parameters used to describe the statistical properties of items. An item bank produced or used by IATA should contain the variables listed in table 8.2.

Table 8.3 presents examples from an item bank data file containing information about five science items named C1Sci31, C1Sci32, C1Sci33, C1Sci34, and C1Sci35. Note that item C1Sci35 does not have any data in the columns labeled \(a\), \(b\), \(c\), and **Level**. As indicated in the table, the only data fields that are mandatory are **Name** and **Key**. If \(a\), \(b\), or \(c\) parameters are missing, they will be

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>(MANDATORY) the unique name of each test item</td>
</tr>
<tr>
<td>Key</td>
<td>(MANDATORY) the information used to assign a numeric score to each item response, which is either the single code corresponding to the correct response or a delimited array of values that defines a variety of acceptable responses and their corresponding numerical scores</td>
</tr>
<tr>
<td>(a)</td>
<td>(OPTIONAL) the first of three parameters that describes how performance on a test item relates to proficiency on the performance domain, referred to as the <em>slope</em> or <em>discrimination parameter</em></td>
</tr>
<tr>
<td>(b)</td>
<td>(OPTIONAL) the second item parameter, referred to as the <em>location</em> or <em>difficulty parameter</em></td>
</tr>
<tr>
<td>(c)</td>
<td>(OPTIONAL) the third item parameter, referred to as the <em>pseudo guessing parameter</em>&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Level</td>
<td>(OPTIONAL) a previously assigned proficiency level for an item based on the initial item specification and expert review (values should be natural numbers, beginning with 1)</td>
</tr>
<tr>
<td>Content</td>
<td>(OPTIONAL) a code or description used to describe the subdomain of the curriculum, also known as a <em>strand</em> or <em>thread</em>, to which each item is most strongly aligned</td>
</tr>
</tbody>
</table>

<sup>a</sup> Use of the \(c\) parameter to describe items may cause certain functions, such as equating, to not work properly. For most purposes, the items are most useful if the value of the \(c\) parameter is equal or set to zero. The three-parameter model should be used only by expert users who are aware of its shortcomings. Estimation and use of the \(c\) parameter is provided by the advanced functionality of IATA. Registration of IATA, which is free, provides access to this advanced functionality. For registration instructions, see the installation guide on the accompanying CD.
estimated during the analysis. Many situations may require entry of an item data file into IATA that is missing these parameters. The most common scenario occurs when the response data for the items have never before been analyzed; in that case, the item bank data file is simply being used as an answer key. Another scenario occurs when some items have parameters that have been estimated in a previous data analysis and one wishes to fix the values of these items instead of having IATA reestimate them; in this scenario, one would leave a, b, and c values empty only for items for which one wished to estimate new parameters. Values for Level and Content may be either manually entered in the IATA interface or left empty.

An item data file may also include additional variables. For example, additional information typically stored with item data includes the question stem in the item bank, statistics describing the number of times the item has been used, and a list of the tests in which each item appears. However, IATA will not use any variables other than the seven required data fields listed in table 8.3.

The national assessment team can use information from any source if it has the required item data in the format presented in table 8.2. For example, national assessments may obtain permission to use items from a large-scale assessment such as one administered by the International Association for the Evaluation of Educational Achievement, which includes TIMSS (Trends in International Mathematics and Science Study) and PIRLS (Progress in International Reading Literacy Study) (http://timss.bc.edu/). If items from existing large-scale assessments are included in a national assessment, the item parameters from the

<table>
<thead>
<tr>
<th>TABLE 8.3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample Section of an Item Data File</td>
</tr>
<tr>
<td>Name</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>C1Sci31</td>
</tr>
<tr>
<td>C1Sci32</td>
</tr>
<tr>
<td>C1Sci33</td>
</tr>
<tr>
<td>C1Sci34</td>
</tr>
<tr>
<td>C1Sci35</td>
</tr>
</tbody>
</table>
existing assessments may be used to create an item data file that IATA can import.

**Answer key formats**

In the column with the header **Key** in an item bank data file, one must provide IATA with information it can use to score each item. In the simplest case, for multiple-choice test items with a single correct option, the value in each column should be the alphanumeric character corresponding to the correct option. The value is case sensitive, which means that, for example, if the correct response is coded as an uppercase A, then the uppercase letter A must be provided in the answer key; if a key value of a is provided, then any item responses with a value of A will be scored incorrect.

In rare cases, during a process of test review, a test item may be determined to have more than one correct response. To assign more than one key value to an answer key, one must enter a list of correct values, separated by commas with no spaces between any values or after the commas. For example, if the responses of A and C are acceptable as correct for a test item, the value of the key for the item should be A,C.¹

**Item data formats for partial credit items**

Partial credit (or graded-response) items are test items that have more than one score value. For example, instead of being scored as 0 or 1, an item with varying levels of correctness may be scored as 0, 1, or 2, where 0 represents an attempted response, 1 a partially correct response, and 2 a perfect response. To accommodate different score values, the answer key must be entered for each score value that is greater than 0. If the marking scheme used for partial credit items uses scores that are all greater than 0, then answer key information should not be entered for the lowest score value. For example, if the possible item scores are 1, 2, and 3, the answer key should provide scoring information for only scores 2 and 3. The format for a partial credit answer key is `<score1>:<value list 1>;<score 2>:<value list 2>; ... <score n>:<value list n>`. For example, for a partial credit item with three scores, coded as A, B,
and C, with scores of 1, 2, and 3, respectively, the answer key should be entered as 1:A;2:B;3:C.

If a partial credit item has already been analyzed, it will have a greater number of parameters than a regular test item. Each score value will have a distinct value for the $b$ parameter, although the $a$ parameter will have the same for all score values. These item data must be entered in a special format. In addition to providing the main entry with the full answer key, one must add a new entry for each score value (except for the lowest score value) as if each item score were a separate test item. The parameter fields for the main item entry should be left blank. For example, if an item has scores of 0, 1, and 2, then a total of three rows would be required in the item data file: one row for the overall item, which would have only the item name and answer key; and two score-specific entries for 1 and 2 that have name, answer key, and parameter information.

The value of the name field for each new score-specific entry is the original item name followed by @<score value>. For example, if the original item name is TestItem, the name for an item score of 1 is TestItem@1. IATA uses an item response model that requires the values of the $b$ parameters to be in the same order as the scores. Therefore, if there are two score entries, 1 and 2, the $b$ parameter value for score 2 must be greater than the $b$ parameter for score 1 (see table 8.4).

When a new row is entered for each partial credit item score, the values of the answer key field must also be specified differently. The analysis of a partial credit item assumes that a student achieving a particular item score has also mastered the level of skill associated with a lower score on that item. In other words, if each score is treated

<table>
<thead>
<tr>
<th>Table 8.4: Sample Section of an Item Data File for a Partial Credit Item</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Name</strong></td>
</tr>
<tr>
<td>PCItem001</td>
</tr>
<tr>
<td>PCItem001@1</td>
</tr>
<tr>
<td>PCItem001@2</td>
</tr>
<tr>
<td>PCItem001@3</td>
</tr>
</tbody>
</table>
as a separate test item, a student with a high partial credit score is assumed to have performed correctly on the lower partial credit scores. To manage this relationship in IATA, the answer key for each score value should list its own key value or values as well as the values of any higher scores.

An example of proper partial credit item data formatting for an item with scores of 0, 1, 2, and 3 is given in table 8.4. Note that no scoring information is provided for the lowest score (0). The main item entry does not have parameter values or a value for Level. Because each score value could correspond to a different standard of performance, specifying the level for the entire item does not make sense. Even though the responses are already scored, scoring information must still be specified using the proper answer key format. For IATA to score the responses properly, the answer key must provide both the values found in the data and the score assigned to each value.

**DATA PRODUCED BY IATA**

IATA produces several data tables that contain the specifications of the current analysis and the analysis results. In general, all results should be archived for future reference. Table 8.5 summarizes the list of data tables produced by IATA. The tables may be saved individually or collectively directly from IATA into a variety of common formats, including Excel (*.xls/*.xlsx), SPSS (*.sav), comma delimited (*.csv), or tab delimited (*.txt).

In addition to the data tables described in table 8.5, IATA produces several charts, textual summaries, and tables of results that are displayed only in the IATA interface. These results may be copied directly from IATA and pasted into other documents for future reference. The method of copying the output depends on the type of output.

For charts, right-clicking on the chart body will raise a pop-up menu with the options to (a) copy the image to the clipboard, (b) save the chart image directly to a file, or (c) print the image. To
display results in tabular form, one must first select the cells, rows, or columns to be copied, then copy the data either by selecting Copy from the right-click pop-up menu or by typing Ctrl+c. The copied data may be pasted to a text file or directly into spreadsheet software such as Excel or SPSS.

### TABLE 8.5

<table>
<thead>
<tr>
<th>Data tables</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Responses</td>
<td>Original response data (including nontest data) imported into IATA</td>
</tr>
<tr>
<td>Values</td>
<td>Unique response codes for all test items and indication whether each response value is coded as valid missing (valid skip) or missing</td>
</tr>
<tr>
<td>Scored</td>
<td>Response data that have been scored as correct (1) or incorrect (0) using the specified answer key, as well as all summary scores and their standard errors</td>
</tr>
<tr>
<td>Items1*</td>
<td>Item answer keys and statistics related to the current analysis and item parameters</td>
</tr>
<tr>
<td>Items2</td>
<td>Item answer keys and parameters of the reference item parameter file used for linking</td>
</tr>
<tr>
<td>MergedItems</td>
<td>Item-by-item matching of items in both the new and the reference item parameter files used by the linking process</td>
</tr>
<tr>
<td>Eigenvalues</td>
<td>The proportion of variance explained by each of the dimensions underlying the item responses</td>
</tr>
<tr>
<td>PatternMatrix</td>
<td>The proportion of variance of each item explained by each of the dimensions underlying the item responses</td>
</tr>
<tr>
<td>Levels</td>
<td>The thresholds used to define proficiency levels</td>
</tr>
<tr>
<td>LinkingConstants</td>
<td>Scale transformation constants used to adjust the latent trait scale between populations or samples</td>
</tr>
<tr>
<td>BookmarkData</td>
<td>An ordered list of items that can be used to facilitate standard setting or creation of definitions for performance levels</td>
</tr>
<tr>
<td>DIF_&lt;specifications&gt;</td>
<td>The results of a differential item functioning analysis, where the &lt;specifications&gt; portion of the table name summarizes the variable and groups compared in the analysis</td>
</tr>
<tr>
<td>CustomTest&lt;name&gt;</td>
<td>A set of items chosen to minimize error of measurement over a specific range of proficiency; the &lt;name&gt; is a user-specified value</td>
</tr>
</tbody>
</table>

---

a. The Items1 data table produced by IATA following an analysis of response data not only serves as an item bank data file but also generates several additional statistics. The additional statistics are described in later sections on analysis of response data. They describe the behavior of items in a specific sample and are useful for guiding test analysis and construction but are not required to be maintained in an item bank file that will be used by IATA.
INTERPRETING IATA RESULTS

Whenever IATA produces analysis results for individual items, it will also present “traffic symbol” summary indicators that provide a general idea of how to interpret results. IATA uses three symbols (see table 8.6). The symbols (circle, diamond, triangle) appear in color on the computer screen.

In the case of analyses in which multiple pieces of information are considered when interpreting results for a specific item, such as the item analysis and test dimensionality results, IATA generates interpretive statements that attempt to summarize the different statistics. These statements are intended as a useful suggestion on how to proceed. However, in any case where IATA recommends a modification to either the analysis specifications or the test items, you should verify that the recommendation is appropriate by examining the statistical results, the actual test booklets, or both, or by discussing the items with a curriculum specialist.

<table>
<thead>
<tr>
<th>TABLE 8.6</th>
<th>Traffic Symbols in IATA and Their Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symbol</td>
<td>Meaning</td>
</tr>
<tr>
<td>🔄</td>
<td>A circle indicates no major problems.</td>
</tr>
<tr>
<td>🟖</td>
<td>A diamond indicates that the results are less than optimal. This indicator is used to suggest that modifications may be required to either the analysis specifications or the items themselves. However, the item is not introducing any significant error into the analysis results.</td>
</tr>
<tr>
<td>🟢</td>
<td>A triangle appears beside any potentially problematic item. It is used either to indicate items that could not be included in the analysis because of problems with the data or specifications or to recommend a more detailed examination of the specifications or underlying data and test item. When this indicator appears, it does not necessarily mean that a problem exists, but it does suggest that the overall analysis results may be more accurate if the indicated test item were removed or if the analysis were respecified.</td>
</tr>
</tbody>
</table>
SAMPLE DATA

When IATA is installed on a computer, it will create a folder on the desktop called IATA. This folder contains sample data that are required for the walk-through examples in this book. The sample data folder contains seven files including six response data sets, each in Excel format, and an Excel file containing the answer keys for each of the response data sets. The files are in *.xls format to be compatible with older and open-source software (depending on the computer’s settings, the .xls file extension may not be visible). The names and contents of the files are as follows:

- **PILOT1**: a response data set corresponding to a pilot test containing multiple-choice items
- **CYCLE1**: a response data set corresponding to a national assessment administration
- **PILOT2**: a response data set corresponding to a pilot test containing multiple-choice items in a balanced rotated design
- **PILOT2PartialCredit**: a response data set corresponding to a pilot test containing multiple-choice and partial credit items in a balanced rotated design
- **CYCLE2**: a response data set corresponding to a national assessment administration with items common with a previous administration
- **CYCLE3**: a response data set corresponding to a national assessment administration with items common with a previous administration
- **ItemDataAllTests**: an Excel file with multiple sheets containing answer keys and information about the items on each of the response data files

These sample data are fictional data sets developed with the sole intent of providing concrete examples and applications of this software. Although they reflect typical patterns of student response and the relationships in the data are similar to those found in most large-scale assessments, the results and discussions of analysis findings do not represent any actual national assessment.
If any of the sample data files are deleted, they may be recovered by running the IATASetup.exe program on the CD or from the IATA website (http://polymetrika.com/home/IATA).

**IATA ANALYSIS WORKFLOWS AND INTERFACES**

IATA differs from statistical analysis programs that provide a variety of analysis functions that may be accessed individually. In contrast, all of the analysis functions in IATA are accessed through workflows, in which the results from each step in the workflow may be used to inform the specifications or interpretation of results in subsequent steps. Five workflows are available in IATA:

- Response data analysis
- Response data analysis with linking
- Linking item data
- Selecting optimal test items
- Developing and assigning performance standards

Each workflow reflects the needs of specific goals that may arise in the context of a national assessment. The following guidelines relate to some situations associated with workflows:

- If a pilot test has been conducted and detailed information is needed about item behavior to determine the content of the final test, the **Response data analysis** workflow should be used.
- If data collection has been completed for the first national assessment in a planned series of assessments, the **Response data analysis** workflow should be used.
- If new scale scores are being assigned to a sample of students who have been administered the same test that was used in a previous national assessment, the **Response data analysis** workflow should be used.
- If a national assessment that shares items with a previous assessment has been conducted and a comparison is desired of the results
of the two, the **Response data analysis with linking** or the **Linking item data** workflow should be used.

- If one wishes to modify a test and needs to identify the best items to retain in the new test to maintain comparability with the previous test, the **selecting optimal test items** workflow should be used.
- If the national assessment has already been conducted and one wishes to interpret the results in a way that is consistent with curriculum expectations rather than simply compare students with each other, the **Developing and assigning performance standards** workflow should be used.

To perform an analysis with IATA, one must select one of these workflows from the main menu. The main menu is reached by clicking **Main Menu** on the bottom right of the language selection and registration screen that loads with IATA (figure 8.2). You do not need
to enter information on this screen to run IATA. The default language for IATA is English, and registration is optional. Registration allows one to access advanced analysis functions and receive notifications when updates to IATA become available.

Figure 8.3 shows the IATA main menu.

Each workflow is composed of a set of tasks that are completed in order. Most of the workflows share many of the same tasks. IATA performs 10 tasks, each with its own interface, which generally appear in the order listed in table 8.7. Not all tasks appear in all workflows. The workflows are designed so that you are required to perform only tasks that are relevant to your analysis goals. Table 8.7 matches tasks to workflows.

The first two workflows (A, B) are very similar in terms of their tasks. In contrast, the last three (C, D, E) analyze only item data. All workflows require that data be loaded into IATA and allow results to be saved.
After selecting a workflow from the IATA main menu, you will be directed into the set of tasks that make up the workflow. Each task has its own interface that allows you to specify how IATA should perform the task and, if applicable, to view results.

IATA’s instructions box and the navigation buttons are shown in figure 8.4. The instructions box provides a brief summary of the specifications that are required for each task and interpretative suggestions. Clicking the button labeled "<<Back" allows review of a previous task; clicking "Next>> moves to the next task. Note that although IATA does not prevent the user from moving back and forth through

### TABLE 8.7

<table>
<thead>
<tr>
<th>Task</th>
<th>Response data analysis</th>
<th>Response data analysis with linking</th>
<th>Linking item data</th>
<th>Selecting optimal test items</th>
<th>Developing/assigning performance standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loading data</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
<td>•</td>
</tr>
<tr>
<td>Setting analysis specifications</td>
<td>•</td>
<td>•</td>
<td></td>
<td>•</td>
<td></td>
</tr>
<tr>
<td>Analyzing test items</td>
<td>•</td>
<td>•</td>
<td></td>
<td>•</td>
<td></td>
</tr>
<tr>
<td>Analyzing test dimensionality</td>
<td>•</td>
<td>•</td>
<td></td>
<td>•</td>
<td></td>
</tr>
<tr>
<td>Analyzing differential item functioning</td>
<td>•</td>
<td>•</td>
<td></td>
<td>•</td>
<td></td>
</tr>
<tr>
<td>Linking</td>
<td>•</td>
<td>•</td>
<td></td>
<td>•</td>
<td></td>
</tr>
<tr>
<td>Scaling test results</td>
<td>•</td>
<td>•</td>
<td></td>
<td>•</td>
<td></td>
</tr>
<tr>
<td>Selecting optimal test items</td>
<td>•</td>
<td>•</td>
<td></td>
<td>•</td>
<td></td>
</tr>
<tr>
<td>Informing development of performance standards</td>
<td>•</td>
<td>•</td>
<td></td>
<td>•</td>
<td></td>
</tr>
<tr>
<td>Saving results</td>
<td>•</td>
<td>•</td>
<td></td>
<td>•</td>
<td></td>
</tr>
</tbody>
</table>
the workflow, later tasks in the workflow may not provide meaningful results unless earlier tasks have been correctly completed.

Regardless of the workflow in which they appear, the general specifications for each task remain the same. Task interfaces are described in detail in the example walk-throughs in chapters 9 through 14 of this volume.

The examples in the following chapters demonstrate how to use IATA to perform item and test data analyses required for a national assessment using data from the IATA folder on your desktop. To analyze your own national assessment data, create a new folder and assign it a relevant name, such as `NATIONAL_ASSESSMENT_YEAR_1`. (Avoid using spaces or special characters other than _ when naming files and folders.) You should store your student response data in this folder and save any results produced by your analyses to the same folder. Although IATA can analyze data in different data formats, such as Excel and SPSS, the data files you use must follow the structures and naming conventions described in tables 8.2 and 8.3. The names of the data files should clearly identify the source of the data.²

NOTES

1. This format requirement means that commas should never be used as answer key values.

2. Additional information on IATA can be found at http://polymetrika.com/home/IATA.
Use the PILOT1 sample data set to carry out this exercise. The answer key for this test is in the Excel workbook, ItemDataAllTests, in the sheet named PILOT1.

This chapter introduces the analysis of pilot test data using Item and Test Analysis (IATA). The Response data analysis workflow will be used to analyze response data using an answer key file. The stages in the workflow include loading of data, specification of the analysis, item analysis, dimensional analysis, analysis of differential item functioning, and item selection. Scale scores or performance standards are not calculated because the distribution of proficiency in the pilot sample is unlikely to be representative of the population.

Consider the following scenario. A national assessment team and its curriculum experts have created a set of new multiple-choice items designed to evaluate grade 5 students’ mathematics skills. The new test items, which were considered adequate for representing the national curriculum, had been created to reflect the main content categories (number knowledge, shape and space, relations, problem solving, and uncertainty) determined by a national steering committee. The final version of the test, which is intended to contain 50 items, is meant to be administered to fifth-grade students of all proficiency levels.
As a first step, the national assessment team administered an 80-item test to a total of 262 students, sampled from seven schools in each of three regions. The team used a larger number of items than will be included in the final test in the expectation that, as is typical, many items proposed for a test will not function well for a variety of reasons. (For example, they may be too easy or too difficult, or instructions may be confusing.) Indeed, some items may have been rejected by review panels before the pretest. In anticipation of further problem items, at least 50 percent more items than are required for the final test should be pretested. Note also that a pilot test is intended to test the operational protocols for a survey as well as determine the composition of items in the final test.

The student response data file contains each student’s multiple-choice answers to each of the 80 items as well as some school-level variables (region identification, school identification, school type, school size) and some student-level information (student sex, language of the student’s home).

From the main menu, click the first menu option, **Response data analysis**, to enter the analysis workflow (figure 9.1). If, at any stage in the workflow, an error is received or results appear that are different from those expected, return to a previous step or begin the analysis again from the main menu.

**STEP 1: LOADING RESPONSE DATA**

Regardless of the analysis path chosen, IATA must be directed to load previously collected or produced data (for example, national assessment pilot test data, or an item data file). IATA is flexible and
has simple procedures and buttons for loading response data, item data, or both. Regardless of the analysis path or type of data, IATA must be told which data file to import and which data in the file to use. IATA can import data in SPSS (Statistical Package for Social Sciences) (*.sav), Excel (*.xls/*.xlsx), tab-delimited (*.txt), and comma-separated (*.csv) formats. Because Excel data files can contain several separate tables, the table to be imported must be specified.

The first screen in this analysis path requires you to import a response data file into IATA. The data-loading interface is shown in figure 9.2. The instructions begin with the words EXAMINEE RESPONSE DATA to indicate that you are loading data containing responses of individual students to individual items. Below the instructions are two boxes: a file path summary and a drop-down menu for selecting data tables in the selected file. To the right of these boxes is the button labeled Open File. The table at the bottom of the interface displays the data for a selected data source. If more than 500 rows of data exist, only the first 500 will be displayed. If the selected data format selects multiple tables, such as

![Response Data-Loading Interface](image-url)
Excel or Access, then the name of the first table in the data file will appear in the drop-down box. Otherwise, the name of the file will appear. For multitable data files, the desired data may not be in the first table. Verify that the appropriate data are selected by reviewing the contents of the data table, which will appear in the large area at the bottom of the interface. If the active table does not contain the desired data, select a different table by clicking the drop-down menu.

For this example, load the *PILOT1.xls* file.

1. Click **Open File** to select a data file. In the file browser, navigate to the folder on the desktop that contains the IATA sample data.
2. Select (or type) **PILOT1.xls**.
3. Click **Open** or press the **Enter** key.

When the file opens, a pop-up dialog will remind you to confirm that the selected data contain the correct item response data. Click **OK** to continue. Confirm that the sample pilot data are correctly loaded; the interface should look like figure 9.2. The data in the figure show the records for each student who took the pilot test. The first seven variables from the left describe demographic and sampling information about the students.

- **PILOT1STDID**: unique student identification code
- **SCHOOLID**: unique school identification code
- **Sex**: sex (gender) of the student (1 = female, 2 = male)
- **SchoolSize**: total number of students in the school
- **Rural**: location of the school (0 = urban, 1 = rural)
- **Region**: numeric identifier for the geographic region
- **Language**: numeric identifier of whether or not the language of instruction is spoken in the student’s home

The first mathematics test item appears in column 8 and is labeled MATHC1019. Scroll across the entire data set to see that the file contains data on 80 items; the item in the last column is labeled MATHC1041. The item names are arbitrary and do not reflect their position on the test. Most cells have values A, B, C, or D, indicating
students’ choice of options. Cells that show the value 9 indicate that a student did not respond to the item.

As with most pilot samples, the students represent a sample of convenience rather than a probability sample of the total population. Accordingly, the response data file has no sample weights.

After verifying that the correct response data file has been loaded, click Next>>.

**STEP 2: LOADING THE ANSWER KEY**

The item answer keys must be loaded now. As with the response data, the item data are in Excel format in the IATA data folder on the desktop.

1. Click **Open File** to select a data file. In the file browser, navigate to the folder on the desktop that contains the IATA sample data.
2. Select (or type) **ItemDataAllTests.xls**.
3. Click **Open** or press **Enter**.

When the file opens, a pop-up dialog will remind you that IATA will estimate any missing item parameters. Click **OK** to continue. The selected data file contains tables for all the examples in this book. Ensure that the table named **PILOT1** has been correctly selected in the drop-down menu. Confirm that the correct item data are correctly loaded; the interface should look like figure 9.3. To find information on a specific item, sort the items by clicking on the header for the **Name** column.

After confirming that the correct item data have been loaded, click Next>> to continue.

**STEP 3: ANALYSIS SPECIFICATIONS**

All workflows that use response data require certain specifications that will affect the results of all subsequent analyses. These specifications include information relating to the answer key, respondent
identification, sample design weighting, and treatment of missing data codes. The interface for providing these specifications is shown in figure 9.4. The large panel on the left contains a table of the test items in the response data file with the column headers Name, Key, Level, and Content. If an item data file has been loaded, the table will contain only variables that have been identified as test items; otherwise, the table will contain all variables. If you skipped the loading of an item data file, you would need to manually enter the answer key specifications for each item in this table (see “Answer key formats” in chapter 8 of this volume).

In the center section of the interface is a button labeled Update response value list. Click this button to change the answer key specifications, either by manually entering answer keys or by deleting existing ones. When this button is clicked, IATA will populate the two drop-down menus with lists of variables in the response data that have not been assigned an answer key and list all of the response values present for the variables identified as test items. If you loaded an item data file, these menus will already be populated with values.
Below the **Update response value list** button are several controls for providing optional specifications: a drop-down menu for specifying the identification (ID) variable, a drop-down menu for selecting the weight variable, and a table for specifying treatment of missing value codes. Specifying an ID variable may be necessary to merge the test results produced by IATA with other data sources. The ID variable should uniquely identify each student; if you do not specify an ID variable, IATA will produce a variable named *UniqueIdentifier* to serve this purpose. The weight variable is used to ensure that the statistics produced during the analysis are appropriate for the sample design of the national assessment but, as noted, will not be applied in analysis of the pilot data. When no weight variable is provided, IATA will assume that all students receive the same weight, equal to 1.

You can inform IATA that a response value is a missing-response code by clicking one of the boxes next to the value in the **Specify missing treatment** table. By default, IATA assumes that all response
values represent actual student responses. If the box in the Incorrect column is checked, then IATA will treat that value as an invalid response that will be scored as incorrect. If the box in the Do Not Score column is checked, then IATA will treat that value as omitted, and the value will not affect a student’s test results. By default, if the response data contain any completely empty or blank cells, IATA will treat them as incorrect unless you have manually specified Do Not Score.

For this walk-through, the answer key and response data have both been entered, so the list of items shown in figure 9.4 contains only those variables with answer keys in the item data. Reviewing the answer key table to confirm that the keys and other data about each item are correct and complete is a good idea; any errors at this stage will produce even more errors in subsequent tasks in the workflow. In the middle of the screen, specify the additional analysis details. Use the following specifications:

1. Use the first drop-down menu to select PILOT1STDID as the ID variable (the ID that was assigned initially to students; see figure 9.2).
2. Because these data do not have a sample weight, you may leave the second box with a drop-down menu blank.
3. Because the value of 9 will be treated as incorrect, check the appropriate box in the table of values in the Specify missing treatment section. Although the PILOT1 data have no blank entries, one can leave the default specification of treating blank entries as incorrect.

When the specifications have been entered, the interface should look the same as figure 9.4.

Confirm that the specifications are correct, and click Next>> to continue. The data will begin processing automatically. The processing stages are setting up data, scoring, estimating parameters, item response theory (IRT) scaling, calculating true scores, and factor analysis. As the processing continues, the interface will display its current stage. Depending on the speed of one’s computer and the size of one’s
data set, this analysis may take seconds to minutes to complete. When IATA finishes processing, it will display the results in the item analysis interface.

**STEP 4: ITEM ANALYSIS**

When the data processing has finished, the item analysis interface will be updated with the results shown in figure 9.5. Using the interface, one can access these results as well as view and save diagnostic information about each test item. Four types of results are displayed in this interface:

- Statistics and statistical parameters describing each item (on the left)
- A graphical illustration of the relationship between student proficiency and the probability of correctly responding to an item, also known as an item response function, or IRF (at the top right)

![Item Analysis Results for PILOT1 Data, MATHC1019](image-url)
• A contingency table describing the proportions of students with high, medium, and low test scores who selected each item response, also known as a distractor analysis (at the middle right)
• A plain-language summary of the item analysis results (at the bottom right)

The table on the left side of the item analysis interface presents statistical information as well as a symbol describing the overall suitability of each item. The Name of each item is in the column to the right of the summary symbols. Examine the detailed results for an individual item by using the arrow keys or mouse to highlight the row in which the item appears. Use the checkboxes in the Use column for each row to include or exclude items from the analysis. Uncheck one of these item boxes to remove the item from the analysis. Then click the Analyze button to rerun the analysis with the reduced set of items. (Removed items will still be listed but will have red triangles beside them.) Return all items to their original state by clicking Reset Items and Analyze. Note that clicking Reset Items will reset all items; to permanently remove an item from the analysis, delete its answer key in the analysis specifications interface. The Scale button does not reestimate any item parameters; it simply calculates IRT scale scores for the response data using the item parameters that have already been estimated or loaded into IATA from an external data file.

**Item Statistics**

The three columns to the right of the item name in figure 9.5 contain classical item statistics: the item discrimination index (Discr); the item facility (PVal), sometimes referred to as item difficulty, although larger values indicate an easier test item; and the point-biserial correlation (PBis) (see, for example, Crocker and Algina 2006; Haladyna 2004). The final three columns, which may be hidden from view, requiring one to use the scroll bar at the bottom of the table, are estimates of IRT parameters: the slope parameter (a), the location or threshold parameter (b), and the pseudo-guessing parameter (c).
In general, the classical statistics may be interpreted directly. The item facility (PVal) ranges between 0 and 1 and describes how easy an item is for the given sample: a value of 0 indicates that no students responded correctly; a value of 1 indicates all students responded correctly. The discrimination index and point-biserial correlation provide alternate measures of the same relationship, which is how strongly related responses to each item are to the overall test score. For both statistics, the value should be greater than 0.2. These guidelines should not be considered absolute because the indexes are affected by factors other than the discrimination of items, such as the accuracy of the overall test. For example, the item facility tends to limit the absolute value of both the discrimination index and the point-biserial correlation. If the item facility differs substantially from 0.5 (is less than 0.2 or greater than 0.8), the discrimination index and point-biserial correlation will underestimate the relationship between proficiency and performance of students on a test item. Although extremely easy or difficult items tend to reduce the observed relationships with proficiency, they may cover important curriculum content that should be included in a test or they may (in the case of easy items, for instance) be required to sustain student motivation during testing. For these or other reasons, including a relatively small number of very easy and difficult items is often desirable.

The IRT parameters should not be interpreted in isolation. Although each describes a specific behavior of the test item, the relationship between responses to the item and overall proficiency is the result of interactions among all three parameters as well as the proficiency level of individual students.

Most items in the current analysis have a green circle, indicating that they have no major problems and are relatively satisfactory. By scrolling down the item list on the left, you will see 13 items with diamond-shaped caution symbols (MATHC1047, MATHC1013, MATHC1002, MATHC1070, MATHC1034, MATHC1035, MATHC1032, MATHC1010, MATHC1068, MATHC1046, MATHC1024, MATHC1058, and MATHC1030). One item (MATHC1075) has a triangular warning symbol and is considered potentially problematic. The best practice is to examine the results for all items, regardless of the
summary symbol IATA assigns. This walk-through focuses on a few examples.

By default, the results for the first item are displayed in the graph and table on the right. IATA has assigned this item, MATHC1019, a green circle. Each of the results IATA produces for this item is described in the following sections.

**Item Response Function**

In the graphics window on the right side of the item analysis interface, IATA displays the Item Response Function for a selected test item (see figure 9.5). Reviewing the IRF is typically more intuitive than examining the IRT parameters or item statistics to determine the relative usefulness of a test item. A useful item will have a strong relationship with proficiency, indicated by an IRF that has a strong S-shape, with a narrow region in which the curve is almost vertical. The slope of the IRF for MATHC1019 is consistently positive, but the relationship is weak; no region has a notably steep slope. This shallow slope corroborates the low discrimination index ($\text{Discr} = 0.36$) and low point-biserial correlation ($\text{PBis} = 0.35$).

As with any statistical modeling method, IRT is useful only if the empirical data fit the theoretical model. For each item or score value, IATA produces a graphic of the theoretical IRF produced using the estimated parameters as well as the empirical IRF estimated directly from the proportions of correct responses at each proficiency level. The graphic can be used to assess the suitability of using IRT to describe each item. If the IRT model is appropriate, the red dashed line will appear to be very similar to the solid black line, and deviations between the two will be less than 0.05, particularly in the region between $-1$ and $1$, which contains many students. For MATHC1019, the theoretical and empirical IRFs are almost identical, indicating that, although the item itself may have a weak relationship with proficiency, its statistical properties are accurately described by the IRF.

**Distractor Analysis**

In the bottom right of the item analysis interface in figure 9.5, IATA produces statistics for each response value (including missing value
ANALYZING DATA FROM A PILOT TEST ADMINISTRATION 159

codes and incorrect response values) and a textual summary of the analysis. The statistics are estimated separately for groups of low-, medium-, and high-performing students according to their percent-correct test score. The data in table 9.1 represent a distractor analysis of an individual item.

An item may have a low or even a negative discrimination relationship with proficiency for many reasons. These include poor wording, confusing instructions, sampling errors, and miskeying or miscoding of responses. Distractor analysis may be used to detect and remediate some of these common errors by looking at patterns in item responses. A well-functioning item should have the following characteristics:

- The correct column option (D), denoted by the asterisk (*), should have a high percentage for the high group and successively lower percentages for medium and low groups. MATHC1019 satisfies this condition, with values of 47.9, 19.9, and 11.4 for the high, medium, and low groups, respectively.

- For the low-skilled group, the percentage choosing the correct option (D) should be lower than the percentage choosing any one of the other options. All of the incorrect options (A, B, C) for MATHC1019 exhibit this pattern.

- Each of the columns corresponding to incorrect response values should have approximately equal percentages in each skill level and overall compared to the other incorrect response values. MATHC1019 violates this pattern, because option B was selected by a considerably greater percentage of incorrect respondents than either option A or C.

- For the high-skilled group, the percentage choosing the correct option (D) should be higher than the percentage choosing any one

| TABLE 9.1 |
| Distactor Analysis for MATHC1019, PILOT1 Data |

<table>
<thead>
<tr>
<th>Group</th>
<th>9(X)</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D*</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>0.0</td>
<td>14.1</td>
<td>23.9</td>
<td>14.1</td>
<td>47.9</td>
</tr>
<tr>
<td>Medium</td>
<td>3.1</td>
<td>15.2</td>
<td>37.7</td>
<td>24.1</td>
<td>19.9</td>
</tr>
<tr>
<td>Low</td>
<td>8.6</td>
<td>14.3</td>
<td>42.9</td>
<td>22.9</td>
<td>11.4</td>
</tr>
<tr>
<td>Total</td>
<td>2.3</td>
<td>14.9</td>
<td>34.0</td>
<td>21.4</td>
<td>27.5</td>
</tr>
</tbody>
</table>

Note: The asterisk denotes the correct answer column.
of the other options. MATHC1019 satisfies this pattern: 47.9 is greater than the values for options A (14.1), B (23.9), and C (14.1).

- For all groups, the percentage of missing value codes should be close to zero. The proportion of students with missing responses (code 9) was greater for low performers (8.6) than for high performers (0.0), suggesting that the decision to treat the code as incorrect (rather than omitted) was reasonable.

- Missing-response codes that are treated as omitted (denoted by OMIT) should have equal percentages of students at each skill level. This code was not used for these data.

IATA provides a textual summary about the item performance, including warnings if the discrimination is unacceptably low, in which case it suggests what may be done to improve it. For example, IATA will identify distractors that are not effective in eliciting endorsements from respondents (or have statistical profiles similar to correct responses). If IATA does detect any common problems in the data, a textual summary of the results is displayed in the text box beneath the distractor analysis table.

In the results for MATHC1019, the textual summary on the bottom right recommends examining the response option coded as A. Looking at the distractor analysis table, one can see that response A is endorsed by approximately the same proportion of high-performing students as of low-performing students, indicating that it does not function well as a distractor.

The distractor analysis of national assessment data may also be useful to providers of in-service education courses for teachers and curriculum personnel because it may help identify common student misconceptions and errors. Curriculum authorities can also use the data to judge the appropriateness of specific material for a particular grade level.

### Comparing Items

Compared to the previous item (see figure 9.6), the second item on the test, MATHC1027, has a stronger relationship with proficiency, indicated by the steeper IRF and larger discrimination (0.65) and
point-biserial correlation (0.53) values. The theoretical and empirical IRFs are almost identical, indicating that the statistical item response model is appropriate to the response data. The distractor analysis table shows that 73.2 percent of students in the high-skilled group selected the correct option (C), compared to 19.9 percent in the medium-skilled group and 8.6 percent in the low-skilled group. All the incorrect response values (A, B, D), as well as the missing response code (9), were more likely to be selected by low-performing students than by high-performing students.

In contrast to the two items already examined, items with triangular warning symbols are typically poor items, inclusion of which in the test may produce misleading or less useful results. The number of poor items that appear in a pilot test such as this one can be minimized by following item creation guidelines described in volume 2 in this series (Anderson and Morgan 2008). The only item with a warning symbol in these data is MATHC1075 (see figure 9.7). By clicking on the item, you will see that the results indicate an almost nonexistent relationship of proficiency with either correct or incorrect responses.
Although a missing response code is still related to proficiency, the expected pattern was not evident. Students in the low-skilled group were not the ones most likely to select each of the three incorrect options, nor were students in the high-skilled group least likely to do so. The item was particularly weak at discriminating between medium- and low-level students. The discrimination index is low (0.14), as is the point-biserial correlation (0.16). The item may be related to proficiency but, because so few students answered correctly ($P\text{Val} = 0.12$), the relationship cannot be estimated. Because responses to the item are not clearly dependent on proficiency, including it in the test would tend to increase the influence of random factors on test scores. Including the item (and other problematic items) in the analysis might also reduce the accuracy of statistical estimates for other test items, because the item statistics and parameters are analyzed using the test scores.

Items can be removed from the analysis by clicking the box to the left of each item name to uncheck the item. After removal of an item, the results should be recalculated by clicking Analyze before...
removing any other items. The removal of a single item will affect the results for all other items. If many problematic items exist, remove only one at a time, because some items flagged as problematic may appear so only because of the influence of poorer items on the analysis results. If too many items are removed accidentally, recheck each item or click Reset Items above the item list to reset the entire item list. For this example, remove MATHC1075 and rerun the analysis to produce the results in figure 9.8, in which the results for MATHC1075 are highlighted after removal. Note that the Discr and PBis data for this item have been replaced by NaN (Not a Number) or out-of-range values; they will not affect subsequent calculations. For removed items, the distractor analysis table on the right does not appear, and a message in the textual summary says that the test data must be reanalyzed. Because only a single item was removed, the statistics for the remaining items are relatively unchanged.

You may continue to review all the items by clicking on each row in the item list or by navigating with the up and down arrow keys. Note that the textual summaries provided by IATA are based solely
on statistical evidence and are not informed by the content of items. An item that is given a poor rating by IATA may not be a poor item universally. However, a poor rating indicates that the item may not provide useful information when the current test is used with the current population.

In general, the recommendations that IATA provides for editing or removing items should be considered in the context of the purpose of the test and the initial reasons for including a specific item. Although some items might be retained regardless of their statistical properties because, for example, of the need to adequately represent key aspects of the curriculum, all items with negative discrimination indexes should be removed or rekeyed (if the key has been entered incorrectly) before proceeding with other analyses. Such items introduce noise or unwanted variation into the item response data and reduce the accuracy of estimates for other items. Removing some apparently weak items during analysis of pilot data will help increase the accuracy of the statistical results. However, the selection of the final set of items following the pilot or trial testing should be carried out jointly by subject matter specialists working closely with the person or team responsible for the overall quality of the national assessment test.

When you have finished reviewing all the items, click Next>> to continue.

**STEP 5: TEST DIMENSIONALITY**

One of the statistical assumptions of IRT, as well as a requirement for the valid interpretation of test results, is that performance on the test items should represent a single interpretable construct or dimension. Ideally, a national achievement test of a construct such as mathematics or science should measure the single construct or dimension that it is designed to measure and should not measure other constructs or dimensions such as reading ability. The purpose of the test dimensionality interface is to detect any violations of the assumptions that (a) a single dominant dimension influences test performance and (b) the relationships between performance on pairs or groups of items can be explained by this dominant dimension.
In most cases, the second assumption proceeds from the first, but for long tests (with more than 50 items), small groups of items may be locally dependent without having a noticeable effect on overall test dimensionality.

The analysis of test dimensionality determines the degree to which the test measures different dimensions of proficiency and the extent to which each item relates to each dimension. The fewer the number of dimensions that strongly influence the test items, the more valid interpretations of the test scores will be. Although this evidence is insufficient to confirm a test’s validity, it can provide important information on the content of specific items. Other aspects of validity, such as content validity (which is very important in the context of a national assessment), are typically considered more important than statistical data when determining the validity of a test or an item. Anderson and Morgan (2008) provide a description of procedures designed to ensure that a test has adequate content validity.

From a statistical perspective, the estimation of IRT parameters and scores depends on the concept of likelihood, which assumes that the probability of an event (for example, a correct response) is conditional on a single dimension representing proficiency. If items are conditional on different dimensions, then the estimated parameters and scores will be incorrect.

In figure 9.9, the graph on the right illustrates both the scree plot for the overall test and the square factor loadings for the first item, MATHC1019. On the left side of the interface is a table similar to that in the item analysis interface. Summary symbols (as described in chapter 8 of this volume) in the column labeled F next to the item Name column describe the overall suitability of an item in terms of its relationship to the primary dimension common to most other items on the test. To the right of the Name column, the classical item facility (PVal) is displayed, along with the loading of the item on the primary dimension (Loading). The loading, which ranges from −1 to 1, is the correlation between performance on each item and the primary test dimension. For example, the value of 0.34 for MATHC1019 indicates that the scored responses to this item have a correlation of 0.34 with the overall test score (percent correct). There is no ideal value, but loadings close to 1 are indicative of better items.
The results in the table should be interpreted together with the graphical results displayed on the right side of the interface. The main result displayed in the graphics window is the **scree plot**, which describes the proportion of variance (eigenvalue) explained by each potential dimension (for example, reading ability). The circle-shaped markers illustrate the relative influence of each potential dimension (eigenvalue)$^3$ on the overall test results, while the solid line connecting box-shaped markers describes the relative influence of each potential dimension on the individual test items (squared loading). The magnitude of the eigenvalues is less important than the pattern of the scree plot. The scree plot for the overall test should have a single large eigenvalue on the left, followed by all other eigenvalues, which should be relatively small and similar in magnitude (figure 9.10). This L-shaped pattern in the plot, with only two distinct line segments, suggests that a single common dimension is responsible for the **PILOT1** test results. The greater the number of distinct line segments it takes to connect the top-left point to the
near-horizontal line at the bottom, the more dimensions are likely to underlie test performance.

Selecting each item in the list on the left will display the item-specific scree plot on the right. Ideally, the plot for individual items should be similar to that of the overall test; the highest value in the item-specific line in the graph should be on the far left (corresponding to the main dimension of the test). However, item-specific characteristics may introduce different patterns, which are not necessarily problematic. For example, item MATHC1019 in figure 9.9 is not problematic; although some nonzero loadings occur on other dimensions, the strongest loading is on the primary dimension. In general, the item-specific results need to be consulted only if more than one dimension clearly underlies test performance (that is, more than two distinct line segments are visible). In that case, one should identify and examine items for which item-specific plots have squared loading values corresponding to the same dimensions as the problematic eigenvalues.
One caveat in the interpretation of scree plots is the effect of item facility. In tests where most items have similar item facilities, items with facilities much higher or lower than the other items tend to produce artificial difficulty factors, particularly with nonnormal distributions of percent-correct test scores. The items with extreme facilities may appear to define a separate factor simply because some students (for example, high or low performers) will generate patterns of response that appear unusually strongly related compared to the relationships between other test items. However, these difficulty factors are not inherently problematic. Reviewing the item loadings may help determine whether secondary factors are artifacts or actual problems. To determine if a secondary factor is a difficulty factor, one should examine the item loadings of the items with low (< 0.2) or high (> 0.8) item facilities (PVal). If the item loadings of these items have a peak that corresponds to the position of the secondary factor, the secondary factor is most likely a difficulty factor and can be ignored.

**Item Loadings**

The IRT model assumes local independence between items, meaning that responses to an item should not depend on the responses to another item. Ideally, under IRT, a test should have questions that are independent in all dimensions except for the primary test dimension. Significant local item dependency can result in inaccurate estimation of item parameters, test statistics, and student proficiency. For example, a mathematics test that includes a complex problem-solving question might assign a set of different scores for each of the logical steps required to compute the final answer. If the test taker answered step 1 incorrectly, this will influence the probability of a correct response on each subsequent step. This set of dependent test items would be inappropriate for IRT modeling; in this case, the item should be properly treated as a single partial credit item.

Because local dependence is typically problematic only in items that are weakly related to the primary dimension, the most effective way to use this interface is to sort the items in the **Loading** column by clicking on the column header once (see figure 9.10) and comparing
the poorly loading items to identify common peaks in their item loading graphs. If many poorly loading items have peaks in their loading plots that correspond to the same dimension, they may have some local dependency. Because these statistics tend to be sensitive to sampling error, any results from the statistical review should be used to motivate more detailed review of item content rather than to make definitive decisions.

After sorting of the items, the selected item is MATHC1075. Because this item was removed from the analysis in the previous item analysis step, its loading is NaN, and no results are shown for it (the graph displays only the scree plot for the entire test). IATA assigns a triangular warning symbol to any item whose dimensionality may be problematic in terms of affecting the estimation of other statistics. Note that IATA has flagged only one other item (MATHC1035) with the triangular warning symbol (figure 9.10). The item is relatively weakly related to the primary dimension and has a noticeably stronger relationship to the secondary dimension, which suggests it may be measuring a dimension that is distinct from that of most other items. However, these results by themselves are not conclusive evidence to warrant removal of this item from the test. Curriculum experts and experienced teachers should review any statistically problematic items to determine whether a content-related issue might warrant their removal or revision.

IATA assigns a diamond-shaped caution symbol to any item that has a stronger loading on a secondary dimension than on the primary test dimension; these items will not be problematic in subsequent calculations. A typical example is shown in figure 9.11 for MATHC1002. This item is related to several dimensions, but because these dimensions have such little influence on the overall test results, as indicated by the relatively small eigenvalues (dashed line) corresponding to the peaks of the strong loadings (solid line), determination of whether the dimensionality of the item is acceptable or not should be a matter of test content rather than of statistics.

All tests are multidimensional to some extent because all items cannot test the exact same thing. Therefore, if the overall scree plot does not indicate any problems, the effects of any item-level
multidimensionality or codependency will likely be negligible. For this example, all items are retained for subsequent analyses because the overall scree plot does not indicate any problems.

After finishing a review of the items, click Next>> to continue to the differential item functioning (DIF) analysis interface.

**STEP 6: DIFFERENTIAL ITEM FUNCTIONING**

DIF analysis examines the extent to which the IRF of an item is stable across different groups of students. If the IRF differs for two groups (for example, males and females), the scores that are estimated using it may be biased for one group or for students within specific ranges of proficiency in the group. The DIF analysis controls for differences in average group proficiency, meaning that the relative advantages and disadvantages expressed by the DIF results are independent of differences in average proficiency in the groups. For example, if one is interested in the extent of gender bias in a particular test item,
the results of the DIF analysis would indicate if the item is biased in favor of boys or girls after taking into account overall test score gender difference.

The DIF analysis interface is shown in figure 9.12. On the left-hand side is the set of four controls used to specify the analysis. The drop-down menu at the top allows you to select a variable from the list in the response data that are not test items. Once you select the variable, IATA will list its unique values in the Possible Values table, along with the (unweighted) percentage of students with each value. To select the groups to compare, first click on the value for the desired focus group, and then click on the value representing the reference group. The focus and reference group specification determines how the summary statistics are calculated; the estimations use the sample distribution of proficiency of the focus group to calculate average bias and stability statistics. To change focus and reference groups, click on the values in the Possible Values table; the values assigned to focus and reference groups will be updated in the text boxes at the bottom left. The DIF statistics are most sensitive to the
focus group, so the usual practice is to ensure that the focus group is a minority or historically disadvantaged group.

For this example, a DIF analysis is performed using the variable Sex to see if female students (coded 1) are disadvantaged relative to their male (coded 2) counterparts. To specify this analysis and review the results, perform the following steps:

1. From the drop-down menu on the left, select the Sex variable. The table beneath will be populated with the values 1.00 and 2.00, with values of 50 percent for each value, indicating that the sample has equal numbers of males and females.

2. In the table of values, click on the value 1.00. This will cause the value of 1.00 (representing females) to be entered as the focus group in the text box beneath.

3. In the table of values, click on the value 2.00. This will cause the value of 2.00 (representing males) to be entered as the reference group in the text box beneath.

4. Click Calculate, and wait for the calculation to complete.

5. When the calculation is complete, in the item list, click on the header of the S-DIF column to sort all the items by the value of the S-DIF statistics.

After completing these steps, you will see an interface similar to that in figure 9.12. In this example, IATA flags 15 items with either a warning or a caution symbol. For each item, two statistics are calculated, S-DIF and U-DIF. S-DIF describes the average vertical difference between the groups (focus minus reference), and U-DIF describes the average absolute differences between groups. The value of the U-DIF statistic is always positive and larger in absolute value than that of S-DIF. Even if one group shows no systematic advantage (S-DIF is close to 0), an item may have a stronger relationship with proficiency in one group, which would produce a larger U-DIF statistic.

**MATHC1035** is an example of an item with consistent DIF, where the absolute values of S-DIF and U-DIF are identical (see figure 9.13). For this item, the female advantage is apparent across the entire
The consistent difference suggests that females are more likely to perform better on this item than males, even if they have the exact same level of proficiency. The S-DIF statistic indicates that, on average, the probability of correct response for females was more than 23 percentage points higher than for males of comparable proficiency.

With DIF analysis, the statistics and figures tend to be very sensitive to sampling error, which may lead to items appearing to have differences that might not be present in a larger sample. IATA assigns a warning symbol when the coefficient of sampling variation\(^5\) for the S-DIF statistic is less than 0.2 (indicating that the observed difference is most likely not due to sampling error) or where a very large difference occurs in either S-DIF or U-DIF that should be examined even in small samples.

Because of sensitivity to sampling error, the graphical results may occasionally be misleading. When the number of respondents at the upper and lower ends of the proficiency scale is small, the responses of one or two students may dictate the appearance of the graphs at these extremes. As the summary statistics weight the calculation by

![FIGURE 9.13](image-url)
the number of focus group students at each proficiency level, they are not affected as much by random error as the graphs. The graph for the results for MATHC1042 in figure 9.14 provides an example of how graphical results can mislead in some instances. Although the graph suggests a very large disadvantage for females (the lightly shaded region), the actual S-DIF statistic (−2.01) indicates a relatively weak disadvantage.

Evidence of DIF might also be found when item-specific content is not as strongly aligned with the primary test dimension as other items. For example, in mathematics, a common learning objective for younger students is to recognize measurement tools for different units (such as centimeters, kilograms, or degrees centigrade). Students in remote or disadvantaged areas, even if they are strong in mathematics, may not have the same exposure to these tools as students in urban areas. As a result, they may be systematically disadvantaged on test items requiring this specific knowledge. However, this disadvantage is not a property of the test items; it is a consequence of a specific disadvantage in proficiency. Before reaching any conclusions about
bias against specific students, curriculum content experts who are sensitive to possible ethnic, geographical, or gender differences should examine the test items to confirm that evidence of bias from statistical evidence agrees with evidence from an analysis of content.

DIF analysis should be performed for all demographic characteristics and groups that will be compared in major analyses of results. The presence of DIF with respect to one characteristic typically has no relationship to the presence or absence of DIF with respect to another characteristic. Usually, the most important variables to consider for DIF are the sampling stratification variables (such as Region) or variables from the background questionnaire. The PILOT1 data have three demographic variables: Sex, Language, and Region. As an independent exercise, you can carry out similar DIF analyses for Language and Region by completing the same steps as for the Sex DIF analysis, making sure to select the minority group as the focus group and to click Calculate to update the results.

Figure 9.15 illustrates a common DIF result when a test is in the home language of some students but not in the home language of...
other students. The results are from a DIF analysis for item MATHC1006. This item is an extreme example of DIF in that the correct response is strongly related to language proficiency in one population (in this case, Language = 2) and has a weak or nonexistent relationship in the other (Language = 1).

The DIF analysis in IATA can serve as a research tool to determine if specific groups of students have problems with specific subdomains. DIF analysis can also facilitate an understanding of differences that may be introduced in the language versions of a test that has been translated. Statistical evidence of DIF can be used to help translators correct translation errors revealed during pilot or trial testing.

The primary purpose of DIF analysis is to prompt discussion and review of the pilot test items and to guide the interpretation of results. For each DIF analysis that is run, IATA saves the results to a data table. These results, and any particularly interesting graphs, should be copied, saved, and shared with curriculum content specialists to determine possible explanations for the pattern of differences between focus and reference groups. If clear agreement exists that an item is biased, it should be removed from the analysis specifications on page 2 of IATA, and the previous IATA analyses should be repeated. Finally, because the results of DIF analyses are notoriously susceptible to sampling error, any decision about whether or not to include a particular test item in the final version of a test based on the suspicion of bias should have a strong curriculum or content justification. This walk-through proceeds without removing any of the test items.

After performing DIF analyses and reviewing the results, click Next>>.

**STEP 7: SCALE REVIEW**

The technique of developing a numeric metric for interpreting test performance is called *scaling*. IATA reports test results using the following scale scores: PercentScore, Percentile, RawZScore, ZScore, IRTscore, and TrueScore. These scales were described in greater detail in table 8.1. Performance on these default scales is summarized on either a scale of 0 to 100 or the standard scale, which has a mean of
0 and a standard deviation of 1. Use the scale that is most useful to the intended purpose of communicating results. Different stakeholders may prefer different types of scales. In general, the IRTscore is the most useful across the widest range of purposes, but it has the communication disadvantage that approximately half the students have a score less than zero. Because many stakeholders may not know how to interpret negative scale scores, creating a new scale is preferable so that none of the student scores has a value of less than zero.

The interface for reviewing the scale scores and creating additional scale scores is shown in figure 9.16. On the left side is a drop-down menu and a graph window. One can select any of the scale score types from the drop-down menu, which will graph the distribution of the selected scale score. The figure presents the graph for the selected scale score, PercentScore. On the right is a panel presenting summary statistics for the selected score. At the bottom right is a set of controls for rescaling the IRTscore by applying a new standard deviation and mean. The rescaling procedure applies only to the IRTscore, which is the primary score output of IATA.
Test Score Distributions and Test Information

IATA displays score distributions as histograms, in which each bar represents a range of scores and the height of each bar represents the proportion of students with scores in that range. For score types that are expressed on scales with means of approximately 0 and standard deviations of approximately 1 (StandardizedZscore, RawZScore, and IRTScore), IATA also plots the test information function as a solid line. The test information function describes the test’s accuracy at different proficiency levels on the standard scale on which the items are scaled. It is inversely related to the standard error of measurement; if the test information is high, the standard error of measurement will be low. The test information function should be interpreted in relation to the specific testing needs or purpose of the test. For example, if the purpose of the test is to identify low-proficiency students, a test that is most accurate for high-proficiency-level students would be unsuitable and would not serve as an appropriate measure for identifying low-proficiency students. In general, the average error of measurement for all students is minimized if the information function for a test is slightly wider but about the same shape and location as the distribution of proficiency for the students being tested. Comparing the test information function to the distribution of test scores can indicate if the test design would benefit from modifying the balance of items with greater accuracy for high or low performers.

Summary Statistics

IATA produces the following summary statistics for each test score:

1. Mean
2. Standard deviation
3. Skewness
4. Kurtosis
5. Interquartile range
6. 25th percentile
7. Median
8. 75th percentile
9. Response rate
10. Reliability
11. Total number of respondents
12. Number of items in the test
13. Number of items included in the analysis

The first eight statistics describe the distribution of estimated scores. Use the scroll bar on the right of the statistic and value box in figure 9.16 to view the last five rows. These statistics help determine the adequacy of the scale scores for various purposes (for example, secondary statistical analysis or reporting by quantiles). The last five statistics describe the conditions under which the analysis was conducted and provide a holistic rating of the test, which should be checked to confirm that the analysis was conducted on the proper data according to correct specifications. Response rate describes the average number of valid (nonmissing) responses on each of the items. Reliability is an overall summary measure of a test’s average accuracy for the given sample of students. Both response rate and reliability range from 0 to 1 and should be as high as possible. The total number of items included in the analysis reflects the fact that some items may be dropped from the analysis if they have been considered inadequate because of poor wording, because they are confusing to students, or because of other technical inadequacies. For the current walk-through, the number of respondents is 262, the number of items is 80, and the number of acceptable items is 79. (MATHC1075 was removed from the analysis.)

The scaling interface is more useful for final assessment administrations than for pilot testing. Because the unweighted pilot test sample is not representative, the distributions of results should not be generalized to population performance. Furthermore, because no test scores will be reported, no need exists to generate derived scale scores. Further results from the scaling interface are not relevant to the analysis of the PILOT1 data. Click Next>> to continue to the next task.
STEP 8: SELECTING TEST ITEMS

Optimal selection of items using IATA is available whenever an item data file has been loaded or created during an analysis of response data. IATA can automatically select items based on their statistical item characteristics to produce the most efficient test for a given test length and purpose. The basic principle underlying IRT-based test construction is that the test designer has some expectation about the degree of measurement error that a test should have at different levels of proficiency in addition to requirements about the balance of content that must be included in the test.

In general, the more items that are in a test, the more information it can generate about the proficiency levels of examinees. Unfortunately, however, tests with too many items are generally neither practical nor desirable; they can be unnecessarily disruptive in schools and can result in test-taker fatigue and deterioration of student motivation, resulting in less accurate results. Overly long tests are also costly to develop, administer, score, and process. To be most efficient, a test should include only the most informative test items from the pool of available items. IATA can help develop a test with the minimum number of test items required to meet the purposes of policy makers and other stakeholders.

Determining an acceptable level of standard error depends on the purpose of an assessment. Building a test that would provide a high level of accuracy at all proficiency levels would be ideal. However, this would require many items, which would increase the length of time each student spends taking the test. The effect might be to lower the validity of the test results by allowing fatigue and boredom to influence test scores. If a test is norm referenced, detailed information (and lower error of measurement) is required for all levels of proficiency. In contrast, if a test is criterion referenced, information is required only around the proficiency thresholds at which decisions are made.

However, item selection at the pilot stage should not be determined solely by the results of statistical analysis. The validity of the interpretation of results is the most important consideration in constructing national assessment tests. Test scores should adequately
and accurately represent the domain being measured. The most important tools for maintaining test validity are the theoretical frameworks and the table of specifications or test blueprint. A blueprint helps determine the balance of content and cognitive skill levels to be included in a test (see Anderson and Morgan 2008).

The interface for selecting optimal test items is shown in figure 9.17. On the left, a drop-down menu allows one to select a source for item selection from a list of available data sources that are produced automatically by IATA on the basis of the loaded data and any analyses one has performed (see table 8.5). In this example, the Items1 table is available, which contains the results of the current analysis. Beneath the data source selection are fields that allow one to specify the name to be applied to the item selection and the total number of items to select from the item data. The table beneath these fields contains a list of all the calibrated items in the selected data source, along with the proficiency level (Level) and content area (Content) associated with each item. Although the latter two data fields are typically read into IATA in an item data file, the data may also be manually edited.
directly in the table. The statistical selection process does not require **Level** and **Content** specifications, but having detailed information about each item will help optimize the selection of items while maintaining desired content representation. Clicking the box to the left of an item name so that it is checked forces IATA to select the item, regardless of its statistical properties.

Beneath the item table, two sliding controls allow you to specify the range of proficiency within which you wish to maximize test accuracy. The controls are set so that the minimum value corresponds to the 2nd percentile of proficiency and the maximum corresponds to the 98th percentile (the current selected value is displayed to the right of each sliding control). You can specify a narrower range in which to maximize the information by modifying upper and lower bounds to reflect your assessment goals. IATA will select items to minimize the average standard error of measurement in the range of proficiency between the lower and upper bounds, assuming a normal distribution of proficiency in the student sample to be assessed.

The primary purpose of pilot testing items is to determine which items will be most useful in the final administration of the national assessment. If the pilot sample of students is perceived to be above average in proficiency, this expectation should be taken into account in selecting items. Keeping in mind that you wish to create a 50-item final test, enter the following specifications into IATA:

1. In the **Name of item selection** box, type **50Items** (the name is arbitrary; the name is used here so that one may compare the results produced with the results in the IATA sample data folder).
2. In the **Total number of items** box, enter the number 50.
3. Move the slider control for the **Upper bound** so that it has a value of 80. This specification indicates that the item selection will not attempt to maximize accuracy above the 80th percentile in proficiency distribution of the sample. This setting is chosen to offset the possibility of higher proficiency of the pilot sample relative to the general population.
4. Click **Select Items**.
When IATA has performed the task, the interface should appear as in figure 9.17. On the left side in the items list, you can view the actual 50 items that have been selected. (The last one is MATHC1041.) On the right side, the graph displays the collective information and expected error of measurement of the selected items if they were administered as a test. The results indicate that the item selection is most accurate around the proficiency score of zero (average proficiency in the current sample). The table beneath the graph summarizes the distribution of selected items across content areas and cognitive levels (for these data, all items have been given a default value of 1; values may be edited directly in the item table or uploaded in the initial item data file). If the data in this table indicate that the statistically optimal selection does not adequately conform to the test blueprint, you can modify the balance of content by manually selecting and deleting specific items using the boxes next to each item name in the table on the left. Manually selecting items will automatically update the summary of the test properties on the right.

The item selection is also recorded as an item data table in IATA with the name CustomTest50ItemsA. As with all results produced by IATA, you can view and export this data table by advancing to the final interface of the workflow (see “Step 10: Viewing and Saving Results”). The items in the table are sorted in the order of suitability for the selection criteria, with the most suitable items at the top.

Given the small number of items in the current analysis, a user may use IATA simply to sort all of the items in order of suitability to the desired range of proficiency (that is, below the 80th percentile in the current sample). The test development team may then review the item data file and, when selecting items for the final test, use a ranking of the items in terms of suitability while ensuring that the appropriate balance of content is maintained. To create a new item selection, perform the following steps:

1. Click **Clear** to remove all previous selections from the item list.
2. Enter a new name for the item selection, **79Items** (if the name was already used, the previous results would be overwritten).
3. Enter the maximum number of items available (79) as the total number of items. If you enter a number that is greater than the number of available items, IATA will select only from the available items.

4. Leave the upper bound at 80, because the target range of proficiency has not changed.

5. Click Select Items.

Figure 9.18 presents some of the results of the analysis of the 79-item pilot test. A table of results (named CustomTest79Items) has been added to the IATA result set, which may be viewed on the final interface of the workflow. Test developers can use this information to help improve the quality of items by identifying and remediating the least effective items.

The process of item selection depends on the quality of available items. IATA cannot introduce accuracy to specific regions of proficiency if no items are available with information in those regions.
The automated process can help select the best items that are available, but it cannot make the items themselves more accurate.

After reviewing the results, click Next>> to continue.

**STEP 9: PERFORMANCE STANDARDS**

At the pilot test stage, insufficient evidence exists to support the setting of performance standards. Although some information is available about the statistical item properties and the specifications that were used to create the items, no detailed information is yet available about the distribution of proficiency in the student population. Therefore, any attempt to set performance standards at the pilot stage would be unnecessary and potentially misleading.

Because this walk-through example of the analysis of pilot test data does not require any standard setting, click Next>> to continue to the results viewing and saving interface.

**STEP 10: VIEWING AND SAVING RESULTS**

For all analysis workflows, IATA produces a number of different results in data table format. These results can be viewed and saved on the final interface of each workflow. This will allow you to review each of the data tables of results produced during the analysis workflow. The interface displays the data table that is selected in the drop-down menu. To change the data source, select a different table from the drop-down menu, as shown in figure 9.19. (Table 8.5 provides a complete list and description of the available data tables produced by IATA.)

Note that, although the creation of any performance standards was not specified, the table PLevels is created automatically using default specification values.

You may save these tables of results in a single output file or in multiple files by clicking Save Data. You may save a single table or all tables at once to a variety of formats. Two file formats are recommended for saving IATA output: Excel (*.xls/*.xlsx) and SPSS (*.sav).
In general, Excel is preferable, because all data tables may be saved into a single data file. The Excel format may also be opened in free software such as OpenOffice (downloadable from http://www.openoffice.org). However, early versions of Excel are limited to a maximum of 255 variables. If a data file has more variables, IATA will save only the first 255 into the *.xls. To save larger data files, use the *.sav or *.xlsx formats. SPSS files have the advantage that they can store larger data tables efficiently and can store metadata (if they are edited in the SPSS software package). Note, however, that SPSS has one main limitation: each data table will be saved into a separate file.

A file dialog will ask you to specify the file name and location for the results, as well as the output format. Choose the desired data format, and click Save to finish saving the table or tables. The resulting files contain all the tabular results produced during the entire analysis workflow, providing documentation of the analysis.

For reference, the results of this analysis walk-through from the Items1 results table are included in the ItemDataAllTests.xls file. The worksheet containing the data from the Item1 table in the
current analysis has been renamed ReferenceP1. In the saved results, the “True” and “False” values in Column E (OK) indicate which items were included in the final analysis. In these results, only MATHC1075 has a value of “False.”

For a real pilot test analysis (that is, one not using simulated data), the results tables and any graphics that have been copied and pasted during the analysis workflow should be provided to the test developers, who can use the information to modify the test by selecting, ordering, and adding items, as required, to maximize the accuracy and usefulness of the final test form.

NOTES

1. See table 8.6 for a description of the symbols and their meanings.

2. A loading equal to 1 is unreasonable, because this would require each respondent to have the same score on every item. This requirement implies that the test could produce only two distinct score values, which is not very informative.

3. The values displayed in IATA have been standardized to express the proportion of total variance accounted for by each eigenvalue.

4. Clicking on the header twice will sort the column in descending order.

5. The coefficient of sampling variation is calculated as the standard error of the S-DIF statistic divided by the absolute value of the S-DIF statistic.

6. All results from this walk-through are available for reference and comparison in the IATA sample data folder in the Excel table named ReferencePILOT1.xls. The DIF result tables are in the worksheets with names beginning with DIF_.

7. One can copy any of the DIF analysis graphs by placing the cursor on the graph and using Copy and Paste functions from the right-click menu.

8. For analyses that involve linking, select from previously calibrated item data (Items2) or the set of items that are common to two item data sources (MergedItems).

9. If one saves all tables and selects the SPSS (*.sav) output format, each result table will be exported as a separate *.sav data file, with the name provided as a prefix to all the table names.
Use the CYCLE1 sample data set to carry out this exercise. The answer key for this test is in the Excel workbook, ItemDataAllTests.xls, in the sheet named CYCLE1.

The analyses in this chapter are based on the performance of students in a national assessment of mathematics administered to a national sample of students. The final test had 50 items, representing five content areas (number knowledge, shape and space, relations, problem solving, and uncertainty) in proportions determined by the test specifications. The final sample design was a stratified cluster sample, with schools as the primary sampling unit and a target sample of 30 students from each school. The sample comprised 79 schools, selected to be representative of five national regions and stratified by rural status and language of instruction. The total number of students in the sample is 2,242, representing a population of approximately 86,000.

This walk-through follows the same steps as the analysis of pilot test data in chapter 9. However, because the final test is concerned primarily with producing and interpreting scores, the item analysis is typically performed without the exploratory emphasis present in the analysis of pilot test data. Accordingly, this walk-through focuses on
the unique aspects of final test data analysis that distinguish it from analysis of pilot test data. In addition to the analyses carried out with pilot data, the analyses of the full test data in this chapter involve the calculation of scale scores and performance standards. Where the steps of analysis are identical to those described in chapter 9, refer to the information presented there.

Begin the analysis by clicking **Response data analysis** from the Item and Test Analysis (IATA) main menu.

**STEP 1: SETTING UP THE ANALYSIS**

The procedures for setting up the analysis are similar to those described in chapter 9. First, load a response file, then load an item data file, and then specify the analysis. If more information is needed, refer to chapter 9, step 1 to step 3, for detailed instructions. The IATA sample data folder contains the following:

- The response data file for this chapter is **CYCLE1.xls**. (This file has 2,242 records and 58 variables.)
- The item data file is in the Excel file named **ItemDataAllTests.xls** in the table named **CYCLE1**. Ensure that the correct table name is selected in the item data-loading interface. (The **CYCLE1** item data file has 50 records.)

The items in the national assessment test are a subset of the pilot items described in chapter 9.

The specifications for the analysis are slightly different from those for the pilot test data analysis, primarily because of the use of probability sampling in the full administration of the national assessment. The first difference is the name of the identification variable, which is **CYCLE1STDID**. The second difference, which will affect the results of the analysis, is the presence of a sample design weight, which is named **CYCLE1weight**. These variable specifications must be selected from the drop-down menus. In these data, the value of 9 represents missing responses that are treated as incorrect. The completed specifications should look like figure 10.1.
Note that the item data for the final assessment also include data in the **Level** field in the third column of the table on the left. These data are natural numbers (1 or greater) that represent the expected level of performance or proficiency that the curriculum content specialists assigned to each test item: Level 1 represents the lowest level of performance (that is, minimum competence), and Level 4 represents the highest level. Although every item is assigned a level, some students may not achieve even the lowest level.

After verifying that the specifications and data are correct, click **Next>>** to continue. The analysis will begin automatically, updating the interface with the progress periodically. With larger data sets or slower computers, the analysis may appear to be slow at the stage of estimating parameters, which is the most time consuming. Do not close the program; IATA will continue to run and will provide an update when the analysis is complete.
STEP 2: BASIC ANALYSIS RESULTS

Because problematic items were identified and removed during the analysis of the pilot test data, no problematic items remain in the full data set. Confirm that the items are behaving appropriately by reviewing (a) the item analysis (IATA Page 4/10) and (b) the test dimensionality (IATA Page 5/10) results. For instructions on how to perform these steps, refer to chapter 9, steps 4 and 5. Note that all the items listed on IATA Page 4/10 have (green) circles with the exception of MATHC1046, which was identified in chapter 9 as being somewhat problematic but which was left in the test. Proceed to the differential item functioning (DIF) interface (IATA Page 6/10) when finished.

STEP 3: ANALYSIS OF DIFFERENTIAL ITEM FUNCTIONING

Although DIF analysis was performed on the pilot test data, replicating the analyses with the full sample is good practice, because DIF analysis results tend to be sensitive to sampling errors. Other reasons for performing DIF analysis are that variables may be available in the full sample that were not available in the pilot sample and that the sample provides a more satisfactory number of cases for DIF analysis.

Here DIF analysis is performed to examine the possibility of urban bias—if rural students are disadvantaged relative to their urban counterparts. For the CYCLE1 data, a value of 1 for this indicator means that a student is attending a rural school; a value of 0 means a student is attending an urban school. To specify this analysis and review the results, perform the following steps:

1. From the drop-down menu on the left, select the Rural variable. In response, the table beneath the drop-down menu will be populated with the values 0.00 and 1.00, with values of 56 percent for 0.00 and 44 percent for 1.00, indicating that 44 percent of the students (unweighted) in the sample attend rural schools.
2. In the table of values, click **1.00**. This will cause the value of 1.00 (representing rural students) to be entered as the focus group in the text box beneath.

3. In the table of values, click **0.00**. This will cause the value of 0.00 (representing urban students) to be entered as the reference group in the text box beneath.

4. Click **Calculate**, and wait for the calculation to complete.

5. When the calculation is complete, click on the header of the S-DIF column in the item list to sort all the items by the value of the S-DIF statistic.

When these steps are completed, the interface will appear as illustrated in figure 10.2. Most of the S-DIF and U-DIF statistics are less than 5, indicating that, after controlling for differences in proficiency between rural and urban students, the differences in item performance between students in these locations tend to be negligible.

**FIGURE 10.2**

DIF Analysis Results for CYCLE1 Data by Location, MATHC1043
The purpose of performing DIF analysis at the final test stage of a national assessment is to determine if an item should be made ineligible for calculating student scores. At this stage of the analysis, sharing the statistical analysis results with the national assessment steering committee would be appropriate; the committee will determine if potentially problematic items should be removed or retained. If an item is removed, the analysis may be rerun after either deleting the item’s answer key in the analysis specifications interface or unchecking the item in the item analysis interface. For the current example, the assumption is that all items are retained.

After reviewing all items, click Next>> to continue.

**STEP 4: SCALING**

The default scale used to calculate the results for the item response theory (IRT) scale scores is the standard, or Z scale, which has a mean of 0 and a standard deviation of 1. Many stakeholders appear to have problems with scores expressed on this scale, because half the students will have negative scores. Similarly, scores bounded by 0 and 100 have communication challenges; many audiences tend to assume that a score of 50 represents a passing score, which may not be the case, depending on the test specifications.

For communication purposes, reporting test results with an average score less than 50 percent or below 0 may be undesirable. Some large-scale assessments transform their calculated scores into scales that have a mean of 500, 100, or 50 and standard deviations of 100, 20, and 10, respectively. Each national assessment team should select the type of score that is most likely to facilitate effective communication of results.

Two types of scaling can be performed in IATA: setting the scale and rescaling. Setting the scale allows you to specify the desired mean and standard deviation of the scale scores. Rescaling allows you to apply a simple linear transformation to the IRT scores, which is useful if the scale scores must be compared to a scale that has been established from a previous analysis. In this case, item parameters from the previous cycle can be used to estimate test scores or equate results.
from the student data in the new cycle so that the IRT scores that IATA calculates are comparable to the previous cycle’s calculated IRT scores. The calculated results can then be rescaled using the rescale function so that they are comparable with the reported scale from the previous cycle.

In either case, the new scale score is created by entering the name of the new score and specifying the standard deviation and mean in the appropriate boxes. When you click the Calculate button, IATA will produce the new scale scores and display the distribution and summary statistics.

In contrast to the pilot test analysis, whose main function is to inform test design, the primary function of the analysis of national assessment test data is to produce scores. Therefore, this walk-through requires closer examination and specification of the properties of the test scores, both of which make use of the scaling interface. First, comparing the distribution of proficiency scores to the accuracy of the test at each proficiency score (also known as test information) informs the quality of inference that may be made about different ranges of proficiency. Second, creating a reporting scale for the test results establishes a metric for communicating results to stakeholders.

The graph in figure 10.3 indicates that test information, illustrated by the solid black line, is well distributed relative to the distribution of proficiency in the sample. The frequency spike at the left side of the graph at approximately −3 on the proficiency scale corresponds to students who did not answer any items correctly on the test. The test does not have sufficient information to determine precisely how low these students’ proficiency levels are; as a result, they are all assigned the same arbitrarily low score.

To review the distribution of IRT scores, select IRTscore from the drop-down menu at the upper left of the interface. The interface will update with descriptive details about the IRT scores and the test information, as shown in figure 10.3. The mean of the IRTscore distribution is −0.02, and the standard deviation 1.08. These values are not meaningful in themselves because they represent the arbitrary scale on which the items were calibrated.

Compare these results with the statistically ideal shape of the test information function in terms of maximizing the overall test
reliability for a normally distributed population, which is illustrated in figure 10.4. For comparison, the standard normal distribution is drawn with a dashed line. The ideal information for a sample should provide the most information at proficiency levels that represent many students, but it also needs sufficient information to distinguish between extremely high- and low-proficiency students.

These results also indicate that the test was relatively difficult for students. The peak of the information function tends to be located at the region of proficiency where students are most likely to score 50 percent. In figure 10.3, this peak is slightly above the mean score of −0.02, indicating that above-average students tended to score only 50 percent correct.

To produce a more useful reporting scale based on the IRT score, use the **Add New Scale Score** functions at the bottom right of the interface (see figure 10.3). For this example, assume that the national steering committee requested a new scale that requires setting the mean equal to 500 and the standard deviation equal to 100.
This scale will be set in the first cycle of the national assessment and will be used in subsequent cycles to report changes in achievement over time. The name of this score will be NAMscore (National Assessment of Mathematics score). To provide these specifications, perform the following steps:

1. Type NAMscore in the field beneath the Add New Scale Score label.
2. Enter a value of 100 for the standard deviation in the Specify St. Deviation field.
3. Enter a value of 500 for the mean in the Specify Mean field.
4. Make sure the Set Scale option is selected. This ensures that the scale score produced will have a mean exactly equal to 500 and a standard deviation exactly equal to 100 for the sample. The Rescale option will simply adjust the existing IRT score by the specified mean and standard deviation.
5. Click Calculate.
When IATA is finished processing the request, it will update the interface with the summary graph and statistics for the newly created scale score, shown in figure 10.5.

Selecting a derived scale score has relatively few limitations. Any valid name can be used for this score as long as it has not been used in the response data (see chapter 8 for naming conventions and restricted variable names). The mean can be any real number, and the standard deviation can be any real number greater than zero. However, it is important to ensure that the lowest reported student scores are not less than zero. Because the lowest score is usually around three to four standard deviations below the mean, setting the mean to be at least four standard deviations above zero is good practice. The choice of a reporting scale should be discussed with the national assessment steering committee at the initial planning stages so that all stakeholders understand how to interpret reported results.

After the new scale score has been created, click Next>>> to continue.
STEP 5: SELECTING TEST ITEMS

The CYCLE1 data represent the initial cycle of a national assessment program. If the test is to be used in subsequent cycles for comparative purposes, a link to the initial cycle results will need to be established. To do this, select a subset of items that are both accurate and representative of the continuum of proficiency.

A reasonable practice for maintaining a strong link between tests is to keep approximately 50 percent of the items common to adjacent assessments. These are known as anchor items. To facilitate the process of selecting anchor items, use the item selection functionality of IATA to rank each of the items in the current test according to its suitability for maximizing accuracy across the proficiency range. To perform this selection, carry out the following steps:

1. Type the name ItemRanks into the Name of item selection field.
2. Type the number 50 in the Total number of items field to select all items.
3. Leave the lower and upper bounds at their default values of 2 and 98.
4. Click Select Items.

The complete results are shown in figure 10.6. All of the available items have been selected and categorized by content and cognitive level identified in their original specifications.

The results produced by these specifications are added to the set of results of the current analysis as an IATA item data table. This table should be provided to the test developers responsible for modifying the cycle 2 (or next) national assessment so that they can select a set of common items, taking into account information about the content and psychometric value of each test item used in the cycle 1 (or first) national assessment. Ideally, a set of anchor items should have half the number of items as the complete test and should represent the content and cognitive test specifications in the same proportions as the complete test. As a rule, any statistical linking where the link items constitute less than 20 percent of the total test is unlikely to provide
a meaningful link, regardless of the accuracy or content representation of the linking items. A pragmatic method of selecting items would be to begin with the most desirable items and allocate items to the cells of the new test specifications according to their content and cognitive levels until the desired number is reached in each cell or the list of items is exhausted.

Once IATA has completed this analysis, click Next>> to continue.

**STEP 6: SETTING PERFORMANCE STANDARDS**

Most contemporary assessments report results in terms of levels. International assessments such as Progress in International Reading Literacy Study (PIRLS), Programme for International Student Assessment (PISA), and Trends in International Mathematics and Science Study (TIMSS), as well as many national assessments such as the U.S. National Assessment of Educational Progress (NAEP), report student achievement scores in terms of performance or
benchmark levels (see Greaney and Kellaghan 2008; Kellaghan, Greaney, and Murray 2009). TIMSS, for example, uses four benchmarks: low, intermediate, high, and advanced (Martin, Mullis, and Foy 2008). The performance standards must be meaningful rather than arbitrary statistical thresholds such as percentiles, because they are the primary tool used to summarize and report student performance. The process of defining meaningful performance standards is known as standard setting.

IATA facilitates standard-setting procedures by first specifying response probabilities (RPs) of correct response for each item and then calculating the proficiency levels (RP values) associated with the specified RP. For example, if an RP were set at 50 percent, the RP value for an item would be the proficiency level associated with a 50 percent chance of responding correctly. A wide variety of RPs, typically ranging from 50 to 80 percent, is used in large-scale assessments. A common practice is to use 67 percent, which tends to be statistically optimal for classifying items. However, the choice of RP should also be informed by normative definitions of what probability of success constitutes mastery at a particular grade level and knowledge of the consequences of how the standards will be used. For example, in an educational context, where the consequences of reporting failure tend to be greater than those of reporting success, lower RPs may be preferred.

Before analyzing the data, a panel of stakeholders, including curriculum and teaching experts, in consultation with the national assessment steering committee, should decide on the number of proficiency levels to be used. Some national assessments choose two levels, such as acceptable and not acceptable; others choose three levels, such as poor, adequate, and advanced, whereas others use four or more. If the stakeholder panel decides on more than two levels, each level except the lowest should be defined by a set of items that are considered answerable by students displaying that level of performance. Generally, unless an assessment includes hundreds of items (requiring a rotated booklet design), only enough items will be available to adequately define three or four levels.

The interface for performing this analysis is shown in figure 10.7. On the left, a drop-down menu allows you to select a source of items
for item selection. As with the item selection interface, you have the option of selecting any of the item data sources available in the current workflow. For the current analyses, only the *Items1* table is available.¹ The items from the selected source are listed in the table beneath the drop-down menu. The values in the *Level* column may be edited directly in each row. To estimate statistically optimal thresholds based on the current item classification, move the vertical slider in the center of the interface to the desired RP. When the interface is opened, the default RP is 67 percent, indicating that the criterion used to rank items or estimate optimal thresholds is a 67 percent probability of correct response on each item.

When you click and drag the vertical slider to adjust its value, IATA will update the optimal thresholds and produce the results in the graph window on the right side and the table of results beneath it. The graph illustrates the position of each threshold with vertical lines relative to the distribution of proficiency and the test information function. This information illustrates the usefulness of the levels. For example, if a level has very few respondents, any summary

---

1. Reference or note for the availability of the *Items1* table.
statistics describing students at that level will be too small or unstable to be interpretable. Similarly, if the test is not accurate at the threshold of a level, the classification of students at that level will be inaccurate.

The table beneath the graph window in figure 10.7 describes the items representing each level with the mean and standard deviation of item \( b \) parameters. The right-most column in the table contains the threshold that was estimated for each level. For example, the mean and standard deviation of the \( b \) parameters for Level 4 are 0.77 and 0.38, respectively. The value of 0.77 indicates that the average of the \( b \) parameters for Level 4 items corresponds to a proficiency score of 0.77 on the IRT scale. The RP67 threshold for Level 4 is 1.32. These statistics are useful in determining if the assignment of items is reasonable. For example, if the standard deviation of items in a level is larger than the distance between the means or thresholds of adjacent levels, the statistical basis for defining the levels may be weak. In these results, the standard deviation within levels tends to be about 0.35, while the distance between adjacent levels ranges between 0.17 and 0.53, indicating that the levels are reasonably well defined.

Several methods are used for determining the most appropriate cut-points or thresholds between proficiency levels. One is called bookmarking, a procedure based on IRT that capitalizes on having item difficulty and person ability on the same latent dimension. It involves an expert standard-setting panel (such as curriculum specialists and experienced teachers) carefully reviewing all test items in the light of information available to them from test specifications, curricula, student test performance, and normative definitions of what students know and can do at each proficiency level (see Karantonis and Sireci 2006; Mitzel and others 2001).

Procedures governing how panels operate relating to the selection, training, and interactions of panelists and their use of data from diverse sources vary. They are not considered here. Rather, the focus is on how data generated by IATA can contribute to the establishment of proficiency levels.

At the outset, the panel prepares a specially constructed version of the national assessment test booklet containing multiple-choice and constructed-response items presented one per page in order of their
RP values. The task of the panel is to identify items at boundaries between cognitively distinct groups of items or levels. The panel then applies bookmarks or placeholders to the boundaries in the specially constructed test booklet. Items selected for the highest-level group (Level 4, for instance) are ones most likely to be answered correctly by students at that level but much less likely to be answered correctly by students at lower levels. Similarly, items selected for Level 3 are most likely to be answered correctly by students at that level (and at Level 4) but much less likely to be answered correctly by students at lower levels.

Suppose, for example, that the expert panel decided to use an RP of 50 percent to validate the initial classification of items by item developers. To provide the validation evidence, complete the following steps:

1. Set the RP to 50 percent by clicking and dragging the slider as shown in figure 10.8.
2. Click **Save Bookmark data**. IATA will produce a confirmation dialog to notify you that the data have been saved.
3. Click **Next>>** to navigate to the results viewing screen.
4. Select the **BookmarkData** table from the drop-down menu.

The results of the bookmark analysis are shown in figure 10.9. The data include the item name (**Name**), IRT parameters (a, b, c), the original level classification (**Level**), the source file of the item statistics (**Source**), and RP 50 values for each item. For example, the third item in figure 10.9 (**MATHC1025**) has values for the a and b parameters of 0.90 and −0.78, respectively, and was rated initially as Level 1, with an RP 50 value of −0.70 (indicating that it has a 50 percent probability of having a proficiency score of −0.70). (In this case, the item has only a single RP value column, but a bookmark data table may include several RP value columns.) The selected table of results should be exported and provided to the expert panel. The values in the **RP50** column inform the presentation order of items in the bookmark method of classifying items into proficiency levels and defining cut-points.

Using the bookmark procedure, panel members review each of the items in order of their RP value. When they encounter an item that
FIGURE 10.8

Performance Standards Interface, RP = 50 percent, CYCLE1 Data

FIGURE 10.9

Bookmark Data, RP = 50 percent, CYCLE1 Data
they feel represents a higher standard of performance, they add a “bookmark” at that location in the specially constructed test booklet. The RP values immediately prior to the bookmark locations represent the proposed thresholds for determining proficiency levels.

A combination of group discussion and statistical averaging is typically used to combine the thresholds produced by different reviewers to produce final thresholds, even if these are not statistically optimal. For development of qualitative descriptions of each proficiency level, the items are classified by the final thresholds.

Many types of information, including item specifications, curriculum references, and normative definitions of what students know and can do at each proficiency level, should be provided simultaneously to the panel responsible for standard setting. The panel must reconcile the different sources of information to determine the most useful cut-points and assignment of test items to levels. At their discretion, panel members may decide to use item classifications defined ahead of time by the item developers rather than reclassifying items based on the results of the bookmark procedure. In either case, the thresholds calculated by IATA represent the statistically optimal thresholds for the specified item classifications. To generate statistically optimal thresholds, simply adjust the RP to a desired percent, and IATA will automatically perform the calculation using the RP and item level classification and will save the results to the PLevels table in the set of analysis results. By default, unless you manually enter values into the table, IATA saves the thresholds corresponding to an RP of 67 percent. Note that IATA does not automatically update the level to which an item is assigned; if bookmark or other item classification procedure modifies the classification of an item, the new classification level should be entered in the item input data or directly into IATA.

You can manually change the threshold level by editing the thresholds directly in the table of results. After the values are changed, the graph is automatically updated. The most common adjustments performed include making the thresholds equally spaced or assigning thresholds that will, after applying scaling constants, occur at whole increments (for example, 5 or 10). Professional judgment should be exercised when reconciling the evidence from the statistical and
content analysis with the need to communicate results to lay audiences. Simplicity should be balanced with accurately communicating meaningful differences in student performance.

For the current example, assume that the panel, after considering the initial bookmark data and other sources of information, proposes to define the levels using the following set of cut-points: −0.85, −0.25, 0.35, and 0.95. Students with scores falling below −0.85 would be classified as falling below Level 1; students with scores in the −0.85 to −0.24 category would be classified at Level 2, and so on.

Click <<Back to return to the performance standards interface, where these cut-points can be recorded in the results data file and students can be assigned to appropriate levels. Perform the following steps:

1. Enter the recommended values produced by the panel of stakeholders into the appropriate rows in the column labeled Threshold. Press Enter after the final entry to ensure IATA updates the interface correctly.

2. Click the Add Levels button. IATA will assign students to the appropriate level based on their IRT scores.

Figure 10.10 illustrates the assignment of the thresholds for the performance levels. The levels are equally spaced, assigning a reasonable proportion of students to each level. Although no mathematical reason exists for the equal spacing, common practice in most national and international assessments is to use equally spaced thresholds because they appear more intuitive to lay audiences, who are the primary audience for proficiency-level summaries.

In the Scored data table, which can be viewed on the final screen of the analysis workflow, the record for each student also contains a variable named Level. This variable contains the level of performance standard to which each student is assigned on the basis of thresholds shown in figure 10.10. For instance, the first student listed under CYCLE1STDID was rated Level 4, had a percentile score of 85.13, and had an IRT score of 1.06. Note that, in this example, the means and standard deviations of the b parameters for each level in the summary table have not changed. Because these values are summaries of item
Having used thresholds to classify items into proficiency or performance levels, the expert panel should develop qualitative descriptions of the levels, specifying the knowledge and skills indicative of competency at each level. You can examine examples of level descriptions and competencies in other volumes in this series, specifically in volume 1, *Assessing National Achievement Levels in Education* for PISA, figure B3.3 (Ganey and Kellaghan 2008) and volume 5, *Using the Results of a National Assessment of Educational Achievement* for NAEP, table 2.6; Vietnam, table 2.7; and Mozambique, table 6.2 (Kellaghan, Greaney, and Murray 2009).

After setting the performance standard thresholds and applying them to the student scores, click **Next>>** to continue to the interface for viewing and saving results.
STEP 7: SAVING RESULTS

On the results viewing and saving interface, you can view the results produced by the current example walk-through. All tables should be saved both for project documentation and for facilitation of test linking with subsequent cycles of data. For reference, the item data results of this analysis walk-through are included in the `ItemDataAllTests.xls` file, in the worksheet named `ReferenceC1`.

NOTE

1. The `Items2` table is also available for analysis workflows that make use of linking.
Use the \textit{PILOT2} sample data set to carry out this exercise. The answer key for this test is in the Excel workbook \textit{ItemDataAllTests} in the sheet named \textit{PILOT2}.

Designs using rotated booklets allow testing of a large number of items by grouping items into different test booklets. Students are administered tests using different booklets, so that no individual student is administered all of the items. Other than the initial analysis specifications, the remainder of the workflow follows the same procedures that were described in previous walk-throughs.

\textbf{STEP 1: LOADING DATA}

The analysis begins with the \textit{Response data analysis} workflow. In the examinee response data interface, select the \textit{PILOT2} sample data file. These data represent a three-booklet design administered to 712 respondents. The data file contains a total of 107 variables, which include 99 items. Not all of the items are included in each booklet. This type of situation might arise if the national assessment steering committee requested that the test be quite long to cover an extensive curriculum. To reduce student fatigue, each student would be administered a
booklet containing a subset of the items. In the data shown in figure 11.1, the third column contains a variable labeled BOOKLETID. This variable contains values of 1, 2, or 3, denoting the booklet administered to the student. In addition to the alphabetic item responses and missing response code of 9 (not shown in the figure), there is a frequently occurring value of 7, which indicates that a specific item was not in the booklet assigned to a particular student. For example, the data in the figure indicate that the booklet for the student with PILOT2STDID = 2 did not contain MATHC2058. The code of 7 is treated as omitted and does not affect a student’s test results. Click Next>> to continue.

On the item data loading interface, load the PILOT2 item data from the ItemDataAllTests.xls file. The PILOT2 table contains 99 records and four variables. (Note that MATHC2047 has the following values: Key = C, Level = 1.00, and Content = number knowledge item.)

Confirm that the correct response data and item data have been loaded, and click Next>> to continue to the analysis specifications.
STEP 2: ANALYSIS SPECIFICATIONS

Under Select ID (optional), select PILOT2STDID. In the Specify missing treatment (optional) specifications in the previous examples, the box in the column headed Incorrect was checked only when data were missing. Because of the rotated booklet design, you must include in the Specify missing treatment (optional) table an omit code to indicate that some responses are not to be scored. Check the value of 7 in the Do Not Score column, as shown in figure 11.2. When the value of 7 is encountered in the response data for a particular student, Item and Test Analysis (IATA) will ignore the item, so it will not affect the student’s results. Similarly, participants with response codes of 7 for an item will not affect the estimation of statistics or parameters for that item.

After entering the analysis specifications, click Next>> to continue; the analysis will begin automatically. Because the computational time is affected more by the number of test items than by the number of students, the analysis will take longer to run than in previous walk-throughs.
STEP 3: ITEM ANALYSIS RESULTS

When IATA finishes running the analysis, the results will be displayed as in figure 11.3, in which the left table has been scrolled down to display MATHC2003, which has been assigned a warning symbol. The results indicate that this item has a weak relationship with proficiency; students tend to have a 0.61 probability of correctly responding, regardless of their level of proficiency. IATA suggests that this weakness may be the result of misleading requirements, which tend to be associated with poorly functioning distractors. When students do not understand the requirements of an item or when no response is unambiguously correct, they tend to guess. On the right side of the interface, the distractor analysis table displays the data underlying this summary, showing that option D is the only distractor eliciting the desired behavior. (The empty column with the header Omit is a reminder that the code of 7 was not allowed to influence the estimation.) To prevent MATHC2003 from potentially reducing the accuracy of the analysis results, remove the

![FIGURE 11.3](image-url)

**FIGURE 11.3**

Item Analysis Results, PILOT2 Data, MATHC2003
checkmark beside the item name in the Use column, and click Analyze to update the results.

After reviewing the results of the analysis, proceed with replicating the analyses that were demonstrated in previous chapters. The specific and interpretation of the remaining tasks in the workflow are largely the same as presented in previous walk-throughs.

If the test administration uses multiple test booklets containing different sets of items, IATA can perform the dimensional analysis only if the different booklets share a sufficient number of items. For example, with three blocks of test items (A, B, C) and three test forms (1, 2, 3), test form 1 should contain blocks A/B; test form 2, blocks B/C; and test form 3, blocks C/A. Because the blocks are fully rotated, no orphan items occur. Conversely, if test form 1 contains block A/B, test form 2 contains block A/B, and test form 3 contains block B/C, then block C is orphaned from block A, and therefore the correlations between items in block A and block C cannot be estimated. If you do have orphaned item blocks and you wish to perform a dimensional analysis, you must remove the orphaned items from the analysis or perform the dimensional analysis on each test form separately. However, once sufficient evidence exists to confirm that the collective set of items is assessing a single dimension, you may include all items in the analysis for estimates of item response theory item parameters and scores.

If the rotation is complex, the general principle for successful design is to ensure that no test items appear only in a single booklet. If an item does appear in only a single booklet, its item parameter estimates may be subject to increased sampling error because the estimates will be associated with only the specific items in that booklet and the students who were assigned that booklet. As the number of booklets in which an item appears increases, the accuracy of the parameter estimates for that item will also increase.

Repeating the analyses discussed in previous chapters with the PILOT2 data is left as an independent exercise. For reference, the item data results of this analysis walk-through are included in the ItemDataAllTests.xls file, in the worksheet named ReferenceP2.
Use the *PILOT2PartialCredit* sample data set to carry out this exercise. The answer key for this test is in the Excel workbook *ItemDataAllTests* in the sheet named *PILOT2PartialCredit*.

As in the national assessment scenario described in previous chapters of this volume, this walk-through follows the continued development of the test instrument with the introduction of analyses for short-answer test items that have been scored using partial credits. Other than the initial analysis specifications, for the last four items in *PILOT2* (discussed below), the remainder of the workflow follows the procedures that were described in previous walk-throughs. This walk-through focuses on the unique requirements of partial credit item analysis (see Anderson and Morgan 2008).

**STEP 1: LOADING DATA**

The analysis begins with the **Response data analysis** workflow. In the examinee response data interface, select the *PILOT2PartialCredit* sample data file. These data contain 111 variables and 712 respondents.
The data represent a three-booklet design with 103 test items and are the same as those used in the analysis of balanced rotated booklet design in chapter 11, with the addition of four partial credit items. For each of these items, students could have been awarded scores of 0, 1, 2, or 3, depending on the quality of their responses. Item and Test Analysis (IATA) will produce statistics for any score value greater than 0. The names of the additional items are MATHSA001, MATHSA002, MATHSA003, and MATHSA004. Click Next>> to continue.

On the interface for item data loading, load the PILOT2PartialCredit item data from the ItemDataAllTests.xls file, as shown in figure 12.1, in which the data panel has been scrolled to the bottom. The bottom four rows of the data file contain data for partial credit items. The scoring key for these items contains the information required to assign different numeric scores to the various codes in the response file, depending on the quality of a student’s response. In this example, the scoring key reflects the manual scoring of each item: the code of 1 is scored as 1, 2 as 2, and 3 as 3. A score specification for a value of 0 need not be provided.
If a response code is not in the answer key and is not treated as missing, it will be assigned a score of 0.

Confirm that the correct response data and item data have been loaded, and click Next>> to continue to the analysis specifications.

**STEP 2: ANALYSIS SPECIFICATIONS**

Because these data are mostly identical to those used in chapter 11, use the same analysis specifications. Under Select ID (optional), enter PILOT2STDID, and in the Specify missing treatment (optional) section, check the value of 7 in the Do Not Score column, as shown in figure 12.2. Note that all the item scores for the partial credit items also appear in the table of item values.

After entering the analysis specifications, click Next>> to continue; the analysis will begin automatically. Because the computational time is affected more by the number of test items than by the number of students in the data, the analysis will take longer to run than previous walk-throughs.

**FIGURE 12.2**

*Analysis Specifications, Rotated Booklets with Partial Credit Items, PILOT2 Data*
STEP 3: ITEM ANALYSIS RESULTS

When IATA finishes running the analysis, the results will indicate that MATHC2003 is problematic (figure 12.3). Delete the item from the analysis by removing the checkmark beside the item name in the Use column and clicking Analyze to update the results. IATA will ask if it should recalibrate partial credit items; click Yes to proceed.

In the left table in figure 12.4, you can see the rows that were created automatically by IATA for each of the item scores for each of the partial credit items. For rows that represent scores of partial credit items (where the Name column contains the @ symbol followed by an integer), the statistics are estimated as if each score were a single correct/incorrect item where the correct answer is any score value greater than or equal to the selected score. IATA will create an additional set of statistical results for each partial credit score that is provided in the scoring key for an item. For example, with a partial credit item having nonzero scores of 1 and 2, the item facility for the score of 1 (ItemName@1) would describe the proportion of students with
item scores greater than or equal to 1, and the item facility for the score of 2 (ItemName@2) would describe the proportion of students with a score of 2.

In the distractor analysis table in figure 12.4, note that MATHSA001@2 uses codes of both 2 and 3 as keyed responses; for MATHSA001@1, codes 1, 2, and 3 would be used as keyed responses. The item facilities are always larger for lower scores of an item because they include all the students who were assigned higher scores. For example, figure 12.4 highlights the results for MATHSA001 with the score of 2 selected. For this item, the score of 1 (MATHSA001@1) has a PVal (item facility) of 0.90, the score of 2 (MATHSA001@2) has a PVal of 0.77, and the score of 3 (MATHSA001@3) has a PVal of 0.45.

Although the statistics describe each score separately, the item response functions (IRFs) for partial credit items reflect the fact that a student can be assigned only a single score value. The IRF for a partial credit item is represented as a set of item category characteristic curves (ICCCs), one for each item score. When a row corresponding
to a specific score is selected, the graph illustrates the ICCC of the selected score in bold. At each level of proficiency, an ICCC expresses the probability that a respondent with a particular level of proficiency will be assigned a particular score, exclusive of all other scores. As shown in figure 12.4, as proficiency increases, the probability of each score value first increases, then decreases, as students become more likely to achieve higher score values.

Although no simple rules apply for analyzing an IRF for a partial credit item, a useful scheme for partial credit scoring tends to have the property that each score value will have the greatest probability of being selected over a certain range of proficiency. For example, the first score value for MATHSA001 has been assigned a caution symbol. The ICCC indicates that the probability of being assigned a score of 1 is not higher than any other score at any level of proficiency. The score value of 2 for MATHSA001, illustrated by the bell-shaped curve with a peak around −0.5, is the most likely score value for all students with below-average proficiency. These results indicate that the score value of 1 does not provide useful information because it is statistically indistinguishable from the score value of 2. With most partial credit items, the highest category is typically the most useful, because human markers tend to be able to clearly identify completely correct or completely incorrect responses but are less able to accurately distinguish between degrees of partial correctness.

One of the main tasks of analyzing partial credit items in pilot testing is to determine whether all of the score values are useful and how to improve the scoring process. For example, if a score category (such as 1) has a low probability of being assigned, two main possibilities exist: either no (or very few) respondents have produced responses that correspond to that score, or the scorers are not able to identify which responses should be assigned the score. In the first case, the responses associated with the score should be consolidated with an adjacent score category (such as merging scores of 1 and 2 into one score category of either 1 or 2). In the second case, the problem may be remediated with more intensive or standardized training of scorers. Thus, although the results of dichotomous item analyses are mainly relevant to item writers and test developers, the results of partial
credit analysis should also be shared with the teams responsible for scoring student test booklets.

After reviewing the results, proceed with replicating the analyses that were demonstrated in previous chapters. The specification and interpretation of remaining tasks in the workflow are largely the same as presented in previous walk-throughs. The only difference lies in the specification required for partial credit items in IATA’s item selection interface.

To run the selection properly with partial credit items, you must first manually include all the partial credit item scores by checking the item scores for an item and then counting each score value as a separate item when entering the total number of items. Thus, if you wish to select 10 items, and one of those items is a partial credit item with two score categories, you must specify a selection of 11 items and manually preselect the item score values for the desired partial credit item. Repeating the analyses discussed in previous chapters with the PILOT2 data is left as an independent exercise.

For reference, the item data results of this analysis walk-through are included in the ItemDataAllTests.xls file, in the worksheet named ReferenceP2PC.
Use the CYCLE2 sample data set to carry out this exercise. The answer key for this test is in the Excel workbook ItemDataAllTests.xls in the sheet named CYCLE2.

Common items are used to link the item response theory (IRT) and public reporting scales of the CYCLE2 data to the respective scales used in the CYCLE1 data. The linked results are used to scale scores and apply performance standards.

Governments and others are often interested in determining if students’ achievement levels have risen, fallen, or remained constant over time. Interest in changing standards is particularly important in times of curriculum change or substantial reform in the system (for example, a change in funding levels). Governments may also be interested in the effects on student achievement of a rapid increase in enrollment caused by the implementation of programs such as Education for All or the Fast Track Initiative, now known as the Global Partnership for Education. With a strong link, scores from a national assessment may be compared with an assessment that was conducted earlier, allowing one to conclude whether a concomitant change in student achievement has occurred (see Mislevy 1992).
Linking of test results is useful in other situations. In addition to comparing results across cycles of a national assessment, the following scenarios are common. A country with a number of educational jurisdictions may create a test for each jurisdiction that contains curriculum content that is specific to each jurisdiction. If the tests for the different jurisdictions share common items, the scores on the different tests can be used to compare performance across jurisdictions. Test linking can also be used to compare the results of a national and an international assessment if the national assessment includes items previously used and calibrated in an international survey. For example, if a country had previously participated in the Trends in International Mathematics and Science Study (TIMSS), including a number of TIMSS items in a subsequent national assessment may help detect if performance had changed since the earlier TIMSS administration. In this scenario, the linking procedure would use the national TIMSS item parameters in the reference item data file that were estimated using the specific country’s student response data. Alternately, a linking procedure can help identify how the country’s performance on the national assessment compares to the performance of other countries on TIMSS. In this case, the reference item data file should include the TIMSS item parameters that were estimated from the international TIMSS data.

The walk-through in this chapter describes the methods required to implement a follow-up cycle in a national assessment program. After launching the program, you will need to follow this workflow to ensure the interpretations that stakeholders make using the results of a new assessment cycle are consistent with, and comparable to, those made in the first cycle. The CYCLE2 data in this example represent the second cycle of a national assessment, following the first cycle, CYCLE1, which was analyzed in chapter 10. A link between these two assessments is possible because the CYCLE2 test contains several anchor items that were also on the CYCLE1 test. The link allows you to monitor changes in student performance over time. If the different assessment cycles are linked with common items, the permanent unique item names will help track which items are used in each of the assessment cycles for linking purposes.
The contents of the chapter focus on the interfaces and specifications that are distinct to this workflow. Review previous chapters for more detailed explanations of the common workflow interfaces.

From the main menu, click the first menu option, **Response data analysis with linking**, to enter the analysis workflow, as shown in figure 13.1. This workflow requires response data, item data (answer keys) for the response data being analyzed, and a reference item data file that is used to anchor the linked results.

**STEP 1: ANALYSIS SETUP**

The workflow begins with the same initial data-loading steps as the response data analysis.

1. In the first interface in the workflow, load the response data from the file named **CYCLE2.xls** in the Item and Test Analysis (IATA) sample data folder. These data include 2,484 records and 61 variables. The first case has the following values: SchoolID = 2; Sex = 2; SchoolSize = 21; and Rural = 0. Click Next >>.

2. In the second interface, load the corresponding item data. When opening the **ItemDataAllTests.xls** file, ensure that the **CYCLE2** table is selected. The table has 53 records and four variables, including three partial credit items. **MATHC2047** has the following values: Key = C, Level = 1.00, and Content = number knowledge. Click Next >>.

3. Use a new data-loading interface that has not been used in the previous walk-throughs. The interface requests a file containing reference item data. The reference item data contain IRT parameters.
a, b, and c (if an option), which have been estimated from a reference sample, such as a previous cycle of a national or international assessment.

The reference item data file must contain data for at least some items that are included in the current national assessment. For this example, use the results produced from the analysis of the CYCLE1 data file. These results are provided in the IATA sample data folder in the file named *ItemDataAllTests.xls* in the sheet named *ReferenceC1* (the results from the exercises in chapter 10 may have also been saved). When opening this file, ensure that the selected table is the table named *ReferenceC1*. This table, which has 50 records and 11 variables, contains the statistical output describing all the items from the first cycle of the national assessment. These reference data are illustrated in figure 13.2. In the current national assessment example, the CYCLE2 test includes 25 items that have item parameters in the CYCLE1 reference item data file. Keeping the names of items consistent across all data files is
important, because IATA matches items in the linking procedure using item names.

Note that this file also includes several data fields that were calculated during the analysis of CYCLE1 data in addition to the a, b, and c variables (for example, Level, Content, Discr, PVal, PBis, and Loading). These variables may be left on the data file but are not used in the linking analysis. Similarly, although the reference item data contain information for all 50 items on the first cycle test, only the information from the 25 items common to the second cycle test are used to estimate the linkage.

After loading all three data files, click Next>> to continue to the analysis specifications interface (IATA Page 4/12). The analysis specifications are similar to those for the CYCLE1 data. Enter or select the following details on the analysis specifications interface, and click Next>> to complete the analysis and view the results:

- The student identification variable is CYCLE2STDID.
- The weight variable is CYCLE2Weight.
- The code of 9 is treated as incorrect.

Proceeding to the item analysis interface (IATA Page 5/12) will automatically begin the analysis. No problematic items are present in the data. On reviewing each of the items, note that, although partial credit items have multiple scores, some of these may still be “easy,” where many respondents achieve the highest score categories (such as for MATHSA004), and others may be “hard,” where relatively few respondents achieve the higher score categories, such as for MATHSA005 (figure 13.3).

Continue through the workflow to review the test dimensionality results and perform any differential item functioning (DIF) analyses that may be of interest on the available demographic variables (location, sex, language) by following the same procedures described in previous chapters. Although many of the items have warning symbols for one or more DIF analyses, for the purposes of this example, all the items are assumed to be okay. After reviewing the DIF analysis, click Next>> to proceed to the linking interface.
STEP 2: COMMON ITEM LINKING

Click **Calculate** on the left side of the interface (figure 13.4). This will automatically select and list a table of all 25 items that are common to both the reference item data and the new data from the current national assessment. In the table of items, the first column, **Use**, specifies whether or not to include the item in the estimation of the linking constants (by default, all items that appear in the reference data and new data are included). Column **L** contains a summary diagnostic symbol for each item; the default caution symbol (the yellow diamond) is updated after IATA calculates the linking results. The most effective way to use this interface is first to calculate the results with all items, and then examine the diagnostic information to identify and remove any items with anomalous results. Repeat these two steps until the link is stable.

Click the **Calculate** button to estimate the linking constants and to evaluate the statistical quality of the link. When the calculation is finished, IATA displays a summary of the quality of the link in the graph on the right and updates the summary diagnostic symbols in the item
The graph displays three lines: a solid line, a dashed line, and a dotted line. The solid and dashed lines display test characteristic curves (TCCs). The TCC of a test summarizes the statistical behavior of the entire set of items, providing information similar to that provided by an item response function (IRF), but for many items simultaneously. Ideally, the linked and reference TCCs should be identical (if only one line appears to be visible, the two are likely perfectly overlapping), indicating that differences in the magnitude and variability between the link scale and reference scale are accounted for across the displayed range of proficiency. The dotted line displays the absolute difference between the two TCCs, expressed as a proportion of the total test score. The value of the difference typically varies across the range of proficiency, indicating that the link may not be stable for all ranges of scores. For score ranges with large differences, the linked results will not be on the same scale as the reference data and, hence, will not be comparable. However, if the average difference is small (for example, < 0.01), the error may be considered negligible.
In figure 13.4, the target curve (solid) represents the CYCLE1 test items, and the linked curve (dashed) represents the CYCLE2 test items after the application of the link. Seeing both curves in the figure is difficult because the target and linked TCCs are virtually identical, something that is also indicated by the error curve, which has a constant value of approximately zero across the displayed range of proficiency.¹

Beneath the graph, the estimated linking constants are displayed in the two text boxes. The location constant adjusts for differences in the magnitude of the original scales of the new data (the current national assessment) and the reference data (the earlier assessment), and the scale constant adjusts for differences in the variability between the scales. Generally speaking, the two constants can be interpreted together such that dividing any value expressed on the raw CYCLE2 IRT scale (for example, a student’s IRT score from the current analysis results) by the scale constant and adding the location constant will render the linked result directly comparable to scores on the CYCLE1 IRT scale. This comparability means that, after the scale linkage has been applied, any remaining differences between the CYCLE1 results and the transformed CYCLE2 results represent differences in test performance, rather than differences in the tests themselves.

In the item table on the left of figure 13.4, the diagnostic symbols are updated after calculation to indicate any potentially problematic links at the item level. A linked item is problematic if its linked IRF is very different from the reference IRF. Clicking on any item in the item list allows you to view the results of the linking function applied to each test item. As with the overall TCC comparisons, the linked IRF should be similar to the target IRF.² Even if the results for the overall test appear very good, the linking function may not work well for some items. However, because more sampling error occurs at the item level, differences between IRFs are typically problematic only if the error between the linked and reference IRFs is greater than 0.05³ (see example for MATHC1052 in figure 13.5).

A common example of a situation that would cause an individual test item to show idiosyncratic behavior in a linking analysis occurs
when a specific content area measured by a link item is used as the basis of instructional interventions between the two testing periods (such as increased emphasis on using a particular form of graph in mathematics or on an aspect of grammar in language). Because performance on that specific test item is likely to improve in an idiosyncratic way, the linking constants estimated from all items together will not account for the item-specific changes between the first and second administrations.

MATHC1052, flagged with a caution symbol, is a mild example of this phenomenon. The results for the item are shown in figure 13.5. Although the linking constants appear to have successfully adjusted to the difference in location of the item (that is, the difficulty of the item relative to the given sample), some gaps occur between the two lines, particularly at the higher proficiency level. The target IRF and the linked IRF are distinct from each other, and the dotted line at the bottom, which expresses the difference
between the two, ranges up to 0.08 but is generally less than 0.05. These differences are inconsequential, and in most practical situations, this amount of error is not problematic.

Should the differences between the target and linked IRFs be sufficiently large to be problematic (for example, if they are consistently greater than 0.05 across a wide range of proficiency), the offending item should be removed by unchecking the box next to the item name in the Use column and clicking Calculate. Although one or two items may be removed without introducing validity issues, if many items are removed from the estimation of linking functions, the validity of the link may become weak, because the anchor items may not adequately reflect the intended balance of content. Bear in mind that the validity of the link depends on both the statistical stability of items and the consistency in representation of content between the two assessments. If the statistical analysis of results suggests that some items should be removed from the link, the recommendation should be brought to the attention of the national assessment steering committee before making a decision. The fewer items that are common between the two assessments being linked, the weaker is the link. The results indicate a very stable link in the case of the current sample. As a consequence, the national assessment team can be confident that the test used in the current assessment (CYCLE2) is appropriate for monitoring changes in student achievement levels since the earlier national assessment.

Two controls beneath the linking constants—a drop-down menu and a button labeled Update—allow you to apply the linking constants directly to the results of the current analysis. View the results of the analysis on the final interface (IATA Page 12/12) of the analysis workflow. To apply the constants to the item parameters in the current results (IATA Page 8/12), perform the following steps:

1. Select Items1 in the drop-down menu at the bottom.
2. Click Update. IATA will add linked item parameters, whose scale now represents the scale established by the CYCLE1 item parameters, to the Items1 table in the analysis results. The linked item parameters are identified as a_link and b_link. IATA indicates when the results have been updated.
To update the estimated IRT scores in the current results, perform the following steps:

1. Select **Scored** from the drop-down menu.
2. Click **Update**. IATA will add a score variable, identified as **Linked-Score**, to the scored table in the analysis results. This score is expressed on the scale established by the **CYCLE1** item parameters.

After updating both the **Items1** and **Scored** results, click **Next>>** to continue.

**STEP 3: RESCALING LINKED RESULTS**

If you have applied the linking constants to the scored data table as described in step 2, IATA’s **Scale Review and Scale Setting** interface (IATA Page 9/12) will include the name **LinkedScore** in the drop-down menu at the upper left. Select **LinkedScore** to display the graphical and statistical summaries for the linked **CYCLE2** results, which are expressed on the scale established by the **CYCLE1** reference item data. The **LinkedScore** for this example has a mean of 0.10 and a standard deviation of 1.07. Note that the default results may display the summary statistics for the **PercentScore** results; to display the correct **LinkedScore** results, select it using the drop-down menu.

To convert the linked IRT score to a scale score that can be compared to the **NAMscore** variable that was produced during the analysis of **CYCLE1** data, perform the following steps:

1. Enter **NAMscore** in the text box beneath the **Add New Scale Score** label.
2. Enter 100 in the **Specify St. Deviation** box, the value originally set for the **CYCLE1** data.
3. Enter 500 for **Specify Mean**, the value original set for the **CYCLE1** data.
4. Select the **Rescale** option. This option ensures that the new scale score retains the link that was estimated in the previous interface.
5. Click **Calculate**. IATA will create the new scale score and display the distribution and descriptive statistics as shown in Figure 13.6. The mean of 510.36 indicates that the current year’s results show an improvement of 10.36 points on the previous assessment.

Because the current workflow is specific to linking, IATA automatically produces the new scale score using the linked IRT score. Because appropriate link items exist, IATA’s linking procedure has been able to produce **NAMscores** for two separate assessments that can be compared as they are on a common scale.

After adding the new scale score to the results, click **Next>>** to continue.

**STEP 4: ASSIGNING PERFORMANCE STANDARDS**

The majority of the tasks in the **Response data analysis with linking** workflow are specified in the same manner as they were in previous workflows. The item selection analyses on IATA Page 10/12 may be
performed as an independent exercise. However, performance standards are treated differently. After the first national assessment, standard setting should be performed only as a validation exercise. Reviewing the thresholds periodically is useful to determine if new performance standards need to be set (for example, if quality of education is improving), but establishing new thresholds for proficiency levels should typically coincide with major policy changes, such as curriculum reform.

When you enter the **Developing and assigning performance standards** interface in IATA (IATA Page 11/12), two sources of item parameters may be used to guide the standard-setting process: the items used in the current assessment (**Items1**) or the reference items used in the previous assessment (**Items2**). These item parameter sources are available from the drop-down menu above the table on the left. Summary statistics describing item parameters and thresholds are displayed in the table at the bottom right according to the item parameter source selected. The summary statistics shown when entering this interface describe the estimates from the current (**CYCLE2**) data in the **Items1** table. The best practice is to use the **Items1** source, which is selected by default, because the **Items2** table contains all items used to produce the current test scores.

The standard-setting exercise should be performed only if either (a) a standard-setting procedure was not already performed on a previous assessment cycle or (b) a reason exists to suspect that performance standards have been changing, in the sense that normative expectations of the types of skills or quality of learning may have changed over time. Note that student performance differs from performance standards. For example, student performance may be increasing if students have better test performance, but performance standards may be decreasing if stakeholders have lower expectations of what constitutes “acceptable” performance. Replicating the standard-setting process using the different sets of item parameters can help identify whether performance standards are stable across assessments. If the standards are stable, replicating the standard-setting procedure using different sets of items should produce similar sets of thresholds. Keep in mind when replicating
any standard-setting procedure that the item parameters in the \textit{Items1} table refer to the distribution of proficiency in the current assessment before applying any linking, and the item parameters in the \textit{Items2} table refer to the distribution of proficiency in the previous assessment. To compare the thresholds produced by analyzing different item parameter sources, apply the linking constants (for example, shown in figure 13.5).

Because performance has increased relative to the \textit{CYCLE1} data, the means of the item $b$ parameters will appear lower than for the \textit{CYCLE1} data for items classified at the same level. However, you should not estimate new thresholds in the current workflow, because performance standards were initially set and assigned in the \textit{CYCLE1} data analysis. The following thresholds (that were established in the first national assessment) are used for this national assessment:

- Level 4: 0.95
- Level 3: 0.35
- Level 2: −0.25
- Level 1: −0.85

Apply performance standards to the \textit{CYCLE2} data by manually entering the thresholds in the table in the \textbf{Threshold} column as shown in figure 13.7 and clicking the \textbf{Add Levels} button. To update these thresholds, first adjust the response probability using the slider in the middle of the interface; this will cause IATA to generate a table with default values, which can be replaced with the values assigned from the \textit{CYCLE1} walk-through. The \textbf{Level} variable will then be added to the \textbf{Scored} student data table. Although the specifications of the performance standards do not change on the basis of the workflow, this interface, like the scaling interface, will recognize that a linking workflow is being used and will assign students to levels based on the linked IRT score rather than the raw IRT score.

The item data results of this analysis walk-through are included in the \textit{ItemDataAllTests.xls} file in the worksheet named \textit{ReferenceC2}.
1. For more interesting results with greater error, replicate this analysis with PILOT1 and PILOT2 results as an independent exercise; keep in mind that to achieve the goal of minimizing errors in linkages, linked tests should have a sufficient number of anchor items and sample sizes to produce accurate statistics.

2. If you have response data from both assessments, you may perform a more sensitive analysis by analyzing DIF in the **Response data analysis** workflow on the combined response data files using the source data identifier as the DIF variable. The IRT scores that are produced will be automatically linked, but they will not be interpretable on the scale of either test unless item parameters were anchored.

3. One exception occurs only with highly discriminating items. Large differences (for example, where the error line is greater than 0.05) between the IRFs typically indicate problems. If these differences occur only over a small range of proficiency (for example, across a range of less than 0.4 proficiency point—two tick marks on the graph’s default x-axis), they will not adversely affect the quality of the link.
This chapter describes the use of four special applications of Item and Test Analysis (IATA) workflows. Each of the interfaces has been described in greater depth in earlier chapters in this volume. The first three parts of this chapter—“Linking item data,” “Selecting optimal test items,” and “Developing and assigning performance standards”—present an overview of the three workflows that use IATA item data files as their main input data. These workflows are most appropriate when student response data have already been analyzed and test item parameters have been estimated. They allow one to perform some analyses using the item parameters without requiring reanalysis of the original student response data. The ability to perform analyses using item data is useful in situations in which student response data may be inaccessible (such as for security reasons or where data transfer capacity is limited), but item parameters are readily available in sources such as technical reports or small data tables. In the final section, “Response data analysis with anchored item parameters,” anchor item parameters from an earlier national assessment are used to analyze student response data from a subsequent assessment that contains many items common to both assessments.
Because the main IATA interfaces used to access these functions have already been described in previous chapters, this chapter focuses on steps that are specific to each of the following topics, rather than providing complete walk-throughs.

**LINKING ITEM DATA**

Chapter 13 of this volume showed how linking parameters were calculated in the same workflow as the analysis of item response data. However, in practice, these two activities may occur at different times. A ministry of education, for example, might conduct an initial national assessment (*CYCLE1*) in year one and a follow-up national assessment (*CYCLE2*) in year five and then carry out or commission a linking exercise using item parameter estimates from both assessments in year six. Analysts can use the linking constants calculated by IATA and update the *CYCLE2* data by producing new linked scores and parameters. These new scores can show the extent of change in student achievement levels between year one and year five.

To link item data using this approach, you must have common anchor items from two separate analyses of national assessment data. For this example, the results from the *CYCLE1* and *CYCLE2* national assessments will be linked. They have 25 items in common. The process is similar to the **Response data analysis with linking** workflow described in chapter 13.

In chapter 10, the data from the first national assessment (*CYCLE1*) were analyzed. Before you undertake linking the two assessments, you must analyze the *CYCLE2* data following the procedures outlined in chapter 10. These are summarized as follows:

1. Click on **Response data analysis**.
2. Load the *CYCLE2* student data (2,484 records).
3. Load the *CYCLE2* item data (53 records).
4. Select 9 to represent missing data scored as incorrect, and use the drop-down arrows to highlight the values for sample and weight for *CYCLE2* (IATA Page 3/10). Recall from chapter 13 that the
additional numeric values (0, 1, 2, 3) present in the values column represent scores for partial credit items.

5. On the item analysis interface (IATA page 4/10), do not delete any items.

6. Skip Dimensionality and Differential Item Functioning analyses as well as the Scale Review and Scale Setting, Item Selection, and Standard Setting functions (IATA Page 5/10–Page 9/10) for the purposes of this exercise.

7. On the final interface (IATA Page 10/10), to confirm correct completion of the analysis, select Scored from the drop-down menu and scroll across to the end. The first student listed (CYCLE2STDID) received an item response theory (IRT) score of 2.58, a percentile score of 99.37, and a TrueScore of 93.26.

8. Save the complete set of results as CYCLE2_UNLINKED.xls.

Having completed the analysis of CYCLE2 data, select Linking item data from the main menu. Perform the following steps:

1. Load a student data file from the current assessment results (in this example, CYCLE2) that will be linked to the previous assessment results (in this example, CYCLE1). The data file must contain the variable named IRTscore. However, if linking results using results that have been modified, ensure that the variable containing the unscaled IRT score is given the name IRTscore. For this example, use the data file named CYCLE2_UNLINKED.xls and load the table named Scored. (The first-listed student, CYCLE2STDID = 1, has an IRTscore of 2.58 and a Percentile score of 99.37 at the end of record.) Click Next>>.

2. On IATA Page 2/5, load the item parameters estimated from the current or most recent assessment results. The data table must contain item names as well as the IRT parameters. In this example, to link the CYCLE2 results to the CYCLE1 scale, load the Items1 table from the CYCLE2_UNLINKED.xls file. This table contains the item parameters from the CYCLE2 national assessment. These data are also available in the ReferenceC2 table.
from the ItemDataAllTests.xls IATA sample data file. The first listed item, MATHC1008, has the following values: Key = A, Level = 3, a = 0.58, b = 0.20. Click Next>>.

3. On IATA Page 3/5, load the item parameters from the earlier CYCLE1 assessment results. The file must contain item names as well as the IRT parameters. In this example, to link the CYCLE2 results to the CYCLE1 scale, load the ReferenceC1 table from the ItemDataAllTests.xls file. The first listed item, MATHC1005, has the following values: Key = B, Level = 3, a = 0.63, b = 0.21. Click Next>>.

4. Click Calculate to estimate IRT linking constants (IATA Page 4/5). The two linking constants, Location and Scale, describe the linear transformation that converts the CYCLE2 IRT scale to the CYCLE1 IRT scale. Review the quality of the link, in particular, any items with idiosyncratic linked item response functions (IRFs). In this example, most of the items display good links, indicated by green circles. Review problematic items; these have caution or warning indicators (yellow diamonds or red triangles). Examine the graphs for the individual items, specifically the dashed blue error lines. Remove any item listed for a linking estimation (on the left) that is consistently greater than 0.05 (on the right axis) within the range of −2.0 to +2.0 (on the proficiency axis). If more than one problematic item is identified, remove a single item at a time. Recalculate the linking constants after each removal by clicking Update. Removing a small number of problematic items may be sufficient to improve the estimation of links for the remaining items. Conversely, removing many items may improve statistical estimation but may well reduce the overall validity of the link. Apply the linking constants to the item parameters and test scores. For this example, the location and scale constants estimated using the full set of common items are 0.106 and 0.988, respectively. These scale constants can be used to convert an IRT score from the CYCLE2 assessment to the scale of the IRT scores from the CYCLE1 assessment by multiplying the CYCLE2 IRT score by 0.988 and adding 0.106. These linking constants indicate that the test
scores in \textit{CYCLE2} are slightly less varied and also slightly higher, on average, than the \textit{CYCLE1} scores. Click Next>>.

5. Note that the linked a and b parameters for MATHC1008 are 0.58 and 0.20, respectively. The linked parameters describe the statistical behavior of each \textit{CYCLE2} item on the \textit{CYCLE1} scale. This information may be useful if a national assessment team wishes to perform standard setting or develop a new test using the combined set of items. The following sections of this chapter discuss how to perform these two functions using only item data. Save the analysis results (IATA Page 5/5).

### SELECTING OPTIMAL TEST ITEMS

Up to this point, test items have been selected while analyzing student response data. However, a ministry of education might reasonably request that a set of items from a completed assessment be identified for future assessments. This request may come long after the student response data have been analyzed and the national assessment cycle activities have been completed. Test developers and analysts can identify and select anchor items from the previous assessment without having to resort to locating and reanalyzing the entire set of student data from this assessment. These anchor items will form the core or at least a major component of the new assessment, to which new items will be added. The anchor items will be used to link the new assessment results to the existing assessment results. To provide the basis for an accurate test and a stable link between the two tests, anchor items should provide the greatest accuracy possible across the expected range of student proficiency. In this instance, the item selection task requires only existing item parameters from the earlier assessment.

The following exercise demonstrates how to select optimal items for linking two assessments using item parameter results saved from the data analysis of the earlier assessment.

1. Select the \textbf{Selecting optimal test items} workflow from the main menu.
2. Load the item data file, the *ReferenceC1* table from the *ItemDataAllTests.xls* file in the IATA folder on your desktop (IATA Page 1/3). These data must include item names and IRT parameters. The data should also include information on item Level and Content. (The first item listed, the lowest numbered item, should be MATHC1005, which has the following values: Key = B, Level = 3, a = 0.63, and b = 0.21.) Click Next>>.

3. To provide the most useful results for item selection, the specified number of items should equal the total number of items in the item data file. In this instance, because the *ReferenceC1* table item data has 50 items, 50 items should be specified for selection. IATA will produce a table in which all of the available items are ranked on the basis of their suitability for measuring students within the specified lower and upper bounds. The lower and upper bounds describe the approximate percentile ranks of the original sample from which the items were calibrated (0 represents the lowest-performing student, and 100 represents the highest-performing student). In general, keep the default values of 2 and 98. However, if the distribution of proficiency in the population of students in the new national assessment is anticipated to be significantly different from the earlier assessment, the lower and upper bounds may be adjusted to minimize the inclusion of unnecessarily easy or difficult items. For example, if you expect the new population of students to have much higher proficiency than the original population, adjust the lower threshold upward to a value, \( x \), greater than 2 to reflect that the lowest-performing student in the new population might have a score equivalent to the performance of a student at the \( x \)th percentile rank from the original population. Doing so will reduce the chances of selecting inappropriately easy items for the new population. Enter a title such as AncItems50 (for Anchor Items) in the Name of item selection box and 50 for the total number of items. Click Select Items. IATA will produce a content \( X \) level table for the 50 items (IATA Page 2/3). Click Next>>. Note (IATA Page 3/3, shown in figure 14.1) that MATHC1029 has the following values: \( a = 1.33, b = 0.17, \) Level = 4, and Key = D.
4. Click Save Data. IATA assigns item selection tables the prefix of CustomTest followed by the unique name specified for a particular item selection exercise. In this instance, the 50-item data file will be saved with the other IATA data as CustomTestAncItems50.

You may experiment with the number of items selected using this interface (such as 30 or 40). Assign a different name for each item selection (in the Name of item selection box) if you wish to save the results.

DEVELOPING AND ASSIGNING PERFORMANCE STANDARDS

Setting performance standards (such as below basic; basic; proficient; advanced; or levels 1, 2, and 3) is an important step in making the results of the national assessment accessible to a variety of stakeholder audiences. Chapter 10 of this volume described the task of setting performance standards as a relatively straightforward exercise. In practice, however, it typically requires iterative work involving review of both item content and statistical results. This should include input...
from multiple sources (such as curriculum personnel), many of whom may have had little experience with data analysis or statistics. The response data analysis should be fully completed before starting the standard-setting process. The panel should use the final item parameters in carrying out this exercise.

The Developing and assigning performance standards workflow in IATA allows use of results from previous analyses to facilitate the standard-setting process. In this section, performance standards are set for CYCLE1 data, using item parameters from the CYCLE1 assessment. To use this workflow, you must have completed an analysis of item response data and have saved the Items1 and Scored results tables. If the data have not been saved, repeat the analysis described in chapter 10. Preferably, from the perspective of developing performance standards, both item parameters and IRT scores should be loaded into IATA. However, the scores are used only for reference in the estimation of thresholds, and the analysis can be conducted using only item parameters without loading any IRT scores. Examine the distribution of scores by threshold level to determine the proportions of students classified within each proficiency level. Standard setting, as noted earlier (see description of bookmarking in chapter 10), is an iterative procedure and involves review of test items by curriculum experts and experienced teachers in light of the available statistical evidence. As was the case in the discussion in chapter 10, the purpose of standard setting in this chapter is not to justify existing cut-points or force the data into previously established proficiency levels, but to determine the most useful proficiency levels and associated cut-points based on the available items. Several rounds of review and discussion may be required to set threshold levels. After the thresholds have been finalized, loading IRT scores into IATA will facilitate adding proficiency levels directly to the student assessment results.

To complete this workflow, perform the following steps. Note that steps 1–4 may be repeated several times before finalizing the set of performance level thresholds.

1. Select the Developing and assigning performance standards workflow from the main menu. The first page (IATA Page 1/4) requests that you load scored student response data. This step is
optional, because the standard-setting exercise may be conducted using only item parameter data. If scored student data are available, they can inform the usefulness of a proposed set of cutpoints by allowing you to estimate the proportion of students that falls within each proposed proficiency level. However, scored student data are useful only after the main iterations of the standard-setting procedure are finalized.

2. Because the cut-point definitions have not been finalized, click Next>> to skip to IATA Page 2/4.

3. Load the item data file containing the IRT parameters and the preassigned level for each item. In this example, results are used in the Items1 item data table produced automatically by IATA. (The data are identical to data in ReferenceC1 in ItemDataAllTests.xls.) Note that each item has a preassigned performance level. Click Next>>.

4. Perform the bookmark standard-setting procedures described in chapter 10, which include setting the response probability (RP) value, saving bookmark data, and using these data to facilitate review of the test items by teachers, curriculum specialists, and other educational stakeholders. Note that in a real-world scenario, you may experiment with a variety of RP values, typically in the range of 0.50 to 0.80. During the performance-setting process, item-level assignments may (and likely will) be modified during discussions by the national assessment team. Steps 1–4 may need to be repeated several times, using the bookmark procedure, until the participants in the standard-setting exercise are satisfied that the agreed thresholds will facilitate meaningful interpretation of the assessment results.

5. Following agreement among the panel stakeholders responsible for establishing performance standards, repeat steps 1–4, making sure to load student IRT score data in step 1. For this example, load the CYCLE1 student data with IRT scores that should have been saved at the end of chapter 10 (IATA Page 1/4). When IATA saves student results, the IRT score is contained in a variable named IRTscore, which is found in the data table named SCORED. This variable contains scores that IATA has
estimated directly from the item parameters, without applying any rescaling or linking. After loading the data, scroll across student responses to individual items (IATA Page 1/4) to see individual IRT scores. (The second student listed, CYCLE1STDID, had an IRT score of 1.764.) Enter the final cut-points into the Threshold column on the bottom right of the screen (IATA Page 3/4). Verify, through comparison of the vertical cut-points with the area of the score distribution, that each proficiency level includes a reportable proportion of students. In cases where very few students are in the highest or lowest proficiency levels, combining adjacent proficiency levels is more useful for reporting purposes. Optionally, the thresholds to the scored data may be applied through the performance standards interface by clicking the Add Levels button. Click Next>>.

6. Save the results (IATA Page 4/4) using a distinctive file name such as NAMPPerfStand. In general, save all tables containing modified data. These include the PLevels table, which has been updated with new thresholds; the Items1 table, which may have new level assignments for items; and the Scored table, with student performance levels.

**RESPONSE DATA ANALYSIS WITH ANCHORED ITEM PARAMETERS**

In previous chapters, all IRT item parameters were assumed to be unknown and had to be estimated from student response data. Test item parameters were calculated for each national assessment (such as CYCLE1 and CYCLE2) and were used to calculate linking constants.

IATA also provides a facility for importing fixed item parameters from an earlier assessment, which can be used to link the results of the assessment with a subsequent assessment. Initially, it estimates parameters for the current assessment and links these to unadjusted parameters from the earlier assessment. The items used in the linking process are called anchored item parameters.
Anchored item parameters are the $a$, $b$, and (optionally) $c$ parameters that have been assigned values in an item data file for some test items before the analysis of a particular response data file, much like the anchor items used in formal linking. When response data are analyzed using anchored item parameters, the parameters of the new or nonanchored items are calculated, while the anchored item parameters remain fixed at their prespecified values. Newly estimated results, such as the IRT parameters for nonanchored items and student IRT scores, are expressed on the scale that is defined by the anchored item parameters. This method is favored over the formal linking process when item parameter estimates produced using the current response data may be inferior in quality to those produced by the already existing estimates. This situation might occur if the current sample was much smaller or less representative than the sample used to estimate the existing item parameters. This method is also appropriate when the majority (more than 70 percent) of the items in the current assessment already have existing parameter estimates. The only difference between the use of anchored items and the walk-throughs described in previous chapters is that some items will already have item parameters in the input item data file.

Consider a scenario where the national assessment steering committee has decided to use a test from a previous national assessment cycle with only minor modifications to the set of test items. In this case, performing the complete linking procedure that was described in chapter 13 is not necessary. For the relatively few new items used in the current assessment, IATA will automatically calibrate their IRT parameters and place them on the same scale as the anchored item parameters. The final student IRT scores will be based on both the anchored item parameters and the newly calibrated items and will be expressed on the same scale as the anchored item parameters.

Use the CYCLE3 sample data set to carry out this exercise. The item data for this test are in the Excel workbook, ItemDataAllTests. These data represent the third cycle of the national assessment program that has been analyzed in previous chapters. For this third cycle, the national assessment steering committee decided to use the items from the CYCLE2 test after making minor modifications to the content of some items and replacing only eight of the multiple-choice
items and all of the short-answer items. Rather than reestimate new parameters and linking constants, the committee decided to use the item parameters from CYCLE2 to anchor the item parameter estimates for the new items.

To perform analysis with anchored items, complete the following steps.

1. Select the **Response data analysis** workflow from the main menu.
2. Load the **CYCLE3.xls** student response data (containing 2,539 records and 61 variables) from the IATA sample data folder (IATA Page 1/10). Check to see that the first student listed in the data file had the following values: **SCHOOLID** = 30, **Sex** = 2, **SchoolSize** = 21, **MATHC2047** = C. Click Next>>.
3. Load the **ItemDataAllTests.xls** file, and select the **CYCLE3** table as the item data (IATA Page 2/10). The table contains 53 records and seven variables. Note that **MATHC2047**, a number knowledge item, has values of 0.80 and -0.75 for the **a** and **b** parameters, respectively. Unlike the item data files used in previous analyses, values for the **a** and **b** parameters are present for some, but not all, items, as shown in figure 14.2. The item parameters that have assigned values are the anchored item parameters. These values were produced during the analysis of CYCLE2 data and were linked to the original scale established for the CYCLE1 assessment. Several items with specified answer keys that do not have item parameters (such as **MATHC2069**) were assigned new item parameters estimated from the response data. Because the anchored parameters were already linked to the CYCLE1 scale in the previous analysis, the newly estimated parameters in the current analysis of CYCLE3 data were also linked to the CYCLE1 scale. Click Next>> to proceed to the analysis specifications.
4. Set the identification variable as **CYCLE3STDID**, the weight variable as **CYCLE3Weight**, and check the value 9 for incorrect (IATA Page 3/10). Note that the additional numeric values (0, 1, 2, 3) in the values column represent scores for partial credit items. Click Next>> to begin the analysis.
5. The results produced are shown in Figure 14.3. Note that all items now have item parameters, but the anchor items maintain their original values (see the a and b values for MATHC2047 in Figures 14.2 and 14.3). Unlike the test-level linking, you can now see how the anchored item parameters fit the current response data by comparing the theoretical and empirical IRFs for each item. For example, item MATHC2047 used anchored item parameters; the IRF labeled Theoretical in Figure 14.3 is derived from the CYCLE2 data, whereas the IRF labeled Empirical is derived from the CYCLE3 data. In general, the fit of the anchored items tends to be poorer than that of the new items, whose parameters are estimated from current data. If the fit between the individual theoretical and empirical IRFs is poor (that is, the magnitude of the vertical gap between the theoretical and empirical IRFs is consistently greater than 0.05), and the sample of new response data is large, the items should not be used as anchored parameters. If, however, the sample is small (such as fewer than 500), then lack of fit between the theoretical and empirical IRFs may simply be caused by random error and can be ignored.
A number of items (such as MATHC1046 and MATHC2034) have diamond (yellow) flags. These items tend to be weakly related to proficiency but were retained for the purpose of this exercise. Note that the score of 1 for the partial credit item MATHA006 (MATHA006@1) has a triangular (red) flag. Almost 99 percent of students achieved a score of 1 or more on this item. However, retaining the score of 1 does not have much of an effect on the quality of the results. The flag simply indicates that the score is not clearly distinguishable from 0, whereas scores of 2 and 3 on this item (MATHA006@2 and MATHA006@3, respectively) are clearly distinct from the other score values. In general, the best practice would be to revise the scoring scheme for this item to treat scores of 2 and 3 as valid scores (the corresponding key entry in IATA would be “2:1;3:2”). However, in this example, retaining the existing scoring scheme has no disadvantage, because IATA has adjusted the estimation of scores to reflect the low proficiency associated with the partial credit score of 1.
6. IATA Page 5/10 and IATA Page 6/10, which carry out test dimensionality and differential item functioning analyses, may be skipped; both tasks are identical to similar ones carried out in previous walk-throughs. Click Next>>.

7. Because the results are automatically linked to the CYCLE1 scale, the mean and standard deviation of the IRTscore (IATA Page 7/10) of the CYCLE3 data may deviate significantly from 0 and 1 in the current sample (in this case, mean = 0.02, standard deviation = 1.04). One important consideration for scaling results that use anchored item parameters is that because the IRT scores are anchored to the linked parameters from CYCLE2, you must use the Rescale option to produce scale scores, specifying the mean and standard deviation values used when establishing the NAMscore scale in CYCLE1. Insert NAMscore and the original values for the standard deviation (100) and the mean (500). Click Calculate. The rescaled mean and standard deviation values for NAMScore are 501.71 and 103.96, respectively. Click Next>>.

8. Skip item selection (IATA Page 8/10). Because the IRT scores are expressed on the scale established with the CYCLE1 data, apply the performance standard thresholds from CYCLE1 (Level 4 = 0.95, Level 3 = 0.35, Level 2 = −0.25, Level 1 = −0.85) to the CYCLE3 data set (IATA Page 9/10). The RP level need not be set because the cut-points have already been established. This will help ensure that the CYCLE3 students are being rated against CYCLE1 thresholds. Press Enter after entering the threshold values. Click Add Levels to assign students to performance standards or levels. Click Next>>.

9. Click Save Data, and save all results tables for the CYCLE3 assessment. Note on the SCORED data files that the first student’s result has IRT and NAM scores of 1.41 and 641.10, respectively. For reference, the item data results of this analysis (Items1) are included in the ItemDataAllTests.xls file in the worksheet named ReferenceC3.
Finally, note that anchored item parameters are particularly useful in situations where the sample size of the new national assessment is small, the tests have substantial overlap, or response data are available from both tests. In the last scenario, the response data should include all respondents from both cycles to facilitate the analysis of differential item functioning between the two tests; the item data would include answer keys for all items, and only parameters for items used in the previous cycle would be assigned values.

During the course of carrying out various analyses of the type just covered, the assessment team may make modifications, such as removing items or adjusting proficiency levels or curriculum content categories. Where the number or the extent of modifications is substantial, item linking should not be used.

Given the likelihood that analysts and other key personnel will change between national assessments, saving all data tables is important, as is having a clear explanation of key decisions and key changes made to item data files. To aid future national assessments, the analyst should write a short description of any change made to an item data file during the current analysis in a ReadMe text file (see Freeman and O’Malley 2012).

NOTE

1. Loading these data is optional, because estimating the statistical link requires only item parameters. If not loading the IRT scores at this stage, apply the results of the link using a different software package (such as SPSS [Statistical Package for Social Sciences] or Excel).
The walk-through examples completed in part II of this volume provide detailed descriptions of the most common statistical procedures required to create, implement, and maintain a national assessment system. After completing these walk-throughs, you should have learned how to perform the following tasks:

- Loading student response data
- Loading item parameter data
- Specifying item response keys, item performance levels, and item content classifications
- Specifying missing data treatment
- Reviewing classical item statistics
- Interpreting item response functions and item error analyses
- Interpreting test dimensionality results
- Interpreting summary test statistics
- Producing scale scores for reporting
- Generating and interpreting differential item functioning analysis
- Estimating and defining thresholds for proficiency levels
- Selecting subsets of items for specific measurement goals
- Saving results to a computer
These tasks represent practically all of the normal requirements for test analysis in the implementation of a national assessment. However, simply replicating the examples as they appear in these chapters is not the same as mastering the ability to perform these functions with your own national assessment data. The next steps in your learning process should be to review each of the examples several times while following the exact instructions in the chapters.

Once you have mastered the Item and Test Analysis (IATA) interface, you should be ready to experiment with some of the choices that structure the analyses in IATA. Once again you should go through each of the walk-throughs, but instead of following each of the instructions precisely, experiment with the options that are available. For example, what happens to analysis results when the number of test items is very small? What happens to equating results when the set of common linking items includes only very easy or very hard items? Many other technical approaches exist to analyzing test data that are beyond the scope of this volume. However, if you experiment with the IATA sample data and compare your results with those obtained earlier during the various walk-through exercises, you should get a much broader understanding of the choices that are appropriate for different situations.

When analyzing your national assessment data, you should first identify the workflow in the IATA menu that is most appropriate to your situation. More than likely, this will be very similar to one of the walk-through examples presented in one of the previous chapters. Some situations may require combinations of workflows, where the results from one workflow or analysis are used as the input data for another.

Ultimately, as your expertise grows, you will appreciate that unique and perfect answers or solutions to the problems of national assessment rarely exist. At best, the statistical methods that are used in modern assessment serve to minimize the influence of errors that inevitably result from the real-world challenges of educational measurement. How national teams choose and implement these statistical methods depends on their goals. What are the needs of stakeholders? What are the consequences of the decisions based on results? IATA is simply a tool (albeit a useful one) for reducing the burden of these statistical methods and clarifying the trade-offs of the analytical choices available in a national assessment.
Two aspects of the classical test theory (CTT) approach to measuring proficiency, namely item facility (or difficulty) and item discrimination, were described in chapter 9. In this annex, the focus is on an alternative approach, item response theory (IRT), which unifies the concepts of item facility and discrimination. IRT has also been described as latent trait theory. It is the most widely used approach in large-scale assessments.

A good starting point in understanding IRT is to contrast what constitutes a good test item from CTT and IRT perspectives. The classical item statistics of facility and discrimination focus on estimating and comparing the probability of correct response for different students. In contrast, IRT characterizes students by the type of item response they are likely to produce and tries to describe the distributions of proficiency for students who respond in different ways. Although a good test item from a CTT perspective has large differences in probability of correct response for students of different proficiency levels, a good test item from an IRT perspective is one where the distribution of proficiency for students who answered correctly differs from the distribution of proficiency for students who answered incorrectly. Whereas CTT focuses on the probability
of correct response, the focus of IRT is on the estimation of the distributions of proficiency. Although these two perspectives are generally in agreement, the IRT perspective describes items in a much richer and more useful way.

Item and Test Analysis (IATA) software calculates results using a variety of statistical methods. Most calculations use closed-form equations, which means that the calculation uses the student response data in an ordered progression of steps to produce the desired statistic, such as the arithmetic mean. In closed-form equations, even if calculations have several steps, the values at each step are based on the original data and the results of the previous steps. Detailed descriptions of closed-form methods for calculating classical item statistics and other basic summary statistics are available in most statistics textbooks (for example, Crocker and Algina 2006).

Some calculations require IATA to estimate a statistic, \( x \), that is based on another statistic, \( y \), but where the value of \( y \) is also based on the value of \( x \). In these cases, because \( x \) and \( y \) cannot be estimated together, IATA must use an iterative algorithm. Generally, an iterative algorithm first assumes some reasonable starting values for \( y \) and uses these to estimate the values for \( x \). Then, the algorithm uses the results for \( x \) to calculate new values for \( y \). The new values of \( y \) are then used to update the values of \( x \), and the process is repeated until new iterations do not substantially change the values of the estimates. This approach to calculation is used in the analysis of item and test dimensionality as well as for estimating IRT item parameters (see Lord and Novick 1968). Both of these calculations require estimating properties of items, such as loadings and IRT parameters.

The analysis of dimensionality uses a common iterative algorithm known as singular value decomposition, or SVD (see http://en.wikipedia.org/wiki/Singular_value_decomposition), but estimating IRT parameters requires specialized iterative algorithms (see Baker and Kim 2004). These algorithms must first estimate the probability that each student will correctly respond to each item and then find the item parameters that best reproduce those probabilities. The new parameters are then used to update the estimated probabilities, which in turn are used to update the estimates of item parameters, and so on, until the estimates
at each stage do not noticeably improve on previous estimates. IATA’s parameter estimation uses a variation of this general approach that is both computationally faster and statistically more robust than other algorithms. It allows IRT methods to be used with a wider range of sample data than is typically allowed by other software.

With IATA’s algorithm, the first stage (estimating probabilities) requires the calculation of two proficiency distributions for each item: the distribution for correct respondents and the distribution for incorrect respondents. These distributions are assumed to be normal in shape, and, for each item, the two distributions share the same variance but differ in their means and relative magnitudes. For example, if more students respond correctly than incorrectly, the magnitude of the distribution for correct respondents will be greater than that for incorrect respondents. Summing these two distributions at each proficiency level describes the distribution of proficiency for all students, and the ratio of the distribution of correct respondents to the summed distribution produces the estimates of probability of correct response at each level of proficiency. This approach has advantages over other methods for two reasons: (a) it describes probability of response at all levels of proficiency, rather than at an arbitrary sample of proficiency levels, and (b) the distributions for both correct and incorrect respondents can be described using the mean for correct respondents and the proportion of correct responses for an item because the total sample mean is constrained to equal zero and the incorrect proportion equals one minus the correct proportion. In contrast, most other methods describe probabilities for only a sample of arbitrarily defined proficiency levels and may require hundreds of independently calculated statistics to estimate the different probabilities. They also usually require specification of arbitrary constraints or rules to correct for estimation errors.

The two distributions in figure II.A.1 illustrate some fundamental features of IRT. The two curves represent distributions of proficiency for respondents to a single test item. The solid line on the left describes the proficiency of students who responded incorrectly, and the second curved line with dashes (---------) describes the proficiency of students who responded correctly. This item has a facility of 0.50, which reflects the identical height of the two distributions
along the vertical axis; there are as many correct as incorrect respondents. The mean proficiency of correct respondents is 0.10, which is reflected in the graph by the peak of the distribution for correct students being directly above 0.10 on the proficiency axis. Because the overall mean of both populations is 0, and they are equal in size, the mean proficiency of incorrect respondents is symmetric at −0.10. The two distributions are very similar in terms of both size and location, indicating very little difference in proficiency between the type of students who respond correctly and the type who respond incorrectly. If there were no difference, both distributions would be identical with means equal to 0, and the item responses would have no relationship to proficiency.

A much more accurate test item, also with a facility of 0.50, is illustrated in figure II.A.2. This item illustrates the strongest relationship between item response and proficiency, where the mean proficiency of the correct respondents is approximately 1 and the
mean proficiency of the incorrect respondents is approximately −1. No overlap exists in the distributions, indicating that, in terms of proficiency, the correct respondents are completely distinct from the incorrect respondents.

In practice, it is extremely rare for correct respondents to be completely distinct from incorrect respondents. Typically a wide region of proficiency exists in which the two distributions overlap. A smooth transition occurs as students with increasing proficiency become less likely to be members of the incorrect distribution and more likely to be members of the correct distribution. This transition is illustrated in figure II.A.3 for an item with a facility of 0.60 (indicating the distribution for correct respondents is larger than that for incorrect respondents) and a mean proficiency for the correct respondents of 0.40. The broken curved line, which is also known as an item response function (IRF), describes the size of the distribution of correct respondents relative to the size of the distribution of incorrect respondents.
In other words, in regions of proficiency where the height of the correct distribution is lower than the height of the incorrect distribution, the IRF is below 0.5; when the reverse is true, the value is above 0.5. The transition from incorrect to correct is marked by a threshold (the vertical dashed line in figure II.A.3) that corresponds to the point where the distribution of incorrect respondents intersects with the distribution of correct respondents.

The IRF can be interpreted as the probability that a respondent with a given proficiency level will belong to the group of correct respondents. The exact values of the IRF can be calculated by dividing the probability for the distribution of correct respondents by the sum of probabilities of both distributions. For example, at the proficiency value of −1, the probability value of correct respondents is approximately 0.06 and the value for the incorrect respondents is approximately 0.15; 0.06/(0.06 + 0.15) = 0.29. Because the proportion of incorrect respondents is the reverse of the proportion of correct respondents, and the mean proficiency of incorrect respondents can
be calculated from the mean proficiency of correct respondents (given that the overall mean equals 0), the IRF is a function of the item facility and the mean proficiency of the correct respondents.

An IRF can be described using a statistical model with three parameters, $a$, $b$ and $c$:

$$P(u = 1) = c + (1 - c)/(1 + \exp(D \cdot a \cdot (\theta - b)))$$

where $P(u = 1)$ is the probability of a student achieving a correct response. $D$ represents a constant used to scale the item parameters; it is usually set to equal $-1.7$ so that the scale matches the standard normal scale. The $\theta$ variable represents student proficiency. The same model describes partial credit items where $P(u \geq x)$ represents any score greater than or equal to a specific partial credit score, $x$. In the partial credit case, each score greater than zero would have a set of parameters.

Although all of the parameters interact to describe the statistical behavior of an item, the $a$ parameter primarily reflects the distance between the means of the correct and incorrect distributions; the $b$ parameter primarily reflects the facility of the item; and the $c$ parameter reflects the probability that a student from the incorrect distribution will be mistakenly included in the correct distribution (for example, a student guesses correctly).

Because the IRT process is iterative and computationally intensive, different software packages may produce slightly different estimates and require different amounts of time to complete the calculations. IATA’s estimation algorithm tends to be more robust across a variety of sample sizes and is noticeably faster than other IRT estimation programs. Whereas other methods use iterative approximation algorithms to perform the item parameter estimation step, IATA calculates item parameters algebraically using the following equations:

\[
\begin{align*}
a &= -\left(\mu_{\text{correct}} \div (-1 + p^* + p^* \mu_{\text{correct}}^2)\right)/1.7 \cdot (1 + q/(q + q_{\text{correct}})) \\
b &= (\mu_{\text{incorrect}} + \mu_{\text{correct}} - (2^*\sigma^2^*\log(q^*/p^*))/(\mu_{\text{incorrect}} - \mu_{\text{correct}}))/2 \\
c &= q/(q + q_{\text{correct}}),
\end{align*}
\]
where

\[
p^* = (1 - (1 - p)/(1 - c))
\]

\[
q^* = q + q_{correct}
\]

\[
\mu_{correct}^* = \frac{-\mu_{incorrect}^* (1 - p^*)}{p^*}
\]

\[
\sigma^2 = 1 - (p^* \mu_{correct}^* + (q^*)^2 \mu_{incorrect}^* + p^* \mu_{correct}^* + (q^*) \mu_{incorrect}^*)
\]

\[
\mu_{correct} = \text{the average proficiency of students who responded correctly}
\]

\[
\mu_{incorrect} = \text{the average proficiency of students who responded incorrectly}
\]

\[
p = \text{the proportion of students who responded correctly}
\]

\[
q = \text{the proportion of students who responded incorrectly}
\]

\[
q_{correct} = \text{the proportion of students who could not answer the question but who guessed correctly (this statistic must be estimated by approximating the lower asymptote of the empirical item response function). Note that if the c parameter is constrained to equal zero (which is a best practice in a wide variety of assessment situations), then } q_{correct} \text{ does not need to be estimated.}
\]

The new parameter estimates are used in each estimation cycle to produce updated proficiency functions for each student, using methods described by Baker and Kim (2004). Although the algorithm still requires many iterative cycles to produce final estimates, the robustness of the preceding equations at the item parameter estimation step greatly reduces computation time and increases stability of the estimates.

**NOTE**

1. In IRT, student proficiency is described on a scale (often called *theta*) that is similar to the Z-score scale: the theoretical average proficiency level is 0, and the standard deviation is 1. Most students usually have scores between −2 and 2, and less than one in a thousand students will have scores less than −3 (or greater than 3).
REFERENCES


Study at the Fourth and Eighth Grades. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.


The World Bank is committed to preserving endangered forests and natural resources. The Office of the Publisher has chosen to print *Analyzing Data from a National Assessment of Educational Achievement* on recycled paper with 100 percent postconsumer fiber in accordance with the recommended standards for paper usage set by the Green Press Initiative, a nonprofit program supporting publishers in using fiber that is not sourced from endangered forests. For more information, visit www.greenpressinitiative.org.

**Saved:**
- 18 trees
- 9 million British thermal units of total energy
- 1,567 pounds of net greenhouse gases
- 8,498 gallons of waste water
- 569 pounds of solid waste
National Assessments of Educational Achievement

Effective assessment of the performance of educational systems is a key component in developing policies to optimize the development of human capital around the world. The five books in the National Assessments of Educational Achievement series introduce key concepts in national assessments of student achievement levels, from policy issues to address when designing and carrying out assessments through test development, questionnaire design, sampling, organizing and carrying out data collection, data cleaning, statistical analysis, report writing, and using results to improve educational quality.

*Analyzing Data from a National Assessment of Educational Achievement* is the fourth of five volumes in the National Assessments of Educational Achievement series. Other volumes have described the procedures in an assessment up to the point when data have been prepared for statistical analysis, the topic of this volume. The precise analyses that are carried out will depend on the information needs of policy makers and education managers. In most national assessments, these relate to the quality of student learning, factors related to learning, equity issues, and in some cases, change in educational outcomes over time.

Volume 4, which comprises two parts, provides step-by-step details on how to analyze data collected in a national assessment. Part I provides a general introduction to statistical analyses normally carried out in large-scale assessments, measuring central tendency and dispersion of student scores and relationships between variables. Part II describes IATA (Item and Test Analysis) software, which uses classical test and item response theories to establish scales on which to report student scores. Steps in the analysis of pilot and final test administrations are described in detail.

An accompanying CD contains specially designed exercises and supporting data files for both parts of the volume.

This book will be of interest to assessment specialists in national, regional, and local governments; research institutions; and universities.