Improving Testing for English Language Learners

Rebecca J. Kopriva
Improving Testing for English Language Learners
Improving Testing for English Language Learners

Rebecca J. Kopriva
Contents

Preface xi

1 Purpose and Overview 1

2 Changing Demographics in a Testing Culture: Why This Issue Matters 13
   Demographics and School Policy 14
   Demographics Snapshot 14
   An Educational Policy Snapshot 17
   ELLs: A Heterogeneous Population 19
   Culture and Learning 20
   Culture and Assessment 25
   The Educational Assessment of English Language Learners 29
   Tests of English Language Proficiency 30
   Large-scale Academic Achievement Assessments 33

3 The Past as Preparation: Measurement, Public Policy, and Implications for Access 37
   A Brief History of Large-scale Achievement Testing 37
   Development of Measurement Theory Models 42
   Reliability to Multidimensional Item Response Theory 42
   A Changing View of Test Validity 45
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fifty Years of Federal Policy Regarding Equity and the Use of Achievement Testing</td>
<td>50</td>
</tr>
<tr>
<td>ESEA: Full Inclusion, Disaggregated Reporting, and Beyond</td>
<td>50</td>
</tr>
<tr>
<td>Other Federal Influences</td>
<td>53</td>
</tr>
<tr>
<td>Building Assessments to be More Accessible for ELLs</td>
<td>55</td>
</tr>
<tr>
<td>Conceptual Underpinnings</td>
<td>55</td>
</tr>
<tr>
<td>Developing Accessible Tests</td>
<td>59</td>
</tr>
<tr>
<td>Estimating Levels of Access</td>
<td>60</td>
</tr>
<tr>
<td>4 Getting Started: Issues of Participation, Alignment, and Validating Access with Test Specifications</td>
<td>65</td>
</tr>
<tr>
<td>Participation in Test Development</td>
<td>66</td>
</tr>
<tr>
<td>Expert Participation in Test Development</td>
<td>66</td>
</tr>
<tr>
<td>Student Participation in Small Sample Pilots and Large Sample Field Tests</td>
<td>67</td>
</tr>
<tr>
<td>Test Specifications and Content Validity</td>
<td>68</td>
</tr>
<tr>
<td>Components of Test Specifications</td>
<td>68</td>
</tr>
<tr>
<td>Documenting Content Validity for English Language Learners: The Access Specifications Package</td>
<td>70</td>
</tr>
<tr>
<td>Alignment Considerations</td>
<td>73</td>
</tr>
<tr>
<td>Overview</td>
<td>73</td>
</tr>
<tr>
<td>Alignment and English Language Learners</td>
<td>76</td>
</tr>
<tr>
<td>5 Providing the Foundation of Principled Test Construction: Maintaining the Integrity of the Item Targets</td>
<td>81</td>
</tr>
<tr>
<td>Overview</td>
<td>82</td>
</tr>
<tr>
<td>Overview of Access and Item Development</td>
<td>84</td>
</tr>
<tr>
<td>The Centrality of Item Targets: Clearly Defining Intended Knowledge and Skills</td>
<td>88</td>
</tr>
<tr>
<td>Development and Completion of Access-based Item Templates</td>
<td>90</td>
</tr>
<tr>
<td>6 Access-based Item Development</td>
<td>101</td>
</tr>
<tr>
<td>Contextual Factors</td>
<td>105</td>
</tr>
<tr>
<td>Culturally broad experiences</td>
<td>105</td>
</tr>
<tr>
<td>Making Expectations Clear and Explicit</td>
<td>108</td>
</tr>
<tr>
<td>Prior Learning Expectations</td>
<td>110</td>
</tr>
</tbody>
</table>
Structural Factors 113
  Simple Language Structures 114
  Vocabulary 118
  Effective Visuals 121
  Item Format 128
  Text Amount 132
  Tools 134
  Impact of Home Language 135
  A Note about Rubrics 136

7  Tools, Test Forms, and Reviews 141
  Tools 141
    Text Supports 143
    Manipulatives 145
    Content-based Resources 147
    Activities 148
  Form Considerations 149
    Choosing Items 150
    Plain Language Forms 152
    Computer-based Testing 153
    L1 Form Considerations 154
    Using Portfolios 157
  Reviews 160
    Internal Reviews 160
    Fairness Reviews 161
  Putting It All Together 167

8  In What Language Should English Language Learners Be Tested? 169
  Defining ELLs 171
    Variation Among ELLs 171
    Patterns of Language Proficiency 173
    Additional Sources of Linguistic Differences 176
    Limitations of Official Definitions of “ELL” 177
    Section Summary 179
  Classifying ELLs 180
    Schooling Background 180
    Defining and Assessing Language Proficiency 181
    Literacy and Academic Language as Part of Language Proficiency 182
Operational and Prototypical Systems 240
  Guidance Models 241
Research-based Standardized Data Collection and Assignment Systems 244
Implications 251

11 Scoring Considerations for English Language Learners 255
  Issues Which Impact Scoring 256
  Linguistic Considerations Influenced by the Home Language 257
  Cultural Considerations 262
  Language Acquisition Issues 268
Access-based Scorer Training 271
  Scoring Process Considerations 274

12 Selected Technical Considerations 279
  Developing an Interpretable Research Agenda for Documenting Validity 281
  Validation Design Model 282
  Research Design Considerations 286
Selected Quantitative Approaches 288
  Modeling Relationships of Variables: Selected Methods 289
  Modeling Data from Continuous and Categorical Variables 290
  Modeling Frequency Data from Categorical Variables 295
Comparison of Groups and Treatments: Selected Methods 297
  Comparison of Test Score Level Data 297
  Differential Item Functioning 298
  Other Methods for Analyzing Items 304
  Comparison of Model Structures 305
Qualitative Approaches 306
  Tryouts and Pilots 307
  Cognitive Labs 308
Comparability Considerations 310
  Defining Comparability 310
  Analyzing Construct Equivalence 313
  Analyzing Score Equivalence 316

References 319
Index 345
Many advances have occurred over the years in how we systematically assess the academic knowledge and skills of students in the U.S. However, there has not been much movement in how students’ cognitive schemata, schooling environments, and backgrounds impact their test scores and experiences, and hence the inferences one can validly make about their academic achievements and learning. This is troubling because, as we know more about the diverse ways students learn, access knowledge, and develop skills, we know that scores can sometimes mean substantially different things.

This book is meant to systematically address one less-developed area by focusing on the largely un-evolved interactions between test and test takers. Traditionally, the only time the interaction is considered is in estimating the relationship between test scores and student abilities on the targeted constructs. This is fine, when ancillary abilities do not interact with how the students perform on items meant to measure targeted content. However, the ancillary abilities of some groups of test takers routinely impact how they answer items, including many English language learners. We know this may confound the measurement of targeted abilities and distort the results. However, to date, we do not know the extent of the impact for different students because our measurement methods are less than ideal.

Until recently not much thought was given as to how test items, forms and implementation might need to be adapted to consider English learners,
whose experiences and backgrounds (much less language) are often substantially different from typical, mainstream, U.S. students. This is surprising, given that there has been solid evidence which supports the fact that socio-cultural influences can seriously affect how students learn. It is also surprising, given that the number of English language learners is exponentially increasing in many states, academic assessments are routinely used to make and defend resource and placement decisions, and that states are now legislatively required to hold schools accountable for all their students, including this population.

*Improving Testing for English Language Learners* was written to explain in some detail how to construct more appropriate content assessments for English learners and how to interpret the test scores for this population, given certain types of procedural and empirical evidence. The goal of the book is to develop items and tests that, if properly constructed, produce results equivalent to assessments for the general population so scores can be compared across the full breadth of students. Further, the book aims to explain, recommend, and demonstrate when the defensibility of score inferences for English learners can be argued to be consistent with those reported for the larger student body, and when inferences for ELLs are vulnerable to misrepresentation.

Chapter 1 will present an overview of the rest of the chapters, explaining how interventions throughout several of the phases fit together and sum to a testing environment more appropriate for this population. Chapters 2 and 3 provide the larger population and testing contexts on which this work is being built. Chapters 5–8 focus on developing proper materials, Chapter 9 addresses other promising accommodations, and Chapter 10 recommends how students and accommodations might adequately fit together to provide a meaningful testing experience. The last two chapters address scoring and recommendations for evaluating the validity and comparability of the proposed inferences. In order to properly build a content validity argument for English learners, Chapter 4 (Getting Started) presents the Access Specifications Package. This package is a tool to collect access data over interventions, items and student needs. Outlined by test developers before a test is developed or customized, the framework is designed to be filled in after the test forms and reviews are completed.

The book has been written over the last few years as an in-depth follow up to *Ensuring Accuracy in Testing for English Language Learners: A Practical Guide for Assessment Development*, published by The Council of Chief State School Officers (CCSSO) in 2000. The current manuscript was built to respond to a request for a holistic framework for improving content testing for this population, and to a request to coordinate and contextualize the
literature to date associated with different aspects of test construction, implementation, reporting, and validation. This tension has proven to be challenging and I have tried to balance the competing demands to complete this volume. I have also had a great deal of help that I would like to recognize here. First, my appreciation extends to co-authors Guillermo Solano-Flores, Elise Trumbull, Jennifer Koran, David Wiley and Sue Rigney who have done a terrific job focusing on their topics. At my Center for the Study of Assessment Validity and Evaluation (C-SAVE) at the University of Maryland, staff and students contributed to the thinking and literature reviews reflected in the book, and helped me clarify what would be covered and how. In particular, many thanks to Phoebe Winter, Carol Boston, Chen Su Chen, Ryan Monroe, and Bob and Robbie Mislevy, as well as Ming-yi Cho, Jessica Emick, Julie Dahmer and other staff, consultants and grad and undergrad students who worked on specific aspects of various research projects. Current staff from C-SAVE and others at the University of Wisconsin and the Center for Applied Linguistics in Washington D.C. have provided valuable input, as have staff from partner states and districts, especially Therese Carr from South Carolina. My deepest thanks go out to Megan Kopriva for her invaluable editing, table-making and coordinating of the references. Lastly, I want to thank Julia Lara, currently the mayor’s educational liaison in D.C. and formerly Deputy Executive Director at CCSSO, and Richard Durán, UC Santa Barbara, who provided wonderful recommendations for improving the manuscript.

This book is a snapshot at a point in time—as it was being completed and placed in production I was very aware of several exciting new initiatives that will continue to refine what is discussed here. Clearly, computer possibilities are extending how measurement for ELLs can be done. This seems like it may be particularly useful for students with low literacy in English where typical accommodations don’t seem to be enough for many students. Work is ongoing now to investigate systematic methods that might work for this population and are simultaneously comparable to results reported on the general assessments. Related work is also examining how less typical testing methods might be used to provide defensible evidence-centered documentation for various populations of students, including how cognitive advances can be built into assessments so progress in learning as well as achievement can be measured. Finally, an issue being tackled recently is untangling linguistic and cognitive complexity. This is not a straightforward task as more complex concepts in subject areas are typically articulated using more sophisticated discourse and language constructions. Findings from this work will clarify when and how items might be measuring academic content and when they are measuring English
proficiency. As item parameters can be identified which can produce items more clearly measuring one or the other, improved writing procedures will help researchers, test developers, and teachers as they address both assessment genres.

It is my hope that practitioners, developers and researchers will find the book useful, challenging and a source of new insights and inspiration for their work. Most important, however, my hope is that the book will translate into improved assessments with increasingly valid inferences of the academic abilities of English language learners. If English learners benefit from some of the ideas raised here, it seems to me this book can be called a success. And this is what makes all the work worth it at the end of the day.

Rebecca Kopriva
June 27, 2007
Improve Testing for English Language Learners is designed to summarize how academic tests, for instance assessments which measure achievement in mathematics or science, might be made accessible for English language learners, or ELLs. The focus of the book is to address this issue within the mainstream large-scale academic tests that states and districts use. Large-scale is being defined here to include tests that are developed to be used to obtain data about the achievement of students over classrooms, schools, districts, or states, although many of the points may be useful for classroom purposes as well. As this text will document, many researchers and test developers have found that it is not sufficient to focus on only one aspect of test development or implementation, for instance on providing administration accommodations post hoc without considering this population when tests are designed or when studies of technical quality are being conducted. Rather, access needs to be considered at several points, beginning with the conceptualization of what the assessment is intended to measure.

The book is organized to take the reader through several of the aspects of test construction and implementation. After summarizing the heterogeneity of the ELL population and the technical and applied measurement context from which an access-based large-scale test will emerge, the chapters will cover aspects such as the design of a particular test or testing system,
item and form development, the consideration of home language issues in how testing materials are compiled and presented, appropriate accommodation assignment, and the analyses of validity and comparability. To date, the research is not definitive in how the development of such an assessment system is to be accomplished. While every effort was made to cite key work associated with each of the points throughout, undoubtedly some literature was omitted. However, it is probably the case that omissions would not substantially alter the fundamental character associated with the current status of the literature. As such, the volume will attempt to outline the issues and findings to date, and make recommendations about future directions within each aspect of the development and execution of the tests.

Each of the chapters will be quickly summarized below. The goal of this overview is to place each of the chapters within the context of the systemic purpose of the text—that is, to build accessible testing systems that work for many students, including English language learners. This is not a book about building large-scale content tests for ELLs that are separate from those built for other students. The contents are intended to illustrate how assessment systems, as they are designed today, can be augmented to handle a broader range of students who are schooled in the U.S.

It is anticipated that the audience for this volume will be those centrally or peripherally related to improving academic test development in ways that are intended to make assessments more accessible for more students. This includes researchers, academics, those in research centers and in test publishing houses, and it covers both those with a stronger expertise in measurement and technical issues as well as those whose expertise is in the population of English learners, language minority students, or other students with various types of needs. It also covers those who are consumers of the assessment systems, that is, assessment, ELL or other staff in state or local agencies and their consultants who work at the national, state or local levels, including teachers and local specialists.

This book will approach the issues from a formal perspective. Readers may find this approach dense, but the direction was taken in order to provide readers with the arguments and literature they may need to defend the building and implementing of such an access-based assessment system. It was written to be used as a reference source of salient work currently completed, and to provide an overview of selected recommendations presented within the test development context as a whole. By locating the range of issues within one volume and explaining how they are related to one another and to the development of an inclusive testing system, the text was also designed to be used as a comprehensive framework and basis for future work in this area. However, because the myriad considerations were placed within one volume, rather than several, it also means that many areas
are just briefly touched upon, and this remains a trade-off inherent in this decision. It is expected that, after completing this chapter, many readers will focus on specific chapters that deal with their immediate concerns. They are encouraged to periodically scan the range of the book’s contents and come back to this chapter, however, in order to place their issues within the total context of an accessible testing system.

At different times in the book, the discussions may be either somewhat detailed or vague as the objective is to provide a rudimentary summary without writing volumes of information and without resorting to more lengthy explanations of commonly utilized methods. Overall, the amount of information written about each portion of the chapters is based on providing evidence and explaining procedures or ideas which have been widely debated, are not widely known but should be, or where there appear to be common misconceptions. Readers are routinely directed to additional literature which may help them more fully explore the issues on their own as necessary.

The intent of the author is that this text should be complemented by “how to” handbooks which will necessarily elucidate the points made here (as well as points based on future work in these areas). Some training materials or manuals have been developed and utilized for some time, and many of these will be mentioned in their appropriate sections. Some new handbooks are underway or are contemplated. For instance South Carolina is completing a manual explaining the process for completing access-based items and test materials across content areas and grades (Carr and Kopriva 2007). The Council for Chief State School Officers (CCSSO) has proposed a handbook which should guide some new work in how comparability might be evaluated, especially comparability when multiple types of testing forms or implementation methods are considered (Bazemore, 2006).

As the chapters will, hopefully, confirm, the issue of improving how the large-scale academic achievement of English language learners is measured is of growing importance. It is not enough to teach these students English and delay their academic schooling, and several volumes have documented how many students may fall further behind if their ongoing academic language acquisition curriculum is not integrated with content learning and appropriate evaluation of both. Neither is it appropriate to measure the academic content of ELLs using tools not designed to provide valid and useful data about these students. In the era of increased accountability, and where resources often follow accountability results, it is important that the academic findings for these students are fairly represented. Further, as Chapter 2 documents, this challenge is ever-increasing as more and more English language learners are arriving in schools across the U.S. This book is meant to provide a basis for integrating what might be done toward reaching this goal.
Chapter Overview

Chapters 2 and 3 will lay a foundation for the promising improvements discussed in the remainder of the book. Chapter 2 explains who the English language learners are: the extent of their influence throughout the U.S., their heterogeneity in terms of racial/ethnic background, an explanation of the range of prior schooling experiences and needs these students come to U.S. schools with, and the sustained and pervasive influence of culture on the way the ELLs interpret and interact with test items and tests in general. It will also summarize where and why their challenges run counter to the commonly used types of text materials and administration and response methods associated with today’s large-scale content assessment. The next chapter is a counterpoint to Chapter 2. It will focus on the history of measurement in terms of advances in psychometrics and how these advances allow for technically rigorous inferences to be made in many circumstances. The chapter will review the limits and trade-offs of the many decisions which have been made to date in order to ensure these technically defendable interpretations. Finally, the role of federal policy and how it has shaped a climate of inclusion (and accountability) will be summarized.

The reason both of these chapters are included here is because it is important for all readers to understand the extent of the complexity of the challenges associated with making progress in validly assessing ELLs. It is also important for readers to grasp the possibilities inherent in the personal strengths of these students built from exposure to more than one culture, and the knowledge gleaned from learning how to compensate for their limits. Further, central to any meaningful improvements are an understanding of measurement considerations, the function of cognitive and linguistic advances, and how the maturing of the psychometric field is setting the stage for sustainable change. Most readers will have more experience and expertise in one or other of these very diverse fields, so the objective is to balance their current knowledge from their area of work with a quick summary of important components in the other. It is argued here that twin apprehension of understanding the students and the intent and applications of measurement is essential for substantial advances to occur.

This book contends that assessments will not be properly accessible if they are not explicitly designed that way. Chapter 4 lays the groundwork for such a system. In particular it emphasizes that three aspects of such a base are key: appropriate participation, independent alignment evaluations, and proper item and test accounting of accessibility. The participation section focuses on the participation of ELL experts at all levels of development, not just in bias or fairness reviews. This is necessary because subtle
distinctions in proper and improper development can be identified and addressed as the test is being constructed. Because of the constraints of scope, time allocation, and number of constituents that need to be present, fairness reviews do not provide the level of detailed dialogue that ongoing participation allows. Additionally, ELL student participation in small sample tryouts and other formative data collections provide vital qualitative evidence about how items are being interpreted by this population. Further, large sample data collections which appropriately sample ELLs allow for empirical evidence to be collected so analyses can be conducted to demonstrate the validity and comparability of testing systems for these students. One section in Chapter 4 outlines the importance of independent content alignment studies (now fairly routine among test developers and their customers). The chapter discusses the issues that alignment reviewers need to address in order to determine, independently from the developer, whether access parameters, among other considerations, have been met appropriately. These parameters will be introduced here.

A routine part of test design is the development of content-based test specification matrices, which, in essence, interpret targeted content standards the test is supposed to reflect. The matrices identify the range of topics to be covered on a test and often a determination of cognitive complexity. Each matrix is generally built to be completed as items are written and reviewed. As tests are constructed, each item is placed in the appropriate cell to identify holes and adjust expectations. The completed content matrices provide explicit and clear verification of the developer’s judgment of content validity and judgment-based validation that the score inferences address the proper material.

Chapter 4 argues that, for items and tests to be accessible, explicit recognition of how each item and administration option is addressing the needs of the ELL test takers is needed as well. All in all, a majority of the access-based options needed by English language learners center around their language and cultural issues. These issues are primarily addressed in the development of proper item and test materials and appropriate response opportunities. Administration options generally support these other accommodations but are not sufficient by themselves. Therefore, this chapter contends that access-based test specification matrices need to accompany the content test specification matrices, and a procedure is presented that explains how this might be accomplished. Just as content test specification matrices are essential at each stage of development, access-based matrices need to be considered and built at the design stage. They also need to be completed during development to identify holes and improve items, test materials, or expand the range of administration or
response accommodations. Once completed, they form an important aspect of validation in that they can be used to demonstrate, at a face validity level, that proper methods have apparently been undertaken to address the needs of the diverse students. The access-based matrices provide essential evidence that the content test score inferences may be appropriate for these students. Finally, the matrices could help guide empirical analyses which can further verify the validity of the scores for this population.

Chapters 5 and 6 implement an important aspect of the design work explained in the access-based test specifications. These chapters address how to write items so that they minimize target-irrelevant interference for students with specific limits and needs, while utilizing their strengths to help compensate for the lack of English language and literacy acquisition. In this book target or targeted content refer to what developers are intending that items or tests measure. Clearly, targeted content at the item level is at a much finer grain, whereas test targets refer to broad explanations of the content (such as content standards) covered in the assessment. Test score inferences flow from the test target explanations. “Target-irrelevant”, “construct-irrelevant”, or “ancillary” are terms which are used interchangeably throughout the text, and which refer to elements in items or tests which are also measured but aren’t part of what developers intend to measure. For instance, a target-irrelevant variable in a science item may be the reading skill that is required for students to understand what science problem the item is asking them to solve.

Chapter 5 focuses on laying the foundation of principled item construction so that the integrity of the target can be maintained over versions of items which measure the same target but require different ancillary skills to access it. Understanding the relationships between the item and condition variations is critical to defending comparability of test scores. This chapter briefly summarizes the research on item structure methods to date which detail how one might produce sound item variations. A method and template for retaining the integrity of the items which vary in their use of ancillary skills are presented and discussed. This method is more stringent than many which have been used to date, as most other template methods focus on changing content only while issues of how and when access might be affected were not considered. However, this chapter contends that the level of constraint in the access-based template is imperative in order to defend the rigor of the comparability. An overview of the type of item writing which could satisfy such a method is outlined and an example is presented.

Chapter 6 presents the specific elements associated with access-based item writing. Relevant research findings are summarized and the elements
are sorted into contextual and structural components based on what parts of the item writing process they are impacting. Contextual factors cover the use of culturally-exclusive and inclusive contexts, the issue of non-consistent prior learning experiences, and the importance of making expectations clear. Seven broad categories of structural elements are discussed, including the roles, methods, and uses of appropriate language structures, vocabulary, visuals, format, and text amount. Even though tools, language aids, and administration and response options will be discussed in later chapters, they are briefly considered here so they can be effectively addressed in and supported by the items they are intending to impact. Finally, the influence of the students’ home language on English items is mentioned. In many of the categories, multiple approaches are explained in order to utilize these structural components to achieve different aims. For instance, purposes of using visuals include the use of graphics to mirror the requirements in an item or support problem solving. Methods of utilizing visuals, such as organizing events in time, replacing text and providing first-person access, are also outlined. A brief discussion of rubric considerations completes the chapter. Throughout this and subsequent chapters, the role of technology is discussed. Because of the diverse needs and strengths of ELLs, technology can provide increased avenues and differential presentations of item options to appropriate students. This seems to be particularly relevant for some students with the lowest literacy in English.

Chapter 7 summarizes salient literature findings about access-based tools, test forms, and internal and external reviews. Benefits and consequences of selected materials and recommendations on selected topics are also outlined within each section. The section on tools reviews bilingual, monolingual English, and picture-word text supports. The use of manipulatives and other concrete and computer-based objects is discussed within the context of increasing access, and content-based resources and the role of activities associated with items or items groups are addressed. The section on forms reviews issues of building access-based forms, including considerations associated with choosing items which anticipate access. Computer-based testing is briefly introduced and home language form issues, such as translation and dual language forms, are considered. Finally, a small section outlines some work associated with considering how to approach the use of other testing methods.

Finally, Chapter 7 summarizes the role of internal reviews and external fairness reviews. Internal reviews of items and test materials are important because they provide opportunities for other ELL and content area experts to have input into whether item options are retaining the integrity of the
targets across items. This chapter recommends that some internal reviews involve the use of proper small sample data obtained from ELLs with a range of English proficiency skills (for instance, data from informal item tryouts and cognitive labs) so these findings can inform the process. The external fairness review process, a common aspect of test development, is discussed as well, and suggestions for improving the scope of the charge and use of these panel experts and representatives are outlined.

Chapter 8, by two authors familiar with native language content testing, reviews the issues critical to making decisions on the language used to assess the academic achievement of ELLs. They discuss the impact of policy and instruction, and the limitations of current approaches used to test ELLs in their home language or in English. They argue that sound decisions need to be based on considering the quality of information about the home and English language proficiencies of this population, the instructional capacity needed to effectively implement testing in each language, and the institutional capacity needed to properly provide accommodations. The chapter presents alternative testing models based on recent research findings.

As Chapters 5–8 document, the task of developing appropriate items and testing materials, in English or in other languages, is not one of following a checklist and proceeding in a manner where all items and forms are handled in the same way. A recently completed project in South Carolina suggests that this process is uneven, iterative and involves a maturing of how diverse methods explained here address unique considerations in different types of items, for different content areas, and different grades (Carr and Kopriva, 2007). The chapters anticipate that future work will shed light on how these methods need clarification or refinement, but to date, the intent has been to give readers a flavor for the complexity of the task and some guidance about useful techniques.

Chapter 9 summarizes the types of administration and response accommodations which appear to be promising for English language learners. Most of the research to date has focused on administration options, and four accommodation options are reviewed. While there has been little work to provide guidance regarding response accommodations, it is argued that these options are also centrally important. For instance, while multiple choice responses are easy to give (just fill in a bubble), this approach does not necessarily mean the item type is providing access for ELLs or that the responses are validly communicating what the students know. Conceptually it makes sense that if students have trouble accessing the item and test requirements because of language, language will tend to be an issue in response as well. The challenge is to find formats which allow these students
to demonstrate their skills. Advances in providing response options via the computer, in particular, are discussed.

The last part of Chapter 9 outlines the concept of pretest support for this population. Two types of pretest support are suggested, the literature is reviewed, and several specific examples of the types of support that could be provided are introduced. Pretest support as it is defined here is distinct from the kinds of test preparation common in U.S. schools. This support focuses on identifying and addressing issues of cultural discontinuities between the students’ and their families’ prior experiences and those which are assumed in U.S. schooling and U.S. testing specifically. The section provides some rationale to suggest that targeted support provided prior to large-scale testing may be able to ameliorate some of the validity problems for ELLs.

Chapter 10 outlines the accommodation assignment issues and presents the work completed to date that addresses how to select appropriate accommodations for the students who need them. As testing systems are improving the ways they are assessing the content knowledge and skills of ELLs, these improvements will not reach the correct students unless reliable and effective student–accommodation matching systems are in place. A series of surveys and studies confirm that improper assignment is commonplace, and it is argued that this oversight has consequences for undermining the possible advances in validity which could be inferred from test scores. The chapter will review key English language learner student variables which appear to impact the assignment decisions, and it will quickly focus the reader on many of the promising accommodations and options for this population which were discussed in earlier chapters. The pros and cons surrounding policy approaches and research-based approaches which have been developed to date are then described, and a number of completed and prototypical systems are outlined. In this chapter, as in several of the others, both English language learner and students with disabilities literature is reviewed and promising next steps are suggested.

Chapter 9 notes that few studies have addressed the issue of access in the responses of ELL students. One point raised in that chapter is that there is a complex interaction between language and the demonstration of knowledge. Because of the nuanced meanings native English speakers ascribe to text or other shorthand answer choices (for instance graphics), having English learners choose between fixed options is not necessarily going to yield a more valid or equitable response. Rather, there is some work which suggests that a number of ELLs prefer to try to explain themselves rather than choose pre-selected alternatives, even if and as they are using language (or related devices such as pictures) to do so. As such,
Chapter 11 focuses on responses to constructed response items, even though these types of items are currently less in favor than they once were. It briefly summarizes how scorers (mostly native, monolingual English speakers) might understand and handle the responses of these students when their responses are presented to the scorers in text-based form. The chapter is largely drawn from a handful of earlier work as there do not appear to have been any more recent information which documents how item scorers might interpret these responses.

One key consideration in Chapter 11 is that linguistic and cultural influences from the students’ home countries or communities tend to influence their responses as they are learning how to express themselves in English at a level that is grade-level competitive with native-speaking peers. Five aspects of their native language (phonology (sounds), morphology (words), syntax (sentences), semantics (meaning), and use) impact how students perform linguistically. Cultural influences, such as different symbols or procedural systems, word choices (similar words in the native language but mean something different than English), and stylistic preferences in their home culture impact responses as well. Further, issues of English language acquisition—immaturity in the use of the language they are learning—interact with what ELLs bring from their home language. This chapter briefly summarizes some of these influences, and provides suggestions for scorer training.

The last chapter addresses selected technical considerations. In particular, it focuses on some issues in validity and comparability in an effort to summarize how the readers might approach the documentation of these two areas. The validity section initially discusses how one might develop an interpretable research agenda related to demonstrating technical rigor of scores of English language learners. Next, the section summarizes research questions focus on modeling relationships and comparisons of groups which can be addressed with data from large samples of students. Selected procedures, and some studies which have used them, are outlined. Two approaches utilizing small sample data collections are highlighted, and then the last section quickly reviews the literature to date associated with demonstrating comparability of test score data for ELLs when some conditions vary (or when they stay the same). Serious documentation of comparability under different conditions is just beginning so this section will focus mostly on providing some suggestions of what data might be used and how.

All in all, the volume covers a tremendously broad expanse of ground. Besides providing some explanation of the individual points where developers and consumers of an access-based assessment system should
intervene, a central goal of the text is to illustrate the systemic nature of this undertaking. Undoubtedly, future research will prioritize these points with greater clarity, but effective measurement of ELL achievement will most likely always involve multiple, interconnected aspects. In a similar fashion, addressing these points will almost certainly have an impact on the system for native English speakers as well. As advances in computer-based testing and cognition continue to inform measurement for the mainstream population, as well as English learners, many of these effects on assessment will be consistent with the efforts described here. Therefore, within the general field of measurement, this book could be seen as partially reflecting the future for all assessment. That is, it appears that future testing will increasingly take advantage of the capabilities computers have to offer. Further, future testing will continue to be more diversified. It is hoped that this text will provide some measure of guidance toward achieving these ends.

Endnote

1. U.S. students who do not speak English as their first or home language are sometimes referred to as bilingual students, limited English proficient (LEP) students, English for Speakers of Other Languages (ESOL), or English as a Second Language (ESL) students. In this book, we tend to use the term “English language learners” because it accurately describes the common task facing them without making any assumptions about their proficiency in other languages.
Changing Demographics in a Testing Culture: Why This Issue Matters

In reviewing figures from the 2000 U.S. Census, a reporter at *The New York Times* reported that “the increase in the immigrant population (a 57% increase from 1990 Census figures), which many state officials believe was undercounted, surpassed the century’s greatest wave of immigration, from 1900 to 1910, when the number of foreign born residents grew by 31%” (Scott, 2002, pp. A1, A20). Further, “for the first time immigrants moved far beyond the big coastal cities and Chicago and Denver and Houston, into the Great Plains, the South and Appalachia . . . ‘These numbers represent an enormous social experiment with very high stakes,’ said Steven A. Camarota, director of research for the Center for Immigration Studies . . . ‘and the experiment is not over’” (U.S. Census Bureau, 2002a, p. A20).

Immigration is having a profound effect on schools across the United States. Educating English language learners, particularly within the context of today’s standards-based instruction and frequent testing, sparks great debate around such issues as the following:

- whether and how children should be instructed in their home languages;
- how children can best learn English;
- how children can gain English proficiency while also making grade-level progress in other academic subjects; and
- when and how English language learners should be tested.
This chapter will describe the demographics and characteristics of English language learners and explain why educators need to evaluate how they are tracking the academic progress of these students. Historic, current, and future demographic trends and policy issues will be outlined, and a summary of the complex cultural considerations educators face in properly instructing and assessing these students will be presented. The chapter closes with an overview of two types of assessments English language learners (ELLs) face: English language proficiency assessments and academic achievement assessments.

Demographics and School Policy

In 1970, the foreign-born population of the United States was 10 million. In 2000, this population was estimated to be 28 million (U.S. Census Bureau, 2002a). The leading five countries of origin for legal immigrants in 2003 were Mexico (115,864), India (50,372), the Philippines (45,397), China (40,659), and El Salvador (28,296) (U.S. Citizenship and Immigration Services, 2003). The Hispanic population alone increased 57.9 percent between 1990 and 2000 (Guzmán, 2001).

By 1997, nearly 20% of all American children, approximately 13 million, were either immigrants or the children of immigrants (Harker, 2001). Based on 2000 census results, 18.4 percent of children between 5 and 17 speak a language other than English at home. More than 20 percent of their households may be described as linguistically isolated, that is, no one in the household age 14 or older speaks English exclusively or very well (U.S. Census Bureau, 2002).

While Spanish speakers make up the greatest number of immigrants, overall, the immigrant population is linguistically, culturally, and racially diverse. Further, ELL programs are not limited to immigrants. While many students in ELL programs in U.S. schools were born in other countries, 40 percent were born in the United States to immigrants (National Center for Education Statistics, 2001). Therefore, states are actually facing significantly higher numbers of students who have limited proficiency in English than the Census Bureau findings suggest. This fact dramatically swells the already enormous adaptations that school districts and states have to make to provide schooling for students who are English language learners.

Demographics Snapshot

Based on reports from state educational agencies, the number of English language learners in public schools has been growing almost everywhere in the country. In the 2002–03 school year, ELL students in public schools in grades pre-K through 12 numbered approximately 5,044,400, or roughly 10 percent of the total public school population. This figure represents a
45 percent increase over the reported ELL enrollment five years earlier, and nearly a doubling of the 2,736,000 ELL students enrolled in 1992–93 (National Clearinghouse for English Language Acquisition 1, 2004). It is possible that the actual number is higher because states differ in how they determine ELL status.

The states with the largest numbers of ELL students in 2002–03 were California (1,599,542), Texas (630,148), New York (302,961), Florida (292,077), Illinois (169,414), and Arizona (149,354) (National Clearinghouse for English Language Acquisition 1, 2004). These six states together accounted for 62 percent of all ELL enrollment, with California alone accounting for 32 percent. In the states where they are most concentrated, ELL students also tend to be enrolled primarily in large urban school districts and the metropolitan areas surrounding them. For example, six urban school districts—Chicago, Dallas, Houston, Los Angeles, Miami-Dade and New York City—reported ELL enrollments in excess of 50,000 students. More than 80 percent of these students spoke Spanish (Council of Great City Schools, 2004).

Schools in urban areas in densely populated states are heavily impacted by ELL students, but they are not alone. While the density of ELL students—that is, the percent of overall enrollment they comprise—is greatest in Florida and that portion of the U.S. from California to Texas, the greatest increase in ELL growth has occurred in western, midwestern, and southern states. Georgia, for example, saw its population of ELL students burgeon from 7,955 to 61,307 in the decade ending in 2001–02—a 671 percent population change. North Carolina’s ELL population increased 652 percent, from 7,026 students in 1991–92, to 52,835 in 2001–02. Nebraska went from 1,856 to 12,451 ELL students over the same period. The ELL student population more than tripled during that decade in South Carolina, Tennessee, Alabama, and Kansas, and more than doubled in the states of Kentucky, Idaho, Nevada, Arkansas, Indiana, Oregon, Minnesota, and Iowa (Meyer, 2004).

As a group, ELL students speak more than 400 languages. Across the country, the most common native language of ELL students is Spanish. More than three-quarters of ELL students (76.9 percent) speak Spanish. Vietnamese is spoken by approximately 2.4 percent of ELLs, Hmong by 1.8 percent, Korean and Arabic by 1.2 percent each, Haitian Creole by 1.1 percent, and Cantonese by 1 percent (Hopstock and Stephenson, 2003). The remaining language groups are spoken by fewer than 1 percent of ELL students. Of these, several, including Armenian, Chuukese, French, Hindi, Japanese, Khmer, Lao, Mandarin, Marshallese, Navajo, Polish, Portuguese, Punjabi, Russian, Serbo-Croatian, Tagalog and Urdu, are spoken by at least 10,000 students (Kindler, 2002).
Currently, ELL students tend to be concentrated in the early elementary grades; an estimated 53 percent are enrolled in grades K-4. In addition, they tend to be poor. More than three-quarters qualify for programs that receive federal funding based on poverty and low academic standing (Macias, 2000). And poverty typically brings another set of schooling challenges, including less experienced, less qualified teachers, larger class sizes, and fewer educational resources.

Any discussion of trends and statistics must also include caveats about the diversity of ELL students. While the profile of a typical ELL student would be young, poor, and Hispanic, there is much variability across major ethnic/linguistic categories by country of origin, cultural background, socioeconomic status, and level of education prior to immigrating to the United States. For example, in *Improving Schooling for Language Minority Students* (August and Hakuta, 1997), the National Research Council panel reported that 35 percent of families that spoke Asian/Pacific Island languages had incomes below $20,000, compared with 57 percent of families that spoke Spanish. Differences also exist within groups. The first wave of Southeast Asian refugees consisted of highly educated people who, even though they lacked English proficiency, could support their children’s learning at home and ease their transition to a new environment. Subsequent immigrants, for example, the agrarian Hmong, refugees from Laos who lacked a written language until the twentieth century, were less educated in their home country and perhaps less equipped to advocate for their children’s education (Vang, 2004–05). The children of immigrant families fleeing war-torn countries are likely to be dealing with multiple stressors that affect academics. These include seeing loved ones assaulted or murdered and homes and schools destroyed (Osterling, 2001). Many children of migrant workers face academic deficits and learning problems related to interrupted schooling and compounded by poor nutrition, substandard housing, and inadequate health care (Green, 2003).

Aside from differences in the circumstances that bring immigrant students to this country, important linguistic differences can arise based on the regions from which ELL students’ families originate. For example, the Spanish spoken by children from El Salvador can differ from that spoken by children from Mexico in seemingly slight, but significant, ways. Some ELL students come from countries with an official language and a variety of regional dialects. For example, students who have been schooled in Chinese are proficient in Chinese script regardless of their province or country. However, there are two versions of the script: simplified and traditional. Students in Hong Kong and Taiwan tend to use traditional script, while the simplified version is more common in mainland China. Further, immigrants from mainland China traditionally speak Mandarin,
while those from Hong Kong speak Cantonese, and those from Taiwan speak Fujianese. Teachers who are trying to support their ELL students will find it helpful to become aware of these differences. It can be limiting to make broad generalizations about Spanish or Asian students and their languages when considering the curriculum, instruction, and assessment issues of ELLs.

*An Educational Policy Snapshot*

The Civil Rights Act of 1964 stipulated that states and school districts receiving federal funding could not discriminate in their operations on the basis of race, color, religion, or national origin. Educational policies in recent decades have consistently maintained that all students be held to the same rigorous academic standards, with achievement measured through testing (Linn, 2000). In practice, however, states and districts have traditionally resisted including many English language learners in their mainstream academic assessment and/or accountability systems (Hakuta, 2001; Council of Chief State School Officers, 2000), in part because many educators charged with schooling this population remain skeptical about whether large-scale academic assessments are providing accurate information about their academic mastery (Figueroa and Hernandez, 2000). In the past, large numbers of English language learners were waived from the large-scale exams given to most students (Lara and August, 1996). As a result, these students were excluded from most district-wide and statewide estimates of academic achievement, for at least two to three years and often much longer (Lara and August, 1996; Wong-Fillmore, 1994). The exclusion of ELLs from regular testing kept them from benefiting from many of the positive educational outcomes of the education reform movements (Huebert and Hauser, 1999; August and Hakuta, 1997). Recent federal policies are an attempt to encourage states and districts to include these students so they can reap the benefits of the school reforms.

The Elementary and Secondary Education Act, the central federal law governing K-12 education which was first enacted in 1965 and last reauthorized in the No Child Left Behind Act of 2001, has increasingly focused on raising academic achievement for students (No Child Left Behind Act of 2001, 2002). As the next chapter describes, the No Child Left Behind Act demands that school accountability include academic progress data from English language learners as well as other subpopulations. It encompasses both the Title III program, which addresses language instruction for limited English proficient and immigrant students, and the Title I program for improving the academic achievement of disadvantaged students, many of whom are also ELLs. Initially, the thrust of these policies has been the inclusion of almost all students, including students in special
education programs and ELLs, in statewide accountability systems. Currently, as part of the No Child Left Behind Act, schools are required to collect yearly achievement data in mathematics and reading for students in grades 3 to 8 and one year in high school. A requirement for a science assessment at least once between grades 3 and 5, once between grades 6 and 9, and once between grades 10 and 12 will be added for the 2007–08 school year. The legislation further requires that appropriate progress be made by each of a school’s sub-groups, including English language learners, special education students, and students from minority groups, in order for schools not to be identified as failing and subject to state and federal sanctions.

Under the current legislation, English language learners, regardless of English language proficiency, are expected to be fully included in the assessment of academic subjects other than English/language arts, where a minimum waiver is offered for some students.

Regulation changes announced in February 2004 specify that ELLs in their first year of enrollment in U.S. schools must take an English language proficiency test and the mathematics assessment, but may choose whether to take the English language arts assessment. Because English language learners exit out of programs after they have gained proficiency, and new ELLs coming in perform at lower levels by definition, it is difficult to show annual progress for this subgroup. Therefore, another regulation change was put into place to allow ELLs to be included as a subgroup for up to two years after they have exited their programs.

Additional sections of the No Child Left Behind legislation that have been significant for English language learners are the Title III requirements mandating yearly progress reports on English language proficiency as well as on academic content for ELL students served under Title III. Two major consequences at the state level appear to be:

1. the development of statewide English language learner standards and assessments measuring language proficiency; and
2. the requirement to measure academic content progress longitudinally for ELLs as they gain proficiency in English.

Specifically, the Title III requirement mandating the establishment of standards and the reporting of language proficiency by state has prompted most states to take this opportunity to upgrade how English language proficiency is defined and measured on large-scale tests. While English proficiency tests are not the subject of this book, some issues related to the assessment of English proficiency will be discussed briefly below because such tests can have an impact on the development of achievement tests for ELLs.
Further, the legislation makes explicit the link between Title III services focusing, to a large extent, on the development of English language proficiency, and academic content that may or may not be delivered by Title III personnel in states and districts. This thrusts ELL students and their issues squarely in the middle of the considerations associated with mainstream state and district curriculum and accountability offices. In holding states and schools accountable for tandem progress in language proficiency and subject area achievement, the legislation moves the conversation for ELL educators beyond the most appropriate types of programs for teaching English to a broader discussion about increasing academic achievement across subject areas. Collaboration among curriculum, assessment, accountability, and ELL/Title III offices is required to design tests to enable ELLs to demonstrate what they know academically as they concomitantly progress in learning English. As such, the requirement impacts how these understandings are translated into the construction and delivery of tests that measure academic constructs as ELLs progress through school.

**ELLs: A Heterogeneous Population**

As noted in the previous section, demographic and policy issues are driving efforts to improve instruction and assessment of English language learners. ELL students are increasing in number, and more states and school districts are facing the challenges of helping them reach high academic standards. In 2001–02, nearly 2,000 school districts were estimated to have at least 100 Spanish-speaking English language learners, and more than 150 school districts were estimated to have at least 100 Vietnamese English language learners. Approximately three-quarters of the English language learners in these school districts were estimated to come from families with incomes low enough to entitle them to free or reduced-price lunches (Hopstock and Stephenson, 2003). These numbers and general facts, however, tell only part of the story.

Certainly, language and poverty influence the educational experiences of children. In addition, the diverse backgrounds of English language learners, their parents, and their communities provide challenges to educators and test developers, especially to the extent that those backgrounds are very different from the majority U.S. culture. Most would agree that assessments in U.S. schools are a reflection of this majority culture or culture that is “in power”—college-educated, western European and Judeo-Christian, where independence, risk-taking and creative thinking is valued within a basic organizational framework steeped in logic and Darwinism. Language differences are part of what make some of the
experiences of diverse peoples distinct, but values, priorities, traditions, and opportunities explain the differences more completely (Marin, 1992; De La Rosa, 2002).

The cultures of students in U.S. schools, to the extent that they are dissimilar from the majority culture, provide rich context that can help test developers understand how to evaluate academic achievement for this heterogeneous population. The following section will summarize some of the cultural considerations, including language, that contribute to the identities of English language learners and illustrate some of the key issues related to the proper assessment of each student.

Culture and Learning

Anthropologists tell us that culture is an extremely broad and complex concept. In 1952, anthropologists Alfred Kroeber and Clyde Kluckhohn provided more than 200 definitions of culture (Kroeber and Kluckhohn, 1952). Today, the use of the term culture is complicated by globalization, whereby the patterns of that complex whole are blurred by an accelerating pace of change brought on by the rapid transmission of people, capital, goods, and ideas across political boundaries (Mintz, 2000). In addition, the popularization of the word culture further obfuscates its meaning and in the process simplifies it “. . . to describe just about everything” (Mintz, 2000).

The challenge of defining and recognizing the impact of culture in the classroom has also been addressed in educational research (Eisenhart, 2001). The difficulty with addressing the cultures of students in the classroom stems from vague and porous cultural boundaries. As students engage more and more in the complex network of information brought on by technology and migration, cultural influences are less defined by their households and communities than in the past. Eisenhart writes that today:

everyday life, including life in schools, seems to be faster paced, more diverse, more complicated, more entangled than before. The kinds of personal and social relationships, exchanges, and networks we participate in seem to be taking new forms, tying together otherwise disparate people, and demanding some new ways of thinking about what to research and how to do it (2001, p. 24).

It is more useful to discuss the concepts that define culture than to state a specific definition of culture that, because of its complexity, would either be too broad or too specific to use at a practical level. Incorporating culture in education involves understanding the student and the world around her. This understanding involves the interface between anthropological and
psychological theories. While anthropology focuses on how culture and society work, psychology focuses on how individuals perceive, process, and interact with culture and society.

The anthropological and sociological theories of Pierre Bourdieu provide a framework for understanding culture in education by specifically addressing the bridge between the structures of society and the behavior of individuals. Bourdieu’s perspective accounts for the structural nature of culture and society but also emphasizes the ability of individual acts to affect the structures of culture and society (Bourdieu and Passeron, 1990; Bourdieu, 2003). Individual people are considered to have agency, that is, the capacity for exerting power, and with this agency, they reproduce the learned structures and patterns that make up culture. Because people reproduce, but do not necessarily repeat, what they have learned, structures are open to change. So culture is reproduced as it is passed from generation to generation, but it is also changeable.

The learning theories of Reuven Feuerstein, an Israeli psychologist, may also be brought to bear on the question of culture in education. Feuerstein’s ideas were significantly shaped by his work with a culturally diverse group of immigrant children with widely disparate educational backgrounds from Asia, Africa, and Europe. His job involved testing these children to place them in appropriate educational settings in Israel. Feuerstein’s fundamental concept in learning is the mediated learning experience (Feuerstein, 1980). Mediated learning experience refers to the idea that children learn to understand and interpret the world via the mediation of their family and local community. This concept explains one of the crucial aspects of culture—that culture is reproduced at the level of the individual. The concept of mediated learning experience rests on the intergenerational link that generates cultural continuity. The mediating persons in a child’s life organize and elaborate the surrounding world—a process that facilitates the child’s ability to use these learned mental processes as she grows. People learn culture and are the reproducers of culture.

It is through education that much of culture is transmitted, making education a cultural mechanism (Bourdieu, 2003). The breadth and complexity of the concept of culture contributes to the difficulty of using it in a methodologically sound manner in the classroom and in educational research. Cultural codes include subtle traits and ways of communicating that often go unrecognized by people of other cultures (italics mine). The Soviet-trained psychologist Lev Semyonovich Vygotsky recognized the dynamic quality of individual development as well as society and culture and contributed to educational research the understanding that culture influences what and how students learn (Vygotsky, 1978). Vygotsky’s perspective views
culture and society as driving forces in individual perception and understanding. Psychological tools such as literacy link individuals to the broader spheres of society and culture.

Many studies that investigate immigrant groups and American-born students from minority cultures have shown clearly that culture influences what children learn and when and how they learn it (e.g., Heath, 1982, 1997; Tharp 1994; Tobin et al., 1989; Wang et al., 2002; Ward 1971; Weisner 2001). With regard to education, there are two significant cultural distinctions: that of the home and that of the school. Home culture refers to the culture of the family and local community or community of origin. The dominant culture of U.S. schools is that of mostly Anglo, middle- and upper classes.

Students experience the home culture before beginning school. In the home, much of what children learn of culture comes from their mothers. For biological as well as social reasons, mothers are commonly the primary cultural mediators of children. This process is apparent in families whose parents are culturally distinct. In the Sierra Tarahumara region of northern Mexico, the pattern in families with Tarahumara mothers and mestizo fathers is that the children are culturally Tarahumara. They speak Tarahumara and adhere to Tarahumara cultural norms. In contrast, the pattern in families with mestizo mothers and Tarahumara fathers is that the children are culturally mestizo and do not speak Tarahumara nor adhere to Tarahumara cultural norms (Merrill, 2004). Reay’s (1998) study of mothers’ involvement in their children’s primary schooling addresses those characteristics of culture stated above. Mothers tend to be the primary supporters of their children’s education. Often, they become the connection between home and school through their involvement in advocating for their child’s needs, overseeing homework, and sometimes participating in parent–teacher organizations. In supporting their children, women draw on their own knowledge of, and experiences in, education. Because schooling changes through time as social values change, and immigrant children attend different schools from their mothers, the culture experienced by primary school students becomes a combination of what they learn from their mothers and what they learn in school—a process that results in continuities and discontinuities and reflects the dynamic nature of culture.

The influence of the home culture in general weakens but continues throughout the students’ schooling. Students whose home culture is more similar to that of school will have a pedagogic advantage over those students whose home culture is different (Grenfell and James et al., 1998). Studies have shown how students with cultural backgrounds that are distinct from
their teachers are placed at a disadvantage. Many of these have focused on cultural differences within the United States, particularly invoking American Indian and African American cultures. In their study of American Indian education, Erickson and Mohatt (1982) found that some teachers identified students as poor learners based on behaviors such as avoiding direct commands and deflecting attention from themselves, when in fact these behaviors and attitudes are culturally appropriate for them. Yazzie (2000) identified the problems of unrecognized cultural traits in teaching American Indian students, and Delpit (1988) found similar issues between white teachers and black students. In reality some of the students were not comprehending the cultural codes or implicit meanings of teachers’ behavior and language.

Numerous contrasts characterize learning across cultures, and many of these have implications for assessment. Major learning differences among cultures are found in means of communication, for example, particularly in the child and adult interaction (see Ochs and Schieffelin, 1994). Learning language involves more than acquiring vocabulary and understanding grammar, it involves incorporating a worldview. An appreciation for this larger impact of first language learning has measurement implications far beyond the evaluation of language or language arts *per se*. For example, in an ethnographic comparison involving white middle class mothers and children from the United States and mothers and children from Papua New Guinea and Samoa, Ochs and Schieffelin (1994) show how the patterns of early language use are consistent with the cultural patterns of adult language. Most white middle class language interaction is between two individuals, and communication with babies follows this pattern, focusing on the primary caregiver and child. In contrast, Kaluli mothers in Papua New Guinea construct language interactions that involve more than two people and are often focused away from the mother—a pattern that becomes very important in adult interactions where status and rank are reflected in language use and social contexts. As the researchers note, “What caregivers say and how they interact with young children are motivated in part by concerns and beliefs held by many members of the local community” (Ochs and Schieffelin, 1994, p. 496).

In her 1971 study, Martha Ward shows how the children of Rosepoint, Louisiana, a community that identifies mainly with the black ex-slave culture, communicate in an entirely different way with adults than what is expected of children in typical U.S. classrooms. Heath (1982) compares the mother–child discourse patterns between white middle-class teachers and black working-class women. She found several important differences in how questions are used in the two cultures. Wang et al. (2002) study shows the difference between Chinese and American Indian cultures with
regard to literacy. Chinese mothers focused on print-based literacy and supported their children’s learning in explicit, event-specific, and elaborative ways. In contrast, American Indian mothers focused on oral narratives and personal experiences and supported their children’s learning in implicit and contextual ways. In her work with Native American students, Swisher (1991) has shown how children in some cultures learn by watching, without the constant verbal instruction and interaction with the person from whom they are learning typical of U.S. classrooms. Swisher and Deyhle (1989) report that common cultural learning patterns of American Indian and Alaska native students include a visual approach to tasks, preference for learning by careful observation before attempting performance, and preference for learning experientially in natural settings. Swisher emphasizes that these differences impact how some children demonstrate progress in learning.

Students’ learning orientation is culturally influenced as well. Is the individual or the group the focus? Trumbull et al. (2000) summarize this difference, noting that the focus on the individual characterizes U.S. middle-class culture, which encourages independence and individual achievement, self-expression, and personal choice. In contrast, a group orientation encourages interdependence, group success, and agreement among its members. Some children experience the contrast between the cultures of home and school, where the home culture is group focused (e.g., Ward, 1971); in other cases, both individual and group orientations may comprise culture, depending on the situation. Swisher and Deyhle (1989) found that Native American children they studied were individually focused and competitive in sports, but group focused in school. In an investigation about why Chinese American students outperformed their African American counterparts in mathematics courses, Fullilove and Treisman (1990) observed that part of the answer lay in the fact that Chinese American students worked cooperatively, exchanging information and correcting each other’s homework. In contrast, the African American students in the study worked independently and did not benefit from input and exchange from other students.

As mentioned previously, today’s cultural influences on students stem not only from home and school but from a wide sphere of linkages. Pitman et al. (1989) point out that “…the major forces ‘shaping’ children and young people in the process of culture acquisition are the same as those that shape or direct all learners, namely, the structures and processes of the entire socio-cultural life going on around them”. The extent of cultural sources widens as children get older, spending a greater amount of time away from home and school environments. Eisenhart (2001) points to Shirley Brice Heath’s comments on this subject in the 1996 Epilogue to
Ways With Words. She compares the black communities from the 1970s with the 1990s:

Present day households and communities of children and youth lack the easily described boundaries of their parents . . . In many households, weeks go by when no two members of a household eat at the same time, share any chore, or plan work together . . . Youngest children are in daycare centers. School-aged children go inside friends’ houses to watch television or play video games; they crowd into the vans of community youth athletic leagues and move from sport to sport by season . . . Older youth . . . race to their cars . . . and head for fast-food restaurants or malls . . . On they go, they listen to car radios or wear headphones and throb to the muffled beat of their compact discs or cassettes (pp. 370–372).

That the larger youth culture is an important influence on today’s students is further apparent in studies of acculturation involving immigrants and first- and second-generation students. For second-generation immigrant adolescents (e.g., those born in the United States of immigrant parents), greater English language proficiency predicted a stronger American identity; for the first-generation immigrant adolescents (e.g., those who relocated to the United States from another country), greater ethnic language proficiency predicted stronger ethnic identity (Phinney et al., 2001). Several studies have found that maintaining ties to their original culture appears to enhance immigrant adolescents’ academic achievement (e.g., Fuligni, 1997; Buriel et al., 1983).

Culture and Assessment

Culture, learning, and testing are inextricably linked (Cole, 1999). Culture plays a role in the development of tests and also influences how children understand tests. In discussing the need for culturally based measures of cognition, Cole emphasizes that care must be given to every aspect of assessing cognitive learning for results to be reasonable indicators of student knowledge and skills. This includes the integration of cultural considerations throughout test development as well as investigation into how testing is experienced by students.

Developers of large-scale achievement tests have their own cultural expectations and values and tend to assume students have grown up in the United States, share a common U.S. experience, and assume a common U.S. value system (Kopriva, 2000). However, evidence does not necessarily bear these assumptions out. For example, arbitrary large-scale academic testing, disassociated from classroom learning and occurring once or twice
a year, is not customary in many American Indian communities (Swisher and Deyhle, 1992). Studies have shown that within their communities, American Indian children do not expect to be tested until they have mastered the task (Swisher and Deyhle, 1992). Further, while in some cultures, adults structure discourse with children in a question-and-answer format similar to the structure of a written test, in other cultures, discourse occurs within quite a different format, creating a hurdle for some students unaccustomed to U.S. mainstream ways of testing. As Ward (1971) points out, in Rosepoint, African American parents usually do not ask questions of their children when the answers are obvious. Questioning children in many American Indian cultures is also incongruous with mainstream U.S. learning and discourse patterns (Swisher and Deyhle, 1992).

In a comparative study of cultural groups from Micronesia, central Washington, and Alaska, Sexton and Solano-Flores (2002) showed how different cultural backgrounds affected how students understood and solved math and science problems. In this study, researchers categorized students’ problem solving strategies for each question according to whether they drew primarily on the academic information given in the question, their everyday life experiences, or other formal school learning. Although most students interpreted items using the information given in the item (general academic skills), Latino students from Washington relied more on home culture experiences than the other two groups. Winter et al. (2006) found that how students interpret what test questions ask them to do, whether they have appropriate resources to solve the problem, and the extent to which the test and item structure allow them to demonstrate what they know all influence their ability to access math questions. Solano-Flores et al. (2001, 2003) demonstrated that language and dialect in items impact the test performance of Chinese, Haitian-Creole, and Spanish English language learners, and that this performance is inconsistent across languages and dialects. One language or dialect (English vs. the home language or home language dialect vs. home language standard) is not always better for understanding the intent of the item or solving the problem.

Issues related to culture and language are important considerations both during test development and in the evaluation and scoring of responses. Kopriva and Lowrey (1994) found that distractor choices in multiple-choice items differed significantly between native English speakers and English language learners. Kopriva and Lara (1997) illustrated how constructed responses from English language learners can be misinterpreted by scorers who are native English speakers, but their interpretations can improve if scorers are trained to recognize key areas of confusion. Kopriva and Saez (1997) and Kopriva and Sexton (1999) illustrated many of these areas associated with scoring both mathematics and science items.
Several researchers suggest that test developers, as well as educators, need to address culture, regardless of whether students are English speakers (e.g., Swisher and Deyhle, 1992; Heath, 1983). When assessing English language learners, the issue of culture is more apparent; however, it can be equally complex to address cultural differences of non-mainstream, English-speaking students.

Speakers of different languages automatically signal exposure to an entirely different culture. It is important to keep in mind, however, that while language is a part of culture, a shared language does not mean a shared a culture. Language is culture’s systematic means of communication. Language is learned, transmitted from generation to generation, and, as an integral part of culture, changes through time. For example, the Spanish speakers of North America, Mesoamerica, South America, and Spain can understand one another because one of the languages they speak originated in Spain and spread through the processes of history. However, these people compose a tapestry of distinct cultures. A teacher or test administrator cannot assume that because a student speaks Spanish and emigrated from Mexico, she is culturally mainstream Mexican. In fact, Spanish is likely the second language the student has learned, with the first one being one of the many indigenous languages spoken in Mexico.

Attitudes and values toward testing often reflect cultural divisions as well. For example, American Indian parents and students often regard tests as unimportant, an attitude derived from the effects of taking tests developed in mainstream U.S. culture (Chavers and Locke, 1989). Some immigrant students may not understand the high value that U.S. schools place on testing. Overall, test results become invalid if students do not honestly attempt to show what they know (Trumbull and Nelson-Barber, 1995). On the other hand, for some parents and children from some cultures, formal testing occurs only when passing scores are necessary for students to be allowed to progress in their education. The pressure the students and parents associate with testing in their own cultures is not consistent with many of the purposes of large-scale testing in the U.S. As such, Monroe (2004) argues that these expectations frequently have unintended side effects, particularly for young ELL students whose exposure to U.S. culture is limited, and students who are recent immigrants.

So, how to proceed? While teachers and educational researchers can know what is typical of a culture, they must be careful not to reinforce stereotypes based on group membership (Eisenhart, 2001). The ability of students with foreign cultural backgrounds to understand the cultural codes in U.S. schools varies at the outset and is likely to increase over time. Considering the combination of issues that affect a student’s understanding of U.S. culture, such a student may fall into one of the following four broad categories:
1. Students who have arrived in the United States recently but have had good exposure to U.S. culture through mass communications, Internet, and previous travel to the United States. World youth cultures are a phenomenon brought on by the effects of globalization in which geographically distant students share a cultural affiliation.

2. Students who were born in the United States or who have lived there for years may also have a good understanding of U.S. culture because they have been exposed to it for a long time. They will have at least two cultural identities generated from the cultures of their home and those outside.

3. Students who have just arrived in the country and have had little prior contact with the influences of U.S. culture.

4. Students who may have been in the United States for years or were even born there but have been influenced mostly by their home culture, which is the same as that of their community.

Regardless of where a student stands in relation to U.S. culture upon arrival, he or she will begin to acculturate, a process that involves changes that take place as a result of continuous and direct contact between individuals having different cultural origins (Berry et al., 1986). Two dominant models—unidimensional and bidimensional—have emerged to explain this process. The unidimensional model uses the cultural continuum as the measure of cultural change. The assumption underlying this model is that immersion in the mainstream culture takes place along with the diminishing of an individual’s identity with the original culture and vice versa. In other words, it is assumed that adaptation to the host culture is inevitably accompanied by a weakening of ties to one’s culture of origin (Sue et al., 1998). According to proponents of the bidimensional model, however, the original and mainstream cultural identities are relatively independent of one another. Individuals can adapt to and incorporate American customs, beliefs, and values while also maintaining a strong identification with their native culture (Berry et al., 1986; Laroche et al., 1998).

This discussion is not an exhaustive description of how cultures differ and how children from different cultures are oriented toward learning or how they may be assessed appropriately. Instead, it is intended to show that culture does involve patterning in behaviors, values, and expectations, and these considerations impact learning and evaluation of learning in U.S. schools. The influence of culture from home lessens as students age, as their interactions with individuals from varied backgrounds increases over time, and as the influence of media, chain stores, and restaurants grows. These external factors are significant for both English language learners
and native-born students. However, it appears that students whose community is more distinct from the majority culture or who recently came from more distinct cultures are more significantly affected by the home culture and have different relationships with other influences and cultures than their peers from cultures more similar to the mainstream one.

While distinctions such as cultural distance or extent of acculturation over time certainly oversimplify the interaction between culture and education, they are reasonable starting points from which to begin to understand how cultural considerations may be addressed in academic assessments. The remainder of this book will reflect many of the understandings and work to date associated with improving how large-scale assessments can address the diverse views of this heterogeneous student population.

The Educational Assessment of English Language Learners

As noted earlier in this chapter, the one area in which English language learners have traditionally been assessed in U.S. schools is in English language proficiency. When ELLs enter school systems, they are generally tested individually to determine their skills in reading, writing, speaking, and listening in English. Depending on whether bilingual instruction is available, they may also be tested in their home language. Among the commonly used English language proficiency instruments have been the IPT, LAS, and others. Recently, a new set of tests have been developed which more fully measure academic language proficiency as well as basic English skills. Results are used to determine student placement in programs and their eligibility for various language support services. After the initial placement, tests of English proficiency may continue to be used diagnostically at least annually to provide information about student language skills and competencies, as well as needs and areas of difficulty. Ongoing monitoring of student language proficiency helps support placement decisions and forms the basis for determining when a student is ready to exit from special programs. Annual testing, specified under No Child Left Behind Title III legislation, requires states to report percentages of students who progress and are exited from English language services.

Like their counterparts who are proficient in English, ELLs are also assessed individually in various content areas such as mathematics, reading, science, and social studies to measure their academic progress. Such assessments may take the form of projects, quizzes, and tests assigned in the classroom, district- or statewide assessments, or other large-scale tests such as the National Assessment of Educational Progress (NAEP). Subsequent test results are used for a variety of instructional and accountability
purposes, including to improve teaching and learning at the classroom level, as well as to evaluate the effectiveness of the instructional program at the school, district, state, or country level. NAEP, administered to a sample of U.S. students by the federal government, provides a snapshot of the quality of state educational systems.

The inclusion of English language learners in large-scale assessments (NAEP began offering accommodations for LEP students on a trial basis in mathematics in 1996) and the reporting of their disaggregated results, as is mandated by the No Child Left Behind legislation, has prompted more focus on what large-scale assessments can actually tell us about ELLs. If an ELL child is unable to read a mathematics test question in large part because of language barriers to understanding, what can we then determine about his or her mathematics ability, areas of strength or weakness, or effectiveness of instructional program? In other words, how can the tests and the specific items within have construct validity?

Ways that content-area tests can be designed, administered, and scored to provide accurate information about the academic achievement of ELLs will be detailed in forthcoming chapters. The next section touches on the testing of English language proficiency before moving to an introduction of issues related to the development and use of large-scale tests as well as testing options for ELLs, including variant forms and accommodations.

**Tests of English Language Proficiency**

It is necessary to evaluate the English language skills of students whose first language is not English in order to place them in appropriate language programs, monitor progress, and guide decisions related to moving from one type of program to another or exiting from language services altogether. While the assessment of English language proficiency is outside the scope of this book, how these skills are measured impacts the measurement of academic achievement to the extent that language proficiency tests provide information about when and where students lack the proficiency in English skills to adequately understand test questions or demonstrate knowledge and skills in content areas. This information is needed to match students to proper testing options. Additionally, for accountability and validity purposes, understanding how English language arts assessments and English language proficiency assessments intersect has implications for the development of both types of tests.

Traditionally, large-scale tests of English language proficiency have focused primarily on beginning literacy (reading and writing or editing) and oral (speaking and listening) language skills (Gottlieb et al., 2003). They
were not built or intended to be used to gauge the full range of a student’s language acquisition. However, educators have become increasingly frustrated that there appears to be a gap between language proficiency as defined by tests used to exit students from special language programs and the language performance required for students to thrive in mainstream academic classrooms. Lily Wong-Fillmore (2002), among others, began to encourage the field to upgrade large-scale language proficiency tests to measure a broader range of more complex skills, particularly those associated with the full range of oral and literacy skills that were assumed in the teaching and learning of academic subject matter. Many states and test publishers have responded to this challenge, and the first generation of these new English language development tests have been built. They should provide a more accurate picture of whether English learners have attained the skills that will enable them to perform successfully in the classroom.

As Cummins (1981) notes, ELLs display two different types of English language proficiency—one in the type of language used for social interaction (which Cummins dubs basic interpersonal communication skills, or BICS), and one in academic English (which Cummins terms Cognitive Academic Language Proficiency (CALP)). He argues that students can have well-developed proficiency in social English before they have developed proficiency in academic English, a process which generally takes several years. While the new tests vary in how they are approaching the measurement of English language proficiency, most are expanding their assessment of academic language proficiency, or the more complex language skills associated with formal educational settings. Bailey and Butler (2002) have defined academic language as the language that is used in the classroom, or other academic contexts, that is tied to learning. Gottlieb et al. (2003) contrasts social language proficiency with academic language proficiency, noting that social proficiency was the primary focus of the previously published K-12 language proficiency tests. She suggests that social language proficiency is associated with language acquisition outside of school and informally in school, is representative of the language of social interaction, and is not grounded in academic standards. Academic language proficiency is associated with language acquisition that, in large part, is tied to schooling, is representative of more formal social as well as academic contexts, is driven by the language of content-based curriculum and instruction, and is grounded in a blending of general language proficiency and academic content standards. In general, assessments are being built to measure the language proficiency range from beginning to more sophisticated in reading, writing, speaking and listening, in such a way that
more proficient is defined by increasing linguistic complexity and use of technical language.

Wong-Fillmore’s vision (2002) is that English language learners exiting English language development services on the basis of performance on the new language proficiency tests should be able to compete linguistically in academic classrooms with their native-born peers and be able to pass the literacy portions of the state’s English/language arts tests at the same rate as native speakers.

Wong-Fillmore suggests, and Figure 2.1 illustrates, that students are English language learners until they reach parity with native speakers in terms of grade-level equivalence in English literacy. Literacy could be measured by English/language arts tests and other school indicators, with the understanding that the assessments in the earlier grades are more literacy bound than the language arts assessments of older students (S. Saez, personal communication, 1999). Under this system, services over the period during which academic English was being acquired would certainly evolve, and formal services might stop altogether for some period, perhaps, except for routine check-in points. Grade-level literacy for ELLs should be reframed to take into account the more complex discourse and other linguistic expectations that come into play as students get older (August et al., 2004). Acknowledging the demands of academic English means that assessment of English language proficiency should not be static; rather, it should develop in sophistication at each level of acquisition as students age.

Figure 2.1 Illustration of suggested relationship between English language proficiency and English literacy, as measured on English language arts assessments.

Source: From Kopriva et al., 2004a.
Making explicit the convergence of English language acquisition and English language arts skills has implications for both the English/language arts standards and their academic assessment. Kopriva et al. (2004a) suggest that it is probably a good idea to explicitly separate grade-level literacy and literature skills expectations within English/language arts standards and assessments so that the standards, measurement domains, and assessments of English language proficiency and English literacy can be explicitly and effectively aligned and linked. Flexibility could also be offered in how literature skills are measured in English language arts tests, in how reading and writing items, as compared with the reading passages, are phrased and presented, and in how students are allowed to demonstrate their responses to items that measure reading.

As noted above, it is critical to measure the full range of English language acquisition, including academic language proficiency, to be able to obtain data to properly include all English language learners in academic tests. Accommodations assignment methods usually use data about English language proficiency, as well as other variables, to determine the best testing options for the full range of English learners (see Chapter 10). It is clear from preliminary studies of the taxonomy (Kopriva et al., 2006b; Koran and Kopriva, 2006) as well as findings from Kopriva et al. (in press) that types of appropriate options change substantially as students become more proficient in English. Further, in Chapter 8 Solano-Flores and Trumbull discuss some issues associated with the limits and strengths of old and recently developed English language proficiency tests and how well these tests inform decisions for academic achievement assessment. Further, data from language acquisition tests are important for the robust evaluation of student achievement and for the defensible inclusion of ELLs for school accountability purposes.

Large-scale Academic Achievement Assessments

Determining how ELLs are performing in content areas at the national or state level is problematic because current assessments have limitations that affect the accuracy and reliability of results. For example, as noted above, if ELL students are not able to decipher the questions they are asked due to the way those questions are presented, they will not be able to demonstrate fully whatever knowledge and skills they do have in the content area, such as mathematics or science, that is being tested. If students are required to convey an answer in writing, and a scoring rubric is heavily weighted toward how a student expresses himself or herself rather than the content of the answer, the true ability of ELLs in a subject area may be underestimated. (Of course there are times, as in writing assessments, when
how students express themselves is what is being assessed.) Given the limitations associated with today’s assessments, it is sometimes incumbent upon educators and policymakers to gather multiple sources of information about English learners in order to make sound judgments about their educational status.

It is also important to note that the overall academic achievement of ELLs has tended to lag behind that of other students. On average, elementary students receiving language services score below their English-speaking counterparts on tests of reading and mathematics (National Center for Educational Statistics, 2003). Gonzalez and Darling-Hammond (1997) reported that 27 percent of immigrant high school students and 19 percent of immigrant middle school students were assigned at least two grade levels below their norms. Bhattacharya (2000) indicated that the low level of proficiency in English is a critical factor in low achievement and school failure. In many cases, reports illustrate that this population is at substantial risk of not completing high school (Heubert and Hauser, 1999; Elmore and Rothman, 1999). The dropout rate for Hispanics, for example, is very high. Based on 2001 data, 27 percent of Hispanics between 16 and 24 in the United States were out of school and did not have a high school diploma or a General Educational Development certificate. (This figure included immigrants who didn’t complete a high school diploma in their home country.) Comparable figures for other groups were much lower: 11 percent for African Americans, 7 percent for whites, and 4 percent for Asians/Pacific Islanders (National Center for Educational Statistics, 2002).

There is encouraging evidence to suggest that improvement on this front is occurring. The ELL population is not monolithic; Portes (1996) reminds readers that there is variability across the population. For example, more detailed, in-group analyses of the performance of English language learners suggest that ELLs born in this country outperform their immigrant counterparts on tests. ELLs who’ve exited programs in some parts of the country, including Chicago, Texas, and New York, score very well on tests of mathematics and reading—in some cases better than native English speakers.

Typically, most large-scale academic testing provided without accommodations favors students who have the following characteristics:

- read American English well and easily with about grade-level expertise;
- express themselves well in written form;
- do not rely heavily on cues other than what is presented in the text of the assessment to access or solve problems;
• have had substantial experience with contexts consistent with mainstream U.S. culture;
• comfortably use the multiple-item per page, multiple-choice format with separate bubble sheet;
• perform well for sustained periods of time;
• have realistic expectations about the purposes and consequences of the test and test results;
• adjust easily and can conform to the various testing schedules and formats required for classroom, district, and state tests.

Many English language learners lack one or more the characteristics associated with test-taking ease. They may struggle with reading and writing in English, for example, or be unfamiliar with U.S. culture, testing formats, or testing purposes. These factors can become barriers to the ability of ELL students to access test items properly, and, to that extent, they can lead to test results with the potential to incorrectly reflect the skills of many in this population. Researchers (for example, see LaCelle-Peterson and Rivera, 1994; Lara and August, 1996; August and Hakuta, 1997) tend to agree on several specific ELL-related factors that appear to affect ELL students’ ability to access unaccommodated large-scale academic testing, including the following:

• English acquisition
• English literacy
• home language literacy
• language of instruction of tested content
• arrival time in U.S. schools
• consistency of schooling in U.S. schooling and in home country
• cultural proximity to U.S. mainstream practices in terms of daily life and schooling experiences
• personal experiences, such as anxiety, frustration, or feeling overwhelmed, associated with standardized testing experiences.

On the other hand, these educators as well as others (see Farr and Trumbull, 1997) also identify student strengths and compensatory strategies that can mitigate language, cultural, schooling, and contextual challenges. All in all, this chapter presents the complexity of the challenge of designing large-scale assessment systems that can properly address the needs of this heterogeneous population. However, the chapter also points to many avenues of intervention, which, when appropriately applied, suggest that the validity of test score
inferences for this population can be improved. This book addresses some of these intervention points.

Endnotes

1. Hispanic refers to people of any race whose origins are from Mexico, Puerto Rico, Cuba, Central American, South America, Spain, or any other Hispanic origin (U.S. Census Bureau 2002b:1). Hispanic refers to the cultural linkages to Spain which began in fifteenth century explorations of the world.

2. Children from families with incomes at or below 130 percent of the poverty level, currently $24,505 for a family of four, are eligible for free meals. Those with family incomes between 130 percent and 185 percent of the poverty level, currently $34,873, qualify for reduced-price meals.

3. Maria Malagon Sprehn contributed substantially to this section.

4. Mestizo refers to people who are of mixed Spanish and indigenous ancestry whose culture is non-indigenous.
The design, use and evaluation of today’s achievement tests are the result of a century of development and debate. Substantial change has occurred in three broad areas: the tests and measurement models crafted to assess student learning, the ideas of how to verify the technical quality of these tests, particularly validity, and the public policy directing the use of achievement tests to evaluate students or schools. These three aspects of achievement testing are intimately intertwined. At various points, the technical aspects of testing appeared to be paramount, advancing procedures designed to evaluate defensibility but limiting the uses to which a test could be put. At other times public policy has driven innovation in test design and changes in the meaning and use of results. Each perspective has also impacted the notion and implementation of equity practices in academic assessments.

A Brief History of Large-scale Achievement Testing
Before World War I, educational testing was a fragmented enterprise, decentralized and the responsibility of individual teachers. College admissions were also fragmented. Individual colleges had separate tests until 1901, when the first College Board tests were administered to standardize college admissions. These tests required applicants to complete essays in specific subject areas.
The first large-scale test was the Army Alpha, an aptitude test created by Robert Yerkes and Carl Brigham to assess the qualifications of army recruits in WWI. Brigham is credited with the creation of the multiple choice item for this test (Carroll, 1990). In the early 1920s, Brigham also administered his own multiple-choice version of the Army Alpha test to Princeton freshmen. As a result, the College Board established a committee headed by Brigham to develop a test—to be called the Scholastic Aptitude Test (SAT)—which was used after 1926 by a wider group of schools for college admissions.

Multiple-choice tests—the so-called “new style” testing—were viewed as a great improvement over essay tests. Multiple-choice tests were valued for their fairness and objectivity because scoring was preset and not subject to the biases and preferences of the scorers. In 1933, IBM machines were used to score achievement tests for the New York State Regents and Providence, Rhode Island public schools. Mechanical scoring made scoring costs considerably lower than “subjectively” scored tests and large-scale achievement testing arose in the 1930s and 1940s to provide information over very broad grade-level content areas. By the 1950s multiple choice tests were firmly established as the preferred achievement test format because of their objectivity and efficiency. The introduction of automatic scoring machines permitted the implementation of “testing programs on a scale that had previously been unthinkable” (Office of Technology Assessment, U.S. Congress, 1991). Overall, remarkable growth of the testing industry during this period is reflected in the increased revenues from sales of commercial tests, which increased from about $40 million in 1960 to about $100 million in 1989. As a consequence an enormous jump in testing ensued and standardized multiple-choice achievement tests became widely used in U.S. schools from the 1950s through the 1980s. Successive waves of test-based reform were all intended to improve teaching and learning, e.g. the minimum competency tests of the 1970s and 1980s. In general, results were reported as percentile ranks and individual student scores were interpreted relative to a national sample or norm group.

In 1988, Cannell’s criticism of norm-referenced achievement test results appeared. Often referred to as the Lake Wobegone Report, Cannell’s observation that all states were reporting results that exceeded the national average provoked controversy and many follow-up studies. The fall 1990 issue of Educational Measurement Issues and Practice was devoted entirely to the discussion. Shepard suggested that “authentic, or performance-based, assessment tasks, if properly developed and implemented, held promise of measuring more complex student knowledge and skills of students than usually possible with the multiple choice item type.”
Widespread state exploration of more complex forms of assessment occurred in the early 1990s. The departure from the multiple-choice-only format was motivated by several concerns in addition to those raised by Cannell. An increasingly urgent press for educational reform coupled with advances in research on learning and cognition sparked interest in a test-driven curriculum and fostered the use of item formats that required students to demonstrate deeper levels of knowledge and skills. Such items often asked students to show their thinking and how they arrived at their answers, and their responses often took the form of written text or other types of open-ended responses. In general, there was dissatisfaction with multiple choice tests because they were viewed as emphasizing recall and basic application, rather than in-depth cognitive skills. Aschbacher (1991) argued that multiple-choice-only tests provided only limited information about the complex thinking and problem solving skills of students and, as such, gave a seriously truncated view of the students’ level of understanding. Further, Aschbacher suggested that these types of tests told teachers, parents and students that the more in-depth but untested skills are not as important. Resnick (1987) noted that recent cognitive science findings challenged the notion of a progression from “basic” skills to “higher order” thinking skills and argued that all students, “even non-English speakers, even the poor,” were entitled to a rich curriculum designed to cultivate thinking skills as well as knowledge of basic facts. In order to foster this curriculum, assessments more suited to the nature of these abilities were required. Some state assessments combined multiple choice items with a wide variety of innovative task formats, including open-ended items that could be answered in multiple ways, extended constructed response items or activities that included multiple steps and might require 15 to 45 minutes to complete. Some focused on revamping testing entirely, by creating portfolios that included a variety of student products.

In a study of state testing practices, Aschbacher (1991) found that, as of 1990, approximately half of the states were implementing, piloting or exploring the use of these new types of item formats, generally known as performance assessments.\(^1\) The national enthusiasm for development of challenging content standards and new approaches to testing seemed to coalesce around Lauren Resnick and the New Standards Project. More than twenty states contributed funds and teacher expertise to the creation of content standards based on research and practitioner knowledge about student learning in language arts, mathematics and science. Teachers from member states collaborated in the development of interesting performance tasks and field tested assessment components and instructional materials. Several states, including California, Vermont, Kentucky, New York, and
others, built on the work of the new Standards Project in the development of their assessment systems, incorporating on-demand performance tasks or investigating the use of portfolio assessments. Additionally, during the early to mid 1990s, several state consortia coordinated by the Council of Chief State School Officers developed constructed response and performance activity item banks in science in social studies which were based on models similar to those being used by New Standards and some states.

The use of performance assessment results was hotly debated. The merits of assessments which relied heavily on complex performance tables were sometimes presented as a trade-off between validity and reliability—more meaningful and authentic assessment tasks despite the challenges of judgment-based scoring. Generally the same expectations for validity and reliability were applied to performance-based assessments as for traditional test formats (Moss et al., 1992; Messick, 1995). Moss et al. (1992) asserted that performance-based assessments, including student portfolios, could generate important system level information in a comprehensive system of accountability to provide “an important supplement to the standardized sorts of performance assessment typically used in the system level to suggest directions for curricular reform” (p. 20). Critics, however, cited several concerns. For example, because performance tasks are memorable and therefore difficult to keep secure, new tasks generally needed to be developed annually. This resulted in increased costs and made cross-year comparisons difficult when some of the more extended tasks or activities were used. Also, variability across tasks presented challenges in scaling and equating. Finally, because complex tasks required more time to complete, the limited number of tasks per student represented limited sampling of the target domains (Mehrens, 1992). A solution addressed by many states was implementation of a matrix design to increase the number of tasks that contribute to a school score. Although individual students did not all complete the same performance task, the matrix model allowed more in-depth coverage of the curriculum at the school level. Gong and Reidy (1996) discussed the tensions between the benefits of performance-based activities and the challenges of using the results for accountability purposes.

Besides the benefits to curriculum and the range of skills which could be measured, another important advantage of the move to complex standards-based assessments was the increased involvement of content specialists in the development of assessments as well as the scoring and interpretation of results. Complex assessment tasks were designed to elicit demonstrations of student understanding of essential curricular content.
and those tasks needed a scoring rubric that described qualitative differences across performance levels. It was the content experts who were best able to articulate the difference between, for instance, a “basic” and a “proficient” understanding of the concept. Over time, measurement and content experts learned how to build good rubrics that applied consistent criteria to different tasks. This advance allowed an assessment in which students could demonstrate understanding of the same construct or skills in different ways. Good rubrics also communicated to teachers and students the central features of the concepts being tested and, so, could be used as effective tools for classroom learning. It was not uncommon for the rubrics to be taught and applied as part of daily lessons.

Because of the psychometric challenges that these new item types presented for large-scale testing, the development activities generated lively discussions between curriculum specialists, psychometricians, and between measurement researchers, some of whom believed it was possible to improve validity while retaining a reasonable level of reliability across students. Development of these types of items and pioneering test formats explicitly started from the concept to be tested and worked back to items and appropriate measurement models. In response to the psychometric challenges, scoring and scaling procedures were developed that supported valid and reliable results, particularly for certain types of tasks (Linn et al., 1995; Wiley, 1997; Young and Yoon, 1998; Young, 1998). The field also progressed in the types of criteria and scoring that were required to warrant comparable results across portfolios and other types of assessment activities. A notable example is the underpinnings for the Advanced Placement Portfolio in Studio Art from the Educational Testing Service, which allowed students to document achievement using a portfolio model that allowed different types of work products for different students (Myford and Mislevy, 1995).

Three issues persisted, however. Some complex items tapped multiple targets, and measurement models based on an assumption of unidimensionality were not always adequate. Second, practitioners discovered the difficulty of producing generalizable inferences among tests when a target skill was found to operate differently across content subdomains. For example, a student who could demonstrate inquiry skills in a biology task may or may not be able to demonstrate the same level of inquiry skills in a chemistry task. Finally, comparability from year to year among selected types of complex performance tasks was sometimes troublesome. Kentucky, for example, was forced to drop “performance events” (lengthy hands-on activities) from its assessment program when results proved to be unstable
from year to year. Still, a few assessments that incorporated a range of open-ended items or tasks made adjustments and survived for a decade or more. These include the California Golden State science exam, the Kentucky writing portfolio, the Vermont mathematics portfolio and the New Standards Reference Exam, which was used as the statewide assessment in two states until 2005.

The use of complex constructed response items and innovative test formats declined or diminished in importance within many state assessment systems in response to concerns regarding the technical qualities of specific assessments, the cost of continuous development and scoring, and the difficulty of developing extended tasks that were comparable from year to year. Recently, the pressure for quick release of assessment results for school accountability decisions has encouraged increased reliance on machine scoreable test formats.

### Development of Measurement Theory Models

Measurement models—the theoretical underpinnings used to evaluate the stability, credibility, and meaningfulness of test structures and results—have changed dramatically in the last 100 years. Increased sophistication of measurement models has been aided by computers which permitted increasingly rapid and accurate completion of complex calculations. In this section the first part will summarize the major focus in educational testing to date, which has been to address the stability and generalizability of test inferences. The credibility of scores could be evaluated by some of the more sophisticated models identified in this part as well. The second part will outline salient advances associated with validating the “truthfulness”, credibility and meaningfulness of the score inferences, by briefly reviewing the history of how some major sets of models have developed. While most measurement models have not been tied to equity, **per se**, it is useful to understand how the technical underpinnings are conceptualized, so that advances which are meant to improve access can address the assumptions and other critical constraints associated with the technical framework of educational testing to date.

**Reliability to Multidimensional Item Response Theory**

The dichotomous item scores of multiple choice tests required only simple indices of reliability. Calculation of a reliability statistic (an estimate of the consistency of results over different test administrations or stability over students with like abilities) was the central method of evaluating the quality of large-scale achievement tests (for example, see Gulliksen, 1987, for a
historical explanation). Spearman (1904, 1910) originated the theory of reliability by assessing the association between two alternative measures of the same thing. This led to the concept of test scores, or observed measurements, as true values plus error of measurement, which is the foundation of classical test theory (CTT). Although the original notion of measurement error in classical test theory came from discrepancies in scores based on parallel test forms, it is usually described today as random variation in scores caused by factors that are unrelated to the content tested. As the theory of reliability developed, internal consistencies (or inconsistencies) among items within a test became an alternate way of conceiving of reliability (and error), as in Kuder and Richardson’s statistic (1937).

Cronbach expanded the early notion of reliability by focusing on multiple sources of error and this resulted in his alpha coefficient (Cronbach, 1951) and ultimately in generalizability theory (Cronbach et al., 1972). Generalizability theory considers both test scores (performances) and the sources of variation (facets) that affect them, including differences among persons and test items. It attempts to explain variations in test performance by partitioning the variations among the facets. With this innovation, Cronbach and his colleagues initiated a change in psychometrics from emphasis on correlations to consideration of variance components and random effects ANOVA. This formulated and made practical multivariate generalizability theory, that, for the first time, allowed analysis and understanding of multiple abilities and their interrelationships and laid out effective ways of estimating universal (i.e. “true”) scores for single and multiple abilities. Generalizability theory uses the same type of data modeled by CTT, but permits analysis of different facets of the test or test situation that may influence the consistency or stability of results. For example, one can examine the influence of different types of tasks, the number of raters, or different administration procedures. Results from these types of procedures could be used to improve the reliability of an existing test or to suggest desirable design elements for a new test.

The constraints imposed by the dichotomous item scores led researchers to question the use of common statistical models for test data which treated all variables as continuous and ignored the discrete nature of item scores. Item Response Theory (IRT) emerged as a solution. Tucker (1946) initially created the item characteristic curve (ICC), which expressed the probability of a correct item response as a function of ability, to adequately account for the dichotomous scores. This concept led to the contemporary way of analyzing item functioning as a more definitive and invariant account of an item’s difficulty and discrimination. Importantly, the invention of
the item characteristic curve shifted the focus of measurement models away from the total test score to the characteristics of individual items.

Lord (1952) fully formalized the concept of the ICC into the “normal ogive” item response model that formally integrated test and item analysis, allowing test performance to be viewed as a summation of item performances in a common parametric framework. Coming from another direction, Rasch (1960) formulated a product-ratio model that led eventually to a logistic alternative to the normal ogive. These models were formally integrated by Birnbaum (1968) using new developments in mathematical statistics. Ultimately, logistic models were extended from dichotomous item responses to the graded score scales of constructed response items (e.g., Wilson, 1992). Within various assumptions, these models allow test and item analyses to be population invariant, yielding generalizations across populations. Further, the procedures make the equating of scores from multiple tests feasible even when the test forms are not formally parallel.

An important limitation to both the CTT and IRT models is that they provide only very limited diagnostic information, because they permit inference based only on the number of right/wrong responses. More complex models have been developed to examine the relationship among responses within a carefully developed set of items designed to tease out the aspects of a concept that the student does not yet grasp. Pellegrino et al. (2001) point out that the recent development of more sophisticated measurement models make it possible to test or evaluate several aspects of student achievement simultaneously, or to consider variation in subsets of test items. These models could provide more detailed diagnostic information but they are seldom used.

Some of the models referred to by Pellegrino et al. included multidimensional IRT procedures. The notion that a single item within a test can measure diverse content has prompted this work in which the logistic model for a single latent trait has been generalized to multiple latent traits. In 1980, the multi-component latent trait model was proposed (Whitely, 1980) to model the components underlying item response processes within an IRT framework. This model denotes the probability of success for a person on an individual item as the product of success probabilities for each of the underlying components. Due to the multiplicative, rather than additive, nature of this conjunctive, non-compensatory model, a deficit in proficiency on any of the components will lead to a smaller probability of getting the item correct. In this model, unless the person parameter is materially larger than the item parameter for all m components, the probability of a correct response to the item will be relatively low. There is another class of multidimensional IRT models which represent the
measurement components in a compensatory, or disjunctive, manner. Reckase (1985) proposed the model in which the undimensional trait level is replaced by a weighted sum of the composite traits. In this configuration a lack of one trait can be made up for by the presence of another. These advances provide insights that may help resolve some traditional problems in test theory and provide tools for scaling and equating tests in new ways. Further, while multidimensional IRT models do not directly address the issue of validity, these types of procedures begin to consider the degree to which student performance is the result of mastery of the target knowledge or skill or the result of other influences. Evaluating these different influences on item scores could be an important indicator of some aspects of validity.

A Changing View of Test Validity

Within the testing community, the concept of validity has changed substantially in the past five decades. In 1950, Gulliksen’s *Theory of Mental Tests* focused primarily on reliability and essentially viewed validity as the correlation between the target test and an external criterion, such as a parallel form. Cronbach and Meehl (1955) transformed the concept of measurement validation, reformulating it as analogous to theory validation as the philosophy of science understands it occurring in scientific fields. This perspective greatly expanded the array of evidence useful in establishing validity, but also attempted to bind it all together using the concept of a “nomological network” of potential empirical relationships as a way of encapsulating the theory of the measurement concept being validated. Textbooks from the mid-1970s defined validity as the degree to which a test or evaluation tool serves the purposes for which it is intended (e.g., Ahmann and Glock, 1975). Three types of validity were generally described: content validity, criterion-related validity, and construct validity, and practitioners were advised to collect data to support the type of validity consistent with the intended use of the test results.

Messick (1989) extended the Cronbach and Meehl framework by de-emphasizing the formal nature of the empirical network and by expanding the kinds of evidence that could be used in validation. Messick rejected reliance on a single kind of validity evidence tied to the test purpose and advanced the view of validity as a unitary concept that requires consideration of content, criterion, construct and consequences, with construct as the unifying force. He speculated that practitioners were reluctant to address the need for construct validity “perhaps because the process of construct validation seems complicated and vague” and because construct validation is viewed as a never-ending process. Messick continues
But just because a process is never-ending does not mean that it should not have a beginning. And a good beginning, at the least, is to attempt to discount plausible rival hypotheses about construct-irrelevant variance in the test. For example, a plausible rival hypothesis for a subject-matter achievement test is that it might, by virtue of its vocabulary level, be in part a reading comprehension test in disguise . . . A variety of correlational, experimental, or logical approaches could be taken to render such rival hypotheses much less plausible (p. 41).

Pellegrino (1988) and others have argued that the traditional approaches to documenting construct validity, such as correlation and factor analysis, are based on trait theory rather than the discoveries of cognitive psychology that have documented the thinking strategies that individuals use in response to a wide variety of items. He suggests that the correlational patterns between test results that we have observed in the past may be less a function of a relationships between the tests themselves and more the result of consistent thinking strategies required across similar types of tests, so that validity might be better defined in terms of models of human thought, particularly its dynamics and limits. Moreover, he argues that instead of examining large numbers of items to determine the cognitive factors involved in a test, we should design the test to include items that systematically reflect the characteristics identified by cognitive science as relevant. Observation of the resulting pattern of performance on the test items could reveal more useful information about student understanding, or misunderstanding, of the construct.

It is important to recognize that, past and present, psychological views of learning and ability have strongly influenced the development of measurement models, the interpretation of assessment results and the validation of the inferences based on those results. It is often the case, however, that these viewpoints are present only as unstated assumptions in discussions of test development and analysis.

Mislevy (2003) summarized the four major schools of thought that have influenced views of testing and validation during the last century. First, trait psychology, which was associated with intelligence and aptitude testing at the start of the twentieth century, presumed that a mental trait or underlying ability is consistently evident across settings or activities and that observations across a variety of situations can provide evidence about a particular trait. Mental measurements focused one or more indicators correlated with competence in a domain. Assessments included multiple observations in the form of a diverse collection of items selected to “spread” students out along a performance distribution. This perspective
Mislevy notes that the psychological perspective shapes validation in terms of the questions that agree with or counter the inference we wish to make. Therefore, rival hypotheses for trait-based assessments seem to focus on the scope of the trait in question: do the items reflect the hypothesized relationships between performance and error or would another set of items suggest a different outcome? The second school of thought, the behaviorist perspective, regards knowledge as the “organized accumulation of stimulus–response associations, which serve as the components of skills.” Measurement focuses on a target behavior from the perspective of the assessor. Success over multiple settings (tasks) is the goal. This is consistent with the notion in testing that small bits of knowledge, such as those represented by multiple-choice items, can be assembled to form a repertoire of complex behaviors. Assessments designed from this school of thought, such as the traditional criterion-referenced tests, estimate the probability of success in a domain to be commensurate with the amount of bits a student has mastered. Items are selected to represent essential knowledge in the domain and inferences at certain levels of percent correct suggest increasingly complex concepts of mastery.

Mislevy identifies the third school as the information processing perspective. This school emphasizes the procedures people use to acquire and strategically use knowledge in order to solve problems within the limits revealed by psychological experiments. The assessment design considers both task features and student performances. Inferences are made in terms of concepts and problem solving strategies rather than indirectly in terms of features of problems as an expert sees them. In this case, “the task a student solves is not the problem as the investigator poses it, but the problem as the student perceives it” (p. 18). Within a domain, tasks can still be grouped by features that are similar from the assessor’s point of view, but the target of inference is the student’s perception of features that make them similar from her point of view. Mislevy suggests that validation would consider an alternative explanation for results such as “might some students be solving, say, purported spatial reasoning tasks using non-spatial strategies?” Fourth, is the socio-cultural school which views knowledge as “conditioned and constrained” by the social context in which it is acquired and employed. Contextualizing assessment exercises decreases the assessor’s control and increases the burden of specifying the salient features of performances and performance situations. Validation questions to be raised and rejected may include “Do the relationships hold for some examinees but not others? Can the shared cultural experiences of low scoring students explain the results?”
As cognitive scientists have developed frameworks for the scientific understanding of learning, there has been increasing emphasis on the small-scale processes by which task performance capabilities are acquired and, consequently, on the fine-grained structures of the specific abilities of which they are constituted. As these processes and structures are elaborated, their combinations better represent the kinds of complex learning that takes place in schools. Individuals construct ever-evolving frameworks that juxtapose current states of ability with new experiences. The interaction of ability patterns with learning experiences results in new learning, e.g., transitions to new ability states. Current large-scale test programs seem ill-equipped to address this view of learning, relying instead on measurement models grounded in a view of achievement as evidence of latent traits, and behaviorist relationships between knowledge of bits of information or skill and the inferences about mastery that they engender. As a consequence, test results represent “averages” of many abilities whose fine structure is not taken into account. In addition, the interrelations which constitute “structure” are correlational, not contingent or prerequisite in nature. This may be the reason that it is so difficult to see most aggregate abilities as anything other than hazy collections with arbitrary boundaries.

In practice, test development procedures, including control of content coverage through the item and test specifications, internal review procedures during development, and the use of standardized administration procedures for all students were traditionally considered to be sufficient to ensure that the test was measuring what was intended and that the same inferences could be made with confidence about all students tested. In the 1990s, Messick’s work, the advances in cognitive science, and the push for other types of test items and formats brought a new focus to how validity might be enhanced and evaluated. This included reconsiderations of what validity means and new expectations of what constitutes adequate validity evidence. Methodology was developed to produce rigorous, independent judgments about the degree of alignment between content standards and tests, and the process became a popular type of validity evidence (for instance, see Hansche, 1998, for an explanation of alignment considerations). The revised Standards for Educational and Psychological Testing (1999) included a substantial redrafting of how validity is conceptualized and guidelines for how it might be demonstrated. In the Standards, validity of score inferences is viewed as a unitary idea under the broad umbrella of construct validity.

The development of a priori validation arguments forms the basis for the evaluation of evidence collected to support the arguments. Mislevy (1994) and others (e.g., Wiley and Haertel, 1995; Shepard, 1990; Popham,
1994; Kopriva, 1999; Rigney and Pettit, 1995) called for more explicit links between what constructs and objectives the items are targeting, the processes that are used to develop, implement, and produce the scores on the assessments, and the evidence which was needed to demonstrate whether the items and tests were functioning as intended. While these researchers agreed on the foundation of validity, their studies focused on different aspects of the measurement process in an effort to provide specific guidance about how score validity might be considered and verified in different situations, and for different purposes or students.

Mislevy et al.’s (2003a) Evidence Centered Design (ECD) provides a blueprint for the operational elements of an assessment, that includes the elements of identifying warrants, specifying procedures throughout test development and making sure that the evidence is generally commensurate with the inferences proposed by the test developers. The researchers maintain that

One cannot simply construct ‘good tasks’ in isolation . . . and hope that someone down the line will figure out ‘how to score it.’ One must design a complex assessment from the very start around the inferences one wants to make, the observations one needs to ground them, the situations that will evoke those observations, and the chain of reasoning that connects them. (p. 2)

ECD describes multiple components and the ways in which they interact. One component is the task model. In addition to describing the materials presented to the student and the desired work products, it also includes features of the task as “a structure for understanding and controlling evidential variation.” The researchers discuss the idea of controlled variation of task features that preserves the target construct, provides comparability, and supports a common inference despite planned differences in the appearance of the task. They say that the “family of tasks” from a specific template are not particularly meant to be interchangeable from a psychometric perspective (presumably where the item parameters remain constant), or necessarily from an access-based perspective. Rather, the model templates understand and control evidential changes and often intentionally manipulate the evidential value and item parameters.

This work forms a foundation for advancing how measurement increases the sensitivity of testing to student learning independent of other influences. It also clarifies and raises the expectations about the kinds of validity evidence considered to be necessary and sufficient.
Fifty Years of Federal Policy Regarding Equity and the Use of Achievement Testing

The Brown v. Board of Education decision in 1954 focused attention on equity in education and debate soon arose regarding racial inequality in testing and the adverse impact of some test-based decisions. The Civil Rights Act of 1964 advanced the federal commitment to equity in education and in 1974 Lau v. Nichols spelled out the educational rights of language minority students. In 1965, Title I of the Elementary and Secondary Education Act (ESEA) initiated federal funding to support educational improvement for economically disadvantaged students. Although education remained the responsibility of each state, federal funds through ESEA were initiated and continue to be distributed as a national commitment to the education of children with disadvantages, including high poverty students, English language learners, students with disabilities and migrant students. Jennings (2000) summarizes the historical shifts in political support for the ESEA and associated changes in funding through the early 1990s when political leaders from both parties called for increased education funding accompanied by a shift in the accountability focus from consideration of “inputs” such as funding to improved student “outcomes” as represented by achievement test results.

ESEA: Full Inclusion, Disaggregated Reporting, and Beyond

The movement to standards-based instruction and assessment in the early 1990s represented a significant paradigm shift for educators and testing professionals. The 1994 reauthorization of the ESEA, known as the Improving America’s Schools Act (IASA), produced profound shifts in both testing policy and practice. First, under Title I of IASA, all states were required to develop a system of challenging academic content standards and assessments with clearly defined performance standards. Moreover, these standards applied to all students. The shift from reliance on norm-referenced tests to criterion-referenced tests changed the logic of test development as well as how test results would be interpreted. Each state was now faced with defining not only the essential content to be tested but also the target performance on the test that represented “how good is good enough?” The challenge of test design became the creation of items that were supposed to represent the content critical for all students rather than simply a set of items that could sort students along a normal curve. Second, state assessments were required to include all students, including students with disabilities and students with limited English proficiency. For ELL students, states were advised that:
The appropriate form of assessment might be assessing the student orally or in writing in his or her native language; providing accommodations such as a bilingual dictionary, extra time, or simplified directions; using an assessment that has been stripped of non-essential language complexity; or administering an English language assessment orally. (U.S. Department of Education, 1999b)

Third, the assessment results were to be disaggregated and reported publicly for the following groups: all students, gender, major racial/ethnic groups, ELL, students with disabilities, economically disadvantaged, and migrant. The publication of disaggregated assessment results was a significant change in testing practice. The requirement was intended to ensure that all students in schools receiving Title I funds received the instruction they needed to reach the achievement standard, including additional support as needed. Public reporting by disaggregated subgroups was the method of holding schools accountable for services provided under Title I.

Despite the changes associated with the adoption of standards-based assessments, the most profound change was undoubtedly the federal requirement of full inclusion. The commercial norm-referenced achievement tests routinely used as measures of student and school accomplishments for more than 50 years did not emphasize the inclusion of most English language learners in the norm group. Since most states relied on these tests as the basis of their assessment program prior to 2001, ELLs were generally exempted from state and district testing. In 1998 all but nine states permitted most ELLs to be exempted from statewide assessments, usually on the basis of the amount of time in the U.S. or length of time in a school program where they learned English (Heubert and Hauser, 1999).

The inclusion of all students, including those that had previously been exempted from testing by state directive, was required in the 2000–01 school year. As the deadline approached and it became evident that few states were implementing this requirement, the U.S. Department of Education released a policy letter that included the following direction.

State policies must guarantee that each LEP student is included in the State assessment system. LEP students are to be provided an individualized determination of the most appropriate language and form of assessment for that student, based on English language proficiency, native language proficiency, language and format of
their current instructional program, or other relevant factors. Whether an LEP student should be tested with the State assessment, the State assessment with accommodations, or (to the extent practicable) a native language assessment will depend on which assessment most validly and reliably measures her or his knowledge and skills. In no instance may a State assess an LEP student against content or performance standards less rigorous or less demanding than the standards applicable to all other students. Accordingly, a blanket State exemption policy for LEP students for Title I purposes, whether permissive or mandatory based on time in U.S. schools or time in English instruction, would not meet the Title I requirements (Cohen, 2000).

However, by January 2001, only eleven states had received full approval for their assessment system as required under IASA. Many of the remaining states were criticized by the U.S. Department of Education for failure to include students with disabilities and limited English proficiency in the assessment system. All in all, the “guidelines and criteria for the inclusion of LEP students in assessment were inconsistent across and within states” (U.S. Department of Education, 1999b). Despite what appeared to be slow progress in the implementation of fully inclusive state assessments from 1994 to 2001, it was within the last two years of this period that virtually all states had changed the statutes and procedures that had previously required or permitted the exemption of LEP students.

The U.S. Department of Education’s peer review of state assessment systems under the 2001 reauthorization of ESEA, No Child Left Behind (NCLB), moved beyond simple participation and reporting requirements to emphasize the use of accommodations or alternate test formats for students with disabilities and for English language learners. The peer review included questions about the validity of scores based on accommodated test administration, including changes to the test format such as translations or use of simplified English. In addition the review considered whether states actually monitor the delivery of accommodations on test day. In the initial review, few states provided such evidence. The accommodations available to English language learners were often the accommodations developed for students with disabilities and the selection of accommodations for an individual student was usually left to the student’s teacher or the person administering the test. Many of the accommodation options were not suitable for ELLs and most states provided little or no training to support selection of appropriate accommodations based on individual
student needs. This raised concerns about whether students were actually receiving the accommodations or other options needed to “level the playing field”, thereby permitting them to demonstrate their reading and math skills independently of a disability or limited English proficiency.

Other Federal Influences

In addition to the ESEA, the Individuals with Disabilities Education Act (IDEA) has also had a direct influence on states’ large-scale achievement testing. IDEA required all states and districts to include all students with disabilities in their testing programs. For most students with disabilities, this was to be accomplished by using accommodations specified by their local Individual Education Plan (IEP) teams that were consistent with what was being used in the classroom. In addition, implementation of alternate assessments for all statewide and district-wide tests were required by July 1, 2001 for those students who were not able to participate in state- and district-wide tests even with accommodations. Most states complied by implementing alternate assessments for their most cognitively disabled students, but the result from these assessments generally represented a less rigorous achievement standard than the general assessment. Subsequent regulations under NCLB have clarified that alternate assessments for a small group of students with disabilities may be less rigorous than the general assessment, but are expected to satisfy professional standards for validity, reliability, accessibility, and fairness. A few states have proposed to extend the idea of an alternate assessment for use with English language learners. In order to meet the requirements of NCLB, such an assessment must cover the same content as the general test and, unlike most alternate assessments developed for students with disabilities, must hold ELL students to the same rigorous achievement standards as the general assessment for the grade in which the student is enrolled.

The National Assessment of Educational Progress (NAEP) is another important federal initiative that has influenced state assessment programs. NAEP was introduced in 1969 to provide a national view of student achievement over time based on a nationally representative sample of students. NAEP, which samples students across the nation in a few grades per content area every few years, has not directly influenced state assessments. Rather, it has primarily served as a model, providing an example of rigorous item types, sophisticated matrix design, and complex statistical modeling and analysis of results. NAEP permitted the exclusion of most ELL students until 1996 and most states followed NAEP’s example. Ironically, an evaluation of NAEP reported that while many students with disabilities
and English language learner students had been excluded from previous NAEP assessments, many of them were actually capable of participating in the assessment (Mazzeo et al., 1996). This was found to be true particularly if certain types of adaptations and accommodations were made available to the students. To prepare for a possible change in policy, the NAEP program pilot tested accommodations in 1996 for special-needs students, including English language learners, and revised its inclusion criteria to make them clearer and promote uniform application across participating states. By 2002, NAEP provided accommodations on all assessments for both English learners and students with disabilities, although some constraints for inclusion still remained.

Two elements of the peer review under NCLB—requiring evidence of validity for scores based on accommodations, and emphasis on the technical quality of alternate test formats—have encouraged state testing programs to expand technical work to document that test results are stable, credible, and meaningful. This includes demonstrating that all students are being held to the same standards, regardless of how they are tested. New approaches and investigations may be needed to demonstrate adequate technical quality for assessments that incorporate variations in test or item format designed to minimize the influence of extraneous factors such as language. These will take time to create or adapt from other literature, and the cost implications of procedural development and additional analyses may be substantial.

Some researchers, including Solano-Flores and Li (2006), Abedi (2006b) and Kopriva and others (Emick and Kopriva, 2006, 2007; Kopriva et al., 2007) argue that the current efforts do not yet reliably translate to equitable and valid outcomes for some English language learners. These researchers maintain that scores for students who can fully access the test items represent information about the targeted constructs, but for students who can not fully access the items, the scores still represent a mix of information about the intended information and information about irrelevant factors unrelated to the targets. Since the bias begins at the item level for these students, it accumulates as items are aggregated to produce a test score and consequently the results represent a distortion that potentially masks where an individual’s content knowledge problems lie (and where they don’t). Some would argue that when test results are used for accountability purposes, this is tolerable because at the school level it results in a cautious application of accountability consequences that favor students. Under NCLB, for example, depressed results for ELLs could result in the opportunity to attend a “better” school or gain tutoring services. Further,
because of the misunderstanding of student needs, there is a very real possibility that depressed scores at the ELL subgroup level leads to an inequitable allocation of school resources and valuable instructional time.

**Building Assessments to be More Accessible for ELLs**

Wiley (1990) distinguishes between the target construct or *intent* of a measurement and the *ancillary* abilities of individual students necessary for a successful task performance. For example, test taking abilities are well known to contribute to performance on multiple-choice tests such as the SAT. Individual differences in such skills constitute part of the “reliable” variation in scores on any test. However, such abilities are not part of the measurement intent of these tests; they detract from validity or, to put it the other way around, they add to the invalidity of a test. In statistical terms, both the intent of measurement and the ancillary abilities are dimensions of measurement. He asserts that ancillary abilities, their distribution over test tasks and their relations—both among themselves and with the characteristics intended to be measured—are critical to an adequate conception of test validity. From this perspective, test validity must be assessed by examining all of the dimensions which contribute to test performance.

Certainly, accessibility is a factor of both reliability and validity. However, it seems that improving the validity of score inferences for ELLs is of particular interest because, without confidence that the scores reflect the intended meaning, consistency in measurement is rather meaningless.

The heart of intended measurement is a clear understanding of what is being measured. This includes explicit statements about targeted knowledge and skills not only at the test level but at the item level as item scores are generally aggregated to produce test scores. As noted above, unlike random error which assumes a cumulative net effect of approximately zero, systematic error over items increasingly adds distortion to test scores. Therefore, when accessibility is a question, it seems prudent to minimize the effect of irrelevant systematic influences by focusing on the components of testing that impact the item level, including item writing, internal test development procedures, implementation, etc. The following section will briefly outline some of the aspects of accessible test-making.

**Conceptual Underpinnings**

In general, today’s item and test performance models focus on how test takers interact with items, *under the assumption* that the interaction is only
or mostly about the targeted content of the item and the student’s knowledge in response to the item’s problem.

Beyond this, the test taker by item interaction seems to be a neglected area of focus. When items or tests are inaccessible for some students, it can be said that the effects of irrelevant and ancillary factors interfere with the measurement of targeted knowledge and skills. In all cases, some amount of irrelevant information always accompanies the communication of any item to a test taker. In most cases, the function of the irrelevant information, such as the fact that the item’s requirements are written on a page in English, is to facilitate the communication of the item’s problem, not inhibit it.

It is suggested that the interference, when it occurs, is actually a function of aspects of the person by item interaction, when irrelevant characteristics in items interact with the academic skills of particular students.

Recent advances in cognition and measurement provide a basis for understanding how students approach, address, integrate and retrieve concepts and skills. Further, research indicates that students move through identified sequences of learning in different ways and at different rates, depending on a multitude of attendant individual and environmental factors. However, much of the work on task processing and learning seems to have been focused on qualities of the tasks and student competence regarding the desired task targets, rather than on identifying differential ancillary characteristics of individual students and how these interact with processing in a variety of tasks. For instance, Embretson and Reise (2000) identify task-specific cognitive sub-processes in tasks and models how students respond to the different sub-processes so that their varying scores are a function of their degree of mastery in the target abilities. Lohman and Bosma (2002) point out that both experimental/cognitive and differential/measurement psychologists frequently array their data in a person by task matrix, and that both sets of psychologists have tended to emphasize the main effects in this matrix. While experimental/cognitive psychologists emphasize differences among tasks/treatments and differential/measurement psychologists emphasize differences among persons, both desire to minimize the interaction between persons and tasks without focusing on the quality of the differential interactions per se.

There is some work that attempts to explore the person/task interactive space and consider what is causing the differences across students. In a major review, Pelligrino et al. (1999) focused on intelligence and aptitude tasks, explaining how the cognitive components approach worked to develop differential models of task performance by exploring components
of performance that varied across individual task takers. The approach assessed performance strategies, executive routines, and how targeted declarative and procedural knowledge interacted with the varying processing capabilities of task takers.

Some researchers explore the targeted and construct irrelevant aspects of the task/construct space while generalizing across students. Snow and Lohman (1993) appear to draw a distinction between component skills and strategy adoption, suggesting the separability of perception, memory, verbal and spatial abilities (as well as strategy) from the targeted construct. Glaser and Baxter (2002), among others, define a content-process domain space for school science within which targeted constructs can be classified and defined in terms of task types. The four quadrants of the space (content—rich to lean and process—constrained to open) provide a context for differentiating targeted construct aspects of task performance from construct irrelevant aspects, if such item characteristic differentiations are specified. Mislevy et al.'s (2003) ECD provides a basic architecture for developing assessments that permit variations while preserving the inferences we wish to draw regarding student competence in the domain. However, in this document and elsewhere, while the researchers have spelled out the key elements which are addressed in assessment development, how target task elements may change, and what these two elements mean for the test development architecture, the work does not focus on how student variations due to irrelevant influences may impact the testing work. This seems to be the key area of focus for accessibility purposes.

Work such as Snow and Lohman, Glaser and Baxter, and Mislevy et al. opens the door for developing more individually tailored items and assessment approaches keyed to measuring the same target abilities while minimizing the effects of the irrelevant factors through active manipulation of their environments. In this case, the focus would be on understanding particularly how items and tests need to be designed and developed to maximize the measurement of targeted knowledge in the person/task interaction space for students with certain challenges. In most cases, the objective is to develop assessments that can defensibly yield common inferences and comparable scores across persons with and without sub-par ancillary challenges. The work also involves understanding the types of evidence that will be needed to make sustainable arguments about comparability and common inferences.

For instance, studying accessibility (and inaccessibility) investigates under what conditions target skills are properly conveyed in the person/item interaction, and when communication about targeted information
becomes systematically contaminated, misunderstood, or distorted. This contamination or distortion is known as systematic error. A detailed focus within the encounter would revolve around how the same assessment tasks differ from individual to individual. A less detailed approach could look more broadly at groups of students sharing certain low skill levels on key variables, and groups of items that share certain active characteristics.

The argument for developing accessible testing systems is based on the essential notion that variations in procedures and materials could be designed to produce common inferences. The grounds for common test inferences are traditionally found in a *procedural argument*: common content in items and a common approach for synthesizing and summarizing items and response data over items. The latter part of this procedural argument required standardized conditions of observation as a key aspect of synthesizing item data. However, based on developments in measurement and cognitive science, one can make an alternative conceptual argument for common inferences. As in the procedural argument, the measurement of common substantive content is important. But rather than requiring standardized conditions of observation, the *conceptual argument* can be built on evidencing appropriate interrelationships between target inferences, the knowledge and skills of interest, the properties of tasks or items designed to elicit the observations, student characteristics that impact testing and items, necessary observations, and the assessment situations where students interact with assessment requests. This approach suggests that data may be collected under alternate conditions. By minimizing the influence of irrelevant input on student performance without adding significant additional sources of error test developers can increase the validity of the task inference without sacrificing reliability. At its crux, Kopriva (1999b) suggests that, when common inferences from a robust conceptual argument are applied to assessment results for which there is sufficient evidence of minimizing systematic error across testing procedures and materials, this should provide the grounds and be the basis for determining the validity and comparability of scores.

Whether access is considered from a detailed or more superficial level, two directions for further work seem to be central. First, there needs to be work completed which is aimed at ameliorating the interference of inaccessible stimuli in order to improve the quality of the students’ performances so they more accurately reflect their targeted knowledge and skills. Second, it seems prudent to develop some “barometers” or other algorithm-based estimates which would calculate the level of access that is available for different students in different testing systems. The more
detailed estimates can also be used to adapt future flexibly built computer-based testing systems in order to maximize the ability of individual students to receive particular items and formats that are aligned to their needs and strengths.

*Developing Accessible Tests*

The following chapters in the book will explain in detail how tests might address the issue of accessibility throughout the processes of item/test development and implementation. As the reader attends to the information in the various chapters it will be important to keep in mind that the person/item access interactions seem to occur primarily at three intervention points:

1. For the interaction between the student’s targeted knowledge and skills and the item’s request for targeted information to successfully begin, the student must have sufficient access to how the meaning and the requirements are conveyed in the task.
2. For the student to initiate and sustain problem solving activities relative to the task requirements, students must be able to access their procedural skills and other ancillary content, and have the tools necessary to implement the activities. (The ancillary content and skills would be considered construct irrelevant and within the domain of access to the extent that they are not part of the target requirements.)
3. For the student to represent their solution to the task requirements, students must be able to access their representation skills commensurate with the representation constraints in the particular task.

Thus, the first intervention point for access could be seen as understanding what the task is asking. Lack of prior instruction, or a confusing graphic that is part of the task, or reading skills are examples of how access might be blocked for some students at this stage. The second point focuses on the students’ access to tools or procedures or language which would help them solve the item problem but which are not a part of the targeted objective. For instance, if the solution requires access to language about wintertime in a cold climate to be able to compose a poem, a student’s inexperience with snow may be the cause of failure to solve the problem. Third is the generation of a response to the item. A student who understands the science but lacks the academic language skill or learning
opportunities to differentiate properly between possible answers in a multiple choice item may not be able to convey his solution properly. All of these access barriers may be classified as characteristics of the task, and should be considered when identifying challenges and strengths for the student. In many cases elements at each of the intervention points can be varied to remove or reduce barriers that prevent the individual who understands the concept from generating a correct response.

It is argued here that two models or templates are at the heart of any test which purports to be access-based. The first is the *Access Specification Package* explained in the next chapter. This set of specifications is broad-based and a macro-indicator, specifying how access should be considered throughout test development. It calls for evidence of access at the item level (including item writing, reviewing, and implementation procedures); at the test form level (for instance, what accommodations are allowed and whether these are sufficient, defensible, and parsimonious set); at the testing implementation level (e.g. are the proper accommodations assigned to the proper students); and at the consumer level (are the scoring and analytic procedures adequate). The model is designed to be used to guide development and also to evaluate the level of access that can be assumed if components are implemented as planned.

Second, is the *Access-Based Item Template*. The template is the nucleus of the *Access Specifications Package*. It is explained in Chapter 5 and is relevant particularly for activities proposed in Chapters 6–8. For comparability purposes, the target and other information required in the template core is of a finer grain-size than those specified by Mislevy and others. This is because it includes an analysis of ancillary characteristics of each item as well as specifying the substantive target. It is argued that this detailed level of specification is required for score equivalence to be defended. The goal was to construct a template with reasonable demands on the item writer, but one that is detailed enough to specify key characteristics of students who would appropriately benefit without providing an unequal advantage for some students.

*Estimating Levels of Access*

The types of multidimensional models of item performance identified by Whitely, Reckase, and others would seem to be useful as a starting point for modeling how access factors interact with one another in an item/testing process for a particular student (Kopriva *et al.*, 2004b). In some ways, access is a compensatory exercise. For instance, a student with a low level of reading skill may be able to compensate for that deficit if the item includes
a visual that makes the meaning of the question clear and the student can decode the visual in a meaningful way. However, when items and tests are not accessible enough, the overall effect seems to be conjunctive as well. In this case, inaccessibility appears to have an inversely related multiplicative effect on target performance, especially when certain non-target abilities are seriously below the threshold level required for the student to properly interact with the tasks. Kopriva et al. (2004b) suggest that, for some students, it is necessary to model score performance as a function of access as well as target ability, and that access factors in this approach need to be modeled to address both conjunctive and compensatory elements.

This work, though preliminary, suggests that for a testing program with certain characteristics it is possible to model overall level of accessibility for subgroups of students, such as students at different language proficiency levels. To date, key student and item/testing characteristics have been identified and judgments have been made about the kinds of relationships between item/test and person variables for various subgroups of students. A preliminary prototype modeled some of the item access-variables, but person variables were not at a sufficient level of detail and test conditions were not considered (Wei et al., 2005). A related study (Hipolito-Delgado and Kopriva, 2006) suggested that ameliorating some of these key testing variables seems to differentiate performance for students with particular needs, and it appears that this work, as well as other research, can be used to advance the model development. Some introductory work has begun on formulating a differential item functioning statistic as well that may be able to minimize a key problem for ELLs in how DIF methods are currently conceived (see Chapter 12 for an explanation of DIF).

For research purposes and for use with future access-based electronic item/test procedural banks, the authors and other researchers have started to formulate what a person/item model might look like. The model would include not only relevant item and person data but also information about what testing conditions would be appropriate for a student with particular needs and strengths. They observed that student and item/condition matches seem to follow a series of procedures and rules that include:

1. the identification of the target and ancillary factors evoked by particular items/conditions across students to form a target and access profile vector for each item/condition combination;
2. the identification of interactive rules within the item profile vector for specified factors across students with different student profiles;
the identification of factors in student profiles across items/conditions to form a vector profile by student;

4 the weighting and interaction of target and ancillary factors within student vectors to provide prioritized matching expectations that differ by student.

Identification of target and non-target factors and interactive rules for persons and items/procedures includes determining the presence or absence of specific components in test components or in students. The model also needs to specify the threshold amounts of each component and weights across components. In part, interactive item rules would both explain when the item requirements or problem solving mechanisms can be accessed by more than one factor or when the presence of one factor could lower the threshold for another factor. For students, some factors may be necessary (although not sufficient), or certain alternative student profiles may apply to specific item configurations.

In general the research approach is to create a target/access profile vector for each student that will be used across items/condition combinations, and an item/conditions target/access vector for each item/condition that explains what factors are evoked in the particular testing situation. In each case, as noted above, the student and item vectors are “customized” by additional rules of compensatory and conjunctive interaction and weighting that differ over students and items, respectively. Once these profiles are identified and the rules programmed by student or item, it appears that the matching of individual students to a task variation or accommodation can be completed electronically. The recently completed computer-based method to assign appropriate accommodations to particular students has successfully tested out some of these objectives (for instance, see Kopriva et al., 2006b, and Chapter 10 for more information). It is likely that this system could be used to move toward such a person-specific access model. One application may be that, when items are delivered electronically, algorithms can be built to identify the best match among like items and testing procedures can be found for each student for each identified targeted construct element that a test is intended to measure.

This chapter represents a great breadth of material about forces which helped influence what today’s tests look like. Over the last 100 years, changes in the specifications of valued academic content, theories of learning and cognitive acquisition, consumer appetite, and public policy, have modified and constrained the content, format and purposes of student achievement tests. Over time, technical models of measurement were developed
and adapted from other literature to specify how assessments might be structured, and what components were needed or assumed in order to defend scores based on concepts of stability, consistency, generalizability, “true” mastery, and meaningfulness. However, only recently has there been a more widespread acknowledgement that ancillary components as well as targeted objectives impact the scores of test takers. Language, culture and school context all present challenges to ELLs as they face a typical achievement test, forming a web of small blockages to access. Past a point, this interference seems to be impervious to traditional equity approaches, such as readability and bias reviews, and to easy post-development accommodation attempts.

It will be a substantive challenge to integrate current research on learning, emerging concepts about test design, advances in multidimensional models, and work that specifies components of tests and students, with a growing understanding of how test takers interrelate with academic testing. This is particularly the case for English language learners who have been regularly excluded from large-scale academic testing until just recently, and whose challenges (and strengths) are complex and multi-faceted. Given the technological advances, and groundwork in understanding some key elements of student/testing interactions for this population, the rest of the book is designed to explain how this challenge might be advanced.

Endnotes

1. “Performance assessment” typically referred to the use of an open testing format (such as portfolios) or incorporation of some type of open-ended items or prompts on a standardized achievement test that requires a student product generated in response to a test question or task. Products from the open testing format or constructed responses on a standardized exam are subsequently evaluated by means of a scoring criterion consistently applied.

2. English language learner (ELL) is the preferred term to indicate a student who is not yet fluent in English; however, federal statutes and publications employ the term limited English proficient (LEP) to refer to these students. These terms are being used interchangeably in the book. For ease of reading, ELL will be used here although the statutes use LEP.
Building a large-scale achievement assessment system that properly recognizes the diverse population that will be taking the tests requires a complex, comprehensive approach. Issues such as for whom the test is intended, what groups will participate during the development of the assessment, and how they will be included, should be an integral part of the test plan. Further, it will be important to provide evidence that reasonable levels of alignment have been achieved among the assessment system, the content standards that drive the assessment, the performance levels that codify the levels of achievement associated with various scores on the assessment, and the enacted curriculum with which the students engage. This requires the identification of alignment procedures and alignment targets, followed by the development of test specifications or blueprints outlining what will be tested and how. Test specifications provide the construct framework for operationalizing the test design through subsequent item development.

Considerations relevant to English language learners (ELLs) need to be addressed at each of these points in order to construct a system that is appropriate for this population. This chapter focuses on some of the relevant points. In particular, plans for how to include ELL experts and students will be summarized. This section is followed by a discussion about the development and use of an access test specifications package,
which would be built during planning, and utilized as the test is developed. Finally, some considerations associated with evaluating alignment are introduced.

The field’s measurement standards (AERA/APA/NCME, 1999) summarize how the testing discipline interprets its responsibilities to appropriately measure the targeted content in K-12 test takers. As Chapter 3 outlined, federal legislation and many state and district policies have driven the field to develop testing mechanisms which assess academic learning of a broader range of their student population than was true in the past. Further, in order to hold all students to the same standards of performance and be able to place them confidently on a common performance scale, it is advisable to build these large-scale assessment systems that yield comparable scores across virtually all English language learners (Haertel and Wiley, 2003; Kopriva, 1999b).

Participation in Test Development

To increase the likelihood that a coordinated assessment system represents that achievement of all children with comparable accuracy, English language learner involvement should be represented during all stages of test development. ELL experts need to be included during test specification and item development, as well as during reviews. A full range of ELL students needs to be involved in item pilots and field tests.

Expert Participation in Test Development

Typically, experts with substantive knowledge of English language learners have been used primarily in bias reviews, where the charge has been very narrow and incomplete (see Chapter 7 for an explanation). They have not been included in the planning, item development, and decision-making processes to the same extent that mainstream teachers and content experts have been in recent years. This includes involvement throughout the design, construction, and technical phases of development. The tasks in which it would be appropriate for them to actively participate include the following:

- designing the comprehensive testing system;
- developing test specifications;
- writing and reviewing content items and rubrics that are appropriate for ELL population;
- providing training to other item writers and developers;
- trying out items in classes;
- evaluating forms for coverage and accessibility;
• making decisions about inclusion or exclusion of items, all testing materials, and administration/response options based on data from pilots, field tests, and other technical data collections; and
• scoring, reporting, and making decisions about test use for accountability and program evaluation.

English language learner experts who bring the most to the test development process have a deep understanding of content standards, experience with adapting academic teaching environments for these students, and knowledge of their students’ strengths and challenges. Test developers should ask for evidence of these characteristics. Because few experts will have experience with the full range of English language learners, it will be necessary to include a variety of educators who bring diverse capabilities to the entire test development process, from designing the system to scoring, and making accountability decisions. Future chapters will, hopefully, make clear how this range of expertise can be used to make development decisions. Relevant expertise encompasses the following:

• educators from classrooms in which students are learning English as well as grade-level academic content;
• educators from mainstream academic classrooms in which ELLs are placed after they have reached a certain level of English proficiency;
• educators working with students who are newly arrived to the United States;
• educators working in classrooms in which the students’ primary language (also known as their first language or L1) is the language of instruction or in bilingual (L1 and English) classrooms;
• educators with urban experience and educators with rural experience;
• educators working with migrant students; and
• educators who come from the primary language and cultural backgrounds of the students they teach.

Student Participation in Small Sample Pilots and Large Sample Field Tests

Just as the diverse perspectives of ELL experts should be included during test development, a full range of English language learners should also be involved in all item and test data collections. Several researchers suggest that students respond differently based on their proficiency levels and adequate accommodations (e.g., Abedi, 2006a; Rivera and Collum, 2006; Emick and Kopriva, 2007), and so participating students should range from new arrivals through former English language learners that have successfully transitioned. Further, Laitusis et al. (2004) found that ELL
students appear to differ in how they respond to items based on their racial/ethnic background and these distinctions should be examined. As such, it is recommended that enough students from pre-identified language proficiency and racial/ethnic strata be included to be able to analyze the data by these subgroups. Depending on sample size, these analyses can be either qualitative or quantitative, and recent evidence (Emick and Kopriva, 2007) supports that the data should be evaluated by at least grade cluster (elementary, middle, and high school) if not by grade.

In assessments that include constructed response items, ELLs should be purposefully sampled so that their work can be inspected at all item score points. Some researchers have noted that item considerations based on language or culture will sometimes come to light for those with different levels of targeted ability (Abedi et al., 2000; Kopriva and Lowrey, 1994; Kopriva, 2005b).

Test Specifications and Content Validity

Test specifications, sometimes called test blueprints or frameworks, outline the content and constructs each assessment in a testing system will cover. Created for each content area and grade level tested, test specifications prioritize and weight the areas to be covered and generally include information about test length and format, item types, and reporting requirements. Each assessment consists of one or more forms that are intended to be parallel, and each of these may or may not have separate specifications associated with them. The test specifications form the bridge from the content standards to the assessment and provide a framework that identifies what the test will measure and what items will be included in the test. This section discusses the various components of traditional test specifications and discusses the addition of a new component—access—to provide information about the appropriateness of the test for ELL students.

Components of Test Specifications

Test specifications represent a translation from the curricular priorities established in the content standards to a design for how assessment developers will evaluate what students have learned. As discussed above, assessments may cover a large or slim portion of the content standards. Further, they may either match or understate the range of academic complexity present in the standards, including the sophistication and depth of knowledge and skills the students are expected to learn within content topics.

Developing test specifications from content standards is one of the first procedures in building an assessment. Test specifications are used consistently as items are developed and content and complexity coverage is evaluated. Further, once the test is completed, test specifications are one
of the best indicators of content validity, or the determination of what is being covered on the test and how. While specifications vary in “grain size” or specificity within and across topics, they should be precise enough to retain the integrity of the standards on which they are based without turning into highly constrictive objectives.

Test developers use specifications to summarize the content upon which all test takers will be tested. Test specifications form the basis for content area inferences about what students know and don’t know. These inferences are valid to the degree that items properly reflect the standards they are intended to reflect and that test takers are able to understand the nature of the items, solve the problems to the extent they know the content, and explain their solutions. The inferences would be incorrect to the degree that test items distort what is being measured and that other variables influence student scores. Linn (1993) encouraged the assessment community to develop methods for tightening test specifications because they are the linchpin to ensure that important aspects of the standards are properly measured in the assessments and reported to stakeholders. Part of his emphasis was on ensuring that the blueprints specify the levels of complexity that tests include. He also indicated that differential item functioning and other information about performance by subgroups should be taken into account in test development.

Today, topic coverage continues to be one of the major components of test specifications, often including specifications of item types (for example, multiple choice, brief constructed response, essays) and item difficulties around anticipated points such as achievement-level cut points. Cognitive complexity coverage is frequently a part of test specifications as well. While item difficulty and/or item type are sometimes used as poor surrogates for complexity, many publishers and agencies understand that complexity refers to a range of skills, from basic to more sophisticated, as they are associated with content (see NAEP mathematics test specifications, 2001, for example (Solomon et al., 2001)). In the mid-1990s, the Delaware Department of Education developed a mechanism that defined complexity in items, which it called depth, as a function of four variables:

- approximate time required to complete an item;
- item scaffolding (ranging through three levels from step-by-step task guidance to no hints or guiding questions);
- level of generalization (in three levels, from highly specific items to items that require generalization); and
- complexity of process (in three levels, ranging from managing only a limited amount of information to processing or considering multiple pieces of information and/or procedures simultaneously).
In assembling forms and tests, depth balance was guided by the content standards and each item was coded on all four variables. Delaware’s approach was developed from the work of Rigney and Pettit (1995), who suggested criteria for assessing the depth and quality of student work in portfolios that were assembled as large-scale evaluation tools.

While coverage and complexity are routinely covered in test specifications, issues of accessibility are not typically made explicit, which means that test developers and test users do not gain a clear understanding of the extent to which test takers’ scores reflect variables irrelevant to the targeted construct. Test development procedures such as analyzing differential functioning of some items in some subgroups and conducting bias reviews, along with the use of post hoc accommodations, are seen as ways to minimize problems. However, to date there does not seem to be any systematic way of dictating standards of acceptable access, specifying points where access should be displayed, and summarizing that adequate work has been done. For test score inferences to be used with confidence, it seems necessary that proper access to intended item content should be documented for test takers, particularly those with identified challenges or limitations that run counter to the methods typically used in large-scale testing.

Documenting Content Validity for English Language Learners: The Access Specifications Package

English language learners are among the population of students whose challenges often affect the accuracy and validity of their test scores. This book explains various ways to improve the viability of their scores; here, it is argued that summary documents—referred to here as Access Specifications—also need to be put into place to explicitly insist on and monitor proper access to test coverage over the entire breadth of topics and the depth of cognitive complexity. It is recommended that the documents target access at both the item and test levels, providing a planning and evaluation framework for access-driven item writing and test-level accommodation decisions. The documents would also serve as evidence to demonstrate the content validity of test scores. Such access specifications would be used for the following purposes:

• to regulate planning and development in order to ensure acceptable levels of validity in score inferences across diverse students, including ELL and native speakers;
• to serve as a formative evaluation tool from which to make mid-course corrections; and
• to provide evidence to document the level and quality of access across the test and test development and implementation procedures.

At the core of the access specifications, as with other test specifications, would be the same content standards and targeted constructs, clearly specified. The intent of developing and using the access specifications is that the test score inferences for English language learners will be defensible and comparable to the inferences of test scores for other students taking the assessments.

As envisioned here, test developers should complete an Access Specifications Package. This package would comprise four documents: an item matrix, an accommodations summary, a description of the matching procedure for assigning accommodations, and a scoring, analysis, and reporting survey. Each of the four documents would summarize the expectations, criteria, and objectives to be used to guide test development, and then allow for the presentation of summary evidence that the work was completed. The four documents that make up the Access Specifications Package are described below.

ACCESS SPECIFICATIONS PACKAGE

Access-Based Item Matrix  The Access-Based Item Matrix allows each item in a test form to be rated for accessibility according to the following criteria: language simplification, reduced reading load, graphics, format simplification, and other appropriate item modifications (e.g., problem context); use of tools and resources; administration options; response options; and rubric accessibility (for constructed response items only). The matrices are completed by filling out the appropriate information by item, for each form of the test. Evaluators are encouraged to be detailed and specific about their findings and consider the composite effect of various criteria on the accessibility of each item. While it is unrealistic to expect that all items will have graphics, for example, it is reasonable to expect that attention is paid to broadening accessibility in some ways for all items.

Item Matrix documentation would include a description of criteria that would need to be addressed for each item (e.g., what constitutes language simplification or use of tools/resources) with additional criteria for rubrics for scoring constructed response items (for example, anchor papers and scoring notes that illustrate how students might use a diagram). Agencies would prepare an evidence document by reviewing and analyzing each item and filling in the grid as appropriate.

Accommodations Summary  The Accommodations Summary identifies and describes all accommodations available for the test and explains the
characteristics of the students for whom each accommodation would be appropriate. The accommodations could be grouped under broad categories as follows in Table 4.1. Options such as those listed below are described in several of the subsequent chapters.

The accommodations available and used should be checked off on the Accommodations Summary. If possible, the number of students who received each accommodation should be noted. Where substitutions are made, reasons for them should be noted and an attempt should be made to estimate their efficacy.

Description of the Matching Procedure for Assigning Accommodations This document specifies the procedures for determining how individual students are assigned one or more presentation, administration, and response accommodations. It summarizes the research base for the decision-making algorithms that are identified in the procedures, and indicates the information to be gathered for each student. It is recommended that the document include a form or procedure for capturing what accommodations are received by each student at the time of testing, as well as instructions for completing the information at the school site. Some examples of how

<table>
<thead>
<tr>
<th>TABLE 4.1 Selected forms, tools, administration, and response options</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Forms</strong></td>
</tr>
<tr>
<td>• ELL or low language proficiency English form</td>
</tr>
<tr>
<td>• L1 or side-by-side form in the following languages: ________</td>
</tr>
<tr>
<td>• Other: ________________________</td>
</tr>
<tr>
<td><strong>Tools</strong></td>
</tr>
<tr>
<td>• Bilingual word list/electronic translator in the following languages: ________</td>
</tr>
<tr>
<td>• Picture dictionary</td>
</tr>
<tr>
<td>• Manipulatives or other problem solving tools</td>
</tr>
<tr>
<td>• Other: ________________________</td>
</tr>
<tr>
<td><strong>Administration</strong></td>
</tr>
<tr>
<td>• Small group</td>
</tr>
<tr>
<td>• Oral L1 in the following languages: ________________________</td>
</tr>
<tr>
<td>• Oral English in the following languages:</td>
</tr>
<tr>
<td>• Other: ________________________</td>
</tr>
<tr>
<td><strong>Response</strong></td>
</tr>
<tr>
<td>• Written L1 or code-switching; scorers available for the following languages:</td>
</tr>
<tr>
<td>• Oral English</td>
</tr>
<tr>
<td>• Oral L1 or code-switching; scribes available for the following languages:</td>
</tr>
<tr>
<td>• Modeled/demonstrated response</td>
</tr>
<tr>
<td>• Other: ________________________</td>
</tr>
</tbody>
</table>
developers or agencies might assign accommodations to individual students are summarized in Chapter 10.

**Scoring and Analyses Documentation** These documents ensure that issues related to ELLs are considered as scoring materials are developed and scorers are trained. They also should impact how scores for forced-choice and constructed-response items are evaluated, analyzed, and interpreted. Examples of descriptions for scoring criteria such as those listed below in Table 4.2 are presented, and agencies check off the items that apply. This is explained more fully in Chapter 11. A discussion of empirical documentation which is particularly relevant for ELLs occurs in Chapter 12. This includes a validation design, and explicit evidence linking appropriate inferences to the weakening of alternate arguments which could wrongly infer the knowledge and skills of ELL students.

Access specifications are not meant to substitute for additional documentation required throughout the test development and analytic phases, but rather to summarize some of the work that has been done.

**Alignment Considerations**

**Overview**

The movement toward standards-based education has prompted educators to pay more attention to linkages among content and performance standards, curriculum, instruction, and assessment. State and local educational agencies, with input from subject experts and subject-area associations such as the National Council of Teachers of Mathematics, specify the knowledge and skills students are expected to gain at each grade level and the performance level considered acceptable. These content standards provide the framework for what schools should teach and guide the selection of appropriate classroom learning materials and activities.

<table>
<thead>
<tr>
<th>TABLE 4.2 Examples of scoring documentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Prompts (including descriptions of acceptable response modes) are explicit and clear.</td>
</tr>
<tr>
<td>• Scorers understand what is being measured, and what is not, in each constructed response item.</td>
</tr>
<tr>
<td>• ELL experts are included in item and scoring rubric development.</td>
</tr>
<tr>
<td>• Bilingual scorers are trained and available.</td>
</tr>
<tr>
<td>• Anchor papers representing ELL samples at all score points are provided during training.</td>
</tr>
<tr>
<td>• Anchor papers presenting ELL samples at all score points are provided during calibration exercises once the scoring gets under way.</td>
</tr>
<tr>
<td>• ELL issues are covered in training and in item notes.</td>
</tr>
</tbody>
</table>
Learners are assessed, both informally and formally, to evaluate their progress in relation to the specified content standards. The state or district defines what levels of mastery are adequate and, in today’s parlance, these are referred to as achievement standards. They also distinguish levels of performance that approach the “adequate” level as well as one or more levels that surpass it. These levels might be indicated through designations such as Basic, Proficient, and Advanced or Does Not Meet Standards, Partially Meets Standards, Meets Standards, Exceeds Standards. Large-scale assessment results may be used diagnostically, within the classroom, to modify instruction, as well as for program evaluation and accountability purposes.

In the narrowest terms, alignment may be viewed as how well curriculum standards and tests match. When off-the-shelf or proposed custom-built assessments are examined in relation to content standards, the great majority of test items or tasks—usually 80 to 90 percent and sometimes even 100 percent—match up with what is specified in a state or district’s content standards. When the extent of standards coverage is evaluated in relation to assessment coverage, however, it is not uncommon for 50 percent or fewer of the content standards to be measured on an assessment (Hansche, 1998). Another aspect of alignment involves the appropriate emphasis or weighting of content standards. Is the emphasis placed on various topics in the test the same as the intended emphasis in the content standards? Additionally, the measured skills should be consistent with the desired rigor described at each achievement level. Appropriate alignment of content standards to assessments means that the tests and test forms match the emphasis, depth (complexity of the skills being tapped within topics), and breadth (content coverage) of the content standards which are being referred to in the interpretation of the test results.

The issue of how many content standards to assess in any large-scale assessment system is an ongoing debate. On one hand, assessing more standards means that there is more sampling across the testing domain to which the inferences are referring. If the interpretation of test scores is to remain broad (for instance the inference may be a description of how well a particular student or school is doing in mathematics, broadly conceived), sampling is necessary as not all standards, at all levels of complexity, can be assessed on an exam of reasonable length. One consequence of this approach, however, is that breadth tends to outweigh depth in most assessments with the result that the tests are a more superficial evaluation of student knowledge and skills. Popham et al. (2006) have encouraged the education community to test fewer standards but do so in more depth. They argue that this approach would provide more useful information to educators who have to take these results and subsequently develop learning
priorities and relevant pedagogy to implement academic improvements, as necessary, in the schools. While more detail about this debate is outside the scope of this book, it is important to remember that, whichever method is used needs to be accompanied by test score interpretation explanations that are consistent with the approach. This is an essential step. Therefore, alignment needs to be not only the alignment of content standards and the assessment system, but needs to include an alignment of the explanations in the achievement standards and other score interpretations which are included in the testing materials.

Several agencies have developed methodologies to evaluate the alignment of tests to standards and achievement levels, ranging from the cursory to the more thorough (for instance, see www.ccsso.org/projects/surveys_of_enacted_curriculum). Virtually all of these methods focus on the content standards, assessment systems, and on at least some of the interpretations of test scores (for instance, on the published state achievement levels). To date, it doesn’t appear that any of these procedures directly address the validity of the alignment matches in terms of different populations of test takers. The closest integrated review system which includes an evaluation of the alignment of test forms and procedures for different populations (as well as other evaluations of other validity indicators) is the U.S. peer review process (USDE, 2004). This process was first developed and utilized under the previous authorization of the Elementary and Secondary Education Act (ESEA). It has been adapted and used as part of the current authorization as a review of each state’s academic assessment systems for purposes of school accountability (No Child Left Behind Act of 2001, 2002).

Finally, while grade-level standards are intended to drive curriculum and instruction, it cannot automatically be assumed that all children at a certain grade level are exposed to the same curriculum. Some teachers may not cover the expected material at the expected level of depth. If students are working below grade level, teachers may need to spend significant time providing remediation before starting in on the grade-level material. Students may miss instruction when they are “pulled out” for various auxiliary services such as reading or language instruction, speech therapy, or gifted or special education programming. They may also miss a significant amount of school due to illness, family travel, babysitting, or other jobs. They may attend schools in which there are disruptions to learning due to lack of materials, discipline problems, inexperienced or frequently absent teachers, etc. Researchers therefore draw a distinction between the intended curriculum and the enacted curriculum and some work has been done to attempt to identify and quantify what types of enacted curriculum students have received (e.g., Schmidt et al., 2001.)
the mid-1990s California attempted to collect information on enacted curriculum when it was conducting its statewide examinations. The intent was to be able to interpret the achievement results within the context of differing opportunity (Burstein et al., 1994). However, this innovation was cut short when the state re-vamped their testing program. To date, it does not appear that any other state or assessment system have attempted to integrate the evaluation of opportunity into the measurement of student content. This information, along with other information about the students would, however, appear to be helpful in explaining test results and it is recommended that future work be completed to identify how this might be accomplished.

**Alignment and English Language Learners**

If curriculum standards, performance standards, curriculum, and assessment are aligned properly, it appears that the results obtained from the assessment system *should* be able to be used with confidence to evaluate whether programs deliver necessary services effectively, whether educational agencies are educating students properly, and whether all students learn the content specified as important in the state or district’s content standards. Since the test items are assumed to be representative samples of the knowledge and skills identified in the standards, and all students are assumed to have been exposed to grade-level curricula, it follows that the assessment results would be considered a valid and reliable measure of this content. In other words, the results are taken to be generalizable evaluations of student performance on the content standards (or a tested subset of content standards) for all test takers. However, as noted above, rarely do alignment evaluations attempt to determine whether alignment results actually extend to specific subpopulations of students because results tend to reflect the experiences of the majority group or an artificial average across minority subgroups.

The relationship between systematic error and alignment is related to a careful evaluation of the conditions and testing materials for all populations, including ELLs. Specifically, there can be a substantial effect of systematic error over items and over persons for many English language learners, or any population where systematic errors have been documented (Haladyna and Downing, 2004). Systematic errors are errors that do not occur randomly or by chance, but which are sustained over cases and caused by non-targeted elements in the testing system which consistently impacts an individual or subgroup in a particular way. Unlike random error which balances itself out over items or students, small errors of this type will compound when items are aggregated to form test scores or when test scores are aggregated for students who share the same sets of qualities that produce
the systematic error. This can lead to gross misrepresentation about the ability of these students, which, in turn, has implications for school accountability and for placement or other cases where educators use individual student scores.

Several chapters deal with ways to minimize this type of error but there needs to be evaluations and empirical evidence at several levels to determine whether the conditions and testing materials appear to be sufficient and suitable for the range of students who are taking the test. Specifically, in order to determine whether the content standards and test alignment is adequate for English language learners, evaluators need to determine whether adaptations made to the assessment system in terms of flexible tools, response options, administration options, and forms are appropriate to address the students’ needs. Further, the alignment between these forms, tools, and options and their use in the classrooms needs to be evaluated as well as their relationship to the content standards and achievement levels.

Finally, it is important that the achievement levels fairly and comparably relate to the content standards and the assessment options that these students will be receiving, and that there is no language in the achievement levels that might inadvertently bias attainment of any level. For instance, developers or agencies should be careful to allow for responses that have a minimal language load, as appropriate, for some ELLs, and attainment of any level should not assume or necessitate language except when this is the target of the composite score inference. To ensure the use of appropriate descriptors, evaluators looking at alignment should find evidence that the achievement level descriptors do not unduly bar English language learners with language or other challenges from being classified correctly. Levels in earlier systems sometimes required students to demonstrate performance through a restricted set of communication methods not related to the content areas (see Kopriva, 2000, for examples). While test development and implementation interventions should tend to minimize this problem because they allow multiple avenues of access, sometimes the use of multiple pieces in the assessment or accountability system reintroduces the problem.

Procedures for establishing the levels themselves should also be evaluated as well as evidence of due process for non-mainstream students. Specifically, evaluators should have documentation of the technical rigor associated with the classification procedures themselves. What evidence is there to document that the cutoff scores are appropriate for ELLs or that multiple sources have been combined correctly? When test forms, testing tools, and testing procedures appropriate for ELLs are in place, as well as evidence of rigorous matching procedures or appropriate validity documentation for ELLs, how are agencies determining whether the test scores or other pieces
of information used to classify student achievement are representing the content-area mastery of ELL students properly? When incomplete evidence related to these sources is found, in many situations it is recommended that other sources of achievement evidence should be used to support classification, either initially or upon appeal.

Concerns related to the alignment of assessments and content standards are raised and addressed throughout this book. Evaluators looking at alignment need to focus on whether appropriate evidence exists to determine that items and forms are accessible. Is there adequate evidence that the items and forms, test implementation, scoring, and reporting allow English language learners to demonstrate their achievement properly? The implementation of a reasonable number of the points outlined here can be used to provide evidence of alignment between the assessment system and content standards for this population. Such evidence should include the following:

- documentation of procedures related to item and forms development;
- documentation of matching, or how agencies determined which students got which forms and accommodations and why;
- rubrics, rubric notes, and scoring training that demonstrate that the content of students’ constructed responses is not inappropriately confounded with language mastery where English proficiency is not the targeted construct;
- evidence that achievement levels allow for students with different levels of language to attain proficiency as warranted in the test score inference;
- results of analyses that show that force-choice responses, correct and incorrect, are not functioning differentially across ELL subgroups and the mainstream population; and
- results of other procedures and empirical analyses that determine that this population is understanding, using, and demonstrating test content properly.

In addition to these, it is particularly important for this population that there be some sort of evidence that ELL students are receiving academic instruction, in addition to the instruction in English. While opportunity to learn is not commonplace evidence across the country in the mainstream testing systems, there is precedence that, historically, certain programs for English learners have been focused entirely or more heavily on the learning of English to the exclusion of keeping up with grade-appropriate content. As such, it is important that academic opportunity be documented to
ensure that this source of systematic error is minimized. Several authors (for instance Wong-Fillmore, 2006) emphasize how English can and should be taught within the context of academic instruction and these may be used as a model for evaluators to follow. However the documentation is accomplished, it is recommended here that information about their programs of study be included in any alignment evaluation process whenever ELLs are part of the student groups being tested.
CHAPTER 5

Providing the Foundation of Principled Test Construction: Maintaining the Integrity of the Item Targets

It has been difficult to decide how to separate out the information contained in Chapters 5–8. As a whole, these chapters form a discussion about how one might improve assessment materials which are text-based and developed for audiences of test takers who are literate in English. Since English language and literacy is the largest stumbling block for English language learners, this topic deserves a complex and relatively thorough discussion. The decision was made to focus on foundational principles in this chapter, summarize item-based specific elements in the next, and associated materials, forms and reviews in Chapter 7. While the points which will be made in Chapters 5–7 are mostly relevant for items and tests in any language, issues associated with L1 assessment are multifaceted as well. As such, a decision was made to devote a separate chapter to this discussion (Chapter 8) and ask two experts in L1 issues and content testing to complete it. It is important to note that L1 forms produced using access-based development procedures discussed in Chapters 5 and 6 will be appropriate for the broadest range of English language learners, including those with limited literacy proficiency in their native language. As Solano-Flores et al. (2003) emphasize, significant magnitudes of score variation across L1 and English levels of proficiency demonstrate that access is an important issue for any language.
Overview
This chapter and the next three explain how test materials should be adapted to differentially respond to the needs of English language learners who have achieved varying degrees of proficiency in English. The targeted population focused on here consists of the range of students whose first language is not English and who are not yet grade-level competitive with respect to their English literacy. This group of students includes some “beginner ELLs” who have been in U.S. schools for only a few months and those who were born in the U.S. but have had little exposure to native English speakers and the mainstream U.S. culture. First in the overview is a brief introduction to access-based test materials in general, including a couple foundational points which apply to the four chapters. Next is an overview of item development principles that will be more fully explicated here and in the following chapter.

Several authors (for instance Lee, 2002; Lee and Fradd, 1998; Lee and Paik, 2000; Kopriva and Lara, 1997; Solano-Flores et al., 2001, 2003a) argue that current large-scale academic testing practices are insufficient and in fundamental ways problematic for English language learners. Cognitive psychologists (see Pellegrino et al. 1999a, 1999b, 2001; Pellegrino and Chudowsky, 2003) report that current practices are out of date with current theories of thinking and learning, and that foundational underpinnings of today’s assessments are significantly flawed. Bransford et al. (1999, 2005) and Heath (1983, 1986), among others, emphasize that cognitive processes associated with making inferences about students’ academic abilities are influenced by linguistics, language acquisition, dialect, culture, prior experiences, and current setting, emphasizing that theories associated with these fields have not been properly integrated into assessment practices.

If properly constructed according to the development procedures explained in the next three chapters, the access-based processes provide evidence that the integrity of the targeted constructs has been retained. If these access-based materials are being “reengineered” or adapted from standard items and other text materials (with proper attention to comparability of measurement content and complexity across sources), it can be argued that these materials can be considered parallel to the standard forms and materials. Upon sufficient empirical verification, that items and forms that have been properly constructed, it stands to reason that the student scores from these forms should be considered to be comparable to the scores of students who are administered the standard forms and other standard materials.

The discussion in these chapters is salient whether tests are presented to students in a paper–pencil format or on computers. The recommen-
dations addressed in these three chapters and in Chapter 9 (Other Relevant Accommodations and Pretest Support) are often easier to provide to students in an electronic form. Using capacities of the computer, many of the associated documents or other tools can be more specifically tailored to particular items, and items or forms can be more specifically tailored to students with specific challenges and strengths. For instance, items can utilize the ability of properly developed programs to dynamically present context and item target elements using simulations and interactive stimuli rather than language. Computers can also dynamically capture student work using capacities such as pull down menus, use of modeling toolkits, and drag and drop procedures. These advances allow ELL students to use compensatory strengths in ways that are cumbersome in typical testing situations. Further, prior non-targeted information referred to within the context of items can be stored and retrieved orally (in English or L1) or through videos. This means that, for students whose prior social or learning experiences are significantly different from the mainstream population, they can be provided regulated information that does not interfere with tested content. For ELLs or other poor readers, this information can be presented in a way that does not unduly exaggerate the reading or language load for these students.

ISSUES OF ACCESS

As Chapter 3 summarizes, item and test access allows a student to properly approach, reach and understand the relevant content in items and the problem solving requirements associated with the content. Once the requirements are properly understood, access is also defined as making available to the student the proper resources and tools to solve the problems, and the availability of proper information exchange avenues that allow students to communicate their answers understandably by the scorers or other identified scoring mechanism. Thus, the essential points of access during the student/item interaction are at the apprehension stage of problem solving, the activity stage of finding a solution to the problem identified in the item, and the explanation stage where the student effectively communicates the solution (Kopriva et al., 2004b).

Access for each of these stages needs to be maximized throughout all of the test materials. Besides writing proper items, these materials or accommodations include access-based forms, tools, and relevant resources. Examples of tools are word lists, mathematics manipulatives, blocks, or scientific supplies; examples of resources are video clips, a dynamic problem interactive environment surrounding the presentation of the items, or information links related to non-targeted prior knowledge needed in an
item or series of items. Access-based materials provide broadened opportunities for some students to apprehend or understand the problems they are expected to solve by lightening the language load in the text materials or providing compensatory information or stimuli to support the item requirements. Some options, such as contextual surrounds or providing appropriate tools, are designed to improve access by allowing students to engage effectively in problem solving activities where they would otherwise be meaningfully barred. Carefully designed items and judicious use of supporting tools and resources can be presented in a way that increases the number of possible explanation avenues available to students to communicate their solutions. Since many aspects that constrain administration and response conditions originate within the items and associated materials, attention to the choice of administration accommodations, as well as response options, must be considered when presentation materials (those discussed in Chapters 5–8) are being developed and selected.

Finally, just as packages of one or more accommodations may be appropriate for a given student, it is not unusual that more than one presentation option would be needed for students with specific challenges. Overall, it is clear that this multi-modal approach should be kept in mind at each level of item construction, materials and test development as well as implementation, analyses, and reporting.

Overview of Access and Item Development

In the mid to late 1990s forums were held that discussed how special populations should be meaningfully included in large-scale academic achievement assessments (for instance, see Kopriva, 1996a, 1996b; and Kopriva and Martin, 1998). One of the key points was that the item writing and test development process itself must be adapted to minimize barriers that were mostly linguistic and culturally driven. These barriers can be present for many English language learners, and also students with learning disabilities, hearing impairments, and attention problems. In the ensuing years, educational agencies and others have become more comfortable with considering the textual needs of these students, and training aimed at developing appropriate assessments for English language learners has burgeoned. This has included federally funded training of states under the 1994 IASA authorization, the Council of Chief State School Officer’s (CCSSO) development of training materials (Council of Chief State School Officers, 1999) and state sponsored trainings (for example, in Texas, 1998, and in Kansas, 1999). The National Center on Educational Outcomes (NCEO) out of University of Minnesota and Center for Applied Special
Technology (CAST) have teamed up to provide training materials on test development from the disabilities viewpoint (National Center on Educational Outcomes, 2003), referring to appropriate test development as the application of “universal design”.

Research on the viability and validity of various approaches subsumed under the adaptation of presentation materials, has been ongoing (e.g., Abedi, 2006b; Abedi et al., 2000; Abedi and Lord, 2001; Johnstone, 2003; Kopriva and Lowrey, 1994; Kopriva and Lara, 1997; Solano-Flores, 2002; Solano-Flores et al., 2003a; Tindal and Ketterlin-Geller, 2004; Tindal and Glasgow, 2005; Winter et al., 2004b, 2006). The research has found that linguistic issues, such as confusing vocabulary, complex syntactical structures of items, idiomatic expressions, home language and English literacy levels and degree of bilingualism, impact how ELLs are evaluated. Issues of culture, prior experiences, and the confluence of their home culture and the culture associated with U.S. schooling practices pose significant challenges as well.

Anecdotal reports suggest that, today, there is generally increased awareness of the textual needs in academic assessments for some populations, somewhat less conceptual understanding of the kinds of adaptations that need to be made, and significantly less understanding about how to specifically achieve these goals. Some state agencies and test developers are unwilling to make wholesale revisions to tests that have been carefully developed to demonstrate technical adequacy for the large part of the tested population, albeit not for students with these types of challenges. Growing numbers of states and publishers provide brief training about access issues for their writers. Many seem to advertise that their tests are “universally designed”, which suggests at least a heightened degree of awareness that advocates can build upon. Overall, however, it does not appear that a concerted effort to provide parallel forms and testing experiences in terms of the constructs measured and score meaning has yet been wholeheartedly undertaken. Thus, both differing attempts at revisions and post-hoc accommodations, and a rather arbitrary designation of some accommodations to approved or not approved lists statewide, irrespective of individual student access requirements (Clapper et al., 2006), have resulted in two tiers of test results in states. That is, insufficient attention to access at many of the essential intervention points suggests that the validity of the inferences to date could be different for some populations as compared to others.

As the consumers and the testing industry continue to embrace new developments, research should encourage the field to increasingly design new items and tests from the ground up that embrace features discussed
Frameworks such as Mislevy, Steinberg, and Almond’s Evidence Centered Design (ECD) (Mislevy et al., 2003a), Embretson’s work focusing on differential needs of students (e.g. 1998), research that takes advantage of using technology to vary testing approaches and goals (e.g. Bejar et al., 2006; Kopriva and Samuelsen, 2004), and the references described within this volume are examples of this work. As this information is being integrated into the testing culture, however, it appears that most of the work in the near future will continue to focus on reengineering existing assessment systems. This includes adapting existing items, writing items that are more access-based but still interchangeable with the types of items used in today’s assessments, and modifying review procedures and data collections. This book is being written to encourage new approaches but also to provide guidance in reengineering existing assessments. In this way, adaptations can be somewhat responsive to today’s students, while not inhibiting the development of new generations of assessments more geared to ongoing learning and tailored to measuring the testing targets while satisfactorily meeting the needs evident in the U.S.’s diverse population of students.

**Related Psychometric Work on Item Structures**

Access-based structures that underpin item development need to be constructed in order to systematically build new items that better address the challenges of many ELLs. Structures could include frameworks or rules that govern item development, as well as templates used to produce certain kinds of items within constraints. The idea of creating structures for item development is not new in the measurement field (e.g., see Haladyna and Shindoll, 1989). Even before the publication of Haladyna and Shindoll’s article, measurement textbooks and publisher manuals demonstrate that, for many years, test developers have used rules and guidelines for item development, and delimited how item stems and response options could be built. More recently, research in item structures has focused along two general lines: (1) systematically varying item features to create items with specific psychometric properties and (2) creating items that measure similar knowledge and skills that are expressed in different ways across items. One commonality among the various approaches is that the item structures are construct-based, drawing upon cognitive research in the content area under investigation (Pellegrino et al., 2001).

Bejar and others (Bejar, 2002; Bejar et al., 2006) have developed and tried out procedures for creating item models that allow for computer generation of quantitative items for computer adaptive testing (CAT) purposes. The goal of their research is to develop items that are interchangeable (isomorphic)
in terms of content covered and psychometric properties. Content experts developed models that would supply content variability but equivalent difficulty. From the examples shown in this article and an earlier technical report (Bejar, 2002), their item generation models allowed for rule-based variations in the specific numbers or other variables used in mathematics problems, with the rules used to maintain aspects that could affect difficulty (e.g., numbers are constrained to a specific range; the ratio of one number to another is maintained), and the item structure, format, and context the same across variations within the item model.

Using a cognitive design system approach, Embretson (1998) developed templates based on certain constraints which produced interchangeable abstract reasoning test items. The items were generated based on item structures that specified theory-relevant aspects of the item. The resulting assessment had acceptable psychometric properties and demonstrated that successfully using cognitive theory to generate items provides strong evidence of construct validity. Carrying the idea of using cognitive theory into the achievement-testing sphere, Enright et al. (2002) used construct-driven item development to develop mathematics problems. Using problem-solving theory as a foundation, the authors systematically varied three item features in mathematics problems in the areas of rate or probability to determine if item parameters were affected. The authors were able to explain much of the difficulty of the items, particularly the rate items, based on values of the item features. For rate items only, the item features predictably affected discrimination and guessing. The results of this study indicated that construct-driven item development, at least in mathematics problem solving, has promise, but that better information about the constructs, the learning progressions associated with them, and their manifestations in items needs to be further explored and integrated into item structures.

Researchers have also attempted to use less constrained item models to generate items that measure the same knowledge and skills but differ in crucial ways. In application of Haladyna’s model to performance assessment, Solano-Flores et al. (1999) created task shells for generating science assessments. The goal was to develop comparable items by controlling for level of inquiry, science concepts, and format. While the shells generated tasks that were similar in appearance, their psychometric properties varied and the tasks were not interchangeable. Perhaps the most general framework for item and task development is Mislevy’s (2004) Principled Assessment Design for Inquiry (PADI) model. The model links cognitive psychology and research in science learning as the basis for developing frameworks for assessment design patterns. These comprehensive design patterns are intended to provide a structure for
assessment developers to produce tasks that support desired inferences about student knowledge and skills by clearly connecting the inferences to types of evidence needed to document the inferences and types of situations that are likely to invoke construct-relevant behaviors. However, Stecher et al. (2000) concluded that the field does not yet have enough understanding of how these performance tasks work from a cognitive perspective to vary features in a predictable way. It is clear that more research into the student/task interactions is needed in order to improve the ability of the field to develop comparability in items that take full advantage of the complex differential cognitive schemas students use (Kopriva et al., 2004).

In a more limited sphere, work that investigates the development of interchangeable structures to address access issues has begun. Kopriva and Mislevy (2001, 2005), Kopriva et al. (2004b), Kopriva and Cameron (2007), Solano-Flores et al. (1999), Solano-Flores and Nelson-Barber (2001), and Solano-Flores et al. (2003a) have begun to systematically address this issue. Kopriva et al.'s work shares many of the features of construct-driven line of structure based inquiry through researching how to build items where item-specific targets are appropriately defined and constrained. In that way, non-target access barriers over like items can be minimized through collecting appropriate information about student factors and subsequently assigning proper options. Solano-Flores et al., argue that variation in proficiency across the four domains (reading, writing, speaking, listening), two languages (L1 and English), and sometimes dialect, as well as across items, impacts how items and forms are structured and the number of items to administer to students with specific profiles. Linguistic levels vary by context, language instruction, and various cultural indicators.

The discussions that follow flow from this kind of evolving work and the dynamic knowledge base the work impacts.

The Centrality of Item Targets: Clearly Defining Intended Knowledge and Skills

For the last sixty years or so, most measurement experts believed that, for similar inferences to be assumed across students taking different forms, testing conditions should be standardized for all test takers, and forms should be parallel in nature. Constraining testing by using the same set of testing conditions over students was considered to be an equity issue and fundamental to being able to robustly generalize score meaning. Parallel forms traditionally meant that a similar balance of item coverage should be apparent across forms, difficulty levels should be roughly similar, and
form results should be equated and placed on a common scale to produce standard scores across forms (Crocker and Algina, 1986). Preceded by Messick’s arguments related to test validity and invalidity (see 1989), improvements regarding how to consider and build parallel forms and items have been suggested over the last fifteen years or so. Cognition experts (e.g., Pellegrino et al., 2001; Resnick et al., 1993) and some psychometricians (for instance, Haertel and Wiley, 2003; Wiley and Haertel, 1995; Shepard, 2001; Shavelson and Webb, 1991; and Mislevy et al., 2003a, among others) have stressed the importance of explicitly developing construct-driven tests and forms, where depth as well as breadth is required, content domains are clearly specified, and sometimes item intent is defined at some grain size level, especially for constructed response items or performance tasks.

Recently, creating computer generated items using identified algorithms has heightened the demand for item rules-based templates for forced choice response items (e.g. multiple choice), as well as constructed response items. It became clear that, while the field had spent considerable time improving the explicit descriptions of test-level constructs, the thinking around item level targets was incomplete. This was especially apparent for researchers interested in improving access to test materials for special populations. As studies investigated variations in testing conditions, the issue of comparability of results over conditions rested squarely on the documentation of construct invariance in items and tests (Kopriva et al., 1994; Kopriva, 1996b; Solano-Flores and Shavelson, 1997; Solano-Flores et al., 1999; Tindal et al., 1998; Tindal and Fuchs, 1999; Bielinski et al., 2001). As noted above, Solano-Flores and Shavelson (1997) and Solano-Flores et al. (1999) proposed building item shells, identifying item targets at a rather broad level and outlining key dimensions of item construction to be considered in order to improve access for English language learners. Kopriva (2000), Kopriva and Lara (1997) and Kopriva and Martin (1998) argued that identifying the proper grain size for the item targets, and explicitly identifying non-target elements in items and tests, were key in building access-based items and tests for assessments in general. This has been reiterated lately by Popham, Pellegrino, Berliner and others (Commission on Instructionally Supportive Assessment, 2001).

Defining the targets too broadly led to items measuring different aspects of constructs; defining the targets too narrowly meant that no variation in conditions or item language could occur. Further identifying what aspects of items were not relevant to the target allowed item writers to construct interchangeable items that addressed the non-target features. One of the foci of a study, convened in 2002, was to specifically develop item templates that would address the grain size of item targets and what item and test
elements can vary (Kopriva et al., 2003). The procedures, while applicable to item writing in general, are particularly important for populations with language, literacy or attention challenges, including English language learners, who have only limited access to standard forms and materials. This approach is described below.

**Development and Completion of Access-based Item Templates**

Access-based item templates identify both the measurement intent of an item and what the item measures that is not intended. Access-based item templates, when completed, provide information about each of the components that specify the conceptual parameters of an item. This information includes location within a hierarchy of definition that spells out how items substantively fit within the content standards, as well as providing precise information about the measurement targets of each item. Then, target-irrelevant knowledge and skills are identified that explain how test takers will communicate with and use item contents. Table 5.3 at the end of the chapter is an example of a completed template. Since terminology for the template definitions being used here is not standardized in the field, one set of terms will be identified and explained as they are used. Readers are encouraged to adapt these definitions and terms to fit their own testing vocabulary as needed.

**Template Components**

**Construct Map Targets** Targeting what is and isn’t intended in individual items is initially based on clearly specifying tree-like **construct maps** that progressively unpack broad-based constructs identified for testing and stipulated in the test specifications. As defined here, the “tree trunk” would be the content area. From the trunk are several constructs which specify the nature of the content area, and each construct under consideration would be a broad constellation of concepts, knowledge, and skills within the content discipline. For instance, in mathematics (the “tree trunk”), a construct could be identified as a strand of mathematics (such as number, geometry, algebra, statistics) the developer wishes to focus upon. The strand is defined in Table 5.1 to be the **domain**. As the construct is unpacked the next level of specificity is referred to here as the **standard**, followed by the **indicator**. In Table 5.1, an item is identified within the domain “Number,” the standard “Operations,” and the indicator “Understand visual representations of operations.”

After general construct maps are constructed so that relevant domains, standards and indicators are specified, items need be assigned or built to conform to the map. Below in Figure 5.1 is an example of an item
and Table 5.1 illustrates how that particular item was assigned within its appropriate construct map. This exercise needs to be repeated for each item so that the foundation of the item source is clear and a clear linkage is made for inferential and comparability purposes. Once the assignment is completed, specification of item targets and so on can be undertaken.

Item Core Targets After items have been located within the general construct maps, they are specified by identifying information that becomes the *core*. Each core has four dimensions, of which the first two indicate the targeted knowledge and skills and intended depth of the content, while the latter two specify further limitations. As defined here, any items that share a core are considered to be interchangeable from a construct-driven structural perspective. Items that share the same objective and complexity level (but differ on the other two dimensions) should be considered to be interchangeable at a slightly higher grain-size level. For coverage purposes (including issues of both breadth and complexity of skills), it is expected that more than one core will be identified in those standards and indicators that are being tested. The core consists of the:

1. **Objective**: targeted content knowledge and skills. Targeted objectives are defined within a specific indicator and are usually the same across more than one item (depending on how detailed test inferences are expected to be).
2. **Complexity level**: specifies the depth of knowledge and skills the item is measuring. Low complexity items expect knowledge recall or low level skills, whereas items with high complexity, for example, ask students to conduct multiple steps to solve problems, extrapolate or make inferences, and in other ways demonstrate skills identified with expert mastery.
3. **Item-specific subject matter** (ISSM): content knowledge and skills that define only an *instance* of the objective. ISSM varies idiosyncratically by core and should include an explanation of the content.
4. **Item-specific constraints** (ISC): any additional constraints are specified here. These may be test level constraints that would be invariant over all items, such as not changing the item type or retaining the same numbers over items that share the same core. Other constraints may be core specific and may or may not be content-related. An example of a content-related constraint is requiring knowledge of prerequisite skills (such as multiplication or addition) when target is asking students to compute an algebraic algorithm. If this is not
specified as a constraint, then it is assumed that lack of prerequisite skills could be compensated for in an item.

Decisions about whether to include ISSMs and ISCs in the test template core ultimately rest on the level of comparability one wants to make. With sufficient empirical verification of validity, it could be argued that items which share a core where the four dimensions are specified are equivalent at the raw score level. Items which share cores where the first two dimensions are identified could be said to be equivalent, at least at the scale score level. For instance, Bejar et al. (2006) suggest that some variations have been found to be reasonably trustworthy in terms of not changing the measurement intent of the item. As an example, at higher grade levels it may be appropriate to vary the numbers used when measuring particular operations as student performance does not tend to be as confounded with the specific numbers used and the pool of eligible numbers is much larger. In cases of tests that use only one item type, inferences should be clearly constrained to reflect the range of type of information that this item type can produce about student knowledge and skills.

Figure 5.1 and Table 5.1 provide an example of how a released state assessment item fits within the targeted definition hierarchy.

Target-Irrelevant Knowledge and Skills Irrelevant information in items is always a part of testing because information about contexts, item requirements, use and item constraints must always be communicated back and forth between assessment and test taker. Therefore, the goal of developing access-based items isn’t to eliminate all irrelevant or ancillary information in items. Rather, it is to 1) become increasingly cognizant of the irrelevant aspects of items and 2) deliberately develop items that use specific, non-target ancillary knowledge and skills in intentional ways to minimize the barriers to testing for students with particular challenges and needs. It stands to reason that, in order to accommodate all test takers so the integrity of the intended targets can be communicated properly, a limited number of item and form options, along with form supports, will probably be needed.

After the item’s targeted knowledge and skills have been identified, the preferred non-target ancillary components should be specified. If already constructed items are being evaluated, they should be examined to determine which irrelevant factors are being used to communicate the target information. Kopriva and Winter (2004), Winter et al. (2006), Mislevy et al. (2005) and Carr and Kopriva (2007) identified specific target-irrelevant components in items like those used on today’s achievement tests
by considering structural and contextual factors particularly salient for ELLs and some other students. These researchers investigated the use of narrowly specified and broadly accessible test forms that incorporate the types of items found in today’s typical on-demand achievement tests. They found that broad-based access forms of this type appear to be adequate for many ELL students, as well as other students with language challenges, including poor readers, and students with attention issues. Carr (2004) suggests that these forms may be useful for some students with learning disabilities and

Figure 5.1 Grade 3 mathematics item

TABLE 5.1 Hierarchy of targeted item information

<table>
<thead>
<tr>
<th>Construct</th>
<th>Domain</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maps: Standard</td>
<td>Operations: understand and use operations on numbers (addition, subtraction, multiplication, division)</td>
<td></td>
</tr>
<tr>
<td>Indicator</td>
<td>Choose and use the appropriate operations to solve single-step or multi-step word problems</td>
<td></td>
</tr>
<tr>
<td>Core: Objective</td>
<td>Solve single- and multi-step word problems involving a single operation, including problems about money</td>
<td></td>
</tr>
<tr>
<td>Complexity Level</td>
<td>Medium</td>
<td></td>
</tr>
<tr>
<td>ISSM</td>
<td>Division, ignore remainder, or repeated addition, index number of addends; context—find largest possible number</td>
<td></td>
</tr>
<tr>
<td>ISC</td>
<td>Multiple choice item type, use same numbers, present the unit cost before the total amount of money available</td>
<td></td>
</tr>
</tbody>
</table>
for hearing impaired students as well. In addition to English access-based forms, additional broad-based forms which attend to other irrelevant factors may be important to develop, including forms tailored to students with L1 literacy, and forms for students with some amount of literacy in L1 and English (Kopriva, 2000). In each of these situations, some of the ancillary structural factors identified in the template will probably be completed at the form level (for all items), rather than at the individual item level. Issues of context, prior knowledge, and format still need to be considered for each item.

Once computer-based large-scale academic assessments become routine, algorithms can and should be able to handle more specific access-adaptive item and form options for students with different profiles of challenges and strengths (Kopriva and Winter, 2004). This includes increasing the kinds of item types as well as response options that would be amenable to addressing the challenges of ELLs. The process of matching items to students can be made more interactive as well, depending on a number of factors that can differ by item and student profile (Kopriva et al., 2003; Samuelsen and Carr, 2004). Even when fewer form options are possible with today’s large-scale paper and pencil tests or computerized testing that is not access-adaptive, the standard form, albeit universally designed, will probably not be appropriate for some students. For example, students with little exposure to English, or to U.S. experiences, schooling and testing practices may need a more nuanced approach that addresses their language and cultural challenges more directly (Kopriva and Winter, 2004; Mislevy, R. et al., 2005). Solano-Flores et al. (2003) suggest that language issues may impede understanding or problem solving even for students with a fairly substantial degree of literacy in English and their home language. Further, for tests with limited audiences, or assessments that are more diagnostic and need to rely on finer communication distinctions, variations that address a more limited set of students may be necessary. For instance, if a local test is being developed for recently arrived Hmong students who, within their culture, have a strong oral tradition and have only recently begun using written communication, ancillary factors that are sensitive to their cultural communication style should be incorporated. In all of these cases, the access-based template can handle this level of variability. The ancillary factors identified at the form level will focus on test takers with a narrower range of proficiency levels, and/or specific language and cultural considerations. Item level issues of context, prior knowledge, format and response would be tailored for this narrower band of students as well.
It is always important to remember that, regardless of the scope of the forms, they may need to be complemented by supplementary materials and accommodations that are appropriate for individual students in order for the forms to properly improve how well students can access the measurement targets.

Table 5.2 provides an example of a completed template, listing target-irrelevant information as well. This is an evaluation of the item found in Figure 5.1. The next section will provide one example of how the specification of irrelevant skills and knowledge might be translated into an effective item for some ELLs.

Before the item writing factors are explained in the next chapter, an example will be given to illustrate how an existing item can be improved by applying guidance from the access-based template.

Consider the following fourth grade constructed response item in mathematics (Figure 5.2 and Table 5.3). This item is a released item and a typical example of items that are used today on many achievement assessments across the country.

### TABLE 5.2 Completed core template for Figure 5.1

<table>
<thead>
<tr>
<th>Core: Objective</th>
<th>Complexity level</th>
<th>ISSM</th>
<th>ISC</th>
<th>Target-irrelevant knowledge and skills</th>
<th>Nouns</th>
<th>Context</th>
<th>Verb</th>
<th>Adjective</th>
<th>Sentence</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solve single- and multi-step word problems involving a single operation, including problems about money</td>
<td>Medium</td>
<td>Division, ignore remainder, or repeated addition, index number of addends; context—find largest possible number</td>
<td>Multiple choice item type, use same numbers, present the unit cost before the total amount of money available</td>
<td>“Stuffed animals”—vocabulary and double meaning</td>
<td>“Allowance”—vocabulary</td>
<td>Having an allowance is not a common experience for some students (SES related)</td>
<td>“Is going to spend”—simple future</td>
<td>“Largest”—double meaning</td>
<td>Use of adverbial phrase</td>
<td>Use of conditional phrase</td>
</tr>
</tbody>
</table>
At Jefferson Midlands Middle School, the sixth grade students and their teacher are planning a field trip to the state capital at the end of the year. In the morning they will visit the state legislature, and in the afternoon they will go to the zoo.

There are 33 students in sixth grade. Five parents and two teachers will be coming with the students on the trip. Each of the adults has a car that can hold four students. One of the students says: “There are not enough cars to take all of us!” Do you agree with the student? Explain your answer.

**Figure 5.2** Grade 4 mathematics item

**TABLE 5.3** Target-relevant template for Figure 5.2

<table>
<thead>
<tr>
<th>Domain</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>Operations: understand and use operations on numbers (addition, subtraction, multiplication, division)</td>
</tr>
<tr>
<td>Indicator</td>
<td>Choose and use the appropriate operations to solve single-step or multi-step word problems</td>
</tr>
<tr>
<td>Objective</td>
<td>Solve whole number single- and multi-step word problems with one or more operations</td>
</tr>
<tr>
<td>Complexity level</td>
<td>Medium high</td>
</tr>
<tr>
<td>ISSM</td>
<td>Multi-step; determine how to use information provided, combine two categories of information (parents/teachers) into one (adults); understand meaning of “each”; division, interpret remainder, or multiplication and addition, compare totals, or repeated addition, index number of addends.</td>
</tr>
</tbody>
</table>

This item was revised to be more accessible for English language learners, as shown in Figure 5.3. While the specific decisions about the changes will be explained later, this example can give readers an idea of the kinds of deliberate item adaptations considered when an access-based item is being developed.
1. Information that is not needed to set the context for the problem has been eliminated, reducing the amount of text.

2. Plain language principles have been applied to the item to reduce the semantic and syntactic complexity of the item. The sentences are shorter and straightforward, using present tense and active voice and reducing the use of prepositional phrases and dependent clauses. A visual is used to illustrate the item. Note that numerals have been used consistently throughout. The translation between a verbal and symbolic representation of a number was considered construct-irrelevant mathematics.

3. The formatting has been arranged to provide maximum access to the problem requirements. Each complete piece of information is presented separately, since, for this item, selecting the appropriate information from among relevant and irrelevant pieces of information was not part of the measurement target. The question is clearly separated from the rest of the text, and the two-stage character of the item, answering the question and explaining the response, is evident.

4. While both the base and the variation assume students are familiar with class trips, which may not be the case in all schools, potential cultural schooling bias has been reduced in the variation by having a student’s statement the focus of the question. In some cultures, children are not used to questioning teacher judgments and decisions.2

5. Students are given options for how they represent their response.

6. Students are allowed to use counters to help them represent and solve the problem. The targeted content knowledge and skills do not preclude allowing various methods of representation or solution, as noted in the ISSM. The manipulatives provide students who are ELLs with a way to represent the text that may help them understand the problem situation.

Figure 5.3 Access-based item
IMPLICATIONS

Several implications stem from developing and using this type of access-based template. Two will be noted here. First, the issue of ancillary fit between test taker and item that was just discussed relates directly to the level of error due to mismatch that different types of assessment systems are willing to tolerate. In general, it seems reasonable that the more detailed the inferences, the less error of this type a test should be willing to allow. Research aimed at identifying acceptable levels of mismatch should be undertaken for different types of tests and tests used for different purposes.

Second, from a measurement perspective, this type of item-structure template primarily addresses construct-equivalence across like items. That is, items produced from the access-based template are considered to be interchangeable for particular groups of students because the integrity of the target 1) has been clearly identified and 2) has subsequently not been disturbed by changes that minimize the effect of text based challenges salient for the specific groups. However, item difficulty is not a reasonable parameter upon which to also gauge interchangeability. Recent research continues to suggest that part of item difficulty in standard items is due to inaccessibility. For instance, one study reported that, over six grades and four content areas, item difficulty on access-based items was generally a wash for non-ELLs who had taken the items under random conditions—on several there was little or no change and some were easier or more difficult than standard-based versions of the items. On the other hand, the probability of correct response on about 2/3 of the access-based items was increased over like standard items, for students with learning disabilities and for ELLs who had pre-functional, beginner, or intermediate proficiency in English (Kopriva and Cameron, 2007). Other empirical methods need to be found to evaluate when access-based items might and might not be considered to be equivalent to their standard item counterparts.

The next chapter will focus on the specifics of how readers might produce access-based items. Writers have found that this is a complex challenge. Rudimentary principles such as those discussed next can be learned and applied in several common ways to many items. However, the writers report that a number of the access-based elements are best applied in response to unique contextual requirements dictated by individual items. For example, almost all items will include plain language, and many may include visuals and formatting aspects as well. But more detailed elements, such as the type of visuals which might be most appropriate, and larger principles, for instance how to address prior knowledge issues, should be addressed as
well for some students. In all, building effective access-based items seems to be a learned skill of which the basics and individual building blocks will be considered next.

Endnotes

1. With thanks to all those who contributed to this work, including Chen Su Chen, David Wiley, Jennifer Koran, Ming-yi Cho, Jessica Emick, Carol Boston, Therese Carr, and especially Phoebe Winter.

2. The appropriateness of questioning information from authority figures in various forms such as newspaper articles, scientific reports, and teacher remarks varies across and within cultures. In U.S. schools, questioning and verifying information is a taught skill.
This chapter will provide specific information about three aspects of access-based item development. The first section will explain some overarching considerations associated with context, the second section will focus on language usage and support mechanisms, and the last section will address some issues surrounding the development of access-based rubrics. While the rest of the chapter will explain the particular aspects within each section, Table 6.1 briefly outlines the contextual and structural factors discussed in this chapter, and the tools and resources discussed in Chapter 7. Within the table are general descriptions of each type of intervention, and a list of the aspects within each intervention likely to be salient for English language learners (ELLs). Readers will note that the section categories are rather artificial. Context and culture are pervasive influences item writers need to be aware of in order to make informed choices about language, text and item supports for various students. Rubric considerations need to account for how students from different backgrounds might vary in their interpretations of items and tasks and must provide the ground for effective scoring of content in student responses. Rubrics have to be sensitive to the broad range of communication strategies the scorers will encounter. The task of writing accessible items involves the thoughtful consideration of barriers for students with many different needs, and an appreciation of how this diverse set of students might differentially compensate for shortcomings so the barriers can be minimized. Because
### TABLE 6.1 Summary table: contextual factors, structural factors, and tools and resources

<table>
<thead>
<tr>
<th>Contextual factors</th>
<th>General description of interventions</th>
<th>Types of interventions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Culturally broad experiences</td>
<td>Cultural expectations seem to have an impact on how a student understands the requirements of an item. These cultural expectations become especially problematic when a student’s experiences or home culture values are distinctly diverse from those typically experienced by the mainstream population in the U.S. (Kopriva, 2000).</td>
<td>1. Prior knowledge that assumes mainstream U.S. experiences 2. Expectations that assume a common U.S. value system</td>
</tr>
<tr>
<td>Clear and explicit expectations</td>
<td>In classrooms, ELL experts know that it is not enough to assume that ELLs will understand test expectations and approaches familiar to those in the mainstream U.S. population. For large-scale tests that measure content over diverse types of students, clarity in expectations relative to all aspects of the items and tests need to be explicit and clearly stated (Farr and Trumbull, 1997).</td>
<td>1. Direct, explicit explanation of item requirements</td>
</tr>
<tr>
<td>Prior learning expectations</td>
<td>Two types of prior learning experiences are addressed (1) the prerequisite knowledge related to target content that is required for an examination of more complex skills, or prerequisite content knowledge and skills at older grade levels where knowledge builds on a foundation developed in the earlier grades, and (2) the use of non-targeted content as context in items, especially items that measure processes such as reading comprehension or science inquiry skills (Kopriva and Cameron, 2007).</td>
<td>1. Assumptions of prior learning required for complex skills 2. Use of non-target content as context in items</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Structural factors</th>
<th>General description of interventions</th>
<th>Types of interventions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple language structures</td>
<td>The issue of language organization is particularly salient for ELLs, because text in their home language is almost assuredly structured differently than English text. The basic presentation of text for ELLs involves a conscious and complex re-appropriation of structural conventions explicitly or implicitly learned as part of their home language experiences (Abedi and Lord, 2001; Johnstone, 2003).</td>
<td>1. Use of simple sentences 2. Use of similar paragraph organization 3. Use of present tense and active voice 4. Minimizing use of rephrasing</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>The vocabulary in all items, directions, and supplemental test materials of both academic and social English must be considered when developing access-based items (Farr and Trumbull, 1997; Kopriva, 2000).</td>
<td>1. Use of familiar language 2. Limit use of substitute words 3. Careful use of multi-meaning words</td>
</tr>
<tr>
<td>General description</td>
<td>Access aspects</td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>----------------</td>
<td></td>
</tr>
<tr>
<td><strong>Effective visuals</strong></td>
<td>1. Use of relevant visuals</td>
<td></td>
</tr>
<tr>
<td>ELLs are both learning the language and learning to read at approximately the same time. Visual cues provide help for ELLs as they struggle to learn English and become literate. However, not all graphics are equally beneficial; thus care must be taken when using this type of support (Filippatou and Pumphrey, 1996; Winter <em>et al</em>., 2006).</td>
<td>2. Use of an effective format</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3. Use of illustrations to mirror text</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4. Use of illustrations to replace text</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5. Use of first person visuals</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6. Use of visuals to organize events in time</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7. Use of visuals to clarify textual meaning</td>
<td></td>
</tr>
<tr>
<td><strong>Effective item format</strong></td>
<td>1. Separating key ideas</td>
<td></td>
</tr>
<tr>
<td>Formatting print material to focus the reader, clarify purpose, and otherwise effectively move the reader through the information is central to the work of print media professionals. Emerging work suggests that attending to and clarifying item formats does play a part in making items more accessible for this population (Kopriva and Mislevy, 2005; Winter <em>et al</em>., 2004).</td>
<td>2. Clearly identify item questions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3. Use of titles</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4. Use of mixing symbols and text</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5. Use of examples</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6. Highlighting key words or phrases</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7. Use of boxes or lines</td>
<td></td>
</tr>
<tr>
<td><strong>Text amount</strong></td>
<td>1. Retain complexity of target while using non-interfering contextual cues</td>
<td></td>
</tr>
<tr>
<td>The guiding principle in making decisions about text amount is to retain the content complexity of the intended target while providing enough information in as non-textualized form as possible (Kopriva and Lowrey, 1994; Wong-Fillmore and Snow, 2000).</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Platform for demonstrated response</strong></td>
<td>1. Computerize items in order to present item context through simulation or to incorporate drag and drop options, manipulation of stimuli and other features</td>
<td></td>
</tr>
<tr>
<td>Limited language acquisition suggests the use of alternate platforms for how items are presented. One key reason for items to be computerized is to extend how students can demonstrate their solutions to the items.</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Impact of home language</strong></td>
<td>1. Use of cognates</td>
<td></td>
</tr>
<tr>
<td>Some item-drafting issues for ELLs reflect the influence of students’ native language. For more frequently spoken languages and for those particularly prevalent in certain areas, some actions can be taken to minimize misunderstandings (Kopriva and Cameron, 2007).</td>
<td>2. Reduce use of linguistically confusing words</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3. Using symbol icons consistently</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4. Reviews of text by those familiar with the culture and language</td>
<td></td>
</tr>
</tbody>
</table>
of this, there is no easy checklist of solutions or a clear distinction between what item writers should and shouldn’t do and when. Given this caveat, the next several pages will attempt to illuminate the issue by highlighting some types of activities considered to be useful for students with different impediments. The goal will be to produce items with multiple opportunities and avenues of access.

The access-based illustrations throughout the chapter are included to illustrate particular points salient to writing this type of item. As item writers gain more experience with these items, it is anticipated that improved examples will become available for public release. In many cases, the released or base item which is being re-written to increase accessibility

<table>
<thead>
<tr>
<th>Tools and resources</th>
<th>General description</th>
<th>Access aspects</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Tools</strong></td>
<td>Content tools are objects which are item or content area specific and that can be used to aid the student in understanding the intent of a particular item, solve the problem or express their solution. For ELLs these tools provide interactive compensatory opportunities to cross over the minimum threshold for understanding key item elements or being able to navigate problems (Kopriva and Mislevy, 2005).</td>
<td>1. Concrete materials 2. Computer simulations with drag and drop options and/or graphic/drawing opportunities</td>
</tr>
<tr>
<td><strong>Text support</strong></td>
<td>Text supports are identified as text-based aids that occur over and above the item text and provides supplementary support (Abedi et al., 2001a, 2001b; Hipolito-Delgado and Kopriva, 2006).</td>
<td>1. Bilingual glossaries or dictionary 2. Monolingual glossaries 3. Picture-word dictionaries 4. Side-by-side forms (a.k.a. dual language test booklet)</td>
</tr>
<tr>
<td><strong>Content-based resources</strong></td>
<td>Resources that provide additional information during the testing experience to fill in gaps or provide a common referent that minimizes additional textual load (Monroe, 2003).</td>
<td>Providing prior learning experiences information or primary sources via: 1. Primary source documents 2. Prior experience information</td>
</tr>
<tr>
<td><strong>Activities</strong></td>
<td>Activities, such as a brief (15 minute) interactive discussion or an activity prior to the period of testing that provides context for all students and ELL experts suggest is another important compensatory support mechanism (Monroe, 2003).</td>
<td>1. Brief interactive discussion 2. Brief collection of data</td>
</tr>
</tbody>
</table>
will also be shown. For purposes of space in this chapter, all items are illustrated in 10 point and with the font used throughout the book. In an actual assessment the type should be larger (at least 12 point and often larger, particularly for younger grades) and in an easy to read font.

**Contextual Factors**

This section will focus on the issue of culture and context in items, and on the importance of making expectations clear and explicit. Prior learning expectations are considered last.

*Culturally broad experiences*

Two sets of cultural expectations seem to have a primary impact on how a student understands the requirements of an item: (1) expectations in items which assume prior experiences that are relatively common if a student grows up in a mainstream U.S. community, and (2) expectations in items that assume a value system common to the value system typically endorsed in the U.S. These cultural expectations become especially problematic when a student’s experiences or home culture values are distinctly diverse from those typically experienced by the mainstream population in the U.S.

In some cases, items expect students to recognize certain things or events, or to have had certain experiences growing up. One good example is from *Ensuring Accuracy* . . . (Kopriva, 2000). This example uses a parade as the context for some items which follow it. In some parts of the world, parades are limited to political agendas, and espouse very different sets of feelings and perceptions for viewers than U.S. parades (even though some U.S. parades may be patriotic in origin). Sometimes items ask a question using an object or geographic association common in U.S. or Northern European cultures but less common in other cultures. Examples include vending machines, certain clothing or household items, and assumptions that the month of April (for instance) is in the spring of the year.

It is easy for U.S. item writers to overlook problems that arise from diverse cultural value systems as they may assume the item is affecting a value that is universal. One example is the assumption that, if requested, a student will provide a response about a particular topic which involves the student extrapolating how an audience (particularly an adult audience) might perceive the topic. Some cultures do not value explicit response to opinions of adults, expression of differing points of view, or other types of persuasive writing. In some cultures it is not the custom to provide an explanation of something when they know the reader already knows the answer. Other examples include assumptions that the student will understand the value of competition, or the value of individual rights.
Solano-Flores and others (Solano-Flores et al., 2003; Solano-Flores, 2006) have largely addressed the issue of culture in terms of language and dialect (for an item, which language provides better access for a student), but they certainly acknowledge that experiences and expectations are culturally based and that language considerations address only part of the concern. As such, decisions to use a particular item context is of pressing concern for English language learners. Items prepared for a large cognitive lab clinical trial with ELLs rated items as contextually complex or not and students were asked to explain if the accessible items were contextually appropriate (Winter et al., 2006). These researchers found that their approach of producing access-based items for released items with problematic context was generally on target for the relatively new arrivals. They focused many of their access-based contexts within U.S. in-school experiences, and common foods, geographical stimuli, and stores. Monroe (personal communication, October 2004), who has extensive experience teaching newly arrived immigrants and other students who speak little English, suggests that many of these are poor students who quickly learn about such items as coupons, sales, and videos, regardless of their home country experiences. Mann et al. (2006) reported that context appears to be a factor compromising validity in some released items, but that attempts at addressing these issues are only sometimes successful. Especially in constructed response items, it appears that using a familiar context, as well as addressing language and other item considerations, helped especially those ELLs who had been in the country at least about 8 months. Kopriva et al. (2007d) reported that success in remediating context differs by item, that the addition of visual elements such as those discussed below seem to have an impact, but that unfamiliar item formats can actually increase context dissonance, particularly in multiple choice items.

Emick et al. (2007) found that, beyond language, teachers reported that familiar contexts as well as other non-language compensatory supports were needed for a significantly greater number of newer ELLs than for other current English learners. On the other hand, the need for context support for exited ELLs was substantially lower than for all other groups, even native English speakers. The work of Carr and Kopriva (2007) demonstrates that addressing contextual issues is more difficult in older grades because of the more abstract quality of the context. They also show how visuals can provide contextual support in place of some text in social studies, science and language arts, although they acknowledge that textual demands increase in older grades. Kopriva and Mislevy (2005) concluded that a need for broad context definitely seems to be one of the issues that impact accessible mathematics item development. However, the impact of context interacts with various other access-based item considerations, and with the
age and content demands of older students. This interaction needs to be more thoroughly understood in items like those in vogue today, and in items that may effectively deal with a greater complexity of cognitive skills in the future.

Figure 6.1 is a released fifth grade item from a state’s assessment system. As the reader can see, the context of stuffed animals, proof of owning, and Million Dollar Bear can be considered to be culturally constrained. Figure 6.2 is an example of how mathematics experts rewrote the item for a more culturally broad audience. This access-based item, as well as most of the examples in the next several pages, is not perfect—for instance, the picture could be clearer and more descriptive. Nevertheless, besides a change in context, the example is also useful because it embodies several additional points raised below.

Patty just received a letter in the mail telling about a new promotion with stuffed animals. When Patty has collected and shown proof of owning 125 stuffed animals she will receive the new Million Dollar Bear free. Patty has 79 animals right now. Which of the following equations show how many more animals Patty will need to collect to get her free Million Dollar Bear?

A. \( \square - 125 = 79 \)
B. \( 79 + \square = 125 \)
C. \( 79 - \square = 125 \)
D. \( 125 + 79 = \square \)

Figure 6.1 Base item

A class has 79 stars. They need 125 stars. How many more stars does the class need? Choose the correct equation.

A. \( \square - 125 = 79 \)
B. \( 79 + \square = 125 \)
C. \( 79 - \square = 125 \)
D. \( 125 + 79 = \square \)

Figure 6.2 Access-based item
Making Expectations Clear and Explicit

In classrooms, ELL experts know that it is not enough to assume that English language learners will understand test expectations and rules familiar to those in the mainstream U.S. population. For large-scale tests that measure content over diverse types of students, expectations relative to all aspects of the items and tests need to be clear and explicit. An understanding of expectations cannot be taken for granted as cultural and regional differences can vary a great deal for a number of different reasons. Item writers may think making explicit judgments of the type of information discussed here is unwarranted because it is too obvious and wasteful of time and resources. Reviewers, however, need to critically evaluate these judgments, to avoid common sources of confusion for ELLs.

Farr and Trumbull (1997) and Hiebert and Calfee (1990) discussed the need to be direct, clear, and specific about item requirements, what latitude the student has in solving the problem or responding, and how students can use additional time, tools, or other accommodations. Clarification of non-targeted vocabulary is essential, but should not substantially increase the language. Malcolm (1991) argued that contextual parameters associated with prerequisite knowledge expectations and response constraints should be explicit and clear. For instance, if the student is supposed to anticipate certain consequences or ways of viewing phenomena, these should be explained in a way that does not increase the literacy load. For constructed response items, it should be made clear how students might be able to explain themselves, for instance, through using pictures, diagrams, charts, algorithms, or L1. Use of a universal symbol might be useful here (see next section). If communication via writing in English is expected, this also should be made clear. Certainly it must be made clear if the quality of the students’ writing, and/or the presentation of information in the charts, pictures, and other visuals is going to be evaluated along with the content of their response. In the latter case, evaluation criteria for the presentation of the material should be made clear to the student, either through the directions, as part of the item, and/or in rubrics provided prior to the test. This is over and above the criteria evaluating the subject matter covered in the item. In the 1990s, rubrics and performance levels sometimes interspersed literacy fluency and content skills (see Ensuring Accuracy for more information here). A good example of the evolution of content and literacy measurement expectations can be seen in the comparisons of NAEP mathematics frameworks and levels from the 1990s to 2001 (White, 2003; Allen et al., 1997; Braswell et al., 2003). The approach of implicitly mixing measurement expectations that include literacy evaluations is problematic both from a measurement standpoint and from the perspective of students.
faced with language challenges. As responses are used to measure more than one target, any and all additional evaluation expectations should be clearly made.

Below are two examples of student work where the ELL student responded to items in ways not considered by the item developer. As the reader will note, neither of the items identified any constraints on the types of responses they were expecting. In the first case (Figure 6.3), the question assumed the student would explain an urban water distribution plant, including storage system and pipelines (in the U.S. or another developed country) and the answer was marked incorrect. In the second example, a student is asked to draw and label a food chain, where a response from the plant and animal world was expected. Although the response in Figure 6.4 may or may not be considered correct, it reflects the student’s cultural background as a recently arrived Punjabi student who does not eat beef (but who clearly has had experience with convenience stores in the U.S.).

![Figure 6.3 Laotian](image)

**Source:** From Kopriva and Sexton, 1999.
Prior Learning Expectations

Expectations about prior learning are a huge issue with English language learners. Many ELLs are seasonally or cyclically “migrant”, meaning that they are routinely schooled (or not) in more than one location—either because of the nature of their parents’ work, or because of regular trips to their home country during the school year. Added to inconsistent schooling
in the home country before arrival, this movement tends to create gaps in learning. Additionally, concerns related to opportunity to learn plague many school systems, and are particularly an issue in high poverty schools often frequented by ELLs. The opportunity to learn issue is beyond the scope of this book, but it certainly contributes importantly to the inferences test consumers make in interpreting student test results.

In any event, test developers cannot rescue themselves from considering issues of prior learning considerations. Two types of issues in particular seem to be problematic. First is the prerequisite knowledge related to target content that is required for an examination of more complex skills, or prerequisite content knowledge and skills at older grade levels where knowledge builds on a foundation developed in the earlier grades. Second is using non-targeted content as context in items, especially items that measure processes such as reading comprehension or science inquiry skills.

**Prerequisite Skills**

Kopriva and Mislevy (2001) argued that prerequisite knowledge or skills can be either ancillary or part of the measurement target. If they are the latter, interpretation of other targeted knowledge should include a consideration of the prerequisite skills. If they are determined to be ancillary, then their influence should be minimized, there should be sufficient compensatory avenues so students can go around them, or sufficient information should be provided to students who need it in order to close any gaps in knowledge which may occur for this reason. In either case, there should be an item by item determination of a) any prerequisite skills that are required, and b) whether or not the knowledge and skills will be assumed to be ancillary or part of the target.

A good example of how to address prerequisite skills can be found in the large-scale assessment of mathematics. Many tests include both calculator and no-calculator sections. The section which allows the use of calculators is minimizing some prerequisite skills by providing students the means to compute data with this tool; the other section is typically an assessment of the students’ computational skills. Without a clear measurement of computational skills separate from the more complex skills attended to in the other section, it would not be clear when incorrect answers should primarily be attributed to computational difficulties or to lack of targeted skill capacity.

For English learners, it is important for item writers and test developers to explicitly examine each item and determine what are targeted and ancillary skills. For ancillary skills, other means must be found to address them. Many examples and ideas of what can be done are provided in the next section. If the item writer expects the prerequisite knowledge and skills
to be part of the target, care should be taken to minimize an incorrect inference when the target contains both the skills or knowledge that were originally intended and also the necessary prerequisite skills. If this is done too often, inferences about the more complex target can be seriously compromised.

Recently completed work addressed how test development staff might attend to such issues. One example these developers have grappled with comes from English language arts, specifically, the measurement of comprehension and associated skills in poems and some prose passages. At issue is when the target is focusing on poetic comprehension associated with interpretation of imagery, or skills associated with passage comprehension involving key non-standard and non-targeted use of language exemplified in idioms and colloquialisms. In these cases, the intended measurement targets are the proper interpretation of descriptions involving language use, not the concrete meaning of the words or phrases. State item writing staff have argued that students should be given the concrete meaning of appropriate terms, in order to allow students to focus on the interpretation. Likewise, they have become aware of and tried to minimize the number of idioms and other non-standard uses of language when the focus is on comprehension of the passage (as compared with measuring knowledge of these sayings or words, Kopriva and Cameron, 2007).

Non-targeted Content

Kopriva and Lara (1997) noticed that a fair number of National Assessment of Educational Progress (NAEP) science items used non-grade level material as a contextual basis for their items. One of the challenges that increases as academic tests advance in grade level is how to handle non-targeted context that is academic (as compared with social) in nature (Carr and Kopriva, 2007). In social studies, for instance, many similar concepts are taught in progressively more depth over the years. This occurs in science as well. In the assessment items for these subject areas, it is easy to refer to knowledge gleaned from earlier years while targeting the measurement of current knowledge. Development considerations have included providing CDs that explain the necessary background information to the students (or orally administered linked material if the testing is computerized). Another consideration is to use only non-targeted contextual content from the current year, but then which concepts are really being measured in the items? A third alternative is to address this issue in much the same way as the industry has tended to address using academic content from one subject area in the measurement of another. This will be discussed next. Regardless, it is clear that assessments in older grades regularly contain this kind of challenge, and the measurement of older ELLs who have not successfully
participated in content classes in U.S. schools for some years tends to be compromised unless appropriate solutions are provided.

For the same reason, using knowledge from one subject area in the measurement of targeted constructs from another can be more problematic for ELLs than for other students. Cross-content area items are common in English language arts and science, and to some extent within the content areas that make up social studies. Some science tests regularly use mathematics concepts to measure aspects of science; similarly, it is common in the measurement of ELA for passages or writing topics to come from a variety of other subject areas. The solution that high school comprehension tests use is to provide all the necessary non-reading academic information in the passages, while focusing on the measurement of skills associated with ELA (for instance, see College Board reviews associated with the Scholastic Achievement Test, or SAT). The new generation of English language development tests for English learners is tending to follow this example in measuring academic language proficiency for this population (see Kopriva et al., 2004a). Some science items contain mathematics within their text which is sometimes challenging as prerequisite information in some science is mathematics based. Sometimes links are used to provide necessary background information here. It is important for all students that science item writers consider how the measurement of mathematics can be minimized.

Structural Factors

Structural factors are defined as characteristics that frame and guide how language and other stimuli related to test requirements and demands are presented to the test takers. They also include supplementary materials which are item specific. Specifically, this section covers factors associated with sentence and paragraph structure, word usage, use of visuals, use of support materials, item format, and text amount.

Several researchers have investigated the structural elements of items and tests, and many have studied how certain improvements might impact access to the test materials. They have variously referred to the text improvements as using plain language or language simplification, and these improvements have usually included format modifications, visuals and other adaptations. A review of selected work will be incorporated below as the different elements are discussed. In a meta-analysis of research on the effectiveness and viability of adapting assessment text using plain language and related text improvements, Sireci et al. (2003) found mixed results. These authors, as well as several others active in researching this phenomenon, found that gauging effectiveness and appropriateness was more complex than initially envisioned. The following sections will describe the most promising findings.
In all cases, investigations are continuing to unravel what elements and aspects of these elements are key in accessing item demands. Table 6.1 summarizes the elements identified in the factors below.

**Simple Language Structures**

Abedi *et al.* (2000), Abedi (2004), Brown (1999), Johnstone (2003), Hanson *et al.* (1997, 1998), Gaster and Clarke (1995), Kopriva *et al.* (1994, 2001), Emick and Kopriva (2007), Solano-Flores *et al.* (2001, 2003, in this volume), and Winter *et al.* (2004), among others, have all researched how to organize language in assessments for students with limited literacy in the language of the assessment. The issue of organization is particularly salient for English language learners, because text in their home language is almost assuredly structured differently than English text. Test developers need to understand that basic presentation of text for these students is not just a convenience, but in many cases a necessity that, for them, involves a conscious and complex re-appropriation of structural conventions explicitly or implicitly learned as part of their home language experiences (Prins and Ulijn, 1998). As such, the cognitive load is substantially heavier for these students. Limited language acquisition in English, plus literacy issues, is an added factor for English language learners when tests are in English. Further, other cultural, dialectical, and discourse experiences are important issues as well, especially if students have had only limited experience with U.S. mainstream schooling and testing practices, or limited experience with “standard L1” tests in their home language. The four recommendations below have been widely cited by several of the researchers as being essential components of straightforward syntactic structures in items and associated text materials.

**Use of Simple Sentences**

Most researchers agree that item sentences should be kept brief and straightforward. Additional clauses or phrases in English should be used sparingly. In general Kopriva and Cameron (2007) found that ELLs in middle school with lower language proficiency were less impacted by additional clauses or phrases than were students before grade 6. This appeared to be especially true for students at the intermediate level of English proficiency (and presumably above, although the more advanced English proficiency levels were not studied). Grade 6–8 students at the “pre-functional” and possibly beginner levels, as well as younger students with low or intermediate English proficiency, seemed to still be vulnerable to additional language of this type. When additional clauses or phrases are used for older students or more advanced ELLs at all grade levels it is recommended here that they follow canonical phrase and sentence structures (subject noun phrase with optional preceding adjective, followed by a verb with optional following adverb, and then an object noun phrase or prepositional
phrase with optional preceding adjective). These guidelines should be
applied throughout the assessment. Multiple choice stems and distractor
phrases should follow the same formula as simple sentences and be
invariant in their organization. It is also important to note that language
arts items should be held to this same standard although the passages would
not be affected. Gaster and Clark (1995) advise breaking down compound
complex sentences into a number of short sentences, eliminating any
redundant information and unnecessary words. Brown (1999) encourages
item writers to avoid unclear signals about how test takers should direct
their attention and be consistent in providing like information over
items. For instance, in a multiple sentence item, Gaster and Clark (1995)
recommend that writers state the most important ideas first. When time
or setting is an important part of the item, this information should be put
at the beginning of the sentence. Figure 6.5, a released 5th grade item, was
re-written to improve accessibility (Figure 6.6).

Every year green sea turtles swim from Brazil to a faraway island in the south
Atlantic Ocean and back again. This means each turtle must swim a total of 468
miles. Let $n$ represent the number of miles in a one-way trip. Select the equation
that shows how many miles green sea turtles swim in a round trip from Brazil to
the island and back to Brazil.

A. $2 - n = 468$
B. $2 \times n = 468$
C. $2 \div n = 468$
D. $2 + n = 468$

**Figure 6.5** Base item

David walks $n$ feet from home to school.
He walks home on the same road.
He walks a total of 468 feet.

Which equation shows how many feet David walks?
Select the correct equation.

A. $2 - n = 468$
B. $2 \times n = 468$
C. $2 \div n = 468$
D. $2 + n = 468$

**Figure 6.6** Access-based item
USE OF SIMILAR PARAGRAPH ORGANIZATION

The range of organization in paragraphs needs to be restrained and held as constant as possible throughout the test and associated test materials (Wong-Fillmore and Snow, 2000). This includes directions, but excludes any modification of the reading passages. For instance, paragraph structures should all begin with the topic sentence, include one or two explanatory sentences, and end with a closing or summary sentence. While complexity and variation in structure might represent good writing, they make tests more problematic for some English language learners, as well as some other students. This advice is particularly pertinent for constructed response items and instructions where more lengthy explanations are common.

USE OF PRESENT TENSE AND ACTIVE VOICE

The present tense and active voice should be used as much as possible in sentences, stems, distractors, and paragraphs. The rationale is the same as the rationale regarding simple structures in stems, sentences, and paragraphs. Present tense and active voice typically are learned first, when ELLs are learning English in U.S. classrooms. Other tenses and voices very easily can camouflage the measurement intent in items for these students and others with literacy limitations. Rosenberg (2003) cautions that a substantial percentage of the vocabulary that confuses English language learners are words and phrases associated with tenses and voices. Unlike nouns and other words, these cannot be easily looked up in a glossary or dictionary. The one exception to this recommendation is when an item purposely uses the sequence of time as part of the item context, e.g., contains both past and present information, and possibly references to future events or actions. In this case the timing of the respective activities should be made clear. Figures 6.7 and 6.8 illustrate a released item and its access-based counterpart, respectively. In Figure 6.7, besides using past tense and additional clauses, the reader will note that the writer used the verb “pay” and its past tense “paid”. Using irregular past tense verb forms also increases the language complexity of the item and should be avoided if possible.

MINIMIZING THE USE OF REPHRASING

The temptation to rephrase or reword ideas and concepts should be avoided. Paraphrasing words or ideas should either not be done or done with all the original words immediately following in parentheses (often a messy proposition). While repetition is considered poor writing, use of variant language usually makes comprehension significantly more difficult for students with literacy and language challenges.
Kopriva and Lowrey (1994) and Rakow and Gee (1987) note that a substantial amount of confusion for English language learners appears to arise from inconsistency in the text structure, and that this confusion seems to increase exponentially when varied sentence structures and varied paragraph organization options are utilized. Several authors (for instance, Wong-Fillmore and Snow, 2000; Sharrocks-Taylor and Hargreaves, 1999;
and Prins and Ulijn, 1998) caution, however, that overly simplified and compressed sentences may result in omission of important connecting information, or in the omission of important compensatory cues. Resolving this tension is discussed in more detail below under Text Amount.

Writers also need to keep in mind the dominant text structure experiences of test takers. In some cases, for instance when the use of a test is limited to administration for a homogenous ELL audience, most of the students may be from one particular cultural background or even be from communities that share one dialect. They may share discourse features (e.g., Heath, 1983), as in ways of organizing written or oral narratives, or share common pronunciation, grammar, vocabulary or use of phrases and style (Crystal, 1997; Wolfram et al., 1999). In these situations it may make sense to respect the structural patterns of that culture.

Solano-Flores et al. (2003) suggested that test developers use structural linguistics and graph theory in a formal approach to examining the linguistic features of items which tend to be unnecessarily complex. Through using the tree diagram procedure from structural linguistics, they observed that the syntactical structure of sentences in some NAEP items was tremendously complex. Using comparative bar graphs to depict the reading and writing demands of items in two languages, based on many indicators, the authors isolated several structural and semantic elements. In one instance, items differed by 1) simple question and elaborate response, versus high reading demands (long sentences, complex contextual information and syntactic structure) with very simple and short response, and 2) very simple syntactical reading structure with one uncommon word, versus complex syntactical structure and common words. Often, demands are not symmetrical across languages.

Vocabulary

The vocabulary of both academic English and ordinary English must be considered when developing access-based items (Sharrocks-Taylor and Hargreaves, 1999). In general, everyday, non-technical language should be used in problems (Prins and Ulijn, 1998). Formal, academic vocabulary should be used when the terms are part of what the item is measuring; ordinary vocabulary about subject matter phenomena should be used when feasible and when this language can suitably retain the integrity and specificity of the item’s meaning. When possible, the use of words with different meanings in ordinary English and academic English (e.g., odd) should be avoided, particularly if the context does not clearly convey which meaning is appropriate. Farr and Trumbull (1997), among others, urge test developers to be mindful of vocabulary in all items, directions,
and supplemental test materials. Three sets of recommendations regarding the semantic structures of text follow.

**Use of Familiar Language**

Gaster and Clark (1995), Brown (1999), Abedi and Lord (2001), Abedi (2004), Winter et al. (2004a) and others encourage the use of clear, commonly used words in items, directions and associated test materials. High-frequency words are the best choice, and general English word frequency estimators are published and available. However, few, if any, of these are child-specific, and most do not contain regional frequency information. For English language learners, it is important to remember that high-frequency issues are sometimes mediated by cultural experiences. Additionally, it is sometimes important that frequency needs to be considered in context: for instance, pepperoni is considered a higher frequency word than anchovies when pizzas are the topic. As mentioned above, academic language must be used when knowledge of these terms is being measured by a particular item. Otherwise, common language is preferred except when it interferes with the meaning of the item (see Winter et al., 2004b, for more of an explanation).

In general, most authors agree that concrete words are easier to understand because students can form a visual image of their meaning. This is also true of active verbs where students can visualize the action. Shorter words are usually preferred, as long as they convey the proper meaning. Brown (1999) points out that ambiguous words and irregularly spelled words should be avoided if possible.

**Limit the Use of Substitute Words**

Pronouns should be used in a very judicious manner. The same word should be used repeatedly, rather than using pronouns to refer to people, things, or ideas. The complexity engendered by multiple pronouns is needlessly confusing because it is often difficult for students to be certain of the referent. Further, different language backgrounds handle the concept of pronouns differently and make differential use of these words in reference to animate and inanimate nouns. Gaster and Clark (1995) emphasize that, when pronouns must be used, the noun–pronoun relationships should be clear and proximate to one another. If proximity is not possible, Kopriva (2000) suggests that pronouns be followed immediately by the term they reference in parentheses.

Paraphrasing of words should not be done. The same words should be used to refer to the same phenomenon, concept, person, place, thing,
action, or modifier. Students with limited literacy and language skills often know a restricted range of vocabulary, and these substitutions are at best confusing and misleading. Additionally, Prins and Ulijn (1998) point out that using missing or ambiguous referents should be avoided. An example of this is beginning the sentence with “It is . . .”. Figure 6.10 illustrates how paraphrasing and use of pronouns can be minimized. While changes in this access-based item (as well as many of the others) makes the item more “boring”, it is easier to read for students with little proficiency in English.

<table>
<thead>
<tr>
<th>The third-grade students are going to raise money for homeless in a shelter. They want to buy each person a hat for five dollars. What is the largest number of hats the class can buy if they have $90? Choose the correct equation.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. 90 + 5 =</td>
</tr>
<tr>
<td>B. 90 − 5 =</td>
</tr>
<tr>
<td>C. 90 × 5 =</td>
</tr>
<tr>
<td>D. 90 ÷ 5 =</td>
</tr>
</tbody>
</table>

**Figure 6.9 Base item**

<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A. 90 + 5 =</td>
</tr>
<tr>
<td>B. 90 − 5 =</td>
</tr>
<tr>
<td>C. 90 × 5 =</td>
</tr>
<tr>
<td>D. 90 ÷ 5 =</td>
</tr>
</tbody>
</table>

**Figure 6.10 Access-based item**
CAREFUL USE OF MULTI-MEANING WORDS

Colloquialisms and words with double or multiple meanings should be omitted or defined in the text, contextually or otherwise. Multiple meaning words and colloquialisms are used far more than we realize; as such, it is important that someone skilled in recognizing these words and phrases review all text. A few examples of double meanings in science include earth (the planet) vs. earth (soil); plate (as tectonic structure) vs. plate (hot plate); and fault (geologic feature) vs. fault (error). When words with multiple meanings or colloquialisms are used, it is best if the context, visuals or other cues clarify the meaning (Kopriva, 2000; Brown, 1999). For instance, largest can refer to size or quantity, and this distinction can be made clear through the use of a well constructed picture. In some cases it might be appropriate to use a test-specific dictionary, define the words in the text itself (in parentheses immediately following the word or phrase), or provide the definition below the item. Defining terms is not a good solution, however, because it usually adds language load to the text.

Effective Visuals

Teachers of English language learners argue strongly that simple visuals should be used as frequently as possible to facilitate the understanding of what is being asked/presented in a specific item or group of related items (Kopriva, 1994; 2000). Young children learning to read in L1 and ELL learners attempting to read in L2 or English share certain decoding issues related to recognition of words. During testing where reading is the primary communication mechanism, content words (nouns, verbs, adjectives, and adverbs) rather than function words (articles, prepositions, conjunctions, pronouns, and auxiliary verbs) are usually the most beneficial to illustrate. This is because content words contain inherent meaning while function words do not. Rather, function words signal the organization of meaning. Eskey (1976, 2002) pointed out that successful readers simultaneously interpret meaning and decode as necessary as they read. However, Clarke (1980) explained that insufficient language skills can short circuit this process. Thus, unfamiliarity with words in L2 can interfere with the ability to simultaneously interpret meaning and decode. For English language learners, Eskey argued that, in general, the reading proficiency level in L1 will determine the ease of learning to read in L2, since reading skills transfer across languages. For students who are not literate in L1, visual cues (as well as oral L1 translation) can be useful.

Gough (1983; Gough and Griffith, 1992, 1997) explains that the typical child entering a school for the first time where the language of instruction is the same as the language of the home has a vocabulary of about 5000
words. These words he or she knows from hearing them spoken. As this student is learning to read, each of these words must be encountered for the first time visually (what Gough calls a “novel word”). For ELL students, they are both learning the language and learning to read at approximately the same time. As such, their task is much more complex than native speakers. Visual cues provide help for ELL students as they struggle to learn English and become literate.

However, not all graphics are equally beneficial. Illustrations, diagrams or graphs should mirror or replace the text, and not be so complicated as to distract from the intention of the item. Filippatou and Pumphrey (1996) reviewed 30 research studies looking at the effects of various combinations of pictures and titles on reading accuracy and reading comprehension. They reported that pictures are not uniformly effective in all tasks and not all types of pictures are uniformly effective for students with different reading abilities. Rather they emphasize that the effect of visuals is a complicated interaction of task, type of visual, and characteristics of the student.

Figure 6.11 provides an example of an effective visual from a local assessment in California (from Kopriva, 2000). This visual is effective because of the bendable nature of the lamp and clearly separate beakers with distinctly different textures to represent water, soil, and sand. This provides a visual grounding and sequencing for students that parallels the language and expectations in the text of the item. The thermometers underscore the heating aspect of the question (although they could probably

---

**Where on Earth?**

Mrs. Flores’ class has been learning about the position and composition of the Earth and the interaction between the earth and the sun. The following experiment was used in part of their study.

Using everything you know about light, heat and water, how do you think this activity helps explain the uneven heating of the Earth’s surfaces?

---

**Figure 6.11** Visual in stimulus

be represented more clearly). The words “dark” and “light” are used to describe the soil and sand and to facilitate accurate understanding of the item. It is also important that the picture contains no other information that could distract students.

Winter et al. (2004b) used visual elements in mathematics test items which were administered in cognitive laboratories to a sample of ELL students. They found that visuals not only functioned as text comprehension aids, but functioned in other important ways as well. That is, illustrations or diagrams could be used not only to apprehend the intent of the item, but also to facilitate problem solving and response options. Some of the visuals were helpful for representing or organizing information found in the accompanying text or for presenting new information; some were optional to use and others were necessary to access the item. Some visuals were used passively (only by looking at the visual) and others could be used interactively (by drawing or writing on them).

The following seven sets of recommendations outline key points that are important to remember when using visual representations in assessments.

**USE OF RELEVANT VISUALS**

Sharrocks-Taylor and Hargreaves (1999) outlined the three types of illustrations that are found in items and associated test materials. These are decorative elements that have no relation to particular items or text, relevant illustrations that mirror the text to some extent, and essential illustrations not repeated in the items or associated materials. Most researchers and text editors (for instance, Shriver, 1997; Filippatou and Pumphrey, 1996; Yuill and Oakhill, 1991) recommend that visuals should be used to facilitate the understanding of what is being discussed in the task at hand. In this case, the focus would be on the requirements and demands in an item or a group of items, or in the presentation of supplemental information. As such, decorative illustrations would not be appropriate. Illustrations would include related pictures, charts, maps, diagrams, figures, graphs, and tables.

For item writers it is important to remember why the visual is being used. Shriver (1997) and Schiffman (1995) emphasized that it is important to use visuals that, as much as possible, are appropriate and relevant to the cultures and experiences of the test takers and not offensive.

**USE OF AN EFFECTIVE VISUAL FORMAT**

Visuals are often important, but they can be confusing or highly effective based on how they are designed and presented. Overall, they should be kept simple and to the point. In paper text, simple black and white line
drawings are often the most clear and uncluttered; color can be used in computer graphics and illustrations as long as the visuals are also clear and simple, and as long as effectiveness is not obscured. In either case, no information that is supplementary or unnecessary to the intent of the relevant items should be placed in the visual to distract students from measurement intent. When primary document illustrations are used, key information should be made as clear as possible. For instance important aspects can be labeled outside the illustration, or, if possible, some information can be muted relative to target information in the visual (Kopriva and Cameron, 2007).

Shriver (1997) explained that information in diagrams and graphs should include clear labeling, be uncluttered, and only include superfluous content when discrimination of relevant and non-relevant information is part of the measurement target. Silver (1994) and West (1997) recommend putting relevant illustrations directly next to their associated items. One exception to this is when multiple items are illustrated by a single visual. In this case the illustration should occur on the same page as the relevant items. In general, computer generated images should follow same rules (Szabo and Kanuka, 1998). Recent experience is suggesting that in many cases the “same page” rule will have to be adapted when interactive computer-based items are used. In this case, the challenge is to effectively connect different

Gina collects toy dogs and horses. She sees the advertisement below in the newspaper.

![Advertisement](image)

Suppose that Gina decides to buy EXACTLY THE SAME NUMBER of toy dogs and horses. What is the greatest number of each she can buy for $3.00? Show and explain how you found your answer.

**Figure 6.12** Base item
frames of an item in a way that is not confusing to the test taker (J. Bauman, personal communication, October 2007). Figure 6.12 is an example of a released item with a confusing visual. Figure 6.13 uses two visuals the item writers thought would be more clear. It also uses the ‘Explain’, symbol which is discussed below under Item Format, Mixing Symbols, and Text.

**USE OF ILLUSTRATIONS TO MIRROR TEXT**

When visuals are used to mirror text they should parallel the item statements and expectations. Mirroring enhances the items by providing two sources of grounding and sequencing, not just the textual source. The visual should represent as many major parts of the item as possible. While additional, unnecessary information is confusing to students, omitting important pieces of the item can be misleading as well. Simple text corresponding to important words in the item can and should be used in the visuals. As in the example above, using “dark” and “light” in the illustration is important not only because the words clarify which beaker holds soil and which holds sand, but because they are the same words as in the text and they act as a bridge to connect the item text and the illustration.

**USE OF ILLUSTRATIONS TO REPLACE TEXT**

Sometimes pictures, diagrams, or graphs can be used instead of text or in conjunction with minimal amounts of text in the multiple choice stem or in the answer choices. Figure 6.14 is an example. Here the illustration provides the context and text is used in the answer choices. While the item response is rather obvious and particular pictures that are used are confusing (clip art), an illustrator could make clear back and white drawings of four phenomena tailored to the desired content complexity.
FIRST PERSON VISUALS
A very effective way of pictorially engaging the test taker is to reduce the textual distance of the item through the use of first-person visuals with speech bubbles. Developing countries routinely use this method in print materials targeted for persons of all ages (J.R. Monroe, personal communication, October 2004). Figure 6.15 is a 3rd grade mathematics item that illustrates sequencing of information through the use of a single person and speech bubbles at each point. Figure 6.16 is an example of an interactive visual (not very well drawn) where two people are interacting with one another to stage the item context.

USE OF VISUALS TO ORGANIZE EVENTS IN TIME
In some items the context of time is an important element. This can be illustrated through a series of visual frames, as is pictured in Figure 6.17, below. In this example, there is no speech; in Figures 6.15 and 6.16, visual frames sequence events and arrange speech over time, where text would otherwise be used to explain this transition. Yuill and Oakhill (1991) have explained how using visuals can serve such an organizing function. This method is effective in reducing the language load associated with context and in minimizing language associated with movement through time, e.g., “once ‘A’ occurred, then ‘B’ . . .”. This level of language and literacy acquisition is more advanced than what is typically taught to novice and intermediate English language learners.

Sequence the following food chain.
A. Kelp → Sea Urchin → Wrasses fish → Shark
B. Sea Urchin → Kelp → Wrasses fish → Shark
C. Shark → Kelp → Sea Urchin → Wrasses fish
D. Wrasses fish → Kelp → Sea Urchin → Shark

Figure 6.14 Visual in stem

Figure 6.15

Sequence the following food chain.
A. Kelp → Sea Urchin → Wrasses fish → Shark
B. Sea Urchin → Kelp → Wrasses fish → Shark
C. Shark → Kelp → Sea Urchin → Wrasses fish
D. Wrasses fish → Kelp → Sea Urchin → Shark
I have 9 old shirts.

I buy 6 new shirts.

Then, I give my sister 4 shirts.

How many shirts do I have left?
Choose the correct number phrase.

A. 9 − 4
B. 9 + 6 − 4
C. 6 − 4
D. 9 − 6 + 4

**Figure 6.15** Simple first person visual

I want to buy the cake.

You have $7.00. $7.00 is \( \frac{7}{10} \) of the price of the cake.

How much do I need to buy the cake?

A. $1.75
B. $7.25
C. $11.00
D. $28.00

**Figure 6.16** Interactive first person visual
One additional function of using visuals is their ability to cue or focus the reader on certain aspects of text (Wong-Fillmore and Snow, 2000). This is a problem if the visual signals the correct answer; however, this function can provide clues that frame the item in the appropriate context. A picture in social studies that depicts persons in period dress could mirror the dates or regions that are referenced in the text. Likewise, visuals can clarify the proper meaning of language when multiple meanings are possible. This is often useful because the vocabulary of English is replete with multiple meanings. Given the context, meanings that would be clear to native speakers are routinely confusing to English language learners, who have only learned certain definitions to date.

**Item Format**

Formatting print material to focus the reader, clarify purpose, and otherwise effectively move the reader through the information is central to the work of print media professionals. In assessments, format is often dictated by cost constraints and traditionally has not been viewed by test publishers or agencies as being a major factor in impacting access for English language learners. Advocates for students with disabilities have argued, however, that formatting is important for students with attention and visual orientation issues, as well as students whose processing demands and decoding...
struggles result in literacy and language challenges (e.g., see Filippatou and Pumphrey, 1996; Grise et al., 1982).

While the language and literacy challenges of English language learners do not originate from processing or other deficiencies, emerging work suggests that attending to and clarifying item formats does play a part in making items more accessible for this population (Winter et al., 2004b; Kopriva et al., 2007d). Use of formatting strategies tends to open up items in much the same way that introducing effective visual representations does. Eight sets of strategies will be briefly explained below (Kopriva et al., 2007d).

**SEPARATING KEY IDEAS**

Many of us who read a lot of text would agree that separating ideas aids our ability to easily identify key points. Separating information using bullets, including a line or more of blank space between one idea and another, using focusing frames to set off key points, or interspersing text and illustrations are all effective ways of dividing and handling the presentation of text. One advantage in using bullets or other kinds of lists is that words or phrases can be used, rather than complete sentences. This reduces the language load and minimizes surround words and phrases so that the test taker can focus on the important elements. Blank space helps increase legibility (Smith and McCombs, 1971), and anchors text or portions of the text on the paper within the item (Menlove and Hammond, 1998). Interspersing illustrations within the item text has a similar effect of placing tables and figures near discussions of these elements in books and articles. It makes it easier to move the attention back and forth between the two print forms without losing the train of thought in the dominant form of communication.

Many of the items shown above illustrate a format that appears to be more accessible. Besides sentences being shorter, the context sentences are purposely separated from the questions and information about the question (e.g., “Choose the correct equation”). Sometimes the separation occurs with a larger space, sometimes by inserting an illustration. In Figure 6.19 mathematics experts and item writers used bullets to re-phrase and present the information found in the base item (Figure 6.18).

**CLEARLY IDENTIFYING ITEM QUESTION**

One key part of the item that needs to be set off from other parts is the question or task demand to which the students will respond. Often items do contain two or more sentences, one focusing on the item question. On the other hand, “one sentence items” may combine context and item question into a single sentence. Such sentences need to be split into two to give the context in one and the question in the other. In either case, students with literacy challenges, such as English language learners, could profit from
Your third-grade class is planning a picnic for the end of the year. You are in charge of buying cookies. There are 27 students in the class. Your teacher and two parents will also be going to the picnic. How many cookies will you need so every person at the picnic can have two? Show or explain how you found your answer.

**Figure 6.18** Base item

How Many Cookies?

These people are going to the class picnic:

- 47 students
- 4 teacher
- 14 parents

Each person will eat 5 cookies at the picnic. How many cookies will be eaten?

**Figure 6.19** Access-based item

a very straightforward sentence that contains only the question, and this question should be set off with blank space above and below it. In the situation where other information is presented in an indented form in a “second level” position (for instance with bullets), the question should be presented in the “first level” position.

**Use of Titles**

Yuill and Oakhill (1991) explained that the use of titles on items is a textual and visual organizing element. In reviewing previous studies which investigated the effects of pictures and titles when students are reading passages in the classroom, Filippatou and Pumphrey (1996) found that these elements impact reading comprehension and accuracy to different degrees. Titles can be helpful if they are relevant and focus the reader on the intended key points. Kopriva et al. (2003) found mixed results using titles, and suggested that they may be particularly useful when items are longer than a sentence or two. Typically they suggested that titles are a way to alert the test taker to the context within which they place the item question. This encourages the student to form a picture in his/her mind without as many words in the body of the item as would otherwise be required. Figure 6.19 is an example of an access-based item where a title is used to focus the reader on the intent of the question.
**Mixing Symbols and Text**

Using a discrete set of symbols throughout the assessment to alert students to certain information is one way to reduce the language load (Monroe, 2003). Winter *et al.* (2004a) used an “explain bubble” (seen in Figure 6.13) that is discussed in the directions before testing begins. It represents that students can explain their answers using words, phrases (as compared with complete sentences), pictures and/or diagrams. It has been suggested that including the bubble on each of the relevant constructed response items (including items measuring reading comprehension) reminds the student each time. This can be more beneficial than only telling them at the beginning of the test time, especially if these kinds of options are not typically allowed or explicitly encouraged on other formal large-scale tests. Prins and Ulijn (1998) suggested that mathematics symbols, such as +, =, >, <, etc., may be useful as shorthand ways of providing information across content areas. These symbols are typically well established symbols the students understand and have been used effectively by advertising and media outside of mathematics settings.

Other symbols may include stick figures of males and females, faces that are happy, sad, etc., and the circle containing some element with a diagonal line drawn across the circle indicating that the element is not allowed. Computer symbols may also be useful, and can be used on print tests as well as computerized assessments. On tests that are administered electronically, accommodations may be built into the pull down menus or otherwise made accessible for students on the entire assessment. Likewise, stand alone symbols on the computer, such as a mouth, can trigger an oral reading of the item (Kingston and Ehringhaus, 2005). Other symbols can trigger a desire to hear the item, phrase, or word translated, or the desire to see a picture that corresponds to a given word. Any symbols that are used test wide should be discussed prior to when testing begins.

**Use of an Example**

Monroe (2003) suggested that providing an example on certain items is a technique that teachers of beginning English language learners often use. This appears to work particularly well on “action” items, items requiring a few words but several steps, or items where the explanation can be easily visually reproduced on a paper and pencil test or simulated on a computerized version. Grise *et al.* (1982) explained that examples are also useful for students with learning disabilities.

**Highlighting Key Words or Phrases**

Educators of students with disabilities have long advocated for bolding or somehow judiciously highlighting words and phrases in text items for
certain students. Monroe (2003) and Farr and Trumbull (1997) suggested that this is a strategy used in the classroom for English language learners as well, and that it appears to be useful in focusing these students on key considerations of the item. In general, the guideline is to use highlighting lightly, to either focus on an exception that is the topic of the question (e.g., What student is NOT in class today?), or set off the key words in the question portion of the item. Figure 6.13 uses capital letters to focus the student on the question. While there is no evidence of the effectiveness of this technique in isolation, two research studies (Emick and Kopriva, 2006; Kopriva and Cameron, 2007) have used this approach as well as others in producing accessible items. The studies both proposed that, assuming the item type is accessible, this technique, as well as others, seems to be useful for communicating the item requirements to many ELLs. Kopriva and Cameron also found that older and more literate English learners were able to navigate items with this support even when the questions used more complex language and sentence structures.

**USE OF BOXES OR LINES**

Boxes which frame text or lines which define where students should write their answers should be used in a judicious way. These formatting tools are often useful in organizing item text or response space in assessments. Care must be taken, however, to avoid distraction with fancy borders, and the purpose of the boxes or lines must be clear. For instance, in a cognitive lab study students were asked to “provide your answer below.” The item was framed in a box to draw the students’ attention to it and white space below presumptively signaled where students were supposed to compose their answer. However, several students tried to squeeze their answers into the item box (Winter et al., 2006). Lines can suggest that students must write their responses but then they may inhibit students from answering with diagrams or other visuals. Some students may also interpret response lines as the requirement to fill the entire space. But the line “issue” is a two-edged sword as leaving white space may be too ambiguous. Educators should guide the spacing requirements: the number of lines and/or amount of white space should be part of the review discussions. Figures 6.10 and 6.14 are examples of how boxes might be used.

**Text Amount**

Critics often suggest that, for English language learners who, by definition, do not have the language or literacy to sustain large amounts of text in English, focusing almost exclusively on the reduction of language in a text is the most humane approach. With this, they laud the benefits of mathematics items with all language stripped away (e.g., 2+2=?), and praise
multiple choice items where there is usually less language than constructed response items. However, Wong-Fillmore and Snow (2000) argued that the problem context or linguistic conventions which allow “shorthand” communication may not be enough to provide the meaning of specific words for ELLs.

This issue is not a simple one. Besides advocating for thoughtful use of language in testing, Wong-Fillmore and Snow also discussed the importance of not “dumbing down” tests or otherwise restricting the types of evaluations of content for the ELL school population. The reason multiple choice “word problems” have gained favor in large-scale test construction is because they measure the acquisition of more complex content skills than “no language” items like the example suggested above. For the same reason, most measurement specialists emphasize that constructed response items (extended as well as short) measure different skills or skills with a greater range of complexity than multiple choice items often can. If tests for English language learners include a greater percentage of “no language” items than those administered to the mainstream population, if they use a more restricted set of item types, or if the tests focus to a greater extent on the simpler, less complex (and therefore perhaps more language intensive) multiple choice word problems than those administered to the mainstream population, the assessments will not be measuring the same constructs. This is dumbing down the test and it poses a serious comparability problem. But perhaps more importantly, history suggests it leads to restricted placement in classes where more advanced coursework is offered (Huebert and Hauser, 1999; Lara and August, 1996). Finally, if amount of language is the chief criterion used to determine the access of test content, one is promoting a faulty guide for the use of L1 tests when these are appropriate. It is well known among most ELL specialists (for instance, Rivera and Stansfield, 2001), that Spanish text is about one-third longer in length than text in English that addresses the same concepts. Since Spanish is the most common language used if L1 tests are considered, this bar could have unanticipated consequences for students who are literate in this language or for promoting the standardized use of oral administration of a Spanish form.

Empirically, Kopriva and Lowrey (1994) found that a large percentage of ELLs in California said they would rather have an open-answer format as compared with multiple choice format for providing their responses. They stated that this format provided them with the opportunity to explain what they know. This preference was confirmed by Kopriva et al. (2007b) who found similar relationships between constructed response test scores and a criterion measure for both beginner ELLs and native English speakers. This was sharply contrasted to the relationships between the scores on multiple choice items, where the squared correlation for beginner ELLs at
two grade levels was about .02, while it averaged about .24 for non-ELLS. The authors interpreted the constructed response findings as evidence that similar validity inferences could be made for these two groups on this test, but that the inferences could not be sustained for the multiple choice results. Further, Kopriva and Cameron (2007) consistently found that the distractor frequencies for the multiple choice items were substantially different for ELLs than for non-ELLS for a majority of multiple choice questions on a statewide test which spanned six grades and four subject areas. While the results of these studies suggest that there is clearly a response opportunity issue in multiple choice items, the findings also seem to suggest that language in the text of items or in how an item allows students to respond is not necessarily problematic. Rather, Kopriva and Cameron (2007) found that it is important to retain enough context in the item to ground the question, and enough specificity in the item requirements to focus the students on the intended target. A newly funded project (Boals and Cranley, 2006) is investigating how context might be presented largely through animation and simulation. However, the researchers note that much of the specificity in the item requirements needs to be communicated through supported language (J. Bauman, personal communication, October 2007).

The findings from Kopriva and Cameron and the approach being taken in the Boals and Cranley research seem to reflect the concerns identified by Wong-Fillmore and Snow. Thus, this means limiting use of colloquialisms, idioms, pronouns, paraphrasing or use of vague phrases that relate back to other information—these and other indicators are discussed above. Additionally, it appears to be important to maximize the use of collateral information in items in order to extend the item text or provide context. In this way the item can primarily focus on the question the students are supposed to solve. This suggests increased use of visuals or other media in order to provide context or mirror it, and can be used in other ways as discussed above. Because of their different life and schooling experiences, the amount of text in an item can be minimized if other information can supplement it. Supplementary information is especially reasonable if and as computers can be used in testing these students. The guiding principles in making decisions about text amount is to retain the content complexity of the intended target, clearly and precisely articulate the target in the question (most often using language but supporting it effectively), and then provide enough information as possible without using text, or at least unsupported English text, in order to create a rich set of contextual cues.

**Tools**

Besides item adaptations, some structural elements are supplementary to what is presented in the test booklets, and Chapter 7 will summarize
some of these. These include text supports such as bilingual glossaries, monolingual English, synonym lists, word-picture lists, and other related text-based aids. Issues of generic tools versus test-specific tools, and which tools might be most effective for which students are briefly highlighted. Besides text-supports, the role of objects which can be manipulated by the student to support problem solving is addressed. Also explained is how computerized test designs with this feature might afford a broader range of opportunities and greater access for some students. The chapter outlines the need and possibilities of providing content-based resources for students who have not had full exposure to item content, or the themes on which items are based. Finally, considerations of standardized activities are introduced. These types of activities could provide data or other types of information which are used to respond to items, but do not interfere with the measurement of the intended targets.

Beyond text support, many tools have traditionally been used in testing as stimuli for the evaluation of more complex cognitive skills. They are found most often in more lengthy large-scale tasks and in classroom assessments. For English learners, ELL teachers suggest that the use of content tools serves another function, one that is essential for many of their students. Just as visuals support, extend or scaffold text, tools such as those explained in the chapter provide compensatory opportunities for these students to cross over the minimum threshold for understanding key item elements or being able to navigate problems.

**Impact of Home Language**

Some item-drafting issues for ELLs reflect the influence of students’ native language. After a point, there is little that item developers can do to alleviate some of the problems posed simply by the wealth of languages and cultures represented by ELLs. Students who speak languages which occur less frequently in the state or nation may find some items less accessible to them, as item writers will never be properly fluent in all languages or have a deep knowledge of the cultural discontinuities between the U.S. and other countries. For more frequently spoken languages, and for those particularly prevalent in certain areas, some actions can be taken to minimize misunderstandings.

A few examples of confusing linguistic difficulties include the use of words or phrases which mean something different when translated from the student’s first language. Rules of syntax or word order that differ in English from a student’s home language can be another source, and it is not uncommon that some words will mean something different in English than what is spoken at home. Different symbols used to indicate numbers or numbers operations occur in different parts of the world; for instance,
periods are sometimes used instead of commas in mathematics in writing numbers: 3.001 is three thousand and one, but we would read it as three and one thousandth. In English the “3.5” is often written with the decimal point at the midpoint between two numbers rather than at the bottom. This could be confused with the symbol for “dot product” in the American mathematical notation for multiplication. In another example, a “billion” is translated numerically into 1,000,000,000,000 in some Latin American countries. We would call this number a “trillion.” Finally, there is often confusion related to different monetary systems. Use of monetary words, such as dollar or peso, may mean different amounts depending on the country. More examples of the types of problems that can arise can be found in Chapter 11 and Solano-Flores and Trumbull discuss several related issues in Chapter 8.

For addressing issues such as these, test developers have only limited options. Reliance on cognates that occur across languages is a strength many English language learners enjoy, especially those whose languages are most similar to English as spoken in the U.S. For item writers, knowledgeable staff or ongoing contact with experts is the first choice in alleviating some problems. Training by L1 or linguistic experts can be very beneficial in pointing out some more common language cross-overs, and these specialists can be an excellent source of referrals to more detailed information. Researcher item writers at the University of Maryland found that making frequent reference to the bilingual dictionaries of the most frequent languages helped them to choose specific words or phrases over others which were most similar to those found in the referent languages (C.S. Chen, personal communication, March 2005). Finally, because of the complexity of this issue, reviews of text by those familiar with the languages and cultures are essential. Regrettably, many reviews which involve these participants are narrow in scope and do not focus on an adequate textual evaluation. This oversight will be discussed more in Chapter 7. A more detailed discussion of these important issues is outside the scope of this book.

A Note about Rubrics

Linda Darling-Hammond (1994) encourages the position that all writers of items (for classroom and large-scale uses) need to think about how students should and might respond to item requirements, and, then, determine which responses are appropriate or not, and why. For open-ended, constructed response items, judgments about these considerations should be mirrored in the rubrics and associated rubric notes. (Notes often reflect how scorers might differentiate between responses, and usually
provide a number of examples of acceptable and non-acceptable responses.) Challenges for English language learners generally revolve around not only expectations of writing ability (generally in English), but also issues associated with language acquisition development, differential experiences, and conventions and stylistic preferences associated with the home culture and language (Lachat, 1999). The concern is that rubrics or rubric notes will 1) explicitly or implicitly confound the sophistication of a student’s writing abilities with the student’s response associated with the non-writing target; or 2) attenuate the appropriate response options to conform with experiences or conventions associated with native speakers who have mainstream schooling opportunities. Mabry (1999) cautions that

In cases in which the overall effect of a student performance is achieved by means not anticipated in the scoring criteria, critical analysis of the quality of writing will deflect a scorer’s attention away from the actual writing, and the score may not support valid inferences about the student’s achievement. (p. 675)

Sometimes, even when notes are clear and well prepared, student responses will not conform easily to expectations. In this case, rater access to target objectives of items is important, and care should be taken to separate out literacy and other contextual influences that may inadvertently narrow how scorers evaluate a response. Toward the beginning of the chapter, Figure 6.3 illustrated a response by an ELL student which was unexpected by the item developer and reflects the students’ distinct cultural experiences. As noted above, the rubric notes expected a response depicting a U.S. water distribution plant and associated pipelines and storage facilities. While this response was not considered to be correct, it may be argued that the response may have addressed the intent of the item if the objective was that the student should understand the nature of water sources and the sequential distribution mechanisms a community uses to provide water to its members.

Explicit references to writing ability in non-writing rubrics have decreased as content specialists have thought about how students from diverse backgrounds might address item requirements. Ensuring Accuracy . . . (Kopriva, 2000) reported that the mathematics achievement standards for the 1992–1998 National Assessment of Educational Progress included literacy expectations in the bottom two proficiency levels (Loomis and Bourque, 2001). Sometimes, expectations that impact English learners have occurred explicitly at the higher score levels where students are expected to respond with a certain level of sophistication (e.g., literacy sophistication or evidence of deeper, subtle knowledge of a value, experience, or perspective
which is more common in the mainstream U.S. experience or belief system than in experiences or beliefs in other cultures). More common is the implicit and largely unconscious attitude of many well educated, well intentioned item writers and scorers that less proficiency in language is associated with less proficiency in non-language content areas. Sometimes this attitude finds its way into rubric notes; it frequently can be found in the examples chosen to illustrate acceptable and non-acceptable responses. Why is this important? First, rubrics or notes such as these limit the number and kinds of recognized ways in which some students must communicate knowledge and skills (unless this is explicitly what is being measured). Second, they tie the quality of written discourse, value, or experience to the quality of content knowledge or skills.

Beyond the assumed sophistication of writing skills or cultural nuance is the tendency for rubric developers to restrict the types of responses that receive high scores. Mabry (1999) pointed out, that while writing rubrics are pivotal in operationalizing systems for scoring constructed-responses in large-scale tests, they also standardize and, by definition, constrict, writing styles. This is true for all students but particularly problematic for students with non-standard backgrounds. As such, rubric issues frequently result when answers from students with different values or experiences are not anticipated or judged to be adequate because of the students’ uncommon approach to solving the problem or explaining their answer. Their responses might result from prior experiences that are quite distinct from mainstream U.S. experiences. For instance, a middle school large-scale science item asked students to define elements that would promote a healthy lifestyle. The test developers determined that proper diet and exercise were key to a healthy lifestyle and these variables were identified as necessary elements to include in the responses for students to receive the higher scores. Many English language learners from developing countries independently responded to this item by highlighting such issues as the need to have clean water, enough food to eat, and food free from toxins. Because of the constraints in the rubric and notes, these answers were not given credit. One can argue that those students understood “health,” and that the points they raised are central for promoting health and a healthy lifestyle.

Improving rubrics focuses on freeing them from improper literacy and cultural interference. What is considered proper or not should rest solely on the explicit explanation of the target being measured, and in the notes care should be taken to use this target as the primary reference point. Rubrics and notes should be developed and considered as items are being written, and they, as well as the items, should be vetted. Try-outs of items (and rubrics) should collect responses from representative groups of ELLs;
in-house and external reviews of rubrics and responses should include educators who are familiar with different sets of issues relevant to the predominant languages and home country cultures of the English learners being tested. Lachat (1999) emphasized that, typically, only a small number of educators participate in the design of large-scale scoring rubrics, and they often do not have the experience with the diverse groups who would tend to respond unconventionally. Try-outs and evaluations should make sure that English learners have access to all score points, making a point to include students with the range of content knowledge and skills associated with those points. Finally, examples of work from ELL students at all score points should be included in the notes or other scoring materials. Solano-Flores and others (Solano-Flores et al., 2001) developed an annotated sampler of elementary ELL responses that cover the entire score range. This may be helpful to developers to emulate as they become familiar with preparing materials with this criterion in mind.

Just as there must be multiple avenues for access in how item requirements are presented to students, it is important that the design of the constructed response items and rubrics allow for different approaches to demonstrating mastery. Klein et al. (1998) argue that items and rubrics should allow non-written information from students to augment or replace written responses. This information might include non-text paper and pencil responses such as charts, diagrams, pictures, and algorithms. If possible, students may also respond orally or perform the solution to a task (such as actually producing a chemical solution which would solve a chemistry problem). Further, since improper literacy expectations usually arise when developers rely too heavily on written responses, it must be explicitly pointed out to item writers that language sophistication and content knowledge are separate and independent. The point should be illustrated with examples of appropriate responses. Independence may be true even in evaluating writing, where the scope of the vocabulary, for instance, might be judged separately from rubrics targeted to assess sentence structure or the ability to cohesively build an argument.

The differential impact of student experiences or value systems needs to be anticipated before tests are implemented. Item writers, along with reviewers, need to be alert to the influence of culture on responses. This influence needs to be written into the rubrics and notes so that scorers will know how to give proper credit to ELL student responses.

Breland et al. (2004) examined the influence of prompt type in writing (persuasive vs. traditional SAT prompt) on various ethnic groups (Asian American, African American, Hispanic, and White). The study found that, although there were significantly different scores (differences ranged from
1/3 standard deviation to 2/3 standard deviation) between ELL and non-ELL students, these differences did not appear to be related to the prompt type. The study suggests that careful planning with the goal of improving the design of items and rubrics can affect how English learners are evaluated on constructed response items.

This chapter has addressed specific factors associated with how principled item development might increase access for English language learners and others with language challenges. The next chapter will focus on other aspects of test material construction and provide some suggestions about how developers might combine this information over items.
Developing access-based items is important but not sufficient in providing access for a number of English language learners (ELLs). Certain English language learners need compensatory support beyond that found in the test booklets in order to interpret and respond to the item requirements. As such, support tools and test form considerations are also a part of making tests accessible. No matter how talented the writers or those responsible for the various aspects of test construction are, proper use of in-house and external reviewers is essential as well. This chapter will address these additional aspects.

Tools

As tools are defined here, they include item or content-specific aids used to help the students understand the intent of a particular item, solve the problem or express their solutions. Such tools include text supports, such as a glossary or side-by-side form, physical aids such as mathematics manipulatives or objects associated with a specific scientific query (for instance, paper cups and string). They are sometimes supplemental supports that seek to reduce prior knowledge gaps. Tools can be computer-generated, may sometimes be used in interactive settings, or may be used to demonstrate responses in a non-language format.

In large-scale tests, tools have traditionally been used by test subject area specialists as stimuli for the evaluation of more complex cognitive skills of
all students, and they are found most often in more lengthy large-scale tasks and in classrooms. The literature documents that teachers of English learners routinely use supplementary aids differently than general education teachers, however, because they routinely use them to teach and evaluate English and academic proficiency of subject matter at all levels of cognitive complexity (for instance see Malcolm, 1991; Farr and Trumbull, 1997; Monroe, 2004). Just as visuals support, extend and scaffold text, additional textual L1 information mirrors item text, or tactile interactions with objects allow students to connect language to their learning and to extend their language when they are being assessed.

ELL educators suggest that one reason the tools are used differently in academic assessments for this population is because the students have to focus on both the English and the item requirements and separate out their language issues from the content problem they are being asked to solve. They also point out that different cultures use what we call supplementary supports differently; some use them routinely when learning, problem solving or demonstrating their knowledge. The U.S. culture, on the other hand, relies more heavily on written explanations that assume certain common cultural perspectives and experiences.

Very little research has been done to empirically determine how tools might play a role in large-scale assessment for English learners. In the work mentioned below, most of the findings are based on a model that looks at the benefit of tools (and often other accommodations) for students with a range of English proficiency and backgrounds. The results are mixed, although there is some indication that the addition of tools for ELLs is a viable concept. Mixed results may reflect research design flaws. First, in most of the studies students were not assigned to a group based on need or differential experiences and several researchers have noted this shortcoming (e.g., Hofstetter, 2003; Abedi, 2004). Hipolito-Delgado and Kopriva (2006) investigated differential need and found that it was an important interceding variable in understanding the utility of tools, as well as other accommodations for individual students. Second, Tindal (1998) suggests that, as tools are usually a part of an accommodations package, it is the package that should properly the unit of study. It will be intriguing to understand the usefulness of tools when need, as well as other design elements, such as package selection, is taken into account.

Classroom research (e.g., Farr and Trumbull, 1997) suggests that selected tools are particularly promising for large-scale use and may provide effective compensatory support if they are correctly built, targeted for particular purposes, and matched appropriately with student needs and backgrounds. Because of this, four kinds of tools will be discussed below: text supports, manipulatives, content-based resources, and activities. Some considerations
of empirical findings, construction, use, and student assignment will be briefly noted in each section.

**Text Supports**

Text supports are identified as text-based aids that occur over and above the item text. Three textual aids will be summarized below: bilingual and monolingual glossaries or dictionaries and picture-word lists. Four kinds of related supports will be briefly mentioned also. The primary drawback of word-based text supports (besides the cost and time to properly develop) is their limit in targeting words. Rosenberg (2003), a linguist and former ESOL teacher, explains that many ELL students struggle with derivations of words, for instance tenses of verbs. These derivations *per se* can not easily be located in dictionaries without knowing the “root” word. This makes the supports of limited value. Additionally, word-based text supports typically do not define sets of words or phrases whose meaning supersedes the meaning of the words as they are separately defined. Colloquial or idiomatic phrases, and non-target words and phrases used as metaphors, are also problematic and are usually not clarified in glossaries or dictionaries. Finally, more abstract terms are not easily defined and so it is not always clear to what extent a lengthy definition with its attendant language load is beneficial for ELL students. Abstract words and ideas are most often found in older grades, particularly in some content areas; this makes the use of these supports more problematic for these ages.

**Bilingual**

Bilingual supports include both glossaries and dictionaries. Glossaries are typically a translation of the targeted word while dictionaries include a definition of the term. These supports can be item-specific, test-specific, or general aids used in the classroom. Because of the obvious problem of providing targeted vs. non-targeted test information, classroom or full-scale dictionaries and glossaries are generally not a good idea. Test-specific bilingual supports can be cumbersome but may offer some relief. Various publishers and some states and researchers are studying the use of computer-generated bilingual glosses of identified words within items.

Abedi (2001) and Abedi *et al.* (2001a) looked at the use of a bilingual glossary and bilingual dictionary, respectively. Neither of these studies showed a significant difference for ELL students, although, as mentioned above, student needs were not differentially taken into account. Neither were student limits—that is, if students were not literate in their home language a dictionary or glossary would not be useful to them. Hipolito-Delgado and Kopriva (2006) included computer-generated bilingual glosses words among other accommodations in their randomized study...
which researched the benefit of receiving appropriate vs. inappropriate or no accommodations. Because of a small total \( n \) (272), they were not able to evaluate the effectiveness of any of the particular accommodations for the students who needed them and could benefit by them. However, the study did find that students scored significantly higher when they received accommodations based on need and ability to use, including the bilingual translation accommodation, than when they received accommodations not based on these criteria. This suggests that appropriate accommodations can play at least some role in improving scores, and perhaps improving the validity of the test results.

**Monolingual English**

Abedi *et al.* (2000, 2001a), and Albus *et al.* (2001), studied the use of an English dictionary, while Abedi *et al.* (2000, 2001b) also examined the impact of an English glossary. Abedi *et al.* (2000) and Abedi *et al.* (2001a) showed significant gains for ELL students who received dictionaries versus those who didn’t, while showing less gain for other students; other studies found insignificant differences only. This is an intriguing accommodation for less-frequently used words that can be easily defined. Questions remain about the use of more elaborate definitions, although they may be useful for students with a level of English literacy to take advantage of the additional language load.

**Picture-Word**

This accommodation is designed to provide a pictorial representation of selected words in items. Ongoing work in South Carolina suggests that pictures work best for nouns; abstract concepts, many verbs, or verb or adverbial phrases often are difficult to gloss with a picture (Carr and Kopriva, 2007).

Kopriva and Mislevy (2005) hypothesized that the students who benefit from this accommodation are ELLs with very limited English language proficiency who cannot benefit from English word-based support (because of their limited language knowledge) and also cannot benefit from an L1 word-based gloss as they are not literate in their native language. The second accommodation utilized in Hipolito-Delgado and Kopriva’s study (2006) was the picture-word support where, when Spanish-speaking students clicked on selected words during a computerized mathematics assessment, a picture of the word popped up. Like the bilingual option, these researchers found that when students who needed this accommodation received it, they performed at a higher level. Since the logic of accommodation assignment in this study was the same as in Kopriva and Mislevy, it appears that this accommodation may be most appropriate for students who are not literate...
in their home language. One remaining question not specifically studied in either study is how useful this option is for those students who are clearly limited in their English but literate in their home language and for whom there is no L1 glossary available. Anecdotal evidence in Kopriva and Mislevy suggests it may not be too useful. While probably not optimum, it will be important to continue to find ways to provide support to this population as well.

**Related Supports**

Four kinds of related supports are discussed in later chapters. First, the effectiveness and issues associated with a test form in the student’s native language (probably translated from English or otherwise made comparable to the English forms) will be discussed in detail in Chapter 8 and briefly under *Form Considerations* below. This seems to be an effective support if the student has acquired the appropriate level of academic proficiency in their first language, because usually he or she has been instructed in those concepts in L1, here or in their native country. Second, dual language forms in both English and a student’s home language will also be addressed below. Oral administration is the third related support discussed in this book. Issues associated with oral administrations in L1 and English will be discussed in Chapter 9. Fourth, the use of a bilingual support person will be discussed in Chapter 9 as well. Untrained and often not monitored properly, there is a fine line between when this latter type of support may be helpful and when it may provide inappropriate advantage to students. However, the interest in trained and monitored bilingual support personnel has resurfaced and may be effective for early learners of English, particularly when no L1 text or L1 oral support for the early ELLs is provided. The benefit seems to reside in providing assistance for language that is not glossable, easily located in dictionaries, or cannot be replicated visually in two dimensions.

**Manipulatives**

As defined here, manipulatives are objects used during testing in several content areas to help students solve problems. Further, they sometimes are used to understand the question, and/or demonstrate solutions. The objects can be concrete (e.g., blocks or sand) or they can be computer simulated so students can use them as they would use physical objects. As mentioned above, typically their role in large-scale assessment has been to enhance the opportunities of test takers to solve more complex extended response items. For ELL students they are used in classrooms as compensatory support to supplement or replace text in items with a wide range of skill complexity.
Unfortunately, very little research has been done to determine the importance and effectiveness of manipulatives in large-scale testing for English learners. Emick et al. (2007) found that 36 percent of grade 3 and 47 percent of grade 5 ELL students who had already been receiving language services in U.S. schools for some time were reported by their teachers to need (as compared with prefer or not need) the use of tactile or kinesthetic supports on large-scale assessments. This number jumped dramatically in each grade for ELLs who more recently qualified for services or who lacked testing experiences like those found in U.S. schools. For this latter group, 50 percent of third graders and 60 percent of fifth graders were reported to need tactile or kinesthetic support. These results are compared with 7 percent and 18 percent of exited ELLs in grades 3 and 5, respectively, who teachers stated as needing this type of help. Interestingly, teachers identified 14 percent and 27 percent of the native English speaking population as needing tactile/kinesthetic support as well, though this may have had something to do with the makeup of the sampled schools, which were mostly defined as at risk. The numbers suggest that tools such as those discussed here may be beneficial as compensatory support, not only for the English learners, but for some students in the broader population overall.

Computer simulations provide probably the most reasonable opportunity for identified students to receive this kind of compensatory help. Drag and drop options, simulated manipulations of “objects,” and computerized graphic and drawing tools provide support by allowing students to manipulate non-text elements in order to solve problems. In many cases the graphic results from the computer-based solutions can be captured and then scored, which greatly increases how students are able to demonstrate what they know without using language. Further, additional elements, such as number of solution paths taken to arrive at the final answer, can be captured by computerized predetermined algorithms. This approach has the potential of increasing formative information as well as providing summative data to educators. Hipolito-Delgado and Kopriva (2006) utilized some of these tools in the constructed response questions on their computerized mathematics assessment. To date, however, the analyses on these data have not been completed.

Thinking about which manipulatives to use and how to use them is not a post hoc activity to be designed into assessments after they have been developed. Like many of the other supports, this option needs to be considered as items and tests are being developed. Clearly, research is needed here. However, because of the prevalent use of objects as compensatory supports in the classroom, it appears that this type of tool should be considered for this population.
Content-based Resources

Chapter 2 explains that English language learners are significantly more likely than other populations to have interrupted schooling or come to school with limited past schooling experience. Further, even when students have had sufficient education, their learning may not have covered content assumed in U.S. schools or where they are currently living. Whatever the reasons, these students often have learning gaps, and prior knowledge within or across years and within or across content areas is common. This poses a challenge for test developers, particularly in tests for older grades, and particularly in assessments of content that build upon knowledge from earlier grades or from other disciplines. Tests of mathematics have addressed this issue to some extent when they allow students to use calculators when computation is not being measured. On the other hand, the problem appears to be the most acute in the sciences and in social studies. Besides building on more fundamental concepts learned in earlier grades, science schooling and tests frequently use and build upon mathematics concepts or algorithms. Likewise, individual social studies content areas often utilize knowledge from other disciplines within the larger social studies framework, and from information taught in past years.

English Language Arts assessments are affected as well. Further, assumptions about students’ understanding of idioms and colloquialisms that occur in passages assume a range of U.S. cultural experiences. Even when the skills and concepts associated with idioms and colloquialisms are being learned in particular grades, assuming that the students have been exposed to the breadth and cultural richness of idiomatic language is very problematic. Finally, certain ELA genres, for example poetry, by definition focus on the reader’s ability to understand the metaphoric or implied meaning of words or contexts. For English language learners, this is impossible if they do not know the regular definition of the word or recognize the context. In this situation, there is an assumption that students know the meaning of the word or have experience with the settings; these are not what is being measured. If not careful, lack of language knowledge here can be inferred as not understanding the skills inherent in appreciating and interpreting poetry.

For the most part, understanding how to address the issue of knowledge gaps properly (without adding substantially to the literacy load for these students) has not been considered in large-scale testing. A project underway in South Carolina is addressing the issue of English language arts passages, by conscientiously attending to the targeted reading level, regional sayings, and context of the passages; by minimizing the use of less-frequently used
language and situations; and by otherwise providing compensatory support as appropriate (Carr and Kopriva, 2007). The South Carolina item writers have attempted to address some of the issues of knowledge gaps in the four content areas and over grades, and recent data collections are now being analyzed.

Being cognizant of the problem in the design and construction of items is key. When prior knowledge contexts are unavoidable, besides providing visual cues as possible, it has been suggested that providing background information orally, accessed as a link during a computerized exam, or through an accompanying CD if the assessment is paper-and-pencil, could ease the burden discussed here. Video clips which can be viewed as part of a computer-based test have also been recommended.

Activities

Using extensive real-time activities in large-scale academic testing has been found to be cost-prohibitive and difficult to control and secure over sites. Like its manipulatives counterpart, this approach was largely used to measure more complex skills a few years ago when large-scale assessments sometimes included more performance components (e.g., in California or Kentucky in the 1990s). The New England Common Assessment Program (a test developed by and used in Rhode Island, Vermont, and New Hampshire) is currently considering the use of activities as part of measuring inquiry skills in its new science test. The advantages for including brief activities as compensatory support for English language learners, however, have not been well articulated. Their role and function in ELL class and resource rooms is similar to that of other tools, and the rationale is essentially the same. Brief activities provide another avenue for presenting information or offering movement or interactive support. To date, even though their use in classrooms for ELLs is routine, activities have not been used as an accommodation for this population in large-scale testing.

Two examples of a type of brief activity (15 minutes or so) which might occur immediately prior to a period of testing are outlined here. It is suggested that this type of activity, with forethought, could responsibly become part of the large-scale standardized testing regimen. First, a pre-test activity could include a video or audio clip that can be experienced as a classroom activity. These could include documentaries, primary source material, native music, theater, or a standardized execution of a science experiment or scientific phenomena. Second, students could complete a brief data collection (e.g., measure their desks or fill beakers), where the data are used in a series of mathematics or science questions. The targets of the items would not be on whether students complete the task of
measuring or filling correctly; it would be on how they manipulate the data. Computer-based testing can be useful in delivering some of the support for either kind of activity; it is recommended that the effectiveness of this type of tool be studied.

There are many issues and questions to balance and consider when making decisions about tools and resources. It is suggested that design deliberations be discussed thoughtfully and with a range of stakeholders, including educators of ELL students. Several issues need to be considered. First, testing environments that incorporate some of these recommendations can become unmanageable. Can a balance be found which is richer and more accessible than the environments of today, but not uncontrollable? Second, who pays for the resources and tools, and how extensive a set should be provided? Should there be enough sets to provide one per student, one per classroom, or one set per four (or so) students? Some of the tools/resources are expensive; is it worth the cost when students will probably only need and/or use them for a few questions? Third, if students are allowed to share tools and resources, and/or if testing is done in shifts, how is security maintained? Fourth, are time and scheduling constraints going to be detrimentally affected? Finally, local educators need public relations help from the state agency and publishers to be able to explain the reasons, rationale, and the balance of plus and minus associated with the incorporation of tools and resources thought to be “novel.” Are such resources available? Each of these considerations needs to be balanced against the need to properly develop accessible environments for students who have language and literacy limits, where options such as those discussed above address crossing a reasonable threshold of access (as compared with providing unequal advantage for some students. It is suggested that none of these questions is insurmountable, especially once the capacities of computers are utilized more fully. However, they are important and addressing them is necessary if some of the recommendations illustrated here are to be considered.

**Form Considerations**

The primary focus so far has been improving accessibility at the item level. This section will briefly discuss selected issues associated with addressing accessibility at the form level, that is, at the test level for a particular student. Most topics associated with forms are outside the scope of this book, but selected issues associated with forms that are particularly pertinent to English learners will be addressed here. If more than one form is developed
Choosing Items

The task of choosing items to go on test forms involves a number of considerations. Among these are selecting the appropriate number of items to adequately address the range of breadth and cognitive complexity specified in the test frameworks. Another is determining which set of items “hang together;” that is, which items are measuring parts of the same whole without only being redundant with other items that make up the test or a particular subtest. Various analyses are done to determine whether sufficient numbers and kinds of items make the cut or not. For some purposes these procedures are effective and useful. However, as Linda Darling-Hammond (1994) argues, tests have been built essentially to be purposely biased when it comes to the needs and skills of some students. As such, it is important that items are chosen which are effective for ELLs as well as other students. Two considerations related to the disaggregation of data and use of difficulty parameters will be briefly introduced.

Appropriate Disaggregating of Data

When items are chosen, developers need to ensure that decisions depend on empirical data and expert judgments which take English learners appropriately into consideration. This means quantitative data need to be sufficiently disaggregated to determine whether distributions of results are “consistent enough” across relevant student groups. For instance, if items are “hanging together” does this occur for each group? When item validity inferences are proposed, are the data supportive of this interpretation for each group, including the range of English language learners? The reason for disaggregating data that is relevant for this purpose is to ensure that the same decisions and inferences can be made for various parts of the population being tested. Usually, it is not enough for ELLs to be represented in a representative sample. This method can mask substantial differences which may occur for various minority groups, including English learners.

Depending on certain empirical methods to ensure that items are adequate for selection purposes is sometimes problematic. For instance, differential item functioning (DIF), perhaps the approach used most widely to examine bias, evaluates individual items within the context of students’ total scores. Key ELL challenges, such as reading proficiency, impact most if not all the items. Because DIF is focused on flagging item issues unique to individual items, the pervasive problem of reading level that occurs over items will often not be identified with this measure. Yet, it clearly impacts how students respond to test items. This issue is discussed more in Chapter 12.
Growing in popularity is the use of cognitive labs where important qualitative data can identify when items are performing as intended. This is a useful tool for evaluating ELL performance, and it is common for this approach to be used to support validity assertions that can directly impact item choice. The method collects process information from students as they complete items or shortly thereafter. The main drawback is that, because the clinical trials involve few students, ELLs are often not represented. When they are, information is always limited to the one or two English learners who are participating. It is recommended not only that ELLs should be involved each time cognitive labs are conducted, but that the participants should be low English proficient language learners and that efforts should be made to ensure they vary in background.

Quantitative IRT or bi-serial correlation data from field tests can be disaggregated by group (including ELLs). Other informal feedback about how items are functioning can be used to evaluate items for selection as well. These involve tryout data and group feedback from students. Informal data are primarily useful during item development to refine items, and confirmation of these changes can be helpful when selection decisions are made.

The Usefulness of Difficulty Parameters

The p-values or other parameters associated with item difficulty usually inform item selection processes. This is because developers want to make sure items on each form span the range of difficulty so students at all ability levels can be adequately measured. Item difficulty, as it is typically measured, assumes that all error is random and the only systematic effect of item performance is due to ability level. This means that if students answer incorrectly, they do not have the targeted knowledge or skills to properly solve the problem. However, with ELLs, estimates of ability using item difficulty results are typically confounded with systematic bias due to their language, cultural, and literacy challenges. As discussed earlier, it is argued that making items more accessible should reduce the target-irrelevant difficulty associated with performance for students who need that change and for whom the items have significant accessibility issues. Usually it seems that difficulty on accessible item versions is reduced for ELLs as compared to the difficulty of the standard version of the item. Sometimes, though, guessing behaviors overestimate the skills of students with little knowledge. Kopriva et al. (2007b) noted this effect for ELL students, who seem to overuse this strategy.

Issues such as these mean that difficulty estimates need to be considered within the context of item development. Since the goal is to select items
that measure the range of ability, using difficulty to make decisions about item selection is only useful if irrelevant influences are minimized. Chapter 4 presents an access-based specification framework which should help inform if difficulty estimates are defensible for ELLs.

**Plain Language Forms**

Over the years Abedi *et al.* (e.g., 2000, 2001a) have investigated if simplifying the language on test forms might address important needs of the ELL population. Item construction methods designed to specifically address these students have been explained in detail in Chapters 5 and 6. As discussed above, Abedi sometimes found that the plain language form was more useful for ELLs than for others. In analyzing changes in multiple choice response distractor patterns between access-based and standard forms for six grades and four subject areas, Kopriva and Cameron (2007) found that the access-based forms seemed to be significantly better for not only ELLs, but students with learning disabilities. These forms took advantage of simplifying the language and its structure and also providing compensatory avenues for students to access the items. In separating out proficiency levels of ELLs and reading levels of non-ELLs, Kopriva and Mislevy (2005) found that some reading capacity appears to be needed to take advantage of the language adaptations, particularly as the grades progress. Kopriva *et al.* (2007) reported the types of language structure and density in individual items that appear to be especially amenable to plain language adjustments.

Besides language considerations, Kopriva *et al.* (2007d) also found that some changes in “plain format” seemed to impact how well third grade students with testwiseness limitations performed overall on items. Many of the format methods occur at the item level (as discussed in Chapter 6), but some are more amenable to the form level. For example, Shaw (1997) suggests that providing information in the same item format throughout the test may be useful to ELLs. Other plain format issues at the form level involve simplifying stimuli that might compete with the additional language load experienced by these students. These include reducing the number of items per page, omitting columns, using simple and clear fonts and easy to read point sizes, taking care to design forms so the visual components of items are not overwhelming or dissonant, and being cognizant that little to no additional visual stimuli are introduced to side-track students. These issues are not unique to ELLs, but reflect that attention should be paid to minimizing over-stimulation and unnecessary distractions.

Chapters 6 and 8 note that, wherever L1 forms are used, item and form construction should follow plain language and plain format considerations as well. This is because many of the students who will be using them are
not grade-level literate in their home language. Finally, Abedi and Kopriva have repeatedly suggested that appropriate plain language and plain format forms are necessary but not sufficient. Proper language, tools and associated administration and response accommodations are also needed. It is probable that, just as with items, form-level improvements explained here will be differentially salient for ELLs at different proficiency levels.

**Computer-based Testing**

As noted throughout the book, computers provide an ideal platform for providing flexible options to English language learners. As many states and districts attest, however, the complex logistics associated with large-scale computer testing are still an issue. New concerns of sufficient bandwidth, restricted numbers of computers, security considerations, and resources needed to complete computer (and usually online) assessments properly are compelling, and comparability of pencil-and-paper and computer administrations are presenting new challenges. Some states (for instance, Maine, Virginia, Wyoming and Kansas) are utilizing both paper-and-pencil and computerized methods in testing academic content of their general population. Some studies seem to find only small to no differences between mean scores when both paper-and-pencil and computerized administrations are used (e.g., Fitzpatrick and Triscari, 2005; Poggio et al., 2005; Pommerich, 2004). However, others are finding item level differences even when mean scores are similar (Johnson and Green, 2006; Pommerich, 2004).

For English language learners, the challenges associated with logistics and comparability need to be weighed against the benefits that computer-based methods provide in terms of ease of delivery and range of options that could be made available. Once the assignment of proper accommodations has been made for individual students, a computer screening program can be configured to provide one set of accommodations for one student and a different set for another. Choices can include multiple oral and written language options, opportunities to respond by moving objects or otherwise producing non-language solutions, and different types of background or supplemental information for students with different identified gaps. Solano-Flores (2006) suggests that item language can be differentially selected for students based on identified linguistic strengths in L1 and English.

Accessible computer-based assessments for the purposes of providing access are in their infancy. Kopriva et al. (2007) used a computer-based system in their research study to assign different accommodations to students taking a mathematics test. Boals’ and Cranley’s (2006) recently
funded investigation is developing a large-scale prototype of a computerized test in science which will use item types different from those used on typical state tests plus tools, and various administration and response options important for students with low English language proficiency. The prototype is being built to be comparable to a regular paper-and-pencil core test utilized in the participating state. Siskind (2006) is focusing on designing computer-based science items for students with emerging language and literacy which will make use of the computer’s ability to provide options and capture responses without using as much written language.

As mainstream academic testing moves more and more to online administrations, ELL experts will need to work with test publishers and consumers to learn about and make use of the opportunities available in computer-based testing. Open platforms that allow for programs that address the needs of this and other populations are certainly do-able, but need to be considered up front during the design of the large-scale systems. Research needs to be conducted to guide the development and usefulness of computer-based options that can utilize the inherent flexibility in this medium but that would have been considered impractical when tests are administered in their paper-and-pencil form.

L1 Form Considerations

Chapter 8 is devoted to an in-depth discussion of the complex interplay between language, culture, site capability and test forms for students who are emerging English learners. A few points will be outlined here in order to set the stage for this chapter. First, native language academic assessments for students whose home language is not English are not new. What is more recent are the mandates that the full range of English language learners need to be included in statewide accountability and participate in the mainstream academic testing systems. Further, these tests must be aligned to state curriculum standards. This means that, not only should accommodations be selected for English forms in order to produce accurate data for specific students, but all forms of the assessment systems should align to those standards and yield equivalent scores at either the scale score or achievement standards level. In this case, a parallel form in a language other than English (or a form which is a variation of the dual language form concept) would be considered to deliver comparable data across test forms as long as the language form is administered to the proper students. If, however, academic tests in other languages are not built to be parallel to the mainstream test, they will not be comparable to the regular educational agency academic tests and cannot deliver comparable data without appropriate modifications and technical studies. The technical
aspects of comparability will briefly be summarized in Chapter 12. It is important, though, to understand that any forms built to specifically address the needs of English language learners and/or other students with emerging literacy are subject not only to the same considerations of technical adequacy, but they also need to be tightly connected and demonstrate equivalence to the mainstream state test if the data are going to be considered to be comparable across forms.

Equivalently designed achievement assessments in the primary academic language of students can be very useful, if they are developed, evaluated, and used appropriately. In general, native language assessment forms are the most effective when students have been taught the information to be assessed in that language (Lara and August, 1996). Sometimes it makes sense to administer a native language form to students who are newly arrived to U.S. schools, even if the instruction here is in English. However, Solano-Flores (2006) warns that, just because the student is taught in one language or another, this may be an over-simplification because all instruction (over years) may not have been given in that language (2006). He suggests that this may account at least to some extent for inconsistency of performance across languages, as unfamiliarity with one key term is all it takes for a student to fail an item.

Finally, translating or otherwise adapting tests from one language to another is not a straightforward task. Many authors (for instance, Figueroa and Hernandez, 2000; Goldsmith, 2003) stress the complex issues associated with properly constructing viable assessments in different languages, and explain the potential harm associated with improper translations and shoddy development. All translations need to be of high quality and undergo the same kinds of checks as the English forms to ensure that they are measuring what is intended. Further, sensitivity to cultural variations within the same language is important (Solano-Flores et al., 2003b). A few of the issues will be clarified below.

Translation

This book cannot begin to do justice to the complexity of producing appropriate translations or trans-adaptations. Robin et al. (2003) describe three types of empirical procedures that are generally used to evaluate the comparability across language forms. Solid translation procedures include multiple forward translations, use of back translation in addition to forward translation, and simultaneous development of forms (across languages). Further, the literature agrees that some items are more amenable to translation than other items (see Auchter and Stansfield, 1997 as an example). Just as Chapter 5 discussed, when producing items in a more accessible
mode of English, or in another language, it is imperative that the target intent be retained across versions. This means translators, or those supervising translators, need to be cognizant item writers who understand the subtleties associated with writing good, large-scale test items. As the previous chapter explained, sometimes contextual details or other item content need to be altered; this is true when translations are competently completed as well.

Even when the larger issues are traversed adequately, subtle but potent implications associated with translations need to be considered. For instance, Solano-Flores (2002) and Solano-Flores et al. (2003a) found differences across items in translated tests. He urged that research and comparability be focused at that level, and that even differences by dialect, within the same language, can influence outcomes. Ercikan et al. (2004) found that some items functioned differently across English and French versions even when the two versions were developed to be equivalent using a simultaneous test development model. Gierl and Khaliq (2001) found that differences in items across languages seemed to be due to omissions or additions that alter meaning, distinctions in words, expressions and sentence structure that are inherent or not to each language, and differences in item format.

**DUAL LANGUAGE FORMS**

Side-by-side or dual language forms generally provide the entire assessment text in English and another language. The exact formats can change but the intent is to provide students with the option of reading different items or the entire test in either or both languages. Many ELL experts (for instance August et al., 2004) say that students usually use one language predominantly, and look to the other language for cues or clarification. It is important to remember that this generally means students will take longer to complete the test; there is concern as well that some students may find this approach overwhelming (Kopriva, 2000).

In Oregon, Durán et al. discussed the development and implementation of Russian–English and Spanish–English side-by-side forms (2002). While not widely utilized, the development process and issues the authors considered as they developed the forms provided a foundation for future investigations of the usefulness of this method of assessing ELLs. Duncan et al. (2005) studied the use of a dual language Spanish–English test booklet in mathematics, looking at the effect of this accommodation for eighth grade Spanish speaking students with less than and greater than three years of English instruction (n=127 and n=74, respectively). After controlling for English proficiency, they reported that there was no significant effect
of side-by-side \((n=181)\) vs. English booklet \((n=20)\) on scores, and not a significant difference between scores on the dual language booklet for those who answered in Spanish vs. those who answered in English. As such, they considered the booklets to be comparable. Eighty-five percent of students using the Spanish-English booklet said it was helpful; a majority of these students preferred this format to a Spanish-only booklet or an English-only booklet plus a bilingual dictionary. Of interest, students tended to predominately answer in either Spanish or English. In disaggregating this data, 116 students who answered 90 percent or more of the questions in Spanish were almost all students who had received less than three years of English instruction; 83 percent of the students who answered mostly in English had received three or more years of English instruction.

Duncan et al., suggest that, while students may answer predominantly in one language, they are actively using both forms to read and solve the problems. This seems to occur even when they may not be literate in Spanish but are becoming literate in at least one language. This is consistent with what August et al. (2004) and Solano-Flores et al. (2003a) report. Unlike the findings reported in Duncan et al., however, Solano-Flores found that students tended to answer different items in different languages based on linguistic properties found in the particular items. Further, August emphasizes, that, for the most part, ELL students benefit mostly from the other form when they have achieved at least a critical amount of literacy in one language or the other. Earlier than that, the benefit is less certain. It seems reasonable that identifying this literacy point may be important for future research and development.

Solano-Flores et al. (2002) found that it can be effective to assign appropriate forms to students based on generating generalizability coefficients that are sensitive to linguistic differences. In a recent paper Solano-Flores suggests five models of form types based on the individual needs of students and where different numbers of items can be identified to produce acceptable reliability for appropriate students (2006). While this is obviously impractical in a paper-and-pencil administration, his work has possibilities when computerized administrations are used. He suggests that the parallel language forms should be developed simultaneously and writers should remain cognizant and active in making modifications on both if they are required on one version or another.

Using Portfolios

The concept of using portfolios as part of large-scale academic testing has been considered and implemented since the early to mid 1990s. Portfolios, as defined here, are collections of student work produced in the classroom
throughout the year which, in the aggregate, would demonstrate levels of academic knowledge and skills of students. Myford and Mislevy (1995) built an enviable portfolio system for evaluating the visual art abilities of high school students in Educational Testing Service’s Advanced Placement ‘exam’ for studio art. States, such as Vermont and Kentucky, established portfolio systems in mathematics and language arts in which all students in the state were required to participate. In a strong nod to the advantages of multiple measures, the states estimated the students’ ability levels through a combination of scores from both the portfolios and the on-demand tests. Portfolios were also designed into the assessment systems of other states (for instance, California and Wyoming) but were never fully operationalized. Challenges in building and defending such systems abound, but the field has learned how to address several of these over time.

For instance, as viable portfolio systems developed and matured, it became apparent that the target constructs needed to be well defined, clear, and narrow in focus. Rigney and Pettit (1995) reported that identification of clearly defined characteristics of entries was also essential, which do not specify the nature of the tasks but require that specific representation elements need to be included in order for scorers to properly evaluate the work. For instance, if student work included tables and graphs, certain attributes, such as naming the axes on the graphs and providing a legend, were required. Otherwise, an absence of detail meant that scorers could not determine whether the student lacked the requisite knowledge and skills, or whether they didn’t demonstrate them in a way that could be accessed by the scorers. Completion of a few anchor tasks generally was required as well. These common anchor tasks, which all students needed to complete (Kentucky used released NAEP items), helped to calibrate or anchor the scores of the portfolios across students. Finally, strong, clear scoring rubrics, and in-depth rubric notes were key to building a successful portfolio system, with training of scorers consistent with the standards of the field. Producing a defensible portfolio system that could be systematically scored across students, teachers, and schools requires forethought and planning. However, auditors of the statewide language arts portfolio system in Kentucky (1998) found that criteria could be effectively communicated throughout the state, and that teachers could learn to be accurate scorers, given proper training and rigorous oversight.

Within the last 10 years, states have been experimenting with using portfolio systems to evaluate the achievement of certain populations within the state, particularly students with severe cognitive disabilities. These endeavors generally were distinct from the earlier work in that the collection and evaluation of portfolio entries would only occur for certain students, whereas the achievement of the rest of the test takers in the states would
be measured only through on-demand tests. Lately, some states have attempted to require portfolio systems for measuring the skills of English learners statewide, but results have been poor if sufficient technical rigor had not been designed into the systems up front. Virginia and North Carolina, though, have been more successful in developing defensible portfolio collection and evaluation systems for English language learners (Bazemore, 2006). As in the systems that were used statewide in the 1990s (or nationwide as in the case of the AP Studio Art portfolios), these systems have attempted to explicitly define clear and focused target constructs, and to identify attributes of entries that must be present. Further, the states have set up or are in the process of developing rigorous scoring and oversight methods to evaluate the portfolios (Triscari, 2007; K. Barton, personal communication, November 2007). Gong and Marion (2006) published an article discussing the tradeoffs of building assessments where the items, form elements, and even accommodations are more standardized up front (as in traditional on-demand tests) versus the trade-offs when flexibility in entries is different across students.

On the whole the portfolio systems are not built to be an equivalent form to the on-demand testing system per se. However, if built with sufficient rigor and attention to the technical attributes that have to be met, portfolio systems such as these could be a viable alternative for measuring certain academic abilities of ELLs. The challenge will be to determine how to defend the comparability of this system versus the scores received by the rest of the test takers. If there has been foresight planned into the design of the portfolio systems as to how to produce and document common elements of validity and comparability across the two testing modes, the scores received by students could probably be seen as equivalent at the achievement level (e.g., basic, proficient, advanced) and possibly somewhere between the scale score and achievement levels.

Portfolios provide important and unique information about the abilities of students and can be a useful way to approach the academic measurement of some ELLs. It is not recommended, though, that large-scale portfolio systems be built where scores are ultimately not comparable between ELLs and other test takers. Today, on-demand test scores are used for many reasons, and history has shown us that, if the scores of ELLs or other students are not comparable to those of the majority, their needs will tend to be shortchanged. However, if states or districts put the resources into developing a strong portfolio system that is also built to be comparable at some level to scores received by others on the on-demand assessments, then it is an approach that could be practicable and promising. Earlier chapters provide some guidance as to how to define and clearly explain target requirements and specify task attributes necessary for scorers to see in order
to rate the work properly. Teachers should benefit from examples in the chapters about how tasks might be constructed for classroom and portfolio use as well in order to make them more accessible for English language learners. Because tasks can include classroom projects and other more lengthy constructed response-type items, the chapter information would need to be adapted for these entries. Chapter 11 discusses the scoring of constructed response items and issues such as those explained there should be considered for the scoring of portfolio entries individually or holistically. Finally, interested readers are encouraged to review the literature about portfolios in general, not just those related to ELLs. They are also encouraged to review state and project technical manuals or other documents related to large-scale portfolio systems that have proven defensibility.

Reviews

Several kinds of reviews occur throughout test construction. These reviews mostly occur at the item level and are both internal and external. Internal item reviews involve those who are part of the test development process in some way. External item reviews include fairness reviews and independent alignment studies. Internal and fairness reviews will be discussed here; considerations associated with alignment studies are outlined in Chapter 4. Descriptions of specific procedures associated with the internal and fairness reviews are outside the scope of this book; however, in each case, a few points associated with English language learners will be addressed briefly.

Internal Reviews

Internal reviews are completed typically by item writers or other development staff employed by the test publisher and often area educators as well. These reviewers may be charged with offering ideas for item content, building items, or providing ongoing feedback. They may evaluate the items only (including items and their rubrics or any associated tools), or the items within the context of additional data, including student work, item statistics, or qualitative findings. Ultimately, it is recommended that internal reviewers complete the Access-Based Item Matrix described in Chapter 4. As a framework it guides development; as a completed document it provides robust documentation of the purported content validity of items for ELLs.

Reviewing Item Options

When item options salient for English learners are being created at the item level to be comparable to their traditional item counterparts, the item sets should be evaluated in tandem. Options may be translated items and/or
items in English that address the needs of students with limited literacy and language skills. The item variation(s) should be reviewed to ensure that they measure identical targeted content knowledge and skills as the base item. This is necessary to properly address the comparability of inferences; indirectly it has been found to be very valuable in clarifying measurement intent of the base items (Carr and Kopriva, 2007). It is recommended that item templates be reviewed or completed by reviewers for each set of item versions. Carr and Kopriva found that reviews conducted in this way revealed important aspects of the items and further clarified intent. For example, although the item displayed in Figure 5.1 was written as a division item, it is likely that many students will use repeated addition to solve the problem. In this case, the targeted indicator and objective did not require a specific operation to be used. This information is important as item options are being honed. Evidence of intent and irrelevant influences should be available as part of the documentation of the completed access-based item matrix.

**Reviews of Associated Materials and Accommodations**

The access-based item matrix would also document that other appropriate materials and accommodations are available throughout the test. This includes a cursory review of the breadth of associated materials and accommodations and also an inspection of the interrelationship between each of these elements, the presentation of the item, and the item requirements.

By reviewing the target irrelevant knowledge and skills of whatever item options are available, reviewers familiar with the compensatory needs of ELLs should determine if access for this heterogeneous population is being properly addressed somewhere within each item grouping or if the form is coupled with proper administration or response accommodations to address the range of needs. It is suggested that the evaluation of materials and accommodations per item not be left to the final internal evaluations, but that it is considered as part of the ongoing review process as items evolve. Further, independent reviews can confirm the item level evaluation of relevant materials and accommodations but should not be a substitute for them. This is because participants in fairness reviews will not have had the same level of item training, familiarity with target-relevant and irrelevant nuances found in each item, and knowledge about the interactions between items, materials and accommodations. This level of detail is primarily within the scope of an internal evaluation.

**Fairness Reviews**

Fairness, sensitivity or bias reviews serve an important function in test development because they provide the primary opportunity for representatives
of special populations to review and comment on testing materials. However, as recently as 2002, Thompson et al., noted that most state fairness review committees did not include a representative for English language learners (much less a number of panelists who represented the heterogeneity of the population). Typically, guidelines for test publishers define participants as those who appropriately represent various populations to ensure that the final testing materials will be as inclusive of all racial/ethnic groups as possible. Members may include other educators or experts not usually involved in the development, such as local college faculty or school staff, as well as parents, community members associated with particular special populations or interests, business people, and legislators. However, because of the different types of experiences ELL students bring to testing, more and more test consumers are requesting that participation on bias reviews involve ELL experts as well (Lara and August, 1996). For instance, local parents of ELL students, and independent teachers, college faculty and community members familiar with English language learners could be used to form an effective fairness review panel to properly reflect ELLs issues. Kopriva (2000) proposes that, if members from dominant ethnic or racial ELL groups have been previously represented in other development work, the reviews may be an opportunity to hear from educators and community members from demographically less prevalent linguistic groups in order to receive their input.

Traditionally, these reviews focus on reviewing materials for offensive stereotyping, wording, or subject matter (e.g., Ramsey, 1993). Reviewers typically are asked to review vast amounts of item and related materials in a very short time frame, which necessitates, at best, an evaluation that is well-meaning but superficial. Recently, some review panels have been asked to surmise why certain items flagged by DIF analyses might be problematic. However, this is usually requested within the context of a range of participants (not just those from the group flagged as responding differentially), and little to no training on item writing or reviewing for this purpose is provided. It is therefore no surprise when Linn notes that this approach tends to produce ineffective information about the reasons items were flagged by large DIF scores and salient ideas about how to address the problems (1993). To the credit of publisher and client agencies, the recommendations from these reviews often are taken seriously and incorporated into the revisions of materials.

Are fairness reviews necessary, when other considerations that appear in this book are taken into account? It is suggested that they can be, for a number of reasons discussed here. However, to be useful, their charge needs to be expanded. Geisinger and Carlson (1992), and Popham (2001), among
others, have encouraged that the reviews should take advantage of the expertise of the reviewers and their knowledge about the populations they represent. These participants can offer valuable insight into how students will read the questions being posed on the exams and provide some recommendations for editing. They may also suggest test or form level concerns. Further, the researchers note that the independence of the fairness reviewers provides a fresh eye for errors that seasoned educators may miss.

The current fairness review process has grown and developed since the 1960s when the process was known as the minority review process (Ravich, 2003). With the evolution of the minority review process the definition and conceptualization of bias has changed to reflect the current thinking. The Office of Civil Rights (2000) defines bias in two ways: “In a statistical context, a systematic error in a test score. In discussing test fairness, bias may refer to construct under-representation or construct irrelevant components of test scores. Bias usually favors one group of test takers over another (p. 73).” Fairness reviews are intended to mitigate the impact of systematic error that is caused by factors irrelevant to the target constructs being measured on the assessment. According to the Standards for educational and psychological testing (AERA/APA/NCME, 1999), there are many possible construct irrelevant factors, including linguistic factors, test sampling, unclear test directions, differential scoring of non-standard responses, cultural factors, and opportunity to learn.

Chapter 11 gives several illustrations of the types of responses ELL students might make to answering constructed response items. Included are examples of responses which represent how some students with different cultural and educational experiences interpret items in ways that are unexpected and not anticipated by the rubrics or rubric notes. A variety of researchers and test developers have produced lists that include areas to consider when assessing an item for cultural bias (Wiesen, 2000; Kopriva, 2000; Hambleton and Rodgers, 1995). Further, the governing board for National Assessment of Educational Progress (NAEP) reported that, in the future, assessments “will incorporate awareness of this critical consideration, especially related to students’ previous opportunities to learn and their experience and background both in and outside school” (National Assessment Governing Board, 1999, p. 7). Finally, Ravich (2003) suggests that part of the problem is also that the timing for fairness reviews is off—that is, it occurs when most item writing has been completed. She argues that the independent review needs to happen earlier. However, to date, little work has been done to systematize this process, make it earlier in the test development cycle, or build review procedures that routinely address all items and rubrics.
PROCEDURAL RECOMMENDATIONS

Empirical research on the bias review process appears to be almost non-existent, but on a conceptual and pragmatic level it is possible to see what changes need to be made in order to move the process from its current narrow charge to a level that would be more effective and useful. Some changes to the process have been suggested. Hodgkins (1997) wrote that the expanded procedures would review wording, phrasing, and layout for clarity in addition to offensive content in items, reading passages, contextual materials, test questions, and other assessment text. Also, she reported that panelists should review scoring, directions, and reporting proposals. Below in Table 7.1 is an extension of Hodgkins’s original work. It provides a series of possible questions of the type that panelists might consider when participating in a fairness review where the scope has been enlarged to address additional issues.

Initially, participants should be briefed on all steps taken to ensure accessibility throughout the development and implementation processes. The steps need to be explained in simple, straightforward language with examples if possible. This information should be included to provide an overview that participants can draw on as they are providing guidance.

It is suggested that developing procedures for the expanded fairness reviews should involve systematically moving through the draft test forms and essentially evaluating the materials item by item (including directions and associated materials in the evaluation). The reviews provide an excellent opportunity for test developers and publishers to receive guidance about test formats, wording, rubrics, non-text item accessibility, and administration and response conditions. Many of these are commonplace issues that educators of ELL students deal with routinely in classrooms and, therefore, their insights are valuable. The advice from parents and other public stakeholders about these additional issues should be a welcome reality check as well, and it is suggested that it may be surprisingly informative. For instance, considerations include:

- test format issues, including page layout, item spacing, font size, choice of visuals, and available response options;
- test item wording, including word choices, sentence structures, and paragraph structures in test directions, test items and all contextual information;
- rubrics and rubric levels, including attention to content and sophistication of responses in relation to increasing sophistication in language or literacy;
TABLE 7.1 Expanded questions for the panel to consider

<table>
<thead>
<tr>
<th>Test formats</th>
</tr>
</thead>
<tbody>
<tr>
<td>• What type of test formats would be best for your student population and why?</td>
</tr>
<tr>
<td>• What format problems should the assessment avoid if possible and why?</td>
</tr>
</tbody>
</table>

**Wording in items and directions**

• What general advice about the wording in the items and directions can you give developers to ensure that your student population will understand what they are being asked to do?
• What is your advice about how to structure and incorporate additional contextual information which students will receive as part of the items, including pictures, diagrams, and text?
• Given a subset of items and directions, what changes would you make to ensure that your student population will understand what is required?
• What wording problems should the assessment avoid if possible and why?

**Rubric parameters**

• What special issues associated with your student population need to be considered when the constructed response items are scored?
• Given all or a subset of the rubrics which will be used in the assessment, what changes would you make to ensure that the response from your student population will be scored correctly?
• What issues about how items are scored should the assessment avoid if possible and why?

**Non-text item accessibility**

• Besides wording and scoring considerations, are the requirements of the items you have reviewed accessible for students from your population? That is, do the text and non-text supports allow students to understand what they are being asked to do or know, and give them what they need to process their response?
• Generally, what guidance can you give test developers which would help ensure students from your population are able to effectively access each item?
• What item elements constrain accessibility and should be avoided if possible and why?

Source: Adapted from *Ensuring Accuracy* . . . (Kopriva, 2000).

- non-text accessibility, including all text and non-text supports such as pictures, diagrams, performance activities, and access to specified tools and resources; and
- conditions of administration and response, including advice from participants on strategies particularly appropriate for their populations.

It is unrealistic to assume that the panel can review all items and their associated rubrics and contextual materials unless the test is a one-form
assessment given to relatively few grades. Therefore it is recommended that a random subset of items/rubrics and materials should be selected for review in all subjects tested. This includes all visuals, non-text contextual materials (e.g., videos or computer simulated materials), and examples of tools and resources the students will be able to access. In the case of a content area such as language arts, all passages, items and associated materials might be reviewed. Specific selected items or other materials flagged as problematic by item writers might be included as well. Finally, item target information, and item data, as available, should be included in the review materials.

This approach differs from some current procedures, and time must be made to allow for this level of individual review and group discussion. It is recommended that a checklist or some type of guidance should be provided for the individual reviews, and then the entire panel or subgroups assemble a consensus list that provides evaluation information about each component of the assessment forms as well as some overarching recommendations. Panel findings should be as specific as possible. If enough time is allotted, recommendations should include ideas for how specific items or other materials might be revised to be more in line with the panel’s concerns. Hodgkins (1997) found that participants can contribute important feedback about test questions, other assessment text such as reading passages, directions, and contextual materials, as well as assessment layout, and grading and reporting proposals.

Recommendations by the community fairness review panel should be organized and passed on to developers. Being careful to retain the intent of what the item is to be measuring, the item writers should review the recommendations and make modifications based on the kinds of ideas suggested in previous sections of this guide. As with student feedback, if the modifications require a major reworking of the materials, it is suggested that this version be tried out again in classrooms and be sent again to a future community fairness review. Within this expanded charge, review members would evaluate items, all related materials such as reading passages and background information, sets of directions, and test format and implementation plans.

There are a variety of implications associated with expanding the bias review process. Clearly this type of approach takes more time, and this will translate into more expense and staff time as panelists take longer to complete their work or additional panelists and panels are needed. These resource issues will have to be balanced with the need to upgrade the assessments so the results are sound for English language learners.

Based on the magnitude and importance of the fairness reviews and the narrowness of the charge traditionally given to those reviewers, one must
carefully consider what is the best way to ensure fairness for ELLs. It will be useful for research to document which of these suggestions are essential and which are not. Empirical work is especially important to conduct, considering the huge range of abilities and needs of ELLs. An expanded fairness review process can provide valuable information and insight in any case, but is absolutely key when small minority ELL subpopulations are not adequately considered during development, internal reviews or formative data collections.

Putting It All Together

Chapters 5–7 (and 8) discuss how to increase access for English learners to the text materials of K-12 academic assessments. As the introduction in Chapter 5 emphasized, this is not an easy process. Nor is it something that can be reduced to checklists and one-time training sessions. The difficulties in applying the information presented in these chapters comes about because two items are rarely the same, no matter how similar they look and whether or not a computerized item template was used to generate items with fixed parameters, construct specifications or content information. At this time it does not appear to be possible to produce quick “programmable” access-based item templates that allow users to easily generate items or details of associated materials which effectively meet the myriad needs of this population.

The items in the last few chapters provide some examples of the kinds of access interventions some researchers and states have recently used. In many cases, the illustrations could be clearer, the language or target objectives could be better constructed, and sometimes additional tools could and should be identified to better help English learners. Rarely, however, do any of the figures illustrate only one technique. Rather, they commonly combine a few, and, taken as a whole, it is the hope that they might provide a few ideas of how different methods might be used in tandem. Carr and Kopriva (2007) recommend a multi-session, multi-feedback, iterative approach that South Carolina used to develop items and they offer many examples of items where access characteristics were incorporated differently by the item writing staff and project researchers. As the access-making process matures, other considerations and solutions will certainly arise. It is anticipated this will be particularly true as writers maximize the capacities of technology and how items interface with that mode of presentation. What is presented here, though, should provide a foundation, and help large-scale and classroom test writers to move toward accessibility of their materials for this population of students.
A fundamental notion in test validity is that the scores of a test should not be significantly influenced by factors other than the skills and knowledge that the test is intended to measure (Messick, 1989). This notion is critical to examining language issues in testing. Since language is the primary vehicle for administering tests, academic achievement measures are largely influenced by language variables, even when the constructs of interest are not directly related to language. That tests inevitably assess language skill in addition to the target knowledge domain (AERA/APA/NCME, 1999; Kiplinger et al., 2000; LaCelle-Peterson and Rivera, 1994) appears not to be an overstatement.

Even when students are native, fluent speakers of the language in which a test is administered, their performance on that test is shaped by their ability to understand printed text or spoken language and to express their ideas orally or in writing (Popham, 2001). Hence, one question that must be addressed in determining in what language students should be tested is the status of their proficiency in the languages in question. For that reason, in this chapter, it is necessary to discuss language assessment—even though the book is focused on academic assessment.

Students’ performance on a test is also shaped by the ability of test developers to write test items clearly, the ability of test administrators to provide test directions orally or in printed text properly, and—in the case of open-ended items—the ability of test scorers to interpret the examinees’ responses properly.
When the examinees are English language learners (ELLs), the challenges posed by language as a source of invalidity are greatly magnified. For instance, even when ELLs “know” the vocabulary used in a test item, they are not likely to have the elaborated word knowledge of a native speaker of English (Garcia, 2000; Laufer, 1997), and their understanding of that test item may be highly influenced by contextual factors and their socio-cultural experiences (see Durán, 1989). As a consequence, their interpretation of test instructions or test items may be compromised.

Inclusion of ELL students in national and state assessment systems is driven by the aim of obtaining indicators of school performance based on testing as many students as possible and supported by the assumption (or hope) that effective approaches are available or can be devised for minimizing varying proficiency in English as a source of test invalidity (see Kopriva, 2000). An approach that may first come to mind is to test ELLs in their first language (L1). Another approach consists of testing them in English, their second language (L2), with certain accommodations—such as simplifying the linguistic features of items or providing students with glossaries with translations of potentially difficult words—intended to reduce or eliminate the effects of language proficiency on test performance (see reviews by Abedi et al., 2000; Rivera et al., 2006; Sireci et al., 2003).

In this chapter, we discuss three factors that are critical to both formulating national or state testing policies on the language used to test ELLs and examining the appropriateness of such policies:

1. the soundness of the criteria used to identify students as either ELLs or English proficient;
2. the accuracy of classifications according to their level of proficiency in English; and
3. the effectiveness of testing practices in addressing the English proficiency level of ELLs.

In the first section, Defining ELLs, we examine the adequacy of official definitions of the ELL individual according to current knowledge of bilingualism. In the second section, Classifying ELLs, we examine the appropriateness of current approaches for measuring English proficiency according to current knowledge in the field of bilingual development and second language acquisition. In the third section, Testing ELLs, we examine the limitations of testing ELLs in L1 and in English. We also discuss how, in order to be sound, decisions in favor of one or another language need to be based on consideration of the quality of the information on the language proficiency of ELLs and the institutional capacity to properly implement testing in each language and to provide effective testing
accommodations. In the fourth section, *Future Directions*, we present alternative testing models intended to address the limitations of existing approaches to ELL testing in both L1 and English. Such testing models are based on recent research showing that the performance of ELLs is unstable across languages.

**Defining ELLs**

Evidence from research on language development and second language acquisition shows that the term “English language learner” refers in actuality to an individual who may exhibit any of an almost indefinite number of patterns of proficiency in L1 and English (Brisk, 2006; Romaine, 1995). Hence, it is more useful to think of an ELL as defined by a broad spectrum of language proficiencies in both L1 and L2. This research also shows that, in order to be accurate, any definition of ELL students and any strategy intended to identify ELL students should be based on the recognition that populations (either monolingual or bilingual) vary considerably due to factors inherent to language, culture, and experience, and that bilingual individuals vary tremendously in their development of L1 and L2. In this section we examine these factors. Then we discuss how the interaction of these factors produces different patterns of bilingualism among ELLs. Finally, we discuss the limitations of official definitions of ELL.

**Variation Among ELLs**

We identify six factors as critical to properly identifying ELLs and to making appropriate decisions concerning each ELL individual’s testing: 1) relative language proficiency, 2) proficiency variation across language modes, 3) minimum bilingual development, 4) language dominance, 5) language variation, and 6) educational variation.

**Relative Language Proficiency**

An individual’s proficiency in a language is a relative, rather than an absolute, condition. Strictly speaking, total proficiency does not exist even in the native language, as even amongst the most proficient individuals there are always aspects of their linguistic skills that can be improved. Among the kinds of linguistic knowledge and skill that may continue to develop are word meanings, idiomatic expressions, complex syntax, spelling, ways of expressing ideas in writing, reading fluency, and ways of interacting socially through language (Grosjean, 1989).

**Variation Across Language Modes**

An individual’s proficiency in a language varies across listening, speaking, reading, and writing modes. A person’s proficiency in a language is rarely
the same across these four language modes (Wei, 2000). Moreover, different individuals have different sets of strengths and weaknesses in each mode. Whereas two given individuals could be thought of as having comparable, high levels of proficiency, one could be better at listening and speaking and the other better at reading and writing in L1 or L2 (Fishman, 1965). Although there are practical advantages of using the word “proficient” as an overall condition of an individual’s command of a language, its use unintentionally implies that a person’s proficiency is the same across language modes. In reality, rarely is an individual equally proficient in the four modes.

MINIMUM BILINGUAL DEVELOPMENT

In spite of their limited proficiency in English, virtually all ELLs have at least some minimum level of competency in L2. For example, they may be able to understand some words when they hear or read them, or they may be able to speak or write short phrases—rudimentary skills which result even from a limited exposure to L2 (Mackey, 1962). Virtually all ELLs can be thought of as bilinguals, albeit “incipient bilinguals” in some cases (Diebold, 1964, cited in Romaine, 1995, p. 11). Although the word “bilingual” evokes the image of someone who is equally fluent in two languages, rarely is an individual entirely bilingual (Valdés and Figueroa, 1994). Thus, terms such as “English language learner” or “limited English proficient” may lead to underestimating the proficiency of some ELLs in L2 or overestimating their proficiency in L1.

LANGUAGE DOMINANCE

Bilingual individuals tend to be better in one language than the other. Terms such as “English-dominant” or “Spanish-dominant” are used to characterize an individual in terms of his or her preferred language. Language dominance should be interpreted as a condition that is shaped by factors such as context and personal history (Wei, 2000). For example, even an individual with a high proficiency in L2 may use L1 for tasks that involve automaticity in the use of language, such as counting or performing arithmetic computations. Or, individuals who have developed their academic language in L2 may find it extremely difficult to use L1 in writing about subject matter learned in L2. Moreover, in some cases, bilingual individuals may be better in L1 than in L2 for one language mode but better in L2 than in L1 for another language mode.¹

DIALECT VARIATION

Different communities of individuals within the same broad linguistic group may vary considerably in the implicit language rules they follow and
in the uses to which they put language. Even within populations of native
speakers of the same language, ELLs may vary tremendously in their dialects
of both L1 and L2. Such variation may result from factors such as the
diversity of places of origin (geography) and contact with other linguistic
groups. Many people speak more than one dialect of a language or
languages, and many students learn—at least comprehend—the dialect of
school as well as home, when those differ. Discourse features may also differ,
as in ways of organizing oral or written narrative or responding to questions
(Cazden, 1999; Gee, 1996; Heath, 1983). These pragmatic or social aspects
of language are probably more important than differences in vocabulary,
pronunciation and syntax because they differ from what is expected in
school; more important, instead of being identified as language differences,
they may be wrongly identified as cognitive deficits. Such differences also
have implications for test validity (Deyhle, 1987; Estrin and Nelson-Barber,
1995; Swisher and Deyhle, 1992).

Patterns of Language Proficiency

The linguistic strengths and weaknesses of each ELL student vary with
language mode (oral, written), situation (more formal, less formal), role
of language (thinking, comprehending, producing), and social domain
(e.g., family, school) (Fishman, 1965; Romaine, 1995). Depending upon
the nature of the task, the context, and the conditions, an ELL student is
likely to be more comfortable or competent in one language or the other
(Romaine, 1995; Valdés and Figueroa, 1994). One might better think of an
individual’s “dominance configuration” (Fishman, 1965), or the pattern of
language proficiency represented by the interaction of all of these elements.
In this section, we examine possible patterns of variation in language
proficiency based solely on the language mode in question: listening,
speaking, reading, or writing. Naturally, if we attempted to account for
task demands, social context, and conditions surrounding the task, it would
be possible to generate an endless number of patterns of far greater
complexity.

Figure 8.1 shows the proficiencies in English and in Language X of eight
hypothetical individuals according to language mode. For the purposes of
our discussion, we make the assumption that linguistic proficiency in a
language can be measured accurately. We know that in reality this is
extremely difficult, as proficiency may vary considerably with context. For
instance, measures of reading proficiency may have some limitations
because they use reading materials that are only samples from the universe
of materials that a student could possibly need to read. Moreover, it is well
known that the proficiencies of linguistic minority students in both English
and their native language are seldom assessed properly (De Avila, 1987;
MacSwan, 2000; Saunders and O’Brien, 2006). Despite its limitations, however, the figure is useful for illustrating the imprecision of such terms as “proficient,” “monolingual,” “bilingual,” or ELL. If an individual’s proficiencies in the four language modes were exactly the same across languages, the lengths of the bars for the four language modes would be the same. Also, if so-called “bilinguals” were equally proficient in two languages, the lengths of the bars on all language modes would be symmetrical across languages. In reality, this kind of symmetry would be a rare occurrence because even individuals who are deemed fully bilingual tend to be dominant in one language (Romaine, 1995).

![Figure 8.1 Language proficiencies of eight hypothetical persons. The length of the bars, on the left and right side of the chart represent, respectively, proficiency in L1 and L2.](image-url)
The first two individuals are called “monolingual.” Person A is a monolingual native speaker of English, and Person B is a monolingual native speaker of Language X. But the monolingual condition of each is not absolute. Both have developed minimal listening and reading proficiencies in their second languages (e.g., some words that they can pick up in a conversation or others that they can recognize in printed form). This may have happened incidentally; for example, from their exposure to the second language through the media or through contact with native speakers.

Persons C and D are identified as “bilingual.” However, their patterns of bilingualism are different. Person C could be a native English speaker who was born in and has received formal instruction in the U.S. but has had continuous contact with speakers of Language X through conversation. In contrast, Person D could be someone who has received formal instruction in both Language X and English. Whereas Person C has acquired English through social interaction, Person D has acquired it mostly through formal instruction. As a result of these differences, Person C speaks English fluently, and her reading and writing skills in English are strong. In contrast, Person D has strong reading and writing skills but moderate listening and speaking skills in English. Some would be surprised to realize that Person C has such limited reading and writing skills in Language X after hearing her speak in the two languages. Also, from Person D’s conversational skills in English, many would find it difficult to call this individual “bilingual,” in spite of his having similar reading and writing skills in the two languages.

Persons E–H are students classified as ELLs. They are just examples of the many patterns of language dominance in ELLs that result from different kinds of language development in English and their native languages (see Stevens et al., 2000). Of the four ELLs shown in the figure, E is probably the closest to what many people have in mind when they talk about ELLs—individuals with reasonably strong skills in their native language and poorly developed skills in the four language modes of English. The reality is that this pattern of language dominance is far from being the only one among ELLs. Indeed, E might well be a rare case of ELL.

A more complex pattern can be observed in F. He is more proficient in his first language than in English in the listening and speaking modes. However, he is more proficient in reading and writing in English because it is in this language that he has received formal instruction. His linguistic skills are moderate in both languages.

Overall, G and H are more proficient in English than F. However, there are several subtle but important differences between them in their proficiencies in both languages. First, the proficiencies of G tend to be
symmetrical across languages, with better listening and speaking skills than reading and writing skills. Second, the reading and writing skills in both languages are better for H than G. Third, whereas the listening and speaking skills in English are better for G, H is more proficient in reading and writing in English. H is more like those students who come to school “fluent” in their home language but who have developed little of what is called “academic language proficiency”3 in that language (see Cummins, 2000). It should be noted that academic language is not restricted to the written mode; there are oral forms of discourse that are highly academic, some general and some particular to the disciplines (Schleppegrell, 2004).

Figure 8.1 shows that many ELLs are partially proficient in their native languages and partially proficient in English. It also gives an idea of the enormous diversity that may exist even within the same broad linguistic group of ELLs. Classifications of students as “ELL” or “LEP” (limited English proficient) are generalizations at best, as each category may actually describe a wide variety of types of ELL or bilingual students (see Aguirre-Muñoz and Baker, 1997; Casanova and Arias, 1993). As a consequence, classifications of ELLs based only on criteria such as whether English is their native language or whether they are proficient in one or two language modes in English are flawed. And this is true even if we assume that proficiency is properly assessed for those language modes. As discussed, individuals—especially ELLs—may vary considerably in their linguistic proficiencies across language modes. Moreover, in spite of their limitations in English, some ELLs may be more fluent in English than in their first language for some modes.

Additional Sources of Linguistic Differences

Individuals with similar patterns of bilingualism may vary considerably in their use of L1 and L2 due to differences in cultural background, dialect, and schooling. These factors shape the accuracy with which ELLs can be identified. Variation in cultural experiences encompasses both home culture and the culture of school. Whereas school culture is relatively homogeneous in the U.S., students’ home cultures are quite diverse. Home culture experiences include a student’s opportunities to develop proficiency with the native language of the family (including uses of language in social contexts such as school) and to absorb the values and beliefs of the family and cultural community (including those related to teaching and learning) (Heath, 1983; Pease-Alvarez and Vasquez, 1994; Trumbull et al., 2004). Parallel experiences take place within the school; however, they may not be recognized as cultural in nature (Trumbull et al., 2004).

ELL students may have close ties to family members in their parents’ country of origin and travel with some frequency to that country—not
unusual for Mexican-American students, for instance (Delgado-Gaitan, 1994; Solano-Flores et al., 2002). They may participate in weekend language schools where they continue to learn L1 to a high level of proficiency, as do many Chinese American students (Wang, 1996). Also, they may be part of a multi-generational family in the U.S., with many of its longstanding traditions intact and opportunities to use L1 in a variety of contexts, as is the case with many American Indian students (McCarty, 2002).

Other ELLs are more thoroughly immersed in the dominant culture and have fewer opportunities to maintain and develop the language of home. Language is, of course, a fundamental cultural resource and the most important tool in carrying culture on from generation to generation (Fishman, 1989; Hinton, 1995; Wong-Fillmore, 1991). However, there is strong evidence that even when families of recent immigrants shift to English, often within a generation, they continue to pass on their traditional values—even in the ways they use English (Delgado-Gaitan, 1994; González et al., 2001). So, maintenance of a family’s heritage language is by no means the only measure of persistence of cultural values and child-rearing goals, both of which affect students’ ways of using language and participating in school activities, including testing (Swisher and Deyhle, 1992; Rothstein-Fisch et al., 2003).

Limitations of Official Definitions of “ELL”

Strictly speaking, sound classifications of ELLs for instructional placement decisions should be based on examining the proficiencies of each student in the four language modes in both English and L1. At the very least, the decision about the language to use to test ELLs should be made separately for each individual and should be based on examining their proficiencies in at least reading and writing—the language modes most commonly used to administer tests—in both L1 and L2.

Unfortunately, legislation concerning the identification of ELLs does not reflect the complexity of proficiency across languages and language modes discussed above. Standard definitions of “English language learner” are quite minimalist. The National Clearinghouse for English Language Acquisition defines ELLs as “students whose first language is not English and who are in the process of learning English” (National Clearinghouse for English Language Acquisition, n.d.). The federal government, along with most states and districts, uses the term “limited English proficient” (LEP) instead, which it defines as “those students who have insufficient English to succeed in English-only classrooms” (Lessow-Hurley, 1991). More often than not, the emphasis in legal definitions is on English alone, and attention is not given to the level of proficiency ELLs may have attained in L1—something that must be considered in decisions about the language
of testing (Saunders and O’Brien, 2006). Focusing on English alone in characterizing these students is to overlook other linguistic resources they may well have and may lead to wrong decisions concerning the language of testing or wrong inferences about their ability (Figueroa and Hernandez, 2000).

Another limitation of official definitions of ELLs has to do with the level of analysis used to characterize these students. Take, as an example, the No Child Left Behind Act of 2001 (NCLB), which defines an ELL as an individual

(A) who is aged 3 through 21;
(B) who is enrolled or preparing to enroll in an elementary school or secondary school;
(C) (i) who was not born in the United States or whose native language is a language other than English;
   (ii) (I) who is a Native American or Alaska Native, or a native resident of the outlying areas; and
   (II) who comes from an environment where a language other than English has had a significant impact on the individual’s level of English language proficiency; or
   (iii) who is migratory, whose native language is a language other than English, and who comes from an environment where a language other than English is dominant; and
(D) whose difficulties in speaking, reading, writing, or understanding the English language may be sufficient to deny the individual—
   (i) the ability to meet the State’s proficient level of achievement on (academic assessments such as reading, mathematics, and science);
   (ii) the ability to successfully achieve in classrooms where the language of instruction is English; or
   (iii) the opportunity to participate fully in society.

Only one of the four main components of this definition (D) addresses issues that are critical to the condition of being ELL: Components A and B set the boundaries of the demographic characteristics of the segment of the population of interest, for assessment purposes, and component C refers to attributes (e.g., place of birth, being an immigrant) that are associated with but do not necessarily determine the condition of being limited English proficient. Unfortunately, while Component D recognizes the fact that proficiency in a language may vary across language modes, it does not
actually provide a definition of ELLs but, rather, a description of some consequences of being an ELL. The possibility of failing on tests, not achieving in school, or having a limited participation in society are ambiguous criteria for defining ELLs that may be interpreted in many ways by different people.

A potential danger of the interpretation of (D) is that, in order for districts to determine English language proficiency, they may use achievement tests standardized on monolingual English speakers as a proxy for language assessment, reasoning that such tests represent the performance criteria students should meet. But these kinds of instruments were not intended for such a purpose or for use with such a population (Linquanti, 2001). Nevertheless, appropriate or not, it is these same instruments that will be used to make decisions about students’ academic progress, once they have been designated “fully English proficient” and moved into English-only instruction. This will be the case, even though their performance may look more like that of ELLs than that of English-only peers for some years to come (Katz et al., 2004). Unjustified judgments about students’ abilities may be made on the basis of their language proficiency, and inappropriate decisions about student placement or potential may be made as a consequence (Figueroa and Hernandez, 2000).

Whereas the federal definition above may do a good job of describing the pattern of social conditions of the population of ELLs in the U.S., it does not address the fact, discussed before, that each ELL has a unique set of strengths and weaknesses in L1 and a unique set of strengths and weaknesses in L2. When applied to individuals, the definition may render flawed classifications for testing purposes. Such flawed classifications can be false positive or false negative identification cases—respectively, individuals who are identified as ELLs without actually being ELLs and individuals who are not identified as ELLs in spite of being ELLs.

Section Summary

Official definitions of ELLs provide directions for decision makers and administrators to make placement and testing decisions about students who are or may be ELLs. However, they are not sensitive to important facts about the nature of being bilingual, such as varying proficiencies bilingual individuals may have in listening, speaking, reading, and writing in L1 and the varying proficiencies in those language modes in English. In addition, official definitions of ELLs leave too much room for subjective interpretation. These inaccurate views of the characteristics of ELLs may lead to flawed decisions about who should be tested in which language and when.
Classifying ELLs

According to legislation derived from NCLB (No Child Left Behind Act of 2001, 2002), states are required to assess yearly their ELL students’ progress toward developing proficiency in English. Unfortunately, some of the tests used by states and districts have flaws derived from the fact that the process used to develop them has not adequately addressed important issues relevant to ELLs, such as sampling; the nature of their native languages; and their proficiencies in different language modes (see Rabinowitz and Sato, 2006). In this section, we discuss schooling background as an important factor that shapes the proficiency of ELLs in English but is frequently misused as an indicator of specific individuals’ proficiency. Next, we examine the challenges of defining and measuring language proficiency. Then we describe a modern view of language proficiency, which, from our perspective, characterizes the range of skills that should be measured by language proficiency testing. Finally, we discuss the limitations of tests of language proficiency that do not reflect this modern view of language.

Schooling Background

Participation in different types of programs (e.g., dual immersion, bilingual, transitional, L2 immersion) is known to have a powerful effect on the ELL student’s oral proficiency in L1 and to some degree in L2 (Brisk, 2006; Saunders and O’Brien, 2006; Walqui, 2002). For this reason, students with different program experiences cannot fairly be measured against each other. It is simply unreasonable to expect ELLs who have participated in “very diverse program types [such as] . . . English-only, transitional bilingual, dual language immersion . . . with very different program goals . . . to perform at similar levels” (Huempner, 2004, p. 379).

Programs that foster continued development of L1 also foster improved literacy in L2 as well as better achievement outcomes (Lindholm-Leary and Borsato, 2006; Riches and Genesee, 2006). Unfortunately, classifications of ELL students for purposes of making decisions about their language proficiency based on their participation in educational programs may be flawed because programs vary considerably in type and fidelity of implementation (Torres-Guzmán et al., and Han, 2005), and because their effectiveness is shaped by a multitude of contextual factors (see Cummins, 1999; Greene, 1998; Krashen, 1996).

Among the many education-related factors affecting students’ varying proficiency in L1 and L2 is the number of years students have spent in U.S. schools versus the number of years they have spent in schools in another country (Abella, Urrutia, and Schneyderman, 2005; Rumbaut and Cornelius, 1995); whether the schooling in either country has been continuous or interrupted (Freeman et al., 2001); and the fidelity with
which a given program is implemented during the time a given student participates in it (Collier and Thomas, 2004). Curricula may emphasize different aspects of English language development (see, e.g., Crawford, 2003; Dutro and Moran, 2003; Moschkovich, 2000). Also, the performance of ELLs on academic tasks may vary according to whether they participated in competitive versus non-competitive assessment activities (Rothstein-Fisch et al., 2003).

As important as program design are the social practices that students and teachers engage in as language is learned (Hawkins, 2004). Language-learning is above all a dynamic social process (Nelson, 1996; Ninio and Snow, 1988), thus it is influenced by the ways the social life of the classroom is structured and roles students are expected to take—for example in discussions, in response to questioning, in completing learning tasks (Brisk, 2006; Losey, 1995; Tabors, 1997; Tharp and Gallimore, 1991; Trumbull et al., in press).

Defining and Assessing Language Proficiency

We have discussed varying patterns of language proficiency among ELL students and the need to understand an individual student’s profile in order to make a decision about his or her readiness to move into assessment in English. However, it must be acknowledged that the construct of “language proficiency” is, itself, a troubled one. Two of the most fundamental problems are that a) what counts as “proficiency” is extremely variable, dependent upon the context within which one is operating (MacSwan, 2000); and b) proficiency is in a constant state of flux (Pray, 2005). The term has come to mean so many different things that some researchers avoid using it (e.g., Bachman, 1990, cited in Katz et al., 2004). It follows that different language tests are based on different notions of what constitutes language proficiency, some of them on only implicit notions (Abedi, 2004). It is, thus, no surprise that different tests of oral proficiency would yield different classifications of English language learners: A student may score as “proficient” on one test and “not proficient” on another (Katz et al., 2004).

Pray (2005) conducted an empirical assessment of the validity of three well-known English language proficiency tests by investigating how 40 native English-speaking fourth- and fifth-graders would score on them. All of the students were classified as “fluent English speaking” by the Language Assessment Scales—Oral (LAS—O). However, of the same group, only 85% (34) were classified as “fluent English speaking” by the IDEA Proficiency Test. And none of the forty participants scored in the “fluent” or “advanced fluent” English range on the Woodcock-Muñoz Language Survey. In fact, fully half of them (20) were classified as “very limited in English” by the test.
MacSwan (2000) examined the Language Assessment Scales (LAS) Español (for grades 2–12) and the Pre-LAS (for grades pre-K-1), perhaps the most widely used assessments of oral language in Spanish, and found them seriously flawed. Both assessments have a section requiring the child to answer questions about a Spanish-language story (weighted 30% and 50% of the total score for the Pre-LAS and LAS respectively). Failure to respond requires the examiner to give the child a score of “O” for the section, which will result in a classification of “non-Spanish speaker,” no matter how well the child does on the rest of the test. MacSwan uses an actual case example of five year-old “Gabriela,” who scored 14/15 on a sentence completion task on the Spanish Pre-LAS, but was nevertheless categorized as a “non-Spanish speaker.” Results from such studies as these do not give us confidence in the instruments districts routinely trust to help them make major decisions about ELLs.

Literacy and Academic Language as Part of Language Proficiency

Originally, the construct of language proficiency encompassed oral language alone; but as researchers concerned with educational outcomes tried to operationalize it relative to schooling, literacy (reading and writing) was added (e.g., Cummins, 1981). Now, the construct, as used by educators and educational researchers routinely includes literacy as well as “academic language,” which has been defined as the advanced oral and written forms and functions of a language required to achieve in school, including the vocabulary, grammar, and discourse conventions associated with the different disciplines (Chamot and O’Malley, 1994; Scarcella, 2003). It is this kind of language that requires years of learning, as distinct from conversational/social language, which may be learned by ELLs within two or three years (Hakuta et al., 2000; Katz et al., 2004).

There are both advantages and disadvantages to this broadening of the scope of the term “language proficiency.” Literacy and oral academic language are essential to successful academic achievement (Riches and Genesee, 2006). However, under current definitions of language proficiency, people who successfully use oral language but do not read and write would be characterized as having limited language proficiency. In addition, as academic language is held out as the pinnacle of linguistic achievement and fundamental to the ability to handle cognitive complexity (Cummins, 2000), the implication is that other language registers students may use are inferior, when they simply serve different purposes in different contexts. It is not difficult to see how current conceptions of language proficiency could lead to a deficit view of many students within our educational system and/or to judging groups of people who are non-literate or speak languages that are not written as intellectually inferior (MacSwan, 2000; MacSwan et al., 2002).
Reading and writing are relatively new human skills, having begun within the last 6,000 years (Fischer, 2001). They are not universally acquired through exposure in the way oral language is but, rather, generally develop only as the result of explicit instruction. They might better be thought of as language-based academic skills rather than as components of language proficiency. In fact, literacy is not strictly linguistic: it depends upon other, non-linguistic, cognitive and perceptual skills, as well as content knowledge. Reading comprehension, itself, is confounded with one’s personal experience; it is well-known that many comprehension questions on tests can be answered correctly with reference to mainstream, middle-class experience without full reading of the text (Popham, 2001). Another apparent consequence of the emphasis on literacy is less attention in the research community to oral language. According to Saunders and O’Brien (2006), there has been only one major study in the past 20 years (Hakuta et al., 2000) that has “explicitly addressed the rates of oral English language proficiency attainment” (p. 23).

Having expressed these concerns, we acknowledge that proficiency with reading, writing, and academic language is a fundamental goal of schooling and that a broad definition of language proficiency (incorporating these elements) is appropriate to the task of understanding the linguistic progress of ELLs. We must point out that commonly-used tests of English language do not assess academic language. A perspective on language proficiency is needed that maintains a focus on the complexities and variations in oral language, while incorporating the academic domains associated with language arts programs in schools.

A Modern View of Language Proficiency: Communicative Competence

We believe the most useful way of thinking about language proficiency is as communicative competence, a conceptualization that is grounded in a functional and sociolinguistic perspective (Bachman, 1990; Canale, 1983; Hymes, 1972). Historically, linguists focused on individuals’ mastery of forms and implicit rules—without particular regard for the social purposes and contexts of language use. In recent decades, there has been increasing recognition that language must be understood in context, as it is actually used. This approach has been widely embraced by applied linguists, speech/language professionals, and educational researchers (Cazden, 1999; Cummins, 1980; Gee, 1996; Heath, 1983, 1986). In fact, a communicative competence perspective captures both a focus on proficiency with rules and forms and proficiency with language use in context. What counts as competence is relative to the cultural community one is participating in (Romaine, 1995). So, what is considered communicative competence at home may not be considered so at school and vice versa.
Figure 8.2 below shows three common ways of thinking about the components of language proficiency. Teachers tend to think in terms of the elements of the language arts curriculum, and educational standards are framed in those terms, as we discussed earlier. Traditional linguists have characterized language in terms of the following components: phonology (the sound system), morphology (the word-forming system, including word roots, prefixes, and endings), syntax (the implicit rules of sentence formation—what we usually think of as “grammar”), and semantics (meaning conveyed by words and relations among them). Pragmatics, which concerns language use in social contexts is now included in many taxonomies. As mentioned earlier, reading and writing are not part of the traditional linguistic domain.

Communicative competence entails “grammatical competence” (where “grammatical” refers to the forms and rules of all systems—phonological, morphological, etc.), as well as three other components. “Pragmatic competence” refers to knowledge of the implicit rules of language use in social contexts, which could be thought of as specifying who can say what to whom, when, and in what manner (Hymes, 1972). It entails knowing how and when to speak formally or informally, for instance, or which language or dialect to use. Of course, pragmatic rules or expectations vary from one speech community to another as well as according to social setting. “Discourse competence” refers to knowing how to carry on conversations and compose and interpret spoken or written texts such as arguments, explanations, or narratives. Discourse competence is heavily

<table>
<thead>
<tr>
<th>Language Arts View</th>
<th>Listening</th>
<th>Speaking</th>
<th>Reading</th>
<th>Writing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traditional Linguistic View</td>
<td>Phonology</td>
<td>Morphology</td>
<td>Syntax</td>
<td>Semantics</td>
</tr>
<tr>
<td>Communicative Competence View</td>
<td>Grammatical competence</td>
<td>Pragmatic competence</td>
<td>Discourse competence</td>
<td>Strategic competence</td>
</tr>
<tr>
<td>phonology</td>
<td>morphology</td>
<td>syntax</td>
<td>semantics</td>
<td></td>
</tr>
</tbody>
</table>

Figure 8.2 Three views of language proficiency. Adapted from Trumbull et al., 2005, p. 13.
entailed in using and processing the academic language of the classroom (Adger et al., 2002).

“Strategic competence” means proficiency in communicating one’s message, making corrections or changes as needed, varying one’s volume or tone, or clarifying what has been said. Strategies such as these are a metacognitive aspect of language use (Bachman, 2002). Strategic skills are relevant not only to social conversation, but also to academic communication. Anticipating the kinds of clarifications needed in a formal discussion or essay are more academic aspects of strategic competence.

The communicative competence approach is the most comprehensive and useful approach for evaluating a student’s language proficiency. To operate successfully in a second language, a learner needs to be proficient in all of the domains specified within the communicative competence framework; and to understand a student’s school functioning via a second language, a teacher needs to be able to evaluate the student’s success in all of these domains (Wong-Fillmore and Snow, 2000). Taking a test, itself, calls upon proficiency in all of these aspects of language: using phonological and morphological knowledge to identify, spell, and derive words; understanding vocabulary and navigating syntax to comprehend instructions and test items; grasping pragmatics of language use (How formal should the writing be? What kinds of questions is one permitted to ask the teacher or another student?); understanding discourse patterns that crop up repeatedly on tests; and trouble-shooting *vis-à-vis* potential miscommunication, by proof-reading, putting oneself in the place of the person judging the answer, and the like.

One acknowledged problem with assessing students from a communicative competence standpoint is the difficulty of operationalizing all of the constructs associated with it and the amount of time it would take to get a full picture of a student’s language proficiency (Saunders and O’Brien, 2006).

**Limitations of Mandated English Language Development Tests**

Language testing, no matter what form it takes, is limited in that no standardized test or process can ever fully capture a language user’s competence—something that makes it difficult to operationalize language proficiency levels (Katz et al., 2004, p. 14). The dominant theoretical perspective in the field of language assessment, however, is that assessment that provides authentic contexts for eliciting learners’ language use—one that at least integrates form and meaning in as natural a way as possible—holds the best hope of getting an approximation of a language learner’s language proficiency (e.g., Bachman, 2002; Oller, 1992). This approach mirrors dominant second-language instruction theory that stresses language
use and learning in context (Cummins, 2001; Krashen, 1982, 1991), though it has been tempered with recommendations to include some explicit skills instruction (Dutro and Moran, 2003; Wong-Fillmore and Snow, 2000).

Unfortunately, existing approaches for measuring language proficiency do not appear to reflect recent conceptualizations of the range of skills within the language domain (Kopriva et al., 2004). Moreover, the effectiveness with which tests can measure English development in ELLs may be limited by the fact that no standards are available that specify how language development is supposed to take place in ELLs. As a consequence, the development of many tests of English language is heavily influenced by theoretical frameworks that underlie English language arts standards rather than frameworks that capture more accurately the complexities of the linguistic competence to be developed in a new language. For example, California’s “English Language Development” (ELD) standards (California Department of Education, 2002) are framed in the same terms as the state’s English Language Arts (ELA) standards. This is probably the reason why the California English Language Development Test (CELDT), which was developed from those standards, has proven to be only moderately predictive of a student’s linguistic readiness to handle the content on achievement tests, particularly at higher grade levels (Katz et al., 2004).

The standards proposed by the national professional organization, Teachers of English to Speakers of Other Languages (TESOL) (Gottlieb et al., 2006) are similar to California’s in that they focus primarily on students’ ability to use language in relation to the academic domains (mathematics, language arts, etc). The TESOL standards are educational/academic standards, related to but not explicitly tied to a full developmental characterization of language proficiency in terms a language specialist might propose. Moreover, all of these standards (ELA, ELD, TESOL) are expressed in terms of broad grade-level clusters (e.g., K-2, 3-5, 6-8), guided generally by research on monolingual English speakers but without a basis in empirical research “documenting the rate at which the average [native English-speaking] student is able to acquire [meet] specified content standards” (Shin, 2004, p. 254).

At the time of this chapter’s writing, some states are updating their practices concerning language development testing for ELLs (see Forte, 2007; National Clearinghouse for English Language Acquisition and Language Instruction Educational Programs, 2006). However, there is still a strong need for a more refined timetable, with many yearly benchmarks of progress toward ELD standards—related not only to grade level but to language proficiency level, as assessed by valid and reliable instruments (Shin, 2004).
Whereas a communicative competence model of language proficiency may be theoretically more defensible, it must be acknowledged that efforts to create language assessments based on a communicative competence model have had mixed success as predictors of which students are likely to have difficulty with achievement tests in L2. It appears that both tests that assess language competence in terms of discrete aspects of language (e.g., syntax, vocabulary) and those that assess communicative competence yield some information about students’ readiness for testing in their second language (see citations in Katz et al., 2004). One question that remains is whether a well-developed test of communicative competence might help us understand why it is that many students who do poorly on standardized tests are actually doing quite well in their classrooms (e.g., Katz et al., 2004). Perhaps they are able to negotiate academic language in contexts where they can seek clarifications from teachers and peers (using strategic competence) and use social language skills to scaffold their academic language skills.

Underlying the limitations cited above is the fact that learning English as a second language (ESL) is often confused with learning English as a foreign language (EFL). Of course, there are obvious parallels between these ways of learning English. For example, all newcomers to a language have to learn its systems of pronunciation, grammar, vocabulary, and socially acceptable ways of using the new language. However, research identifies important differences between individuals who are learning English as a foreign language and individuals who are learning English as a second language—and whom some refer to respectively as “circumstantial bilinguals” and “elective bilinguals” (Valdés and Figueroa, 1994, p. 13). The former are doing so voluntarily, not as a social necessity (e.g., González et al., 2001); they are often adults or secondary students taking an elective course. By contrast, ELLs are learning a second, though not foreign language; they are using English as a medium for learning mathematics, science, social studies, and the like while they learn the language; and they are attempting to catch up with a moving target—the continuously developing language proficiency of their native English-speaking peers (Shin, 2004). Moreover, they are still in the process of developing their first language, are thrust into a situation in which they have to learn English in order to survive academically and socially, and often do not have access to a strong program that builds both languages and supports students’ academic success.

Section Summary
Tests of English language development provide for classification of students according to broad levels of English proficiency. However, they are not refined enough to capture gradations of language development. Moreover, because they are aligned with English language arts standards, they are not
designed to yield a full picture of the communicative competence required for proficiency in a second language. In addition, performance on even these tests—designed to assess the language skills needed in school—may not correlate highly with performance on achievement tests in language arts. Due to these limitations, existing tests of English development may produce invalid measures of English proficiency or fail to measure important aspects of language proficiency that are relevant to learning in an English-dominant school context or to demonstrating knowledge on tests administered in English.

Testing ELLs
As discussed in the previous sections, ELLs even within the same broad linguistic group vary tremendously on their linguistic proficiencies across the four language modes in both L1 and English. This heterogeneity is accentuated by important factors usually not considered in the definition of ELL, such as dialect differences, history of formal education in L1, history of participation in bilingual programs, and fidelity of implementation of those bilingual programs. In addition, recall of experiences is influenced by the language in which an experience was first encoded (Altarriba, 2003; Marian and Neisser, 2000). If the student is asked to use the language associated with the original experience, he or she is more likely to recall the experience. What this means is that in order to get at an ELL student’s knowledge, one needs to test in the language in which the original learning experience was encoded.

Because of this heterogeneity, ideally a decision about the language used should be made individually for each student, based on his or her proficiencies in both L1 and English. Moreover, the form in which achievement tests are administered to each ELL student should take into account his or her weaknesses and should take advantage of his or her strengths in each of the four language modes within each language.

We know, however, that such an individualized form of testing for ELLs is impossible to implement because of its high costs. Even if cost were not an issue, this approach would fail because of the fact that accurate information on the proficiencies of ELLs may not be available. As discussed before, even if scores from tests of English development are available, these are suspect because the instruments are limited in their effectiveness to render accurate classifications of ELLs.

Given these circumstances, there is a different set of limitations associated with testing ELLs in either L1 or English. In this section, we discuss the limitations of testing ELLs in English and the limitations of testing ELLs in L1. Then we discuss the fact that ELL performance on tests is unstable across languages.
Limitations of Testing ELLs in English

It would be ideal if ELL students were not tested in English until their English proficiency approached that of their native-English-speaking peers (Katz et al., 2004). Until that point, testing in English is not likely to reveal what ELL students know. They may well be performing at a higher level in the classroom than they can show on a test, which requires them to comprehend language out of context, cope with the language of tests (the “test register”), and read and respond in a timely fashion (Katz et al., 2004). The requirement of the No Child Left Behind Act of 2001 that an ELL student be tested in English after three years flies in the face of considerable research showing that ELLs require from four to seven years on average to attain full proficiency in English, including the linguistic forms and functions mentioned above (Hakuta, 2001; Ramirez et al., 1991). Thus, this effort to legislate educational accountability is driving practices that threaten the validity of academic achievement measures (Abedi, 2004).

Critical to testing ELLs in English is the use of testing accommodations intended to reduce score variance due to students’ limited proficiency in this language. In Chapter 10, Koran and Kopriva provide examples of algorithms and computer software for processing multiple student background variables that can increase the ability of assessment systems to make appropriate accommodation assignments. Unfortunately, these exciting technological innovations are not paralleled by conceptual developments on the nature of the accommodations. To date, there is no theoretical framework that educators, administrators, or even assessment systems can rely on to design defensible testing accommodations for ELLs—a reflection of the fact that testing practices for ELLs have not incorporated current knowledge from the fields of bilingualism and second language acquisition (e.g., see Figueroa and Hernandez, 2000; Hakuta and Beatty, 2000; Solano-Flores and Trumbull, 2003).

Many accommodations used in ELL testing are borrowed from the sets of accommodations used for students with disabilities or are based on misinformed views of language. For example, Ferrara et al. (2004) reviewed surveys conducted by different agencies and programs (including the National Center on Education Outcomes and the Council of Chief State School Officers) on ELL testing practices and found that some of the actions reported by some states as accommodations for ELLs included the use of enhanced lighting conditions or larger font size. Such practices expose a deep misunderstanding of the nature of ELLs’ needs (Figueroa and Hernandez, 2000) and cannot be expected to yield valid measures of academic achievement.

Whereas not all practices concerning accommodations are deplorable, the accommodations used by major assessment systems in the country fail
to fully address student proficiency in the four language modes. For example, as Table 8.1 shows, of the nine testing accommodations permitted by the National Assessment of Educational Progress (2005), four (see Accommodations 1–4) do not appear to be directly related to language and, therefore, their ability to reduce the effect of language proficiency on the performance of ELL students may be seriously limited. Although they may enhance the conditions in which ELL students are tested and they may even contribute to reducing the performance gap between ELLs and non-ELLs, they do not operate on factors related to language proficiency; and their impact on student performance is not specific to ELLs. The other five accommodations address language, but only partially. Three of them (Accommodations 5–7) are effective only for ELLs who have well developed literacy in L1. The other two (Accommodations 8–9) are effective only for students with specific combinations of differences in their listening and speaking proficiencies across English and L1. Since information on ELLs’

<table>
<thead>
<tr>
<th>Accommodation</th>
<th>Related to language</th>
<th>Condition assumed in ELL student</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Small group</td>
<td>No</td>
<td>——</td>
</tr>
<tr>
<td>2. One-on-one (tested individually)</td>
<td>No</td>
<td>——</td>
</tr>
<tr>
<td>3. Extended time</td>
<td>No</td>
<td>——</td>
</tr>
<tr>
<td>4. Preferential seating</td>
<td>No</td>
<td>——</td>
</tr>
<tr>
<td>5. Bilingual version of test (Spanish/English) a</td>
<td>Yes</td>
<td>Student is able to read in L1</td>
</tr>
<tr>
<td>6. Bilingual word lists or glossaries (Spanish/English) b</td>
<td>Yes</td>
<td>Student is able to read in the L1 and is familiar with the academic language in L1</td>
</tr>
<tr>
<td>7. Bilingual dictionary without definitions c</td>
<td>Yes</td>
<td>Student is able to read in L1</td>
</tr>
<tr>
<td>8. Passages, other stimulus materials,</td>
<td>Yes</td>
<td>Student is more proficient</td>
</tr>
<tr>
<td>or test questions read aloud in English or presented by audiotape c</td>
<td></td>
<td>listening in English than reading in English</td>
</tr>
<tr>
<td>9. Passages, other stimulus material,</td>
<td>Yes</td>
<td>Student’s comprehension of academic language spoken in L1 is better than student’s reading comprehension of academic language in English</td>
</tr>
<tr>
<td>or test questions translated aloud into native</td>
<td></td>
<td></td>
</tr>
<tr>
<td>language or presented by audiotape d</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


a Only for mathematics and science.

b Only for science.

c Except for reading.

d Only for Spanish/English, bilingual mathematics and science.
reading and writing skills in both English and L1 is not always collected or available, it is not difficult to see that many accommodations for ELLs have only a slim chance of contributing to the production of more valid scores. In other words, blanket measures not tailored to individual differences are likely to mask their effects on certain students.

Limitations of Testing ELLs in L1
While testing ELLs in English is limited in its capacity to produce valid academic achievement scores for ELLs, it cannot be assumed that testing ELLs in their first language is necessarily a better option, even in cases in which that is their preferred language (August and Hakuta, 1997). Students who have been placed in English-only programs will have been learning content-area vocabulary in English and may not have parallel vocabulary knowledge in L1. The standard wisdom is that such students should be tested in English (e.g., Abedi et al., 1998; Kujawa et al., 2001). However, as mentioned, these same students cannot be expected to have vocabularies in English equivalent to their native English-speaking peers (Baker, 2000; García, 1991). Their teachers may overestimate their English proficiency because of their conversational fluency (Kujawa et al., 2001). Moreover, even ELL students who have been classified as English-language proficient are often unable to demonstrate their content-area skills on English-language achievement tests (Linquanti, 2001) and may perform better on a test in their home language (Abella et al., 2005, p. 141).

Even if it is assumed that limited academic language proficiency in L1 is not an issue in the testing of ELLs in that language, there are serious threats to the validity of tests in L1 that result from the process of their development. Usually, tests administered in L1 are translations of tests originally developed in English. Tests in English have been developed through a more or less careful process in which their wording is refined and tried out with students to identify and correct for misinterpretation of the test items. In contrast, tests administered in L1 are usually translated from English under very tight timelines that limit the possibilities for refining the wording of items (Solano-Flores, 2007b; Stansfield, 2003). Moreover, several mistakes commonly made in the selection of translators jeopardize the validity of translated tests. Those mistakes include judging the translation skills of individuals on the basis of their perceived speaking skills in the target language or assuming that individuals are good translators just because they are bilingual or because they are good writers in the target language (see Hambleton and Patsula, 1999). These are very serious issues, if we consider that even when sufficient time and resources are allocated for test translation, test items may end up measuring different constructs (Van de Vijver and Poortinga, 2005).
Procedures for the simultaneous development of tests in two languages have the potential of eliminating the threats to validity that result from translating tests (Rogers et al., 2003; Solano-Flores, 2007a; Solano-Flores et al., 2002). These procedures are based on the notion that meaning cannot be transferred totally across languages (Greenfield, 1997) and the fact that translated items may function differently across the populations tested in the original language and in the target language (Sireci and Allalouf, 2003). This differential functioning can be extremely sensitive to even subtle variations in the precision of translations (Ercikan, 2002).

The simultaneous development approach is based on the notion that an equal treatment of languages throughout the entire process of test development is most likely to preserve meaning across languages. In this approach, tests are developed in two languages concurrently and pilot-tested with an equal number of students from the two target linguistic groups (monolingual and ELLs who are native speakers of a given language); the two language versions go through the same number of review-revision iterations; also, any modification made in one language version as a result of pilot testing has to be reflected in the other language version.

It would not be reasonable to attempt to develop tests simultaneously in English and in multiple languages. However, it is feasible to develop tests simultaneously in English and in Spanish, the native language of the vast majority of ELLs in the U.S. The simultaneous development approach does not appear to increase test development costs (see Solano-Flores et al., 2002). However, it requires from assessment systems the willingness to modify their test development practices in a way that the minority language is included in the test development process from the very first draft versions of a test, rather than as an afterthought, when few significant improvements can be made.

In the absence of proper mechanisms for examining ELL students’ schooling background and their exposure to academic language in L1 and without rigorous procedures for test translation, testing ELLs in L1 may not produce valid measures of academic achievement.

Major Challenges in the Selection of Language for Testing ELLs
It appears from the discussion above that no perfect solution exists when it comes to deciding what language should be used to test ELLs. Even if student classification and implementation were not a problem in the testing of ELLs, an issue yet to be properly addressed is a fact well known among linguists—that bilingual individuals do not function as two monolinguals in one (Grosjean, 1989). Since bilingual individuals tend to have different sets of strengths and weaknesses in each language, their ability to address the cognitive challenges posed by an item is not replicated across languages.
(Genesee, 1994; Hakuta et al., 1987; Stevens et al., 2000). As a consequence, there is something about an ELL’s academic achievement that is not captured by testing him or her in L1 but also something about the same student’s academic achievement that is not captured by testing him or her in English.

This limitation of testing ELLs in one language is expressed as a considerable amount of error when ELLs are given the same set of items in English and in L1. By using generalizability (G) theory (Cronbach et al., 1972; Shavelson and Webb, 1991), a psychometric theory of measurement error, it is possible to estimate score variation due to student—the object of measurement—and several facets (sources of measurement error). In a series of investigations performed with native speakers of different languages, we have estimated the score variation due to the main and interaction effect of student, rater, item, and the language of testing. Table 8.2 presents results for native Haitian-Creole speakers in the same school district (Solano-Flores and Li, 2006), who were given the same set of grade 4 mathematics items sampled from the NAEP 1996 public release (National Assessment of Educational Progress, 1997) in English and in their first language. The table shows the estimates of the amount of variance (and the percentages of that variance) due to 15 terms that result from the main

<table>
<thead>
<tr>
<th>Source of Variability</th>
<th>n</th>
<th>Estimated Variance Components</th>
<th>% of Score Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>student (s)</td>
<td>49</td>
<td>.0299</td>
<td>20</td>
</tr>
<tr>
<td>rater (r)</td>
<td>4</td>
<td>.0000*</td>
<td>0</td>
</tr>
<tr>
<td>item (i)</td>
<td>10</td>
<td>.0097</td>
<td>6</td>
</tr>
<tr>
<td>language (l)</td>
<td>2</td>
<td>.0074</td>
<td>5</td>
</tr>
<tr>
<td>sr</td>
<td></td>
<td>.0000*</td>
<td>0</td>
</tr>
<tr>
<td>si</td>
<td></td>
<td>.0164</td>
<td>11</td>
</tr>
<tr>
<td>sl</td>
<td></td>
<td>.0100</td>
<td>7</td>
</tr>
<tr>
<td>ri</td>
<td></td>
<td>.0004</td>
<td>0</td>
</tr>
<tr>
<td>rl</td>
<td></td>
<td>.0001</td>
<td>0</td>
</tr>
<tr>
<td>il</td>
<td></td>
<td>.0016</td>
<td>1</td>
</tr>
<tr>
<td>sri</td>
<td></td>
<td>.0000*</td>
<td>0</td>
</tr>
<tr>
<td>srl</td>
<td></td>
<td>.0000*</td>
<td>0</td>
</tr>
<tr>
<td>sil</td>
<td></td>
<td>.0589</td>
<td>39</td>
</tr>
<tr>
<td>ril</td>
<td></td>
<td>.0000*</td>
<td>0</td>
</tr>
<tr>
<td>sril,e</td>
<td></td>
<td>.0166</td>
<td>11</td>
</tr>
</tbody>
</table>

Source: Adapted from Solano-Flores and Li (2006).

*Small negative variance of negligible magnitude set to zero, following Brennan’s (1992) approach.
effect and interaction effect of student, rater, item, and language. The last term (sril,e) is the error term, which cannot be interpreted because the effects of student, rater, item, and language are confounded with error, e, due to other, unknown sources of score variation. The most important source of measurement error in the testing of ELLs is the interaction of student, item and language (sil), which accounts for 39% of the score variation.7

These results indicate that, to great extent, a given student may perform well or poorly on a given item, depending on the language in which the item is administered. Also, these results indicate that, in addition to its intrinsic cognitive and content knowledge demands, each item poses a particular set of linguistic demands in English and a particular set of linguistic demands in L1. At the same time, as we have discussed, in addition to his or her academic strengths and weaknesses, each ELL has a particular set of strengths and weaknesses in English and a particular set of strengths and weaknesses in L1.8

**Section Summary**

The validity of scores from both tests administered in L1 and in English is threatened by the fact that the official definitions of ELLs and the tests used to measure English language development may produce erroneous information on the preparedness of students to be taught or tested in one or another language. Effective testing of ELLs in L1 is limited by the fact that it does not take into account, among other things, the formal instructional experiences that these students have had in English. On the other hand, effective testing of ELLs in English is limited by the lack of an appropriate theoretical framework that allows making sound testing accommodation decisions and the lack of mechanisms and resources that ensure fidelity in the implementation of testing accommodations. Moreover, effective testing of ELLs in either L1 or English is limited by the fact that the performance of ELL students varies considerably across test items and the language in which those items are administered. As a consequence, even under conditions in which ELL classification and fidelity of implementation are not a problem, testing ELLs only in L1 or only in English yields only a partial picture of their academic achievement.

**Future Directions**

Unfortunately, there is neither a simple nor an optimistic answer to the question, “In which language should ELLs be tested?” Asking the question would not be necessary if testing practices were sensitive to the characteristics that are critical to the condition of being bilingual. As a
consequence, sound decisions on the use of one or another language to test ELLs require deep examination of both the quality of the information on the students' linguistic proficiencies and the resources needed for properly implementing testing in each language. Equally unfortunately, the goals of the No Child Left Behind Act—to make districts accountable for the academic success of every student—cannot be met if the quality of assessment information about ELLs is problematic and if schools are not well equipped to properly implement tests in L1 or to provide proper accommodations when they give ELLs tests in English.

Throughout our discussion, it has become evident that, to a great extent, the process of deciding about the language used to test ELLs has to address the uncertainty that results from flawed classifications of ELLs, limited effectiveness of testing accommodations, and limited resources needed to properly implement those few testing accommodations that have been proven to be effective in reducing the effect of language proficiency on test performance. Thus, we conclude this chapter by sharing results from research that examines which language should be used to test ELLs from the perspective of score dependability. Ultimately, this research is intended to allow development of testing models that are resilient enough to respond to the fact that ELL populations are tremendously heterogeneous and that address factors critical to the condition of being bilingual.

**Local Specificity of Testing Models**

In the context of our studies on the instability of the performance of ELLs across items administered in L1 and in English (see above), we have performed decision (D) studies with the purpose of determining the minimum number of items needed to obtain dependable scores with tests in which all the items are administered in English, all the items are administered in L1, and some items are administered in English and some in L1. Key to examining the effectiveness of those testing models are two types of G coefficients: \( \rho \), a relative-decisions, norm-referenced coefficient that refers to the relative standing of students, and \( \phi \), an absolute-decisions, criterion-referenced coefficient that refers to the absolute level of student performance (e.g., the number of items correct) (see Shavelson and Webb, 1991).

We (Solano-Flores and Li, 2006; in preparation) have found that the most effective testing models for ELLs—the ones that give the highest G coefficients with the minimum number of items—are those in which students are tested only in one language. However, as Table 8.3 illustrates, we have also observed that even within the same broad linguistic group (i.e., ELLs who are native speakers of the same language), the G coefficients obtained may vary considerably across groups of students from different localities (schools or school districts). For example, we have observed that
the minimum number of items needed to obtain dependable scores if students are tested in L1 and the minimum number of items needed to obtain dependable scores if students are tested in English may be different for students from communities where the same L1 is spoken.

A consequence derived from this finding is what could be called, the principle of local specificity of testing models: the number of items needed to test ELLs in either L1 or English cannot be generalized across broad linguistic groups (all ELLs) or across localities within the same linguistic group. Therefore, in order to ensure score dependability, any decision about the language in which ELLs students should be tested should be made separately for each locality.

**Optimum Number of Items as Testing Accommodation**

The finding that a considerable amount of score variation in the testing of ELLs is due to the interaction of student, item, and the language of testing supports the notion that, in addition to being samples of the knowledge domain that a test is intended to measure, test items are unintended samples of the language of testing (Solano-Flores, 2006). Thus, a form of testing accommodation can be devised in which the language to be used to test a particular group of ELLs is determined by giving students from that group the same set of items in two languages and then performing a series of D studies to determine the minimum number of items needed to obtain dependable scores when they are tested in L1 and the minimum number of items needed to obtain dependable scores when they are tested in English. With this form of accommodation, different groups of ELLs may be found to need to be tested with different numbers of items in different languages. Also, different number of items may be needed to produce dependable scores for ELL students and non-ELL students.

Figure 8.3 shows the procedure for implementing optimum number of items as a form of testing accommodation for ELLs. In addition to testing ELLs in English, as is now mandatory, the procedure considers the possibility of testing ELLs in L1. As a result of the No Child Left Behind legislation, the national trend is to test ELLs in English after a short

<table>
<thead>
<tr>
<th>Site V</th>
<th>Site W</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho^2$</td>
<td>.21</td>
</tr>
<tr>
<td>$\phi$</td>
<td>.20</td>
</tr>
</tbody>
</table>

Source: Solano-Flores and Li (in preparation).
Figure 8.3 Procedure for using optimum number of items as a form of testing accommodation for ELLs.
time of instruction. However, there may be states or school districts for which testing ELLs in L1—although with some restrictions—may still be an option.

Consistent with the notion that each group of ELLs is unique and each content area poses a specific set of linguistic demands, this procedure should be performed separately for each locality, for each broad linguistic group in the locality, and for each content area.

Some of the terms and directions used in the figure deserve elaboration. First, the term “locality” (see Box 1 in Figure 8.3) refers to either a school or a school district. Whereas, ideally, this form of accommodation should be performed at the school level, that is not always possible because a good number of schools in the country have small enrollments of ELLs.

Second, the sample of items used to test ELLs in the preparatory stage should reflect the types of items (e.g., open-ended, multiple choice) to be used in the operational stage of the test. Therefore, the sample of items should be provided by the assessment system (e.g., they can be assembled with items from a pool of released items from previous test administrations).

Third, when testing ELLs in L1 is an option (Box 2), it is important to keep in mind that translations made by translators in a “standard” version of a language may not be as sensitive to differences in language use across localities and the ways in which curriculum is enacted as are translations made by teachers from the students’ localities (Solano-Flores et al., 2005; Solano-Flores, 2006). Ideally, the translations of test items into the students’ native language should be made by translation teams that include teachers from the same localities.

Fourth, testing students “with the same items in both English and the native language” (Box 3b) means giving each item in the two languages on different occasions, not providing side-by-side formats in which the two language versions of the same items are administered simultaneously (e.g., Sireci and Khaliq, 2002). Appropriate actions must be taken to control for the effect of learning, fatigue, and repetition. For example on each testing occasion, students are given some items in English and some items in L1; on the second occasion, they are given the complementary language versions of the items administered on the first occasion. Also, the order in which the items are sequenced on each occasion is determined randomly.

Fifth, the importance of rater as a source of measurement error should not be underestimated (Boxes 5a and 5b). Even if ELLs are tested only in English (Box 5a), the proficiency of raters in the students’ native language may have an important effect on the accuracy with which they interpret student responses to open-ended items. In our research (see Solano-Flores and Li, 2006), we have found a negligible amount of score variation due to rater. However, we believe that this small effect was due to the fact that the
raters we used were teachers from the same school districts as the students and were native speakers of the students’ native language. In another study in which groups of teachers developed and scored a literacy test for students who were newly transitioned to English-only instruction, one of us documented how knowledge of students’ first language (Spanish) aided interpretations of responses written in English (Beaumont et al., 2002; Trumbull and Koelsch, in press). Unfortunately, reports on the testing of ELLs provide only superficial information on the language proficiency of the raters. To what extent the linguistic skills of the raters in the ELL students’ native language influence the quality of their scoring of student responses when they are assessed in English is yet to be investigated.

In conclusion, a decision about which language to use in testing ELLs depends upon the specific group of students in question. The decision-making process is fraught with difficulties, not least of which are those posed by the limitations in existing language assessments, lack of language professionals who can evaluate students’ language, and the dearth of research-based, theoretically defensible approaches to matching testing accommodations to student needs.

In an ideal scenario, such a decision would be made for each student, with consideration given to many factors. Among these are the student’s particular patterns of proficiency in L1 and L2, his or her educational background, and the cultural group(s) in which he or she has been socialized.

We have shown how one alternative approach to assessment, the “optimum number of items” method, has proven to be a feasible way to obtain better information about ELL student learning and achievement. Research on this method points to the complexity of interactions among student, language, and test item. It also highlights the importance of local involvement in designing and developing assessment systems. Even in the absence of perfect information on every student, teachers and administrators have knowledge of their student populations that can guide decisions about assessment practices within a school or district. And, even in the absence of perfect language tests based on a communicative competence view, an understanding of students’ language development can be constructed through the use of existing tests, teacher observations, and communication with parents about their children and how language is used in the home. If the education community is truly committed to maximizing the validity, utility, and equity of test scores, it must seriously entertain such alternative methods.

The issues we have discussed in this chapter pertain not only to decisions about choice of language for student testing but also to the assessment of ELLs in general. Districts may not be able to craft assessment systems that are perfectly aligned to the dual needs of their student populations and public accountability. Nevertheless, with a deep understanding of the
limitations of current assessment practices for ELLs and an awareness of the kinds of possibilities that hold promise for improved assessment, they are likely to develop better visions of what is feasible within their own contexts.

Endnotes

1. Here, we use the components of language proficiency used by educators. “Language proficiency,” a disputed construct, has traditionally been based on evaluation of a person’s knowledge and use of oral language.

2. In reality, each individual has his or her own linguistic repertoire, based on individual life experience, so even the term “dialect” obscures many individual differences. However, it is often associated with participation in a particular social and cultural community. A dialect is a variety of a language that is distinguished from other varieties of the same language by features of pronunciation, grammar, vocabulary, and use (Crystal, 1997; Wolfram et al., 1999). Cultural groups vary in their beliefs about how language ought to be used by children versus adults, what is appropriate in different settings, and for what purposes language should be used (Bialystok, 2001; Greenfield et al., 2000; Heath, 1983, 1986). Although the term “dialect” is sometimes taken to mean a sub-standard version of a language, in fact everyone speaks a dialect, and all dialects follow their own consistent rules of grammar. The most socially prestigious dialect is usually labeled “standard,” as in “Standard English” (Preston, 1998).

3. The term “academic language” is now used to refer to the advanced oral and written forms and functions of a language required to achieve in school including the vocabulary, grammar, and discourse conventions associated with the different disciplines (Chamot and O’Malley, 1994; Scarcella, 2003).

4. According to Katz et al. (2004), “proficiency” has been historically associated with the “proficiency movement” of the American Council on the Teaching of Foreign Languages and has particular associations. Nevertheless, it is the term most commonly used in legislation and education literature.

5. It should be pointed out that tests in which particular language skills are examined out of context also contribute to an understanding of a student’s language proficiency (see, e.g., Katz et al., 2004), despite their limitations in capturing language in use.

6. Figueroa and Hernandez (2000) argue on the basis of several studies that processing time in L2 is longer and that the difference gets magnified as a task becomes cognitively harder. In fact, a recent experimental study with Spanish–English bilingual adults showed that they took significantly longer to retrieve common words in their non-dominant language than in their dominant language (Langdon et al., 2005). So extending the time for testing would, in that case, address a language need, particularly for more difficult test items. Extra time has also been shown to benefit monolingual students when scores of large groups of ELLs and monolinguals are compared. It could be argued that extra time for ELLs alone would be more equitable, to compensate for the greater ease of linguistic processing by monolinguals.

7. Notice that the score variation due to the main and interaction effect of rater is minimum. In our studies, we were very careful in selecting raters who were native speakers of the ELL students’ L1 and who had experience teaching students from the same localities. A larger effect of rater might have been observed had we had included raters with other language backgrounds (e.g., bilingual teachers whose L1 was English).

8. August and Hakuta (1997) discuss the results of an investigation with translated tests, in which Anderson (cited by the same authors) observed that ELLs performed better on some items administered in L1 and better on some items administered in L2. August and Hakuta suggest that this performance inconsistency across languages was attributable to the fact that item difficulty was not controlled for. In our investigation, differences due to different difficulties across languages would be reflected as a large main effect of language and large interaction effect of the facets, item and language. However, these effects were small and a considerable effect due to the interaction of student, item, and language was observed—which supports our conclusions.
This chapter will address administration and response accommodations which appear to be promising for English language learners (ELLs). It will also present two pretest support options not ordinarily discussed in this context but which are accommodating practices that may have potential. As Kopriva (2000), Abedi (2006a) and Rivera et al. (2006) emphasize, the primary focus for addressing the challenges of ELLs in large-scale testing should be properly adapting standard testing materials (and this has been discussed in the last few chapters). Administration and response accommodations and pretest support are sometimes necessary as well, however, but they are not generally considered to be sufficient if not used in conjunction with appropriate assessments and other linguistic tools and aids. The role of the additional accommodations discussed here seems to be to provide an alternate mechanism for communicating the testing requirements (for instance, oral administration in L1 or English), allowing flexibility in the time or procedures associated with the administration of the assessment, or providing flexibility in how students can explain what they know. The purpose of pretest support, which is distinct from test prep practice in vogue today, is to familiarize students who have different expectations or a lack of some types of experiences associated with large-scale testing.

Recently, two researchers described current state accommodation policies for ELLs. In reviewing data from 2000–01, Rivera and Collum
(2004) reported that seventy-three accommodations of all types were allowed for ELL students across the nation, with forty-eight administration or response options identified. In 2006 Rivera et al., reported that, while seventy-five were more recently reported by states, forty-four accommodations seemed to address the needs of English language learners (these included presentation, administration and response options). In an evaluation of Rivera’s 2004 data, Abedi (2006) described that, of the seventy-three accommodations cited for ELLs, he considered only nineteen to be moderately to highly relevant for this population. Of those, ten administration and two response accommodations were cited as being at least moderately relevant. In Abedi’s work only accommodations which directly address language needs were identified as relevant. Therefore, some administration and response accommodations were not considered to be useful even though, ostensibly, they could and should be used in conjunction with the linguistic accommodations. For instance, extra time, small group, frequency of breaks, familiar personnel administration, audio-administration or response of the assessment in English, and response options including voice recognition or other computer manipulations were not considered.

Pennock-Roman and Rivera (2006) completed a meta-analysis of the experimental accommodation studies from 1997 to 2004 and analyzed both direct linguistic effects and indirect administration effects. They found four studies that investigated extra time and one which investigated small group. It does not appear that any other work has been completed which intentionally evaluated administration effects, although some other work have used administration accommodations in their studies (e.g., Emick and Kopriva, 2007), or time was unspecified to begin with (e.g., Abedi and Lord, 2001; Kopriva and Cameron, 2007). In general it is impossible to disentangle the effect of linguistic accommodations vs. administration ones at this time although we know the beneficial effects are largely interactive and can’t be reduced to additive components. Further, virtually no experimental or quasi-experimental research has been published to date to evaluate the effectiveness of response accommodations, although some qualitative cognitive lab results suggest these may be useful (Winter et al., 2006). To date, no pretest support investigations have been undertaken either. All in all, it seems prudent that future work should investigate how these accommodations interact with students’ linguistic proficiencies, past experiences, and amount of exposure to U.S. culture and schooling conventions. Obviously, these factors are variable among students so the accommodations need to be considered differentially by students with similar profiles.

Promising administration options will be discussed next.
Promising Administration Accommodations

Four administration accommodations are highlighted here: extra time, small group, oral administration (in English and in the home language), and use of a language liaison. In general these appear the more salient administration accommodations which can be used to supplement language options that English language learners may need.

Extra Time

Because of the unfamiliarity of the testing language or context, increased lengths of some native language tests, use of additional language aids, fatigue, or psychosocial factors, English learners frequently appear to need extra time on exams. This is true, of course, when students are given test materials that they can comprehend with enough time—all the time in the world won’t be useful if students can not understand what assessments are asking them to do. Abedi et al. (2001b) found extra time helped all students (not just ELLs) about the same amount, when they received a dictionary and when they only got extra time. Hafner (2001) found similar results in her quasi-experimental study. On the other hand, Abedi et al. (2003) reported that extra time helped English language learners but not non-ELLs. Further, another study (Kiplinger et al., 2000) documented that when extra time was combined with a plain language form in English and then separately with an English glossary, it benefited students with midrange English language proficiency more than those with the highest proficiency or no proficiency in English. While not investigating the effects of time per se, Kopria and Emick (2007) and Kopriva et al. (2007b) found as well that English proficiency seems to impact the validity of inferences of scores that were produced when the test was un-timed. While these studies suggest the benefit of un-timed assessments or provisions for extra time, the studies sampled only one or two grades, and most of them were based on scores from multiple choice items in mathematics.

Of note, Sireci et al. (2003), Pennock-Roman and Rivera (2006) and Kopriva et al. (2007c), among others, are concerned that score increase versus validity has been the focus to date in these analyses. They agree that clear studies need to be completed to determine how and when extra time impacts the quality of the measurements of identified constructs and when it does not. At this time it is uncertain whether getting more items correct on the studies completed to date means the constructs were changed when the time pressure was relaxed, or students were allowed to address the items in a timeframe that worked better for them. As long as this question can be answered during the development of the items and testing systems, the
trend toward un-timed tests would seem to be useful for whoever needs them (including but not limited to many ELLs). Further, there has been no work to suggest that un-timed tests injure some students. On the other hand, if items have been developed to be sensitive to time and this is a clear condition of the subject matter construct, development specialists would need to differentiate between when a student should receive extra time to complete basic processing work versus time which would lead to changing the nature of the construct. For instance, some developers have suggested that increased time may change the nature of the writing piece students are supposed to compose on some assessments. When this is the case developers need to receive a judgment from both subject matter specialists and specialists from populations which might be at risk, in order to conclude which students may need more time for other reasons and which students should not receive more time. It is suggested that this judgment is what should guide the practice and research around this accommodation.

**Small Group**

Most large-scale academic content tests are typically administered in a large group setting. For some ELL students, this setting may be problematic to the point where it interferes with their ability to perform adequately on the test. Participants in a focus group of ELL teachers suggested that this seems to be particularly true for beginner English learners and those students who have recently arrived from other countries (Douglas, 2004). For these students a small group administration, if possible with the student’s ELL teacher or another known adult (for instance a classroom aide), may be a viable accommodation. Only one study (Abedi et al., 2003) has examined the effect of small group administration and in this investigation they researched the effect of the accommodation without combining it with any other. The results were similar for both ELLs and non-ELLs, where students did less well when separated from other students. It seems that, alone, this accommodation may not be useful. On the other hand, when it is paired with language options, it may be useful for students with less English proficiency. No work has been completed to date to understand how the accommodation impacts other groups of ELLs, or to investigate the effect of accommodation packages that included but were not limited to small group. Several studies (see Koenig and Bachman, 2004) have been carried out with packages of accommodations for students with disabilities. While small group was not specifically investigated, the nature of the research designs (single subject) or the combination of disability and research package suggests that students often were tested in an individual
or small group administration. Several of the results favored students with disabilities over students without disabilities but it is not clear what effect the individual or small group administration had on these findings.

**Oral Administration**

Separate sections will briefly consider oral administrations in English and the home language of students.

**Oral English**

This accommodation seems to be one of the most frequently used accommodations for ELLs (for content areas other than reading, Rivera and Collum, 2006), but there appears to be almost no empirical data to suggest that it is useful or effective. Rivera and Collum suggest that most policies assume the test will be read aloud by a school staff person. Several disability researchers (for instance, Tindal *et al.*, 1998) urge districts to use video (or audiotapes if necessary) for a standardized administration because of this problem. Early work with ELLs (Kopriva and Lowrey, 1994) cautioned that the investigators found evidence of cuing students if oversight was lax or non-existent. They also reported that students throughout many districts in California who had developed oral proficiency skills in English, along with some amount of English literacy, preferred oral administration in English as compared with only written English. Recently, classroom evidence suggests that oral administration in English may be appropriate for a subset of ELL students who have achieved a level of verbal fluency in English but not enough English literacy (Kopriva, 2005a). Further, Kopriva *et al.* (2007e) found that, based on similar criteria, the ELL students who received oral administration (as well as zero, one or two linguistic aids that the criteria thought were appropriate) did significantly better than those ELLs who received the accommodation but did not fit the criteria. With a larger sample, Emick and Kopriva (2007) found that, overall, scores for a subgroup of ELLs with intermediate English proficiency for whom the oral accommodation was considered to be an appropriate accommodation (along with other accommodations as appropriate) were still significantly less valid than for non-ELLs and exited students. This was particularly true for the multiple choice items. The literature suggests that some students with enough oral English fluency but not enough English literacy may benefit from a combined oral administration with appropriate linguistic aids. However, evidence regarding similar validity of score inferences for ELLs who receive this accommodation as compared to non-ELLs has yet to be empirically supported.
ORAL HOME LANGUAGE

Administration of the academic test in L1 is an accommodation for which there currently does not appear to be any published documentation. The state of Michigan has recently begun providing video of an administrator reading the assessment text in Spanish and Arabic (J. Martineau, personal communication, February 2007). Researchers are scaling the data from these administrations in order to evaluate the factor structures as compared to the structures from the standard written administration approach (Sireci, 2007).

The same issues of administration would apply here as when tests are read aloud in English. That is, if possible, the administration should be administered in a consistent way across schools, usually by using an audiotape or videotape of the text being read aloud, or oral administration on the computer. In Michigan, the written text given to the students who are viewing one of the videotapes is in English; others recommend that the written text should be in the students’ home language if possible (August et al., 2004).

Lara and August (1996) emphasize that many students who are immigrants are not literate in their home language but speak it fluently. Kopriva and Lowrey (1994) reported that students preferred oral L1 administration if they were now being taught in that language but did not have sufficient literacy skills in either their native language or English. They also found that students who were new to the U.S. were currently being taught in English, and those who were not literate enough in their home language to take a paper-and-pencil test in that language preferred oral L1 administration as well. Likewise, Pennock-Roman and Rivera (2006) speculated that newly arrived students might benefit from oral L1, even though they were being taught in English.

Until research informs the decisions about when to use oral L1, and with whom, common sense would suggest that low literacy levels should dictate if a student receives an oral administration. In general, if the student has reached a reasonable level of oral fluency in English, has only rudimentary literacy skills in English, and is being taught in English, participating in an English oral administration would make sense. If, on the other hand, the student is new to the country (regardless of the current language of instruction), or if he or she is being taught in L1 but does not yet have the literacy skills in L1, an oral administration in L1 would seem to be appropriate.

Language Liaison

At the request of some of the agency partners on two related projects, a recent study (Mann et al., 2006) utilized the concept of language liaison as...
an accommodation provided to a small number of specified students taking a large-scale plain language test in English. The researchers found nothing in the large-scale testing literature to guide them in developing or evaluating the accommodation, although classroom assessment has discussed these concepts (for instance, see Farr and Trumbull, 1997).

Aikenhead (1996) introduced the idea of a teacher as cultural broker or language liaison for students in the classroom environment. In the large-scale testing context, and how it was used in the study by Mann and others, a language liaison is a person trained for this purpose who bridges, links, or mediates between students of a specific language and the standardized testing context. He or she is generally from the student’s culture, and if possible is a staff member in the school system and familiar to the students. Their function is to answer very select language questions in the assessment text from students, individually, that they’ve been trained to answer. Glossary word lists are very limited (much of the language doesn’t appear on them—for instance, tenses of verbs, phases, colloquialisms, or abstract conceptual language that would take more language to explain), and most of their benefit is in translating nouns. Agency partners suggested that this accommodation may be appropriate for new arrivals if the students are not literate in their L1, have little to no fluency or literacy in English, and for whom no oral L1 administration was available. As the accommodation was structured, it is conceivable that two or more language experts could be in the room together to offer assistance with students from more than one language.

Clearly, there are risks associated with trained persons answering queries that go beyond questions about test directions. Two essential components of this accommodation are proper training and proper monitoring and oversight. As developed for the study, training consisted of specific instructions and guidance about what language questions the liaison may and may not answer for the student, role playing, and rehearsal. This training occurred over and above instructions about general test administration, and each person was given hard copies of the language liaison guidance. Language liaison trainees needed to be sufficiently bilingual in English and the L1 of the students and evaluation of their bilingual skills was part of the screening that occurred prior to and during the training. For the project, administrations were taped and monitored for quality control purposes. The focus of the study was not to evaluate the efficacy of the accommodation by giving it to control groups, but rather to get a formative sense of its viability and usefulness. A preliminary review of tapes made during the assessment administration suggested that liaisons did not cue students. Current analyses are determining how, given their English proficiency status, students who received this accommodation
fared relative to ELLs at the same proficiency level who received other accommodations.

**Promising Response Accommodations**

Most standardized testing that is produced today utilizes close-ended items, most often multiple choice. While this item type doesn’t require writing, and the text of the key and distractors often is rather short, the format makes several assumptions that haven’t been well researched. Kopriva et al. (2007) found that beginning ELLs on average were responding to multiple choice items almost at the chance level in grade 3 and not different than chance at grade 5. There was a significantly higher percentage of ELL students (particularly beginners and intermediates) than exited and native English speakers whose teachers reported they had learned and been able to demonstrate specific content skills in the classroom, but whose multiple choice scores on the large-scale test did not reflect this knowledge. This was particularly evident for the multiple choice scores. Further, the validity of scores for beginners and advanced ELLs in both grades was markedly different on multiple choice than on constructed response, where the constructed response validity indices were significantly higher (and on a par with native English speakers in grade 5).

The field has not confirmed that multiple choice items are suitable and valid for some ELLs. This may be because the shorthand language in the key and distractors are read differently by this population. However, as long as this item type is going to be used with English learners it is crucial that work be completed to determine its validity for the population.

Recently, Boals and Cranley (2006) have taken another approach, investigating the use of alternate close-ended formats which are utilizing the interactive and dynamic capabilities of the computer. One of the key elements of this project is that the interactive components are allowing for response options such as manipulation of stimuli elements, and drag and drop. In the long run, solutions such as these may help address the validity concerns seen by Kopriva et al.

In the meantime, it seems prudent to increase the response avenues for ELLs on the current item types. Chapters 5 and 6 have explained how to develop such items that invite multiple avenues of response, such as non-language options and computer simulation possibilities. Work with native speakers has shown that open-ended responses to items which could also be answered via multiple choice produce different results. However, it may be useful to research if allowing open-ended responses for multiple choice type items would increase the validity of the scores for this population.
Outlined below are a few approaches that would work for constructed response and may be effective for items like traditional multiple choice as well.

**Written Home Language or Code-switching**

Chapter 11 discusses scoring issues that arise when scorers are evaluating responses of English language learners to constructed response items. If they are literate in their home language and have had content instruction in this language the written responses may reflect these skills. Whether students are asked to respond in English or not, it is not unusual for English learners to infuse their responses with language or conventions from their home language or to pepper their responses with visuals or other non-verbal representations. Code-switching, or responding with written text which uses both the English and L1 language or conventions, is commonplace among students with lesser amounts of English literacy. Therefore, if at all possible, it is wise to use scorers who are literate in the language of the students as well as English. As one study reported, monolingual English scorers, with training, can be taught to be sensitive to several of the nuances and use of non-verbal instances, score many of them, and, if necessary, pass along a subset of the responses to bilingual scorers (Kopriva and Lara, 1997). To date no work has been done addressing whether this language accommodation is beneficial. Common sense would suggest it is. The main consideration seems to focus not on the allowance of the accommodation *per se*, but on scoring procedures so all papers are scored with the same rigor as English papers. It would be useful if scoring procedures would include an evaluation of a subset of responses written in L1 and/or code-switching in order to ensure that the scores for these students are of the same quality as those of native English speakers. To date, the Kopriva *et al.* (2007) investigation seems to be the only study to include this accommodation in their project. To date, however, the implications of using this accommodation versus not using it have not been evaluated by the researchers.

**Oral in English, the Home Language, or Code-switching**

As in the other response options, no research has been done to determine if and when this accommodation might be useful for ELL students. A couple of studies have looked at the use of scribes for students with disabilities. One experimental study (Fuchs *et al.*, 2000b) reported a differential effect favoring use of a scribe when a combination of accommodations including transcription was used for students with learning disabilities. Johnson *et al.* (2001) found that accommodations (including use of a scribe) did not result in unfair advantage to the students with disabilities.
If oral response, in any language, is handled individually by speaking to a scribe or into a tape player for transcription, the accommodation is time consuming and may invite additional error whenever an adult is involved in representing what the student said. Staff from a research project in the mid-1990s (Kopriva, 1996a) looked into the possibility of using computerized voice recognition software for use with children, but at that time there were no reliable voice recognition programs that were appropriate. Voice recognition software lets people speak into the computer and then “translates” this into written language. For testing purposes existing software would take that written transcription and store it for later printout if scoring is completed by hard copy, or keep it electronic if scorers were accessing materials in that format. It appears that technology now is able to handle children’s responses (particularly with older students). However, results still include a higher rate of misspellings for students than for adults, and the English learners’ accents and use of L1 language more often “confuses” the software program. To date, though, it does not appear that any research into the accommodation has been done utilizing this software.

The oral response option may be particularly important for older students who do not have adequate literacy in English or L1. This option may be useful for this population because it seems that non-verbal responses (for instance use of pictures or graphics) can be used less often as concepts become more complex in the older grades.

**Demonstrated or Modeled Response**

If English writing or speaking is not the target of measurement, the ELL student’s writing is low in both English and L1, and their time in the U.S. is low, consultants to a recent project (Kopriva, 2002) suggested that some students may benefit from showing responses through use of hands-on materials. Although no research findings on this accommodation have been published to date, one study utilized this technique in a computer-based administration where ELL students were allowed to manipulate objects on the screen to answer a few constructed response items (Kopriva, 2005a). This information was computer scored, but, currently, analyses of the results have not been completed. As with oral response, if adults are involved in representing student responses, un-standardized conditions may increase the rate of error. However, the use of technology in presenting stimuli, and storing and scoring the responses makes this accommodation promising. Electronic software could allow students to respond by performing “experiments”, manipulate objects or otherwise provide non-verbal responses which can then be scored through a series of algorithms. The item types being studied by Boals and Cranley (2006) may be useful in computerizing many constructed response or performance tasks and
increasing the ability of test items to measure a larger range of content complexity. Additionally, it seems that the ability to manipulate objects on the screen may be a boon for ELL students who would be allowed to demonstrate their knowledge even on more basic open-ended items and possibly extend how they could respond on the traditional close-ended items as well.

**Pretest Support**

However, while somewhat beneficial, the literature argues that test prep alone is not nearly sufficient in bridging the cultural gap for English language learners. For example, Estrin (1993) noted that students from the dominant culture demonstrate more “testwiseness” as compared with their peers from non-dominant cultures. This testwiseness is not limited to the types of information covered by traditional programs but, rather, includes skills assumed by those familiar with the U.S. mainstream schooling culture. Chapter 2 explains that the significant cultural “disconnect” needs to be addressed for some students who come from other countries or for those whose home communities tend to be rather isolated from mainstream U.S. schooling expectations. A description of what pretest support might include is briefly discussed here. No research has been conducted to date to determine how these supports might contribute to validity of test results, but it is proposed that, prior to testing, identifying and addressing the needs (such as those that are summarized here) may be beneficial in increasing the validity of test scores for this population.

Tindal and Fuchs (1999) reviewed four research studies that involved teaching pretest strategies to students with disabilities. Although the results were mixed (e.g., the effectiveness of the teaching of pretest testing strategies was inconclusive), the studies were the first to recognize a need for a pretest category of supports which are particularly geared to non-mainstream schooling populations. Since that time other work (Lam, 2004; Monroe, 2004; Samuelsen and Kopriva, 2004; Emick et al., 2007) has begun to actively consider the role of pretest support for culturally and linguistically diverse students. For example, Lam (2004) reports on testing workshops for low-income families, including language minority families. In his study, Lam provided two (math and language arts) sessions during the year. As part of the family workshops teachers explained the role of testing in the U.S. schooling culture, and reviewed test forms and types of questions to be expected. Families completed practice tests with sample questions, practiced strategies and learned how to check students’ writing against a rubric. He reported increased test scores and improved parent–teacher communication regarding specific academic goals.
The computer-based system matches individual ELLs to specific accommodations depending on their needs includes pretest support as part of what it recommends for select students (Kopriva et al., 2006a; see Chapter 10). The concept of pretest support for ELLs assumed in the system goes beyond the currently accepted idea of test prep or the approaches considered by Lam, or Tindal and Fuchs. It is based on the notion that pretest support which is specific to the student’s individual needs and strengths could be useful in offsetting other construct-irrelevant problems with interpreting test scores after students have taken the assessment. Emick et al. (2007) have recently explained more fully what this type of aid might look like. Specifically, they identified two approaches geared to ELL students which target specific sets of needs and strengths the students will bring to the large-scale assessment. The approaches are (1) ongoing classroom support, which would be conceptualized and implemented by the teacher and tailored to individual students, and (2) Family Assessment Night, which would address some common misconceptions across families from similar and different cultures, and which is intended to build a common understanding between the family, child as student, and the school.

Classroom Support

Emick and others suggest that a majority of pretest support would occur as part of ongoing learning in the classroom. These interventions would seem to be most appropriate for students who had little or no experience with the types of formative and summative evaluations used in U.S. schools. Through a series of ongoing activities that build on the student’s prior knowledge and experience, classroom support is a systematic way of introducing the student to the purpose and approaches to testing in the U.S., and to the types of questions of value in classroom, textbook, and standardized tests in this country. The overarching concept of classroom support would not focus on the format or procedures of testing per se (e.g., bubbling in answers or eliminating answers in multiple choice tests) although students may also need this as part of test prep. Five facets of support would seem to be of particular interest.

First, support may focus on aid associated with testing experiences or perceptions about testing that are known and are taken for granted in mainstream U.S. schooling but are foreign to some students. For instance, the support may focus on understanding the purposes of evaluation which occurs before mastery is achieved. Conversely, for others, support may address why they are being tested on material in which they may have already demonstrated mastery. Second, an element for classroom support may be to assist students in understanding the assumptions and rationales that underlie testing as this is different across cultures. For example, classroom tests are
routinely used to provide feedback to U.S. teachers and are a regular part of instruction. These ongoing evaluations are a usual and useful way for teachers in the U.S. system to understand where students are having problems, and consequences associated with any one of these ongoing assessments are negligible or relatively minor compared with some other situations. Some other cultures limit classroom testing to summative evaluations only, and the distinctions about what procedures are being used here and why needs to be clarified for the student and addressed over time. Large-scale on-demand tests also have a function in the U.S. that are different than the experiences of some students or their parents. For some, the large-scale tests they are familiar with are more high stakes than how the assessments are used here. For these students, outcomes of tests in their home country may decide if students will be allowed to move forward in their education or they may decide tracks of education which limit students’ future employment possibilities. How and why large-scale testing is conducted here needs to be thoroughly understood by the students, and explanations may need to be provided over time as students become more comfortable.

Third, the focus of items, and the way items are written or scored in the U.S. have cultural conventions that are less familiar to some students, and teachers need to have identified this need and addressed it properly with ongoing practice evaluations. Therefore, as students are becoming familiar with the purposes of testing and being asked questions to which they may or may not know the answer, another facet of classroom support could be one of appropriate exposure to novel items. For some this may involve requests for factual knowledge in formats not typically experienced by students. For others this may involve experiences with items that ask for more than rote knowledge, such as items which ask students to provide interpretations or speculate on causal implications. Sometimes, because of language limitations teachers may overuse formative classroom assessment that is factual-based, even if students would be comfortable with a broader range of questions. However, since state content standards and large-scale assessments often require students to make inferences or provide their opinions, ELLs of all proficiency levels need to be routinely exposed to and practiced in responding to items that involve the range of skill complexity. As an example, students may be asked to make predictions about plant growth in science. This concept of making predictions and explaining the implications of the results is less common in some cultures. Another example would be a test item that asks students to write a persuasive argument. In some cases, what is unfamiliar in this assessment task may be the audience of the intended response. If the audience is an adult or an adult “in charge” this may be problematic, since, for some students, the idea of children telling these adults what they think or believe is not
acceptable. In each case, it is important for classroom support to include
time to explain what these items are useful for and why, and then time to
practice learning about and working with this kind of question or evaluation
approach.

Fourth, it is not uncommon for ELL students (and others) to lack the
proper exposure to the range of tools and language aids that may be
beneficial for them in the classroom and which would hopefully be available
for use on large-scale tests. Some of the tools, activities and aids are
explained in Chapter 7; others are discussed in state standards and other
instructional materials. Students need to have adequate time to practice
these as well. Fifth and finally, pretest classroom support should probably
include familiarizing students with terminology found on tests such as
phrasal verbs or content specific terms.

All in all, the purpose of individualized classroom support is to provide
the rationale practice time, and appropriate exposure to the types of
experiences and concepts summarized above, and to do so early and in an
ongoing fashion. It may seem like explanations of the types described above
are straightforward and quickly made. However, Monroe (2004) argues that
students need to experience the ramifications of the explanations and the
procedures used in the U.S. schooling systems. These experiences take time
and patience. Thus, it is suggested that this type of individualized support
may be key in allowing students with less experience (for cultural, linguistic
or poverty reasons) sufficient opportunities to be exposed to the culture of
U.S. testing and, as such, be able to more validly perform in the classroom
and on the on-demand assessments.

Family Assessment Night

Designed as a family orientation to the role that assessment plays in U.S.
education, as well as to U.S. standardized testing practices, Monroe (2004)
and Emick et al. (2007) suggest that this assessment accessibility strategy
should be developed especially for ELL families who have little familiarity
with the large-scale tests and testing accommodations. The purpose is to
involve the family of the student (especially the parents and perhaps older
siblings), include them in the familiarization process the student is
undergoing in the classroom, and enlist their support by increasing their
understanding.

Lam (2004) and Monroe (2004) suggest that Family Assessment Night
could serve as a systematic approach to the important work of bridging the
cultural gap between the school and the family as related to testing. This
approach has been used in schools with high ELL populations in content-
focused activities, such as Math Night and Science Night. The approach
utilizes the family as a partner in helping students become familiar with
U.S. testing, and acknowledges the importance of supporting students’ need to move successfully between their two cultures. By comparing and contrasting U.S. testing practices with those from their country of origin, parents and their children learn together about testing, thereby strengthening the parents’ ability to support and monitor their children’s education in their new country. It is intended to be part of a comprehensive partnership plan between parents and school. By communicating with parents in two-way partnership the school communicates expectations and the parents provide insights into the students’ culturally and linguistically diverse educational history, language and literacy proficiency in L1, and other important cultural information. This includes their children’s compensatory strengths and challenges for the purpose of supporting instruction and choosing accommodations. It is anticipated that Family Assessment Night can occur in one or two evenings. Although there are not yet empirically supported guidelines, pragmatic planning would suggest that assessment nights typically last only a couple of hours and that they should be held about six weeks prior to the first large-scale testing event.

These authors assert that there are many components of a successful Family Assessment Night. First, it is important to have buy-in and support from the administration staff, and to have administrative staff members (e.g., principal, vice principal, guidance counselor, school psychologist) attend to convey the importance and value of the families. Hanson and Lynch (1992) caution that, when meeting and working with families, it is imperative to understand that school personnel engagements with the family may be construed as welcome or threatening. Some cultures, particularly Hispanic cultures, believe that the school is capable of attending to their child’s needs and often do not tend to interfere with the educational process. Thus the impetus is on the school to work to create collaborative relationships with families, recognizing the fluid nature of these family systems (Esler et al., 2002). Second, Monroe (2004) suggests it is important to provide time to systematically gather information about students’ previous testing and schooling experiences, their levels of comfort with U.S. testing practices, and the level of cultural discontinuity. If possible, prior to Family Assessment Night, cultural brokers—trained liaisons between the school and the home communities of the students—will interact with the community and document relevant cultural, linguistic, and academic knowledge that the families bring to the table.

Third, at the first Family Assessment Night activity, school staff would provide an introduction about the large-scale tests used in the school. After the opening presentation, the attendees would be split into groups by cultural background, and school personnel would be a part of each facilitation team along with the relevant cultural broker. The broker would
guide the discussion about testing in the families’ country of origin and the similarities and differences from U.S. testing. Parents, along with the brokers, would identify areas of discontinuity, and use that information to help explain expectations to the students. Monroe (2004) suggests it is ideal to have parents share their experiences with the facilitators, leading the conversation from broad cultural differences about schooling and testing experiences to more specific questions about the parents and children’s individualized experiences with testing. A scribe should record relevant academic and testing experience information (e.g., the majority of the parents report their children have limited experience with inference-based questions). This systematic data collection can include any educationally relevant cultural factors that would further help the school understand the values, beliefs and expectations of the families. During the evening, each facilitator should acknowledge some examples of cultural discontinuity. Because cultural discontinuities can be a somewhat vague and obtuse construct, it is imperative to have provided to facilitators beforehand a number of generalizable examples that they can use as needed.

Fourth, Monroe argues that some information about U.S. practices should be integrated into this conversation. Fifth, Monroe (2004) recommends that, if possible, the school engage the parents in a discussion about the appropriate role and use of accommodations. This is to obtain the parents’ input on the needs and strengths of their children and to help parents understand that they do not “dumb down” the test but instead provide accessibility to the content being assessed. Sixth, an important part of the evenings is to allow parents and students some time to review different types of test questions of the types which will be present on the large-scale test. If only one evening is scheduled, this may be an informal exchange after the facilitation group period. If two evenings are planned, this could be the focus of the second Family Assessment Night. Lam (2004) suggests providing parents with copies of the writing rubrics and asking the parents to have their children write each evening and then check their response using the standardized rubric. Parents should also be given information about the content that will be covered on the assessment. Depending on time, this may involve introducing the concept of district or state curriculum standards.

Emick et al. (2007) argue that important consequences of Family Assessment Night are to inform ongoing classroom practice in the school, and to improve how large-scale test results are interpreted to the parents and students. Additionally, sponsoring assessment nights such as what is described here is intended to complement the other opportunities the schools have of interacting with the families and their communities.
Even as tests may be developed and accommodated to specifically address the needs of English language learners (ELLs), if there is no technically rigorous mechanism in place to get the specific methods to the specific students who need them, it is argued that these efforts have little effect. Several researchers who investigate accommodation effectiveness for ELLs and students with disabilities point out that consistent and appropriate accommodations decision-making is critical to the validity of standardized academic testing programs and the ability to properly use scores to compare student performance across states and districts (e.g., Kopriva et al., 2006a; Fuchs et al., 2000a; Kopriva, 2000; Hollenbeck et al., 1998). At the individual level when accommodations decisions are not appropriate to meet the needs of individual students, test results misrepresent their knowledge and skills (Hipolito-Delgado and Kopriva, 2006). At the aggregate level, when accommodations decisions are inconsistent from classroom to classroom or district to district, comparisons across classrooms and districts may be unfair and meaningless (Abedi, 2006a; Solomon et al., 2001; Fuchs et al., 2000).

Test accommodations for ELLs are meant to remove systematic measurement error that exists while the students gain proficiency in the English language and knowledge of common U.S. cultural and educational contexts. Elliott and others (see below) have identified sixteen critical access skills that may present problems for students with disabilities when they take tests under typical standardized testing conditions. Several of these
are salient for ELLs and provide examples of the type of systematic measurement error that accommodations are designed to remediate. When choosing the accommodations a student should receive—including consideration of an option in which the student receives no test accommodations—the goal is to choose the accommodation(s) that reduce the overall amount of error due to factors irrelevant to the constructs being measured on a specific test for the student. While it is probable that other sources of error could be added whenever conditions are varied, for some students it is clear that taking the assessment under non-accommodated conditions will result in scores that substantially misrepresent the knowledge and skills of those students. The Office for Civil Rights (2000) at the U.S. Department of Education specified that the important guideline is which alternative provides better validity of scores, taking into consideration a practicable limit to cost and resource allocations that would be needed to address these issues properly.

Ideally, matching students with appropriate accommodations can be viewed as a process of selecting the most optimal package of accommodations for that student. The selection of a package of accommodations is focused on effectively reducing measurement error for a specific ELL. We will use “matching” and “accommodations decision making” interchangeably throughout this chapter.

In the large-scale assessment literature, research on the most effective accommodations for both English learners and students with disabilities is ongoing. Concurrently, movement is also being made to appropriately integrate these students in standardized testing systems designed for the broad range of students being educated in U.S. schools (No Child Left Behind Act of 2001, 2002). It is generally believed that accommodations should be selected to meet the needs of the individual student. However, research confirms that one cannot validly assign accommodations to groups of students based on some broad classification or status (Sireci et al., 2003). Emerging work suggests that systematic methods of assignment may work better than relying on current policy approaches. Further, it presents evidence that using systematic methods to match the particular needs and strengths of individual students to specific accommodations which address these needs and strengths may increase validity and be superior to using educator directed decision-making alone (Koran et al., 2006; Kopriva et al., 2007e; Tindal, 2006).

Key Assignment Considerations
This section will summarize three key factors which impact the assignment process for English language learners. The first focuses on student factors which appear to be the most salient for making differential decisions about
testing options. Next is a brief summary of the primary accommodations or options which appear to relieve access barriers caused by traditional testing methods, and, finally, issues of quality, and capacity and opportunity are outlined. The latter considerations may influence the decisions, or they may influence the inferences of the scores.

**Student Factors**

Butler and Stevens (1997) focused on three critical background factors that they believe impact accommodation selection: English language proficiency, prior formal schooling, and length of time in the U.S. Abedi (2006a) and Rivera and Collum (2006) largely focused on levels of English language proficiency and the language of instruction. In an extensive literature review, cognitive labs, and interviews and focus groups, Kopriva and others isolated relevant student factors that appear to be salient for matching ELLs to proper accommodations (e.g., Kopriva *et al.*, 2007a, Kopriva *et al.*, 2005a; Winter *et al.*, 2006). As illustrated in Figures 10.1–10.4, the variables identified by these researchers were levels of English and home language proficiency (in reading, writing, speaking, and listening), and an aggregate variable named cultural proximity. The authors also identified key information related to teacher judgments of student needs and compensatory strengths, and teacher data related to ongoing classroom experiences with accommodations and other testing forms and formats. Cultural proximity, as Kopriva and others have defined it, is an aggregate variable which refers to the similarity between the student’s native country schooling (if applicable) and selected home school-like experiences, relative to U.S. schooling and testing opportunities. Aspects of this variable were highlighted by several researchers, including Saville-Troike (1991) and Butler and Stevens (1997). Information obtained focuses on native country schooling experiences, including resources, time and pedagogical approaches, and purposes of testing. The time and consistency of the students’ experience in U.S. schools is also obtained under this category as ELL experts know this is a salient and ongoing issue for this population. Selected literature associated with these aspects will be briefly explained below.

![Figure 10.1 Overview of relevant student factors](image-url)
Figure 10.2 Relevant language proficiency student factors
THE INFLUENCE OF LEARNING ENGLISH

In most large-scale testing situations, students are assumed to have reached a level of English language familiarity and reading ability necessary to understand and successfully respond to test items presented to them in a standard format. Writing proficiency in English is also sometimes key. English language learners, however, are typically not at this level in reading, writing or oral communication although they may have reached varying degrees of language proficiency in each of the domains (LaCelle-Peterson and Rivera, 1994; Durán, 1997). Clearly, ignoring the varying levels of language proficiency reached by ELLs impacts the validity of test performance for those students (Kopriva, 2000). Similarly, lumping ELLs into dichotomous “Proficient/Non-proficient” or ELL/not ELL groups does not take into account the important and salient differences in the level of proficiency each individual student may possess. In addition to limiting the usefulness of the test scores for agency use, this has been a large problem.
in research associated with large-scale testing accommodations (for instance see Duncan et al., 2005).

Considering students’ proficiency in reading as opposed to listening, or writing as opposed to speaking, may have an impact on general understanding of information, learning, and test performance as well. Solano-Flores (2002) and Solano-Flores and Trumbull (2003) found that sometimes different linguistic demands are met by the students’ English capabilities in the different domains. LaCelle-Peterson and Rivera (1994) emphasized that different populations of English learners differ in the process of learning English in terms of whether spoken or written expression develops more quickly. Further, as explained in Chapters 4 and 5, aspects of language knowledge appear to be somewhat differentially problematic for ELL students and native English speakers who have problems in reading or one of the other domains. ELL students present compensatory strengths as well that may or may not be the same as those of native speakers with language difficulties. Several (including Butler et al., 2000) argued that students need to learn academic English as well as conversational English and this has recently prompted a revision in English proficiency testing in order to properly encompass measurement of these crucial skills. Chapin et al., (2003) have identified procedures for how academic English is learned within the content instruction in English. This has direct implications for the efficacy of measurement of content knowledge, particularly the more complex academic concepts. Finally, fatigue and frustration based on language load and unfamiliar testing exposure often adds necessary testing time, and will also likely determine that students may need more frequent breaks (Durán, 1997). Because of the psychosocial factors, accommodations such as these may need to be added as well to the overall testing package for identified individual students when they are tested in English.

The Influence of the Native Language

While students may be “limited” in English proficiency, if some English proficiency can be assumed (usually reading), bilingual children often have greater overall linguistic expertise and advantage compared with their monolingual peers who are having trouble reading tests (Kester and Peña, 2002). Unequal proficiency in native language abilities may impact testing and necessitate accommodations to compensate for this strain. Research in this area suggests that using native language proficiency represents a collaborative “meaning-making” process with which to facilitate the measurement of academic content (Knapp, 1995; Ruddell, 1995). Oral L1 proficiency is particularly important if they are being instructed in L1 or if they are not yet literate enough in their native language. Additionally,
oral L1 aid for non-glossed words may be useful for students with no to little written reading proficiency in their native language and if they have been learning content in English for a short period of time (Mann et al., 2006).

If and as students have some degree of written native language proficiency, written aid in L1 can supplement tests in English for those in English-speaking content classrooms. This because they can get clues about English item meaning from L1 cognates and other linguistic components (August et al., 2004). Several studies (for example, Collier and Thomas, 2002) have shown that as students are taught literacy skills in their home language this seems to have an effect on educational outcomes such as test scores, academic performance, etc. Even when ELL students are rather literate in English, Solano-Flores et al. (2003b) suggested that native language aids can continue to be helpful. Finally, several authors (including Rivera and Stansfield, 2001; Duncan et al., 2005) emphasized that native language and dual language forms, as well as use of language aids (particularly if they are written), take more time than if students are not proficient in English and using English forms. Knowledge about these components in ELLs’ home language reveals details about their level of proficiency which offers relevant information about these students’ specific test-taking needs (LaCelle-Peterson and Rivera, 1994; Durán, 1997; Butler et al., 2000).

The Influence of Culture

It has been argued that to better understand children’s learning, one must first appreciate the cultural environment in which such learning occurs (Boykin, 1995; Greeno, 1989; Rogoff, 1990). Research indicates that students whose home culture is similar to their “school culture” (typically based on mainstream, middle-class U.S. culture, according to Farr and Turnbull (1997)) tend to fare better educationally than those whose home and school cultures differ substantially (Boykin and Bailey, 2000; Grenfell et al., 1998). Clearly, language minority students come from many different cultural backgrounds and life experiences that affect their educational needs (e.g., Hakuta and Beatty, 2000; LaCelle-Peterson and Rivera, 1994). While many ELLs share the common goal of improving their English language proficiency, they differ in cultural background, home language, family and educational history and so on. ELLs are from various distinct communities, and research in this area asserts that each of these communities may interact with schooling very differently (Au, 1980; Au and Jordan, 1981; Butler and Stevens, 1997; Jacob and Jordan, 1987; LaCelle and Rivera, 1994; Ogbu, 1987; Phillips, 1983; Trueba, 1989). Therefore, several of these researchers suggest it is important that information be obtained about some of the
cultural experiences of the student, especially as they relate to their schooling experiences that are key to understanding classroom and large-scale testing environments.

Outcomes from this body of research find that careful consideration should be paid to particular components of the experiences of the students’ parents and the expectations of their home communities to the extent that these are different than the mainstream U.S. testing perceptions. Further, knowledge of student strengths and challenges given their home communities and cultural backgrounds seems to be salient. Malagon Sprehn (2004) reported that that, because a large percentage of ELLs were native born or in home communities somewhat isolated from mainstream U.S. culture, the culture of their home communities seems to have an impact in some of the same way that experiences in home countries might and they suggest that this as well as home country experiences should be considered. Butler and Stevens (1997) reported that length of time in the U.S. as well as amount and type of prior formal schooling is a necessary but not sufficient “proxy” for the degree of cultural connections and disconnections a student has with the U.S. schooling and testing culture. Additionally, Lara and August (1996) maintain that consistency in U.S. schooling is closely related and should be considered. The time and school consistency in U.S. are often used to suggest the degree of English language proficiency as well.

The Influence of Current Schooling Experiences

Much of the literature associated with the schooling of English learners discusses that teachers assess the content and ancillary needs of their students on an ongoing basis, and individualize and adjust instruction to address them (e.g., Farr and Trumbull, 1997; O’Connor and Chapin, 2006). Because language plays such a central role in the traditional delivery and assessment of content in the classroom, these teachers have had to learn how to separate out these limits while they attempt to teach or assess other targeted knowledge. As part of this endeavor, particularly for those students with noticeably limited language skills, the teachers are trained to be attentive to the ancillary strengths of the students and utilize these characteristics on an ongoing basis. Kopriva (2005a) and Koran et al. (2006) have noted that teachers of ELLs appear to be able to differentially report some language needs of students, psychosocial concerns, and some compensatory strengths that students have exhibited in the classroom, when they are directly asked about these characteristics. Further, there is some evidence to support that, when teacher reports of these factors are paired with accommodations which should ameliorate the needs and support the strengths, there seems to be some boost in test performance and possibly validity of the test scores for the students (Hipolito-Delgado
and Kopriva, 2006; Kopriva et al., 2007). For all these reasons, it appears that this compensatory experience in classrooms of ELL students should be important to consider when accommodations are being selected, and that teachers seem to be in a position to provide information about student strengths and challenges.

While teachers of ELL students appear to be skilled at identifying student needs and strengths, and can apparently accommodate them in the classroom, there is not much evidence that they can translate their classroom rationale into reliably recommending appropriate large-scale accommodations for their students (e.g., Douglas, 2004). This will be explained in detail later in this chapter, but it becomes one of primary reasons why some type of additional guidance about matching needs to be considered.

As the alignment between large-scale accommodation constraints and teacher skills are being sorted out, the issue of adequate practice needs to also be considered. Most authors (e.g., Durán, 1997; LaCelle-Peterson and Rivera, 1994; Hakuta and Beatty, 2000) agree that the best methods for testing ELLs in standardized testing environments are to mirror, as much as possible, instructional and evaluation techniques which occur naturally in the classroom. Only a subset of these is reasonably feasible, however, and decisions about which large-scale accommodations will be used for students will be, to some extent, externally imposed. Therefore, it remains important that every effort should be made to make sure there has been adequate use of whatever techniques are available prior to testing. The methods include, but are not limited to, compensatory techniques which might be practicable in large-scale settings, but which may or may not have been used by individual teachers. They also include the practice of relevant item types or other relevant test mechanics. As noted later in this chapter, the importance of practice provides a tension somewhat between when the matching recommendations should be collected and occasions for adequate trials. Because many of the salient testing capabilities of ELLs grow over time, often rather rapidly, recommendations made later in the year will ensure that more up-to-date needs are being reflected. However, this may not give teachers the time that is needed for students to learn about and/or practice less familiar options. Because of the heavier burden on teachers of ELL students and on the students themselves to teach and learn the language plus the academic content, time for sufficient practice is in short supply. Adequate practice presents another challenge but one that is an important aspect to the appropriate assignment of accommodations.

**Accommodations**

Based on research findings and practitioner experiences, Rivera and Collum (2006) have recently presented the subset of accommodations that
seem to be the most salient for English learners. Pennock-Roman and Rivera (2006) have also reported effect size benefits in their meta-analyses of relevant research to date. In addition, the promising accommodations outlined in several chapters in this book are consistent with what these authors are reporting. As noted throughout, the research findings are at various levels of consensus on the most beneficial accommodations for this population. However, this is largely because the field has continued to learn about how to evaluate effectiveness in general (Sireci et al., 2003), and particularly how to determine effectiveness for students with specific needs.

Kopriva (2005a) completed a task analysis of the active characteristics for fifteen promising accommodations. Active characteristics were those aspects of the accommodations which define which skills are needed to access them. For instance, oral administration in English requires that a student have sufficient auditory capability. Further, they should be able to understand the language of English at the level of language complexity in the written text which is being recited. The fifteen accommodations were selected based on literature reviews and pilot findings in this project and another related study (Kopriva and Mislevy, 2005). For the most part, they were consistent with what was cited by Rivera and Collum (2006), and by Abedi (2006b). Once the analyses of the accommodations were completed, the options were preloaded into the assignment system which was being built. The list of the accommodations in Table 10.1 has been adapted from this work. The system will be discussed later in this chapter, but the logic essentially begins with a screen which evaluates the student’s capability of taking the assessment under standard testing conditions. These conditions include a standard form as well as the testing procedures typically used in conducting large-scale testing. Therefore, accommodations are activated in the system when minimum thresholds for taking the assessment under standard conditions are not meant. Because the accommodations are discussed at length in other chapters, a discussion of the options, their research base, and their effectiveness will not be repeated here.

Kopriva (2005a) also completed a task analysis of active characteristics for two types of pretest supports, a Family Assessment Night and individually tailored classroom support. Subsequently, they were also preloaded into the assignment system. Pretest support is briefly discussed in Chapter 9 but essentially attends to student needs in the following ways. Family Assessment Night works with the family to understand, explain and clarify previous and current expectations and conventions of testing. Classroom support directs the teacher to a range of support which a particular student appears to need. Many of these considerations native language students take for granted, such as practice with types of items that ask questions and items
that ask for student’s original thinking. In all, pretest support was included because these avenues may be useful in helping to ameliorate certain cultural discontinuities of newer ELL students which, at the present time, are not adequately met even when options such as those identified in Table 10.1 are employed.

**Capacity**

Solano-Flores and Trumbull (Chapter 8 in this volume) argue that decisions related to the language of testing should attend to the institutional quality of L1 and English proficiency information, the institutional capacity to implement the test in L1 (text and/or oral), and their capacity to provide other accommodations. They contend that these variables should factor into the assignment decisions as well. They suggest it is not enough to produce a matching system without addressing the rigor of the information which goes into the assignment. Nor is it satisfactory to develop an assignment protocol without considering issues of agency capacity and district or state level oversight and auditing procedures which should be part of any decision system.

One important variable that has been identified but is not currently part of any matching system to date is opportunity to learn (e.g., LaCelle-Peterson and Rivera, 1994; Durán, 1997; Butler and Stevens, 1997; Hakuta

---

**TABLE 10.1** Promising accommodations for English language learners

- **Forms**
  - Access-based form in English
  - Native language or dual language forms as available
- **Tools**
  - Bilingual word list, general or test specific
  - Picture-word dictionary
  - Problem solving tools
- **Administration**
  - Oral English
  - Oral home language
  - Small group
  - Language liaison
  - Extra time
  - More frequent breaks
- **Response**
  - Written in native language or code-switching
  - Oral English
  - Oral in native language or code-switching
  - Demonstrated or modeled response
and Beatty, 2000). Other than opportunity to practice accommodations and the types of questions being used on U.S. mainstream assessments, understanding the students’ learning opportunities is clearly key to properly interpreting testing inferences. While this does not impact the procedures associated with assignment of accommodations per se, it is a crucial aspect of accurately interpreting results based on appropriate matching decisions. It is hoped that future work will determine how to integrate this information into an effective system for assessing English language learners.

Defining the Task

Studying test accommodations and understanding how they work can be quite complicated. Tindal and Ketterlin-Geller (2004) noted that in accommodations research, “few accommodations can be viewed as single, isolated variables,” and that “. . . accommodations are best thought of as a package in which no one accommodation is ever studied very well in isolation” (p. 7). On one hand, test changes that we think of as a single accommodation, such as reading a test aloud to a student, may have multiple facets that remove or contribute construct irrelevant variance for a particular student. For example, if a test is read aloud to a group of students, it is possible that students may benefit from the effect of having the test administrator pace them through the test, which is different from the benefit derived by having the reading demand of the test eliminated or reduced (Hollenbeck et al., 2000; Weston, 2003).

Furthermore, groups of such accommodations may also be thought of as packages (Hollenbeck et al., 2000), as the multiple facets of various accommodations interact to produce a net increase or decrease in construct irrelevant variance. Linguistic accommodations in particular (such as different forms of the test that incorporate linguistically-oriented changes in the presentation of items or the availability of language aids, for instance a bilingual word list or decreased language complexity) function most optimally as packages. For example, side-by-side (dual language) administrations require extra time because working back and forth between the two language versions of each item takes more time than completing a comparable test in one language (Choi and McCall, 2002). As Chapter 1 discusses, this seems to be particularly true for ELL students with little English proficiency who may be especially affected by their previous schooling environments, past experiences, and current home environments, which may be very different than those associated with mainstream U.S. students.

One may think of multiple possible combinations of accommodations that function together to meet the student’s need. One package of
accommodations may be considered the preferred package for a particular student although other packages compensate for the student’s need nearly as well. For example, an ELL may have a level of English proficiency such that a bilingual word list may not be needed if the student is to receive a form of the test where items are presented in plain or simplified English or in their native language. However, if that student were to receive a standard English version of the test form, he or she would need the bilingual word list to help him or her access the items. In selecting accommodations for ELLs most accommodation researchers are now recognizing that the package must be tailored to the needs of the individual student. Further, over-accommodating can be just as problematic for the student as not providing the needed accommodations at all. Administering a test with all possible accommodations may be overwhelming and possibly counter-productive (Kopriva, 2005a). Administering a test with improper “bells and whistles” (particularly unnecessary additions to the test items or forms) that are not needed can be distracting (for instance, see Sharrocks-Taylor and Hargreaves, 1999).

The need to make individual accommodations decisions is common to both ELLs and students with disabilities. However, less research has been published regarding the processes used to make accommodations decisions for ELLs. In the next sections we discuss policy-based methods and research approaches. Policy methods generally reflect current practice in selecting appropriate test accommodations for individual students in both groups, although the mechanisms associated with these assignments are somewhat different. Research-based methods are emerging and approach the task from the perspective of how the literature might structure the decision process and narrow down choices for students with particular needs. Where research is lacking with respect to ELLs, we have supplemented the discussion with relevant research concerning similar issues with students with disabilities.

Policy-based Approaches

Current guidelines for selecting accommodations for students with disabilities primarily stems from authorizations of federal legislation (Individuals with Disabilities Education Act, IDEA). The practice for assigning large-scale accommodations for students with disabilities typically focuses on the role of the Individual Education Plan (IEP). In addition to developing and evaluating each student’s learning goals and instructional plans, the IEP addresses the proper test accommodations appropriate for each student at both the classroom and standardized testing levels. Regulations or instructions for assigning accommodations to individual ELLs, on the other hand, are generally policy-based, most often at the state
level. Current practices typically used to assign large-scale test accommodations to individual ELLs reflect that decisions are generally made by a single person (commonly the student’s teacher or the school ELL specialist). Reviews of ELL test accommodations policies and practices both across the U.S. and within certain states and districts confirmed that ELL teachers still are the primary decision makers when it comes to deciding ELL accommodations (Rivera and Stansfield, 2000; Kopriva et al., 2007a). Some educational agencies are beginning to use teams of people which are similar to IEP teams for students with disabilities (for instance, involving parents, teachers, specialists, and administrators) to make the decisions (Kopriva et al., 2006a). However these teams differ in major ways from IEP teams, particularly in the scope of their charge. For English learners the scope is much more constrained and standardized test focused than that defined under federal statute for students with disabilities.

Some disability researchers emphasize the importance of the IEP as a decision-making process as well as a document (Shriner and DeStefano, 2003). However, the protocol for IEP as process is often not well defined or straightforward for a team of people to implement, and similar problems may be found in the guidelines for teams making accommodations decisions for ELLs. In both situations guidelines tend to offer broad parameters rather than specific guidance for those who must make accommodations decisions. Both individual teachers and teams making accommodations decisions attempt to work within the policies given to them by the federal, state or local educational agency, but these policies generally do not contain specific recommendations for how to address the needs of specific students. Thurlow and others (e.g., 2002; 2004) routinely collect information documenting a great deal of variation and inconsistency in allowed accommodations across states for students with disabilities and ELLs. A more in-depth look at selected agencies found substantial variation in the explicitness and formality of policies for both populations among partner states and districts (Kopriva and Hedgspeth, 2005).

It appears that, presently, both ELL individuals or teams and IEP teams tend to rely heavily on the implicit and ill-defined sense of knowledge about the student obtained from the child’s primary teacher(s) and/or well-meaning adults. Research on accommodations decisions for students with disabilities shows that, with little additional guidance, they tend to be inconsistent decision makers when it comes to assigning appropriate accommodations for individual students. Ysseldyke et al. (1994) cite vague accommodations guidelines and altruistic motivations (such as assigning an accommodation to lessen emotional distress to the student rather than because the student needs the accommodation to receive a valid score) on the part of the local decision making team as contributing to the
inconsistencies found in assigning accommodations for students with disabilities. Fuchs et al. (2000a) found that teacher judgments about whether to assign accommodations for students with disabilities were associated with some demographic and performance variables where there were no such associations with the accommodations decisions made by an objective accommodations decision making system.

Douglas (2004) found that ELL teachers had difficulty articulating the specific decision making process that they use in assigning accommodations. Koran and Kopriva (2006) found that teacher recommendations, unfortunately, were not statistically different from random assignment of accommodations to ELL students. This was true when teachers were asked to provide guidance without additional training from that provided by their states or districts. However, it was true even when the teachers took the leadership role themselves in a standard data collection process designed especially for this purpose. Similarly, Fuchs et al. (2000a) reported that teacher judgments did not correspond very well to learning disabled students’ demonstrated differential boost using each of three accommodations on different alternated forms of a reading assessment. In another study, teachers predicted with no more than a chance level which special education students benefited from read aloud accommodation on a mathematics test (Helwig and Tindal, 2003). Further, Weston (2003) found that teachers did not do any better than chance at predicting which learning disabled students would gain the most from a read aloud accommodation on a mathematics achievement test. Of interest, however, Koran and Kopriva found that teachers could easily and clearly specify the different needs of students. What they appeared to struggle with was the assignment of accommodations to students based on these particular needs. This was echoed in a recent Plake and Impara report (2006) where educator experts of students with disabilities could understand differential needs of students across a broad spectrum of disabilities but had trouble linking these needs to particular accommodations. It seems that, while teachers may be proficient in identifying needs of individual students, at this time they have problems assigning differential testing accommodations based on these needs.

The focus group results reported by Douglas were revealing of the strategies ELL teachers use when given imprecise guidelines for making accommodations decisions for individual ELLs. In particular, teachers seem to work within the guidelines given to them, such as policies restricting the range of allowable accommodations, but, in general, they assumed that more accommodations were better. This is consistent with researchers’ findings that teachers tend to over-assign accommodations for students with disabilities (Fuchs et al., 2000a, 200b) and that accommodations
decisions tend to be an “all or nothing” phenomenon (DeStefano et al., 2001). The common attitudes seemed to be “When in doubt, give the accommodation” and “If you are going to accommodate, give them everything you can.” In large-scale assessments, the ELL focus groups reported on by Douglas that teachers felt it was better to have a higher rate of false positives (giving accommodations to students who don’t truly need them) so as to reduce the chance of false negatives (failing to assign an accommodation to a student who really needs it).

In general, present policies did not provide the level of explicit instruction necessary to allow teachers to make reliable decisions. Among the policies there appear to be few student-specific guidelines for assigning specific types of accommodations. As such, it does not appear that the policy-based guidance alone would provide appropriate guidance for making specific accommodations decisions for individual students who need them. It is hoped that, with more specific training and additional information to make good decisions, the rates of both false positives and false negatives could be decreased. An early computer-based system, the Minnesota Online Decision-Making Tool, used a policy-based approach for making accommodations decisions with students with disabilities and ELLs (Anderson and Spicuzza, n.d; Swierzbin et al., 1999). This model provided little support, however, for answering the questions of whether the student needed accommodations and which accommodations would meet that student’s needs. It did attempt to systematize a process, though, that to date had been given only scant attention. More recently, Elliott and Roach (2006) identified salient areas of need for students with disabilities and provided recommendations for focusing IEP teams on systematically addressing their diverse needs. Butler and Stevens (1997) developed theory-driven guidance for educators about how they might better approach the assignment of accommodations for ELLs based on linguistic and other student-specific variables. These will be discussed in some more detail in the next section.

Some authors attribute poor educator and possibly parent judgment with regard to assigning accommodations to inadequate training in the areas of measurement, standardized assessment, and accommodations (Hollenbeck et al., 1998). Indeed, DeStefano et al. (2001) found that their intensive and comprehensive teacher training program improved the quality and extent of accommodation decisions for students with disabilities. However, their results were confounded with two substantial changes in the state’s testing program that may have also contributed to dramatic changes in accommodations decisions (Shriner and DeStefano, 2003). While it is important that teachers receive proper training in the areas of measurement, standardized assessment, and accommodations,
teacher training alone may not entirely solve the problem. Fuchs et al., (2000a) found that even when teachers read background information about test accommodations, discussed and reviewed this information with a research assistant, and had the opportunity to ask questions about the material, they still did not do a good job at assigning accommodations. Their work suggests that increased teacher education with regard to accommodations, by itself, may not be sufficient to improve accommodations decision-making.

In other areas of standardized testing, such as grading constructed responses and setting cut scores, human judgment is commonly and successfully used within highly structured, defined, and systematic approaches that involve routine oversight and auditing mechanisms. That is, detailed guidelines are given, judges are trained, and checks are put in place to assure the quality and consistency of the decisions. To not take a systematic approach to the assignment of accommodations seems to represent a breakdown of the systematized chain of evidence that leads to valid inferences (Mislevy, 1994). Such an approach to accommodations decision-making may help ensure not only the quality of the decision but also the consistency of the decisions across students, both of which are needed in order to make meaningful comparisons across schools, districts, and states. There has been a call for more systematized accommodation decision support systems for students with disabilities (for instance, Solomon et al., 2001). There are many critical variables to consider in each ELL student’s background and many potential options for meeting the student’s needs. A structured systematic approach is probably necessary in providing support for the decision makers for ELLs as well who must take so much information into account in making the decisions. More attention needs to be paid to making sure the essential decision-making elements are being included and fairly utilized for this population.

Research-based Approaches

There are a few possible research-based approaches for assigning accommodations to individual students based on their unique needs. Such systems are generally not intended to take the place of the persons who currently make accommodations decisions for ELLs and students with disabilities, but, rather, they are intended to provide a solid recommendation of accommodations which can be used as a basis for the person or team making accommodations decisions. In describing and discussing these approaches, we distinguish between those which are inductive and deductive in nature. Depending on purpose of the assignments, both foundations hold the promise of guiding high quality, consistent accommodations decisions. Each approach discussed below combines a data collection phase with an
accommodations assignment phase. Kopriva and others (Kopriva 2005a; Kopriva et al., 2006a, 2006b) have illustrated that there appear to be four steps to developing a proper method for assigning accommodations, not just two. That is, key variables need to be identified, data need to be systematically collected, data from multiple sources needs to be thoughtfully combined in order to inform the assignment process, and, finally, a logic method of assignment that fairly and sensitively recognizes and matches salient student and accommodation characteristics needs to be built and implemented.

In most approaches (and especially in research-based approaches) where information must be combined to make an accommodations recommendation, building the system on a computer-based platform has been found to be quite valuable. Computerized decision support technology has been used successfully for many years in other fields, such as business, medicine, and law enforcement. For example, systems that produce differential diagnoses for medical personnel or track evidence patterns for law enforcement agencies effectively narrow down the routine work staff do. However, teachers and other team members are not asked to do this large-scale assessment assignment task on a regular basis so it is not surprising that their skills have not developed to produce reproducible decisions over like students or over teachers or districts. For this reason, systems that build research-based algorithms that effectively narrow choices would appear to be especially useful here. Perhaps resource limits have slowed down the development of such systems. It is only very recently that researchers have begun investigating ways to address this problem by utilizing findings and related literature. Most of these are computer-based because of the complexity of the work and the diversity of the populations they are focusing upon, and most of them are still in various stages of development.

**INDUCTIVE METHODS**

In making accommodations decisions for ELLs and students with disabilities, it is possible to take an approach that is inductive, that is, to take specific incidences and make some sort of generalization. One reasonable inductive approach for assigning accommodations to individual students is the direct trial-and-error approach. In this approach different accommodations are tried with a student, one at a time or one package at a time, to see which ones help the student perform better in testing. This approach has been applied both informally and empirically in making accommodations decisions for individual students. In the focus groups with ELL educators, many of them stated that they used an informal trial and error approach to calibrate the appropriate accommodations for a specific
student by trying various accommodations in the classroom and tracking which accommodations worked best with that student (Douglas, 2004). Teachers felt they could use their observations to inform their accommodations decision making for large-scale testing even in the face of a lack of clear, consistent guidelines for assigning specific accommodations to individual students.

Special education researchers have proposed and applied a formalized empirical version of this approach with students with disabilities. In this method the student is administered a short mini-test without accommodations and other parallel forms of the mini-test each administered with a different accommodation (Fuchs et al., 2000a, 2000b). The boost in performance on each accommodated mini-test over the performance on the unaccommodated mini-test is calculated and compared against normative data for regular education students who did not need accommodations. If the student’s boost exceeds a certain cutoff in the normative data, then the student is deemed to need the accommodation. Fuchs et al. found they could successfully predict for which students test scores on a large standardized test would improve substantially when administered with the specific accommodation or accommodations supported by the mini-test data. Recently, these researchers have published their operational approach which uses this methodology. It is known as DATA, the Dynamic Assessment of Test Accommodations (Fuchs et al., 2005) and is discussed briefly in the next section.

The formalized trial and error approach seems to effectively identify students who stand to benefit from test accommodations in research situations. However, limited work has been done to identify how well this works for large-scale testing purposes on a routine basis. Further, in contrast to a formalized trial and error approach, it has not been documented that informal teacher trial and error is effective in the same way. Other limitations need to be considered as well. First, this method is clearly time-consuming. For high school classes or other situations where the teachers or specialists work with many students this needs to be taken into account. Second, while Fuchs et al. (2000a, 2000b) found that their systematic formal approach was effective, small ns suggest caution. Third, unfortunately, the mini-testing can confound itself with some accommodation needs, such as the need for more frequent breaks. If the mini-tests are of shorter length, for instance, it may set up a different situation than what occurs during large-scale testing. Rather, especially in selecting large-scale test accommodations, conditions for the assignments need to be as close to those for which the accommodations are being considered. Several situations, including test length and duration over days are usually constrained during this type of testing. Fourth, as with any approach,
students need enough time to practice any accommodation. This is true whether the students are preparing for a test or for the trial and error assignment process. Fifth, the pure trial and error approach typically does not account for interaction effects when multiple accommodations are needed, unless packages are tried out as well as individual accommodations.

Finally, constant trial and error may be a little like inventing the wheel over and over again. As the field determines that specific steps along the decision process are well-considered and sufficiently differentiated at the present time, it makes sense that systems can capitalize on this work and use the trial and error approach where certain distinctions or tasks are still in doubt. Tindal and others (see below under Deductive Methods) are investigating the use of a hybrid form which combines trial and error and other methods. These and others will be considered next.

DEDUCTIVE METHODS

Deductive approaches based on theory, or generalized notions which have been found to hold promise, also have been proposed. Theory-driven deductive approaches focus on identifying critical factors that are related to the individual needs of students and the active characteristics of test accommodations. Abedi (2006) emphasized that only a small subset of accommodations is most appropriate for ELLs. While he didn’t specify which factors in these accommodations were relevant or for whom, other researchers have addressed this issue (Kopriva et al., 2007a). In reviewing the ELL literature for key student indicators, the researchers who developed the Selection Taxonomy for English Language Learner Accommodations (STELLA) system discussed below compiled a list of 139 different variables important for learning and evaluation in general. For the particular purpose of large-scale testing accommodations, however, their analysis found that only a small number of these seemed to be salient for this type of situation, and these were discussed in the first section of this chapter. The extensive review was useful, however, because it helped to identify some nuances not generally considered in large-scale assignment to date, and provided information which could be used in the development of the logic or research-driven systems. For example, August (August et al., 2004) reported that access to bilingual glossaries have been found to be effective for understanding test requirements in English when ELL students have at least some literacy in their home language. This is because they could transfer cognate cues across languages. The students did not need to be fully literate in their home language to begin using this skill; however, the more literate they were in their native language the more they could avail themselves of this benefit. Her work provided developers of the system guidance about
Direct Assessment of Ancillary Skills

This approach attempts to directly assess selected ancillary skills of individual students that are unrelated to what the test is intending to measure but may block the student’s access to the test if he or she does not have a sufficient command of this skill. Skills related to what the test is intended to measure may also be used provided they are predictive of what accommodations would help the student overcome access barriers in taking the test. Researchers interested in improving accommodations decisions for students with disabilities have attempted to use measures of ancillary skills, in particular reading, to predict whether those students would benefit from a specific accommodation (Helwig and Tindal, 2003). Tindal and others have also included measures of these skills in more comprehensive systems, such as their recent method, the Accommodation Station, which is identifying and evaluating the impact of selected student data related to accommodations and what that might mean for accommodations assignment (Ketterlin-Geller, 2003; Tindal, 2006). Directly measuring the level of access skills within the assignment system may be helpful in selecting some individual accommodations for ELLs, particularly because some of the skills clearly impact student access to large-scale tests in content areas as they are currently conceived.

Like the trial-and-error approach, the direct assessment of ancillary skills involves testing the student directly. Thus, the measurement of the particular ancillary skills must also be accurate to get valid results, and this, in itself, is a large undertaking. For instance, much literature and many testing products illustrate how complex and resource consuming it is to effectively develop measurements of reading proficiency. For those who assess ancillary skills, specific domain and theory-driven variables need to be identified and measured, and logic about how to combine these to obtain proficiency levels would need to occur. Initial attempts at this approach that used an assignment system to make accommodations decisions for students with disabilities were largely unsuccessful (Ketterlin-Geller, 2003). A current project is attempting to build on this work by beginning to identify decision-making logic partially informed by the ancillary skills related to reading that the latest version of the Accommodation Station provides (Siskind, 2007; Tindal and Ketterlin-Geller, 2004). This ongoing work will be summarized in the next section.

While Tindal and Ketterlin-Geller’s method utilizes a direct measure of some skills, they also recognize and have attempted to incorporate other “informant data” in this system. Helwig and Tindal (2003) used reading and
math pretest scores to try to build student profiles that might better predict whether students with disabilities would benefit from having the read aloud accommodation. However, the results were often in conflict and their system did not perform much better than teacher judgment. They suggested that their analysis failed to take into account a potentially critical variable, namely the linguistic complexity of the items in their test of ancillary skills. This complexity is widely recognized in the ELL literature (for instance, see Kopriva, 2000; Abedi and Lord, 2001; Abedi et al., 2000). Student performance with accommodations related to linguistics can be very complicated behavior to model and predict because of possible interactions with the characteristics of the test items. As such, data on the language complexity of the large-scale assessment the student will be using seems to be necessary in determining the effectiveness of some accommodations, or test forms.

Tindal and Glasgow’s (2005) current system uses some of the Fuchs and Fuchs trial and error methodology as well in that they test students with and without a read aloud accommodation. However, recent data related to incorporating this methodology with his other work are inconclusive and disappointing. Finally, Helwig and Tindal (2003) speculated that their system was too simplistic and failed to benefit from the teachers’ intimate knowledge of their own students. They proposed that math and reading screening test scores utilized in tandem with teacher knowledge of individual students might produce better accommodations decisions. Lately, researchers of the Accommodation Station have attempted to address this limitation by including questionnaire data from teachers and the students (Tindal, 2006). This method is also employed by other researchers and will be discussed next. To date, it is unclear how data from the direct assessment of ancillary skills, the informant data, and the trial and error data will be combined in the Accommodation Station to provide guidance to IEP teams. Structuring a system that incorporates different methods, however, is interesting, and it is anticipated that future work will focus on these next steps.

Informant Approach The informant approach relies directly on the insights of those who know the student personally and/or other extant skill information that would be available in the student’s record. In this sense, it has benefits over the other approaches discussed so far because is not limited to skills or needs that can be tested or measured directly. In fact, it may not be necessary for the student to be directly involved in the process at all. The data are indirectly collected, however, which introduces another source of error and caution. Using this approach, information about the student may be collected from one source only (for instance, the teacher)
or gathered from multiple sources, such as the student, the student’s teacher, and the student’s parent or guardian. As defined here, specific information is gathered using standard instruments specifically designed for the purpose of assigning large-scale test accommodations. Further, the intent of this method is to focus instrument questions on collecting clearly delineated information about critical variables that the developers have identified as being salient for assigning accommodations to individual students. Therefore, the informant approach is distinct from the policy-based teacher or team approaches in that method does not rely solely on instinct or sources of information about students and accommodations that are vaguely and implicitly utilized and combined in a non-standard way. On the other hand, depending on the informed source(s) and whether it is collecting data obtained from other test(s) or relying on the judgment of adult(s) who know the student well, the informant approach may rely on information, systematically collected, that is skewed (human judgment of parents or teachers) or outdated (for instance, information in the student record). The quality of the student data which is collected relies on the quality of the information gained from the informants. Further, the quality of the decision is to some extent dependent on the quality of the informant’s knowledge and the acumen of the researchers who interpret theory and build the algorithms. Like any approach, the quality also is bound to be variable for different students.

No literature has been found which systematically links student records information and accommodations assignments for large-scale assessments for students with disabilities or ELLs. Likewise, no research has been reported which combines uniformly collected information about student needs collected from only one “human” source (for instance teachers or parents) with this one human source making assignments based on this information. As noted above, Koran and Kopriva (2006) reported results when ELL teachers, by themselves, were asked to assign accommodations based on a standard collection of data from multiple sources, including themselves. Even when standard instruments were used, teacher assignments were judged to be no better than a random assignment of accommodations for individual students.

STELLA, the Selection Taxonomy for English Language Learner Accommodations, is a newly developed informant system designed to assign individual accommodations for K-12 English language learners (Kopriva et al., 2006b, 2007a). Currently, it utilizes findings from the multiple sources (parents, teachers and records) and systematic student data collection procedures. It identifies critical variables, collects data, combines the data with standard information regarding how accommodations perform, and then uses a standard series of computerized algorithms based in theory
and formative empirical input. The system is built to utilize the latest information about students that appear to be the most relevant for making accommodations decisions about this population. Further, it is designed to be customized to accommodate the policies of different states or districts while also retaining the ability to provide a systematic series of recommendations over agencies which have been found to be meaningful based on best practice and empirical findings. Two initial verification studies found that this system seems to be producing appropriate decisions for individual students. Future work will continue to examine and refine the extensive series of algorithms, update the preloaded aspects of STELLA and possibly add a direct component to the system. This system will be discussed in more detail in the next section.

It is argued that, when direct testing is utilized or multiple sources are consulted, the research-based approaches may mark an improvement over pure policy-based approaches not only by formalizing the content and contribution from each source, but also by separating the role of student advocate from that of primary decision maker. Douglas concluded that ELL teachers felt, in the present policy-based climate, that they were asked to simultaneously take on competing roles as both expert decision maker and student advocate (2004). This may present a dilemma for teachers who are asked to make decisions with minimal guidelines and support while maintaining the best interests of the child. Two sets of work (Koran and Kopriva, 2006; Plake and Impara, 2006) have found that educators struggled to identify proper accommodations even though they could easily identify needs of individual students. Perhaps educators may be better at some roles than others in this accommodation assignment process. Each of the systems summarized below utilizes research and best practice as a foundation of helping educators narrow down choices and provide guidance. As such, they are designed to improve the selection approaches necessary for making informed decisions about accommodations in research and in practice.

Operational and Prototypical Systems
Within the last 10 years there have been some attempts to create tools for improved accommodation decision-making that are individual student based. Research and other information associated with these systems have been mentioned above. This section will explain the methods in somewhat more detail. Three guidance models, one for both students with disabilities and ELLs, one for students with disabilities, and one for ELL students, will be discussed first. Next, three systematized research-based systems, two for students with disabilities and one for English language learners, will be
summarized. The goal of all of these methodologies is to show improvement over current practice, which is used as the comparison.

Guidance Models

Following are three examples of guidance models. Two of the three are hybrids of policy-support and informant research based approaches, while the first system which will be discussed used the policy-support approach only.

MINNESOTA ONLINE DECISION-MAKING TOOL

An early system, computer-based, used a policy-based approach for making accommodations decisions with students with disabilities and ELLs (Anderson and Spicuzza, n.d; Liu et al., 1999; Swierzbin et al., 1999). This tool attempted to improve accommodations decision-making by guiding parents, administrators and others through a decision tree based on one education agency’s test accommodation policies, and was created on the basis of existing state guidelines for the inclusion of these students. However, it did not collect data to make the decisions at the various points. Rather, the users were responsible for collecting the relevant information and making a correct decision at each point. They were expected to address the requirements at each step by relying on their own knowledge of the student and knowledge of accommodations. Feedback from ELL/bilingual educators, coordinators, and administrators reviewing the tool for use with ELLs indicate that they struggle with obtaining the type of knowledge needed to answer the questions posed by the tool at some of the decision points (Liu et al., 1999).

Specifically, the system directed the user through the decision tree to determine whether policy dictates that the student should be exempted, (eligible to) participate with accommodations, or participate without accommodations. The tool then pointed each user to specific policy information, such as what accommodations are available and the tests for which they are appropriate. However, it left the decision about assignment of specific accommodation(s) for specific students to the user. Finally, the tool also facilitated the implementation of accommodations by providing information about whether materials need to be special-ordered in order to provide the accommodation to the student for the state assessment. The developers emphasized that one of the advantages of the online system was that updates to policies on the online system would be easier for educators to manage than updates to printed policies. This assured that teachers are using the most current version of state accommodation policies to make accommodations decisions.
This model provided little support for answering the questions of whether the student needed accommodations and which accommodations would meet that student’s needs. It did attempt to systematize a process, though, that to date had been given only scant attention. A pilot of this system found mixed initial acceptance among IEP team members who used the system in making accommodations decisions for students with disabilities (Swierzbin et al., 1999). However, a pilot with ELL/bilingual educators, coordinators, and administrators found a more positive response for the use of the system with ELLs (Liu et al., 1999). A system such as this may improve the implementation of educational policy, but appears to offer little specific guidance to the user who must decide which accommodations are most appropriate for an individual student.

**FIVE-STEP SYSTEMATIC DECISION-MAKING AND DOCUMENTATION PROCESS**

Stephen Elliott and others (e.g., Elliott et al., 1999; Elliott and Roach, 2006) have continued to provide guidance to IEP teams about how to wisely assign large-scale accommodations for students with disabilities. In the *Assessment Accommodation Guide* (Elliot et al., 1999), *Assessing One & All . . .* (Elliot et al., 2001), and other work, these researchers have identified key information and student needs that teams should know, critical access skills that are particularly salient for this population, and process factors that influence accommodation decision-making. The *Assessment Accommodations Checklist* and associated guidance directs IEP team members through the accommodation selection, implementation planning and documentation processes. The authors encourage members to link any of the sixteen key access skills they have identified as being problematic for an individual student to one or more accommodations that specifically minimize interference between conditions and measurement of target skills. These skills represent elements of typical large-scale standardized testing conditions which could pose a problem for students with disabilities. Some of these are also problematic for English learners as well. The sixteen access skills are:

- Attending
- Listening
- Reading
- Remembering
- Writing
- Following directions
- Working alone
- Sitting quietly
- Turning pages
- Locating test items
- Locating answer spaces
- Erasing completely
- Seeing
- Processing information in a timely manner
- Working for a sustained period of time
- Spelling

**Butler and Stevens Guidance**

In 1997 Butler and Stevens completed a report for CRESST (Center for Research on Evaluation, Standards, and Student Testing). This report provided an overview of ELL student characteristics in general and guidance about which factors are particularly important in selecting appropriate accommodations for individual students. They also provided some direction about how to use this information. While they didn’t address this report to agencies or teachers *per se*, the guidance would be helpful to those making accommodation decisions for students.

The three student factors these authors felt were critical were English language proficiency, prior formal schooling, and length of time in U.S. As they detailed, it is important that appropriate tests of English language proficiency include the measurement of complex academic language, conceptual and discourse skills, not just factual academic vocabulary and knowledge levels. The recent generation of these tests is improving the ability to measure these advanced and more complex dynamics of English proficiency as well as more basic levels of English literacy and fluency. They appear to be closer to what Butler and Stevens were recommending. They felt that the second variable, prior formal schooling, was critical because students with little or no formal schooling tended to have less experience with schooling and testing conventions, especially those used in the U.S. Further, they recognized that not all formal schooling was similar; in some countries the knowledge, skills and methods were quite distinct from U.S. schools. Therefore, it was important to identify both amount of formal schooling and the kind of formal schooling the students had experienced. Those with significant formal schooling in their home countries are more likely to be literate in their L1 as well, and this suggests that students will move more quickly from accommodations to no accommodations as they become more proficient in English. It also widens the types of accommodations which may be appropriate. The third variable, length of time in the United States, was chosen because it suggests exposure to
U.S. schooling and testing conventions and the U.S. mainstream culture in general. While they didn’t include any other variables *per se*, they acknowledged that isolation of home communities from mainstream expectations is an important covariate that should be considered.

Figure 10.5, from their report, suggests how an agency might use this information. Unfortunately, however, it only provides guidance for three choices: no accommodations needed, accommodations needed, or exemption preferred. The report also includes a background questionnaire for middle school students to complete which can elicit information about the variables they have identified. The ten questions include where and when the student was born, how long they have lived in the U.S., if they went to school in another country and, if so, where and for how long. It asks what their first language is, what is spoken in the home and whom do they speak it with, if they ever had formal study in a language other than English and if so which language, and for how long and where did they study (home, school, language school or other). Finally, the questionnaire asks if they have ever learned content in another language (and if so, what), how many years have they attended school in the U.S, if they have learned/are learning content in the U.S. and, if so, which subjects are they studying.

While this report does not guide decision-makers to specific accommodations for individual students it does provide a view into the types of data which would be needed to make those decisions. As such it represents an early attempt at categorization that the *STELLA* system (below) has built upon.

*Research-based Standardized Data Collection and Assignment Systems*

Three systems will be outlined here. In general, the development of the deductive research-based systems has lagged behind that of inductive methods, in part because of the technological requirements and development costs associated with them.

**DATA**

*DATA*, the Dynamic Assessment of Test Accommodations (Fuchs *et al.*, 2005), is a systematic method for assigning test accommodations to students with disabilities in grades 2–7. *DATA* uses a pure empirical trial-and-error approach as described earlier in Fuchs *et al.* (2000a, 2000b). In this method the student is administered a short mini-test, without accommodations and other parallel forms of the mini-test each administered with a different accommodation. The boost in performance on each accommodated mini-test over the performance on the unaccommodated mini-test is calculated and compared against normative data for regular
education students who did not need accommodations. If the student’s boost exceeds a certain cutoff in the normative data, then the student is deemed to need the accommodation. DATA assesses student need for accommodations in tests of reading and mathematics. Research has indicated that the system is successful in predicting which students will...
benefit from specific accommodations on full length standardized tests of math and reading (Fuchs et al., 2000a, 2000b). Unlike the next two systems we will discuss, DATA is now a fully operational product published by PsychCorp (Fuchs et al., 2005).

**ACCOMMODATION STATION**

The Accommodation Station is a computer-based system currently being developed by researchers at the University of Oregon which is primarily a tool for students with disabilities (Tindal and others, e.g., Tindal, 2006; Tindal and Fuchs, 1999; Tindal and Ketterlin-Geller, 2004). Rather than being available now or soon, the Accommodation Station is an approach around which a flurry of related ongoing research studies were and are being conducted. The central large-picture concept of the system is focused on examining aspects of assignment for future computer-based systems that will not only provide information about accommodation assignment but also deliver the accommodations on line, and test students in the content areas of interest (Ketterlin-Geller, 2003). Early studies in this line of research focused on directly assessing student access skills (such as reading and math skills) and using these scores to predict whether students would benefit from certain test accommodations (Helwig et al., 1999; Helwig and Tindal, 2003). More recent studies have added student and teacher questionnaires (Alonzo et al., 2004; Tindal, 2006) and have continued to investigate student characteristics that predict whether a student will benefit from a particular test accommodation (Tindal and Glasgow, 2005).

Virtually all of the research to date has been done on students with disabilities although there has been movement lately associated with adding Spanish proficiency as an ancillary skill to be measured at some point. A trial and error approach similar to Fuchs et al.’s DATA system has also been incorporated into a recent prototype of the Accommodation Station. Student performance on a subset of math items in read-aloud and self-read conditions are compared. Unlike, the Fuchs and Fuchs method, boost is assessed under these two conditions for a student rather than compared against normative data as in the DATA system. Early research flagged a number of problems with using a purely direct approach though (Helwig et al., 1999; Helwig and Tindal, 2003). The Accommodation Station has now added an informant aspect although initial forms and data being collected in this way have been problematic (Tindal 2006). Its expert-driven decision-making algorithms, focused on students with disabilities only, are in development (Siskind et al., 2004; Tindal, 2006).

Variables considered in recommending accommodations for individual students include the results from the direct method of assessing ancillary information, such as the ability to use a mouse to enter responses on a
computerized test, reading comprehension, and level of math skills when these are considered to be ancillary in nature. It also includes input from the student on his or her preferences for accommodations and teacher’s input on factors such as classroom use of accommodations. True to its direct approach roots, data collection in the Accommodation Station is centered on the involvement of the student, both in measuring the student’s ancillary performance and in using the student as a knowledgeable informant of the types of accommodations that would be helpful. Recent findings suggest, however, that the student is remarkably inconsistent with what they perceive to be useful accommodations for themselves (Tindal, 2006). Tindal also reports that some teacher information appears to be inconsistent as well.

Beginning in 2005 an expert panel of participating special education state specialists was convened. A systematic consensus-building method for assigning accommodations to different kinds of students using a case-study approach was employed in late 2005 and 2006. A final report (Plake and Impara, 2006) explained the progress in utilizing such a method and provided guidance about how to continue toward the assignment goal. It is anticipated that future funding will augment the work so far and subsequently complete this part of the Accommodation Station’s development.

STELLA

While some attempts had been made to look at how to match students with disabilities with appropriate test accommodations, very little prior research had been conducted to look into the testing needs of individual English language learners and establish systems for matching ELLs with appropriate test accommodations. The Selection Taxonomy for English Language Learner Accommodations, STELLA, has been developed over the last three years and is in a draft final form (Kopriva, 2002, 2005a, 2005b; Kopriva and Hedgspeth, 2005; Kopriva and Mislevy, 2005; Kopriva et al., 2006a, 2006b). STELLA is intended to be used with K-12 students to assign accommodations to the range of ELL students for use on large-scale academic assessments. It is a computerized system, research-based, and empirically developed and verified. It has been built, most recently, from research conducted as part of two large, multi-year projects. As more evaluation of the system is completed, it is expected that states or other interested users will soon be able to utilize the “beta” version of the system.

STELLA utilizes the informant approach to making test accommodations decisions for individual ELLs. It has been built on a flexible platform to adjust to the needs and obligations of different users, and has been designed to be compatible with and flexible enough to accommodate information from the direct approach if future research suggests this is a
viable addition to the system. At the present time this system uses three forms to collect the most salient data about the needs and strengths of each student from records, parents, and teachers. The records form is based on information that is in the student’s file at the school, such as English language and native language proficiency test scores and other information about the student’s schooling in the U.S. The parent/guardian form is based on an interview with the student’s parent or guardian regarding the student’s proficiency in the home language, prior schooling opportunities, and prior experiences with classroom and standardized testing. The teacher form is based on observations the teacher has made about the student’s language proficiencies on an ongoing basis in the classroom as well as their apparent preferences and accommodation experiences in the classroom. In the STELLA system teachers are not asked to recommend large-scale accommodations at the present time as the authors suggest that research says this currently appears to be difficult for these educators to do.

These data are compiled and then subjected to an extensive set of conversion, consolidation and decision-making algorithms. Output includes an individualized student profile and recommended accommodations for each student. State or district-allowed accommodations that are recommended are highlighted but additional STELLA recommended accommodations are also shown. This allows users to utilize the system under current legislation and regulations and provides some guidance about how these policies may be improved upon for individual students. As noted above, the system is developed so that highlighted accommodations change by educational agency, and decision-making algorithms can be added to allow for additional accommodations. At this time specific accommodations that research and practice have suggested are promising for this population have been preloaded into the system. Likewise, since part of the system is designed to convert different proficiency test scores to common scales, certain tests are preloaded and others can be added as needed.

Another project offered some early insights into salient student and accommodation characteristics and how a matching model using an informant approach might work as an operational system. In this project some relatively simple matching systems were developed, each designed exclusively for the specific accommodations study at hand (Kopriva, 2005a; Kopriva et al., 2007d; Emick and Kopriva, 2006). Since the focus of this project was elsewhere, a later project built on this work and focused exclusively on developing a rigorous assignment system for English language learners. Once funding for this second project was retained, the forms were conceptualized, designed and developed after extensive developmental work that reviewed current large-scale and instructional accommodation
literature, conducted focus groups, obtained recommendations from targeted practitioners, and subsequently narrowed down the most salient student variables. Key variables were selected and interviews with teachers and parents and feedback from partner educational agency staff refined the questions on the forms. Concurrently, project staff, with guidance from educational agency participants, built on knowledge gleaned from the earlier project, identified the most promising accommodations, and completed an evaluation of the active characteristics of the final set of accommodations. Algorithms were developed by project staff and consultants and reviewed and edited by an expert panel (see Kopriva, 2005a; Kopriva et al., 2006, 2007a, for more of an explanation of the formative studies and related development work).

In its present form STELLA does not involve the student directly. Thus, it can be used with students who are quite young. The accommodations matching system initially included a guidance component that asked the teacher(s) who have worked with the student in the classroom which accommodations the student needs. However, this information is not utilized at this point in the STELLA decision-making, but was included for research purposes only. Unfortunately, Koran and Kopriva (2006) found that teachers’ understanding of student needs was no different than random and so this part will probably be deleted from the system. As mentioned above, the system utilizes direct skills information to the extent that it includes test score data in the decision-making algorithms, and considerations in the algorithms that flag data which are too dated. STELLA takes into account student performance on previous standardized testing, the ELL’s experiences with testing accommodations in the classroom, the teacher’s own intimate knowledge of each student’s needs and preferences as evidenced in the classroom, as well as many other variables that contribute to the recommendation of a specific package of accommodations for an individual ELL at a given point in his or her learning.

Once STELLA was completed, two initial validation studies were conducted in 2005. One study collected data on the feasibility of the system (Kopriva et al., 2005). It also researched how the STELLA findings compared with teacher recommended assignments (Koran and Kopriva, 2006). Nineteen teachers who spanned the range of K-12 participated from three states, each selecting six beginner to advanced ELL students. A total of 114 sets of files were completed with each set comprising three completed forms per student. Feasibility results showed that teachers, who were the coordinators of the data collection on each student, could clearly identify needs of diverse sets of ELL students across schools and states. Findings suggested that the system collected consistent data from like questions over
forms, to be used for triangulation and confirmatory purposes, and that each form also appeared to contribute unique data about each student from the unique vantage point of each source.

Koran and Kopriva reported that, in addition to completing the STELLA protocols for their students, teachers were also asked to provide their own accommodation recommendations for each student at three points during the form completion process: before the forms were completed, directly after the data collection and based only on information STELLA had collected, and then later with any additional information they considered to be important about each student (but which was not represented on the STELLA forms). Subsequently, a blind panel of three ELL practitioners and an ELL researcher were convened and asked to independently rate how each of the teacher accommodation recommendations, the STELLA recommendations, and a randomly generated set of accommodations compared. These raters did not know the students; rather, they made their ratings after reviewing data from the forms for each of the 114 students. Rater findings were subjected to goodness of fit analyses. Results indicated that STELLA was found to consistently and significantly be the best fit, over students and over all recommendation alternatives. On the other hand, however, teacher recommendations disappointedly did not prove to be different as teachers learned more about each student from the data collection process. Nor were any of these teacher recommendations significantly different from the random assignment.

The purpose of the second study was to investigate if students who received selected STELLA recommended accommodations performed better on the test relative to those who received improper accommodations or no accommodations whatsoever (Kopriva, Emick et al., in press; Kopriva et al., 2007e). It was designed to provide information about the validity and effectiveness of the particular STELLA assignments utilized in the study. Initially, 276 third and fourth grade South Carolina ELL students who spanned the range of English language proficiency completed a computerized mathematics test under randomly assigned accommodations that were implemented electronically as the students took the test. Three accommodations were used (oral English, bilingual word translation, and picture-word “translation”) and students randomly received one, two or three of them. One group received no accommodations. Afterwards, additional data about the students was used to assign students to one of the three groups (proper accommodations, improper, or none) as per the STELLA framework. Findings showed that students who received proper accommodations performed significantly better than either students receiving improper accommodations or no accommodations. It also showed that students who received inappropriate accommodations (as per
the STELLA assignment) scored no better than those who received no accommodations. This study not only verifies the reasonableness of the affected STELLA assignments but also suggests how important it is for students to receive proper accommodations vs. improper ones.

It is anticipated that future work with STELLA will continue to investigate its usefulness and effectiveness. Developers are interested in studying if the system works well for ELL students of all backgrounds and at all proficiency and grade levels.

Implications

At this point it should be clear that options in the ways test materials are presented and administered is very important for students who come from diverse cultural backgrounds and have English literacy and language challenges. However, no matter how adroitly and carefully these options are selected and created, if they don’t get to the correct students who need them, they are not useful. Emerging research suggests that use of improper accommodations may result in scores that are not substantially more valid for these students than if they received no accommodations. This is very troubling for many reasons but particularly two. First, individual performance on large-scale academic tests and scores may misrepresent the skills of students who do not receive proper accommodations. This can have profound consequences for future learning opportunities for these students, including improper placement or other student level decisions, as well as inappropriately impacting educational agency level decisions and possibly resource allocations. Second, improper accommodations may help explain why the findings from accommodations research have, to date, been so mixed. It may be suggested that only when researchers employ rigorous means of matching students to accommodations they are empirically testing will their work provide clearer results about which options are useful and which are not. Currently, many of us are left with the vague notion of which accommodations are helpful but we’re not sure how or when. The real question may well be “Useful for whom?” It is argued here that this is the question researchers should be investigating.

Many student factors are critical in understanding how to properly instruct ELL students as they become more proficient with the language and conventions of U.S. schooling. This chapter suggests that a subset of these key indicators appear to be salient in determining which large-scale testing options a student should receive, especially as long as standard assessments of academic content remain similar to those which are currently in use today. Test purposes, environments, standardized conditions that assume few if any procedural or material adaptations, and reliance
on written language (most typically in English) mean that some student issues as compared with others take on more importance in the current testing climate. Several of these have been identified here and, it is argued, should be used to identify and determine accommodation need for individual students. Likewise, certain accommodations or testing options have been highlighted as being especially useful for ELL students with certain needs or strengths. These have been discussed in the previous chapters. In tandem, the recognition of these characteristics appears to set the stage for developing ELL matching systems which can be useful for both practitioners and researchers.

To date, the form that these systems should take is not entirely clear. Simple policy guidelines do not appear to be sufficient. It is possible more specific guidance may be helpful and effective, with some type of systematic training and rigorous oversight procedures in place. Certainly, teachers often know the students well, and IEP teams utilized with students with disabilities suggest that other personnel or even the students themselves provide important and unique insights. Relying exclusively on the knowledge and time associated with compiling materials, educating and utilizing overtaxed educators, parents and students, however, may not be practical or reasonable even though in less demanding circumstances they may be “up to the task”. Also, research suggests teachers or other educational specialists are particularly well positioned to identify needs but, to date, appear to struggle with how to assign accommodations. This may or may not extend to parents, other school personnel and students. Computerized systems that guide personnel through the policy process provide standardization and, in the future, may also provide selected educational opportunities as the needs arise within more complex programs.

Electronic systems that provide differential boost information or that directly test the levels of students’ ancillary skills are interesting because they provide up to date needs information that offsets records or other data which may be dated or educator judgments that may be skewed. However, to date, they appear to be time consuming for students. Informant approaches should help streamline the process; however, unless stringent evaluation is part of development and ongoing implementation, they may provide a too-simplistic set of recommendations which would not properly address the nuanced and ever-changing needs of these populations. All in all, it appears that theory-driven systems need to be computerized as long as the algorithms appropriately address the complex realities associated with accommodations assignment.

As system designs mature, any of them will need to have the ability to handle relevant data in a way that sensitive decisions can be made not only for each student but consistently and even-handedly across students. It
is also anticipated that most future systems will be hybrids in some form or another. Such systems will need to have the capacity to prompt, include and combine large numbers of discrete pieces of information in complicated ways with ease, and be able to update frequently and store algorithms and output for use by different educational agencies. All in all, methodologies would probably represent variations of expert (or knowledge-based) systems used in other disciplines (for instance see Wright and Bolger, 1992) where data collection sources include humans as well as other knowledge-based data about students and characteristics of testing practices.

Besides continuing to develop more sophisticated data collection methodologies, continuing attention to the algorithms that are used to convert and combine data, and to the decision-making rules which match student factors with accommodation characteristics, will be important. Additionally, there should be more investigation of the teachers’ role as student advocates and how this affects their input in systems. The effect of stakes on the accuracy of teachers’ responses on questionnaires and the effect of accountability in the decision making system on teachers’ judgments may be useful areas of inquiry as well. Third, implementation issues need to be identified and evaluated. There is currently little published work on these implementation considerations of utilizing these matching systems but there are likely to be practical issues associated with them. For instance, optimal timing to use the systems in order to provide a recommendation should be considered. Accommodations decisions must be made far enough in advance to allow for any practice and to allow sufficient time to set up the logistics of implementing the accommodation for the large-scale testing sessions. Relative to many students with disabilities, the needs and abilities of English learners are constantly and quickly changing as they become familiar with U.S. culture and testing practices. Research should consider what amount of time before testing would allow for the accurate assessment of student ability and need while still allowing sufficient time to implement the recommendations appropriately. Further, while more attention is being given to getting the right accommodations options matched with the right students based on individual student need, the next link in a valid inferential chain is making sure that the accommodations actually get administered to the student, and get administered properly. Some limited research literature (e.g., Jayanthi et al., 1996; Solomon et al., 2001) and extensive practitioner experiences suggest that there appear to be issues with implementing test accommodations that have been recommended through current operational processes.

Finally, some researchers envision accommodation assignment approaches that are integrated into computerized content area testing
systems. Kopriva et al. (2006b) discuss future testing systems that are sensitive to a number of nuances about the ELL students and alter accommodations by item. These systems would not only allow for matching algorithms to operate in them, such as that produced by STELLA, but they would include algorithms that address item characteristic data which are more finely tuned to student needs. For instance, Solano-Flores et al. (2002) suggest that ELL students differentially use native language or English items when presented with dual language forms. While others (e.g., Duncan et al., 2005) suggest students use primarily one form or another, Solano-Flores suggests that his more fine-grain research points to small preferences depending on a number of linguistic factors. In a similar fashion, Ketterlin-Geller (2003) predicts that Accommodation Station may one day implement recommended accommodations for students with disabilities directly within a computer-administered academic test system. That is, similar to the thinking of Kopriva and others, she suggests that the accommodations matching system could be embedded in a larger assessment system that also administers the accommodations.

Endnotes

1. In some cases, the academic L1 is actually a student’s second language, if their first language was a local dialect or oral language not used to teach academic concepts to students. In this section, L1, native languages, home language, and academic ‘L1’ will be used interchangeably as it is generally the degree of non-English academic language proficiency that is taken into consideration when assigning accommodations.

2. Depending on what is being measured, reading and spelling could be target skills and, if so, they should not be accommodated.

3. The Valid Assessment of English Language Learners (VAELL) project completed in 2005. Some results from the project can be found in Winter et al., 2006; and a recently completed article (Kopriva et al., 2007d) has been submitted for publication. University of Maryland.
When responding to open-ended or constructed-response assessment items written in English, English language learners (ELLs) are required to undertake multiple steps: reading and interpreting a question, devising a solution or strategy for responding, recording their answer, and communicating their reasoning. The response usually includes some language, and the response is often scored by monolingual native English speakers. Yet, for many of these students some of the linguistic and cultural features of their responses may be unfamiliar to the scorers, and a review of National Assessment of Educational Progress (NAEP) science items and student work concluded that this unfamiliarity may contribute to inaccurate scoring in typical large-scale situations (Kopriva and Lara, 1997). Time pressures also compound the problem as it is not unusual for scorers to evaluate responses at the rate of one per minute for short answers or three to five minutes per essay question for up to eight hours per day.

The language requirements associated with responding to constructed response items may suggest to measurement researchers and development staff that this type of item is not preferred by ELLs. However, as Chapter 6 explains, two studies have reported that this kind of item is favored over multiple choice (Kopriva and Lowrey, 1994), and the results appear to be more valid for ELLs than are the results from multiple choice items (Kopriva et al., 2007b). Therefore, properly utilizing this item type, and offering training and other guidelines to help scorers address the linguistic and
cultural issues appears to improve scorer accuracy. Improved accuracy in scoring, in turn, seems to contribute to assessment validity.

First, this chapter outlines the linguistic issues and cultural influences that affect ELL student responses. While these considerations are complex, it appears that brief training and adequate support during scoring may be able to ameliorate a reasonable number of these problems. Therefore, guidance about training will be reported in the second section. Most of the examples in this chapter come from the *Guide to Scoring LEP Student Responses to Open-ended Science Items* (Kopriva and Sexton, 1999), and the *Guide to Scoring LEP Student Responses to Open-ended Mathematics Items* (Kopriva and Saez, 1997). For a more detailed discussion of this topic and additional samples of student work, readers are directed to these two documents and to a student work sampler (Solano-Flores *et al.*, 2001), available online from the Council of Chief State School Officers.

**Issues Which Impact Scoring**

As noted in Chapter 2, English language learners are not homogenous. While all ELLs are in the process of acquiring English, they are typically at different stages of acquisition with respect to the use of academic language. Research indicates that the development of ways to structure academic explanations and arguments may take several years. These skills involve a complex process of learning academic language while they are also learning content and acquiring subtle language skills which are more familiar to native English speakers from the mainstream U.S. culture (Collier, 1987; Scarcella, 2003; Wong-Fillmore, 2000). Further, English reading and writing levels of English learners tend to vary considerably and mature at different point as they acquire the language.

As they are learning English, ELLs will interpret and respond to academic assessment items with differing degrees of success, depending upon their academic background, socio-cultural experience, and proficiency in handling the academic language of the subject in English. Even if they are allowed to respond in their home language, a great number of these students are not literate in this language, or their literacy lags significantly behind their grade level. Therefore, responses in L1 are often rudimentary, and are compounded by many of the same problems which face these students if they respond in English. Further, cultural issues of newer ELLs can considerably impact their responses. Because of some or all of these considerations, the students will often articulate their reasoning, but in ways that are difficult for the scorers to grasp, especially under the pressure of a rapid, high-volume scoring situation. Some of the more pressing linguistic, cultural, and language acquisition issues will be summarized below.
Linguistic Considerations Influenced by the Home Language

Student responses to test items reflect five levels of language in the English writing system—phonology (sounds), morphology (words), syntax (sentences), semantics (meaning), and use (pragmatics). Their responses also reflect interaction between prior knowledge or experience and new knowledge. ELLs, by virtue of their knowledge of a primary language and a second language, have more linguistic resources to draw on. While this is largely a plus, initially it tends to lend additional complexity and more confusion than it would for native English speakers. As such, several linguistic features may be apparent in an ELL response. Lily Wong-Fillmore (2004), among others, discusses that ELLs with little English language proficiency face the most challenge in how they express themselves on constructed response items. More advanced ELLs, though, also tend to go through a phase where they are “hypersensitive” about inconsistent linguistic features in text. Unfortunately, many of these features are not obvious to native English speakers, often because of colloquialisms or other common ways that language is used. These inconsistencies are typically not taught to ELLs and only a maturing in the language will lessen their consideration of them. However, during this phase, their concerns sometime result in incorrect usages of linguistic conventions.

Native Language Influences in the Interpretation of English Sounds

Students’ writing may be influenced by sounds in their native language or dialect that differ from English sounds. ELL students may, for example, omit some final consonant sounds (writing “thin” for “thing” or “teacha” for “teacher,” for example) or transpose certain sounds (as when a Spanish speaker writes “sciencie” for “science”). Spanish speakers may also substitute one sound for another (for example, “becows” for “because” or “mathed” for “method”), or reduce consonant clusters (writing “cientist” for “scientist”). Examples that follow in other sections will illustrate how difficult it is to sometimes understand what students are writing. This is sometimes true even for native speakers of the students’ home language because, unfortunately, often students will mix native influences of sound with English phonetic influences as they complete their response.

Code-Switching

Code-switching is defined as the alternate use of two languages at a word, phrase, clause, or sentence level. This is a common phenomenon for English language learners, and may be present in their oral or written samples. A Spanish speaker might, for example, write “también” instead of “also” or “hijos” instead of “offspring,” even while otherwise using English in a response.
**Transposition, Substitution, and Reduction of Words**

English learners may follow the rules of syntax or word order used in their primary language, as in the following examples in Table 11.1.

Sometimes ELLs substitute common words for precise scientific terms and concepts. For instance, the use of “fattest” for “greatest,” “smallest” for “fewest,” or “plus” for “added” are commonly found substitutions of quantitative terms. Depending on what the item is supposed to measure, the student’s substitution may or may not be acceptable.

Reduction of words is depicted in a Czech student’s reflective journal entry in Figure 11.1. In this instance the student uses “inof” for “enough,” “solf” for “solved,” and “tim” for “team.” The student’s attempt to respond to the reflective prompt adequately conveys understanding by communicating the limitations and strategies in the task analysis.

<table>
<thead>
<tr>
<th>Transliteration</th>
<th>Translation</th>
<th>Original Syntax</th>
</tr>
</thead>
<tbody>
<tr>
<td>“a rabbit black”</td>
<td>“a black rabbit”</td>
<td>Spanish: “un conejo negro”</td>
</tr>
<tr>
<td>“a garden vegetable”</td>
<td>“a vegetable garden”</td>
<td>French: “un jardin Potager”</td>
</tr>
</tbody>
</table>

**Table 11.1**

![Figure 11.1](image_url)

What was difficult about this project? Any problems or opportunities?

At first wi had a problem for information bot after wi found inof information

“At first wi had a problem for information bot after wi found inof information.”

What worked? How did you solve problems, if any?

wi solf problems as an tim.

“wi solf problems as an tim.”

Intended meaning: At first, we had a problem finding information. Afterwards, we found the information. We solved problems as a team.
PHONETICS SPELLING BASED ON NATIVE LANGUAGE SOUNDS

English has thirty-eight phonemes (sounds) and only twenty-six letters, meaning that one letter can represent more than one sound. Some ELLs use sounds from their native language while learning to differentiate between that sound system and the less familiar sounds of English. Native language phonetic forms may be present in both speaking and writing. The following chart in Table 11.2 illustrates this phenomenon for native Spanish speakers.

Also, students will sometimes use spelling conventions from their home language to write English words. For instance, in Spanish, words beginning with “s” and followed by a consonant may be written as “es,” as in “eschool” or “escul” (school).

In the Spanish oral language tradition, the ending vowel sounds of one word are usually connected to the initial sound of the following word, making them sound like a “run-on” word. In Figure 11.2, the student is connecting and interpreting phonemes: “earths serfers pressias so hard that it expresis heiat” to represent “earth’s surface’s pressure is so hard that it expels its heat.” The student attempts to explain volcano formation by the interaction between the earth’s surface and the earth’s internal heat, in terms of pressure being released.

TABLE 11.2

<table>
<thead>
<tr>
<th>Native language influence</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>“b” for “v”</td>
<td>“today was bery hot”</td>
</tr>
<tr>
<td>“y” for “ü”</td>
<td>“we can yuse the thermometer”</td>
</tr>
<tr>
<td>“ch” for “sh”</td>
<td>“the horsechoe crab has a long tail”</td>
</tr>
<tr>
<td>“d” for “th”</td>
<td>“de eard is tilted”</td>
</tr>
</tbody>
</table>

How are volcanoes formed?

I gessa That earth’s serfers pressias so Hard that it expresis heiat.

Figure 11.2 Spanish
INVENTED SPELLING

Just as native English speakers learning to write make their best guess about the spelling of English words based upon their knowledge of phonetics, so ELL students take this developmentally appropriate step, though it may not be grade-level appropriate in relation to native speakers. ELL students might, for example, write “wen” for “when” or “in” or “an” for “and.”

Some scorers have noted that interpreting invented spelling seems to be more difficult in scoring papers of older students because often the scorers make more assumptions about the base literacy levels of the test takers. Unintentionally, they appear to connect levels of writing skill and knowledge of conventions with levels of targeted math or science ability which, of course, would disadvantage some students with literacy challenges. By viewing examples and discussing the need to disentangle the targeted skills of older ELL students (as well as others) from their communication methods during scorer training, Kopriva and Lara (1997) noted that trainees reported changes in how they viewed responses of several kinds of students whose literacy seriously lagged behind most of the test takers. For these students, the scorers reported that they spent more time attempting to decipher the responses than they would have ordinarily. Further, they believed that the training and extra time, in general, resulted in what they felt was a more accurate score than they would have otherwise given to the responses of students with more rudimentary literacy skills.

Figure 11.3 shows two students’ responses to the same prompt. Response 1) is an observational response, whereas response 2) attempts to include a comparative and quantitative approach to the task in the prompt. Their interpretation of English phonemes is evidenced by their inventive spelling, such as:

“theacker” vs. “thicker”
“proble” vs. “probably”
“hevear” vs. “heavier”
“obsorbs” vs. “absorbs”
“dident” vs. “didn’t”

Closely related to innovative phonetic spelling in one response is the abbreviation of words and the condensation of them into one word, as in “papertall” vs. “paper towel.”

OMISSION OR MISUSE OF TENSE MARKERS, PUNCTUATION, ARTICLES, PLURALS, PRONOUNS, PREPOSITIONS, OR OTHER WORDS

Omissions and misuses may occur if an English language learner lacks understanding of English conventions or has no native-language equivalent. For example, a West African student whose first language is Creole might
How did you know from the experiment which paper towel holds, soaks up or absorbs the most water?

Student 1

The blue papertall soaks up the most because it was proble more absorbs.

Student 2

Well because the blue one was bigger and theacker. And it dident have that much water and it was hevear.

1. “The blue papertall soaks up the most because it was proble more absorbs.”
2. “Well because the blue one was bigger and theacker. And it dident have that much water and it was hevear.”

Intended meaning:

1. The blue paper towel soaks up the most moisture, because it was probably more absorbent.
2. Well, because the blue one was bigger and thicker. It didn’t have that much water in it and it was heavier.

Figure 11.3 Spanish

omitted the “s” at the end of verbs with third-person singular pronouns or use a past-tense verb only at the beginning and the end of a writing passage, using present-tense in between. Figure 11.4 demonstrates the omission of prepositions, pronouns, plurals, capital letters, and punctuation, as well as the misuse of pronouns in a Hmong student’s response.

When ELL responses are poorly written, it can be difficult to understand what they intended to say. A scorer who has to read a high volume of responses might give a very low score to this type of response; however, a scorer who is aware of the linguistic factors affecting ELL students can readily notice that the response makes more sense when the misuse of language or the missing words, letters, and punctuation marks are recognized. Further, mistakes observed in ELLs’ responses may reflect the structure of their native languages. Besides attending to inventive spelling more carefully, the scorers in Kopriva and Lara’s study (1997) also noted that they were more careful to read for meaning when “words seemed strange,” and when plurals were not used or non-targeted words were
omitted. In the example above, the student did not write the word “they” because in her native Hmong language, the pronouns are omitted under various circumstances. With a little mental editing, the same response might be read as shown below (editing in bold letters):

Intended meaning:

Sue is not good at planting her plants. That is why she only got one. Linda’s and Arletta’s plants grow better because they grew three and seven bigger plants, respectively.

**Cultural Considerations**

Items open to the charge of cultural bias may be flagged during the fairness review sessions that are part of the item development process. However, it is also the case that items reflect the conventions of the cultures under which they are developed, ranging from the symbols, characters, and markings
used to the style in which questions are broached. Of course, no discussion of cultural differences can be exhaustive, but a key part of scorer training is to discuss that individuals from other cultures may not conform to typical cultural patterns. Cultural conventions may include meaning, symbols and other language or character choices. They also sometimes include responses which are unexpected or that appear atypical to a scorer. Such background characteristics may prompt students to focus on certain types of information over others or to emphasize aspects of a problem or its solution that would not be accorded a similar priority by the scorer. Their responses may nevertheless be understandable and defensible, and the solution presented may be acceptable and consistent with the measurement intent of the item. General knowledge of the following points may help scorers make better sense of constructed responses submitted by ELL students. If possible, scoring notes for rubrics should be expanded to reflect potential linguistic, cultural, or language development issues through providing examples of considerations from the item pool.

**DIFFERENT SYMBOLS OR SYSTEMS**

Learning a new language also involves relearning writing conventions and meanings of words, as well as symbol usage specific to the new culture. In some countries, the native language is written from right to left and/or top to bottom, for example, and English language learner responses in English may mirror this. Although subtle, some of these new conventions may alter or even reverse the intended meaning. Table 11.3 below gives examples of some differences in language conventions.

**WORD CHOICES**

ELLs may use words from their language that sound the same as English words but mean something different. They may also create new words in English adapted from their primary language (neologism). A few examples of English and Spanish words that are homologous or from the same root but have very different meanings can be found in Table 11.4.

Figure 11.5 shows how a student creates a new word, *descompounding*, which resembles *descomponiendo*, or “to break apart” in Spanish. The “iendo” ending in Spanish is equivalent to the “ing” ending in English. The student has a basic understanding of how food is broken into smaller components during digestion, but this may not be clear to a scorer without an understanding of the Spanish word.

**STYLISTIC PREFERENCES**

Cultures differ in the kind of writing and discussion style they value and promote. Some cultures put a premium on concise explanations, while others allow for more digressive approaches. Some students favor a
<table>
<thead>
<tr>
<th>Type</th>
<th>Explanation/examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithms</td>
<td>The algorithms taught and used to perform operations such as chemical equation balancing or division may vary by country. This may create the wrong impression that a student’s response is incorrect whereas he/she may be using a different but correct approach to solve some problems.</td>
</tr>
<tr>
<td>Dates</td>
<td>In some countries, dates are not written in the same order as in U.S. English. For example, “12/5/98” may mean “May 12, 1998” rather than “December 5, 1998”.</td>
</tr>
<tr>
<td>Meaning systems</td>
<td>Many students from abroad are not familiar with the U.S. standard system, which may affect their understanding of prompts involving distance, volume, weight, and temperature. Except where the item specifies, calculating miles to the gallon should be considered interchangeable with calculating meters and liters, etc.</td>
</tr>
<tr>
<td>Number reading</td>
<td>In some countries “billion” is interpreted as “one million millions” (1,000,000,000,000), not “one thousand millions.”</td>
</tr>
<tr>
<td>Punctuation</td>
<td>In some languages, the rules for punctuation are not necessarily the same as in English. That involves the use of commas, colons, semicolons, periods, quotation marks, m-dash, etc.</td>
</tr>
<tr>
<td>Number writing</td>
<td>In some European and Latin American countries, periods are used instead of commas to separate hundreds, thousands, and so on. A student might read “3.001” as “three thousand and one” rather than “three and one thousandth.” In England, the decimal point is written at the midpoint between two numbers, not at the bottom. This could be confused with the dot that is sometimes used to indicate multiplication in U.S. mathematical notation.</td>
</tr>
<tr>
<td>Computation symbols</td>
<td>Mathematical operations are indicated with different symbols in some countries. For example, the diagonal symbol used for long division in some countries (e.g., 3/90 means 90 divided by 3) could be confused with U.S.-style fractions. (43/2) means 43 divided by 2.</td>
</tr>
<tr>
<td>Familiarity with animals and plants</td>
<td>Many animal and plant species that are very common in the U.S. may be unknown in other cultures. For example, some students have never seen boysenberries and hedgehogs and their languages may not even have words to refer to them.</td>
</tr>
<tr>
<td>Misleading phonetic similarity</td>
<td>Some words are phonetically similar to English words but may refer to different things. For example, in Spanish “ruibarbo” is not the same plant as “rhubarb,” “lima” is “lemon” (not “lime”); “limón” is “lime” (not “lemon”), etc.</td>
</tr>
</tbody>
</table>
deductive style of reasoning, in which they move from general principles to specific examples, while others argue from examples to general principles via inductive reasoning. An ELL student who is a proficient writer or speaker in his or her primary home or academic language may demonstrate some cultural and stylistic differences that could potentially be scored unfairly if they are unfamiliar to the scorer. Stylistic preferences may not be evident, however, if students lack proficiency in the language. In that case, their response is likely to be an extremely limited one, perhaps even a verbatim copying of some elements of the item.

Figure 11.6 illustrates a circular or digressive style in which a Spanish-speaking student appears to be talking around the primary issue, introducing an example that does not immediately appear relevant and then eventually connecting the example to a conceptual understanding of the concept. Such a response may actually be fuller and richer than a simple

<table>
<thead>
<tr>
<th>English word/meaning</th>
<th>Spanish transfer and new contextual meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insulated: thermal protection</td>
<td>Insolado: being ill with sun stroke</td>
</tr>
<tr>
<td>Library: book repository—study hall</td>
<td>Librería: stationery and bookstore (students will use library as bookstore)</td>
</tr>
<tr>
<td>Saturated: type of fat</td>
<td>Saturado: highly concentrated, relating to solubility</td>
</tr>
<tr>
<td>Fabric: a type of material/cloth</td>
<td>Fábrica: a factory</td>
</tr>
<tr>
<td>Carpet: a type of rug</td>
<td>Carpeta: a binder</td>
</tr>
</tbody>
</table>

**Table 11.4**

Figure 11.5 Spanish

Explain why food stays so long in the stomach before being allowed to pass gradually into the small intestine.

"because the stomach is descompounding the food for it to be ready to pass to the small intestine."

Intended meaning: Because the stomach is breaking down the food, for it to be ready to pass it to the small intestine.
definition; however, it may strike scorers as wordy and off-topic, particularly if they are working under tight time constraints.

Conversely, because of their native language discourse structure and independent system of visual communication (e.g., characters or syllables), some students prefer a compact, abbreviated response style in which every sentence in a paragraph is a topic sentence. Scorers in high-volume situations, who often do not read every word carefully, may miss some important arguments that are presented by students using this style. These responses may thus be scored lower than is appropriate.

In Figure 11.7, the two topic sentences provided by a Laotian student are different points that are called for in the rubric. One refers to the physical property of materials and the other refers to the property of light. The response could be dismissed as limited, even though it matches two of the criteria set out in the rubric.

---

**Define Inertia**

*When you try to push a piece of furniture probably you can't. Then you try harder and maybe you can move it a little bit. The reason is that these heavy things they don't really want to move. You put your force and sometimes it moves and sometimes you have to make more force and when you make more force then maybe it moves a little bit. All matter tend to do this; it does not want to move and this is inertia.*

“**Figure 11.6** Spanish

*Intended meaning: When you try to push a piece of furniture and you probably cannot, then you try harder. Maybe then you can move it a little bit. The reason is that these heavy things don't really want to move. You just put your force and sometimes you have to (give it) more force and when you have (added) more force, then maybe it will move a little bit. All matter tends to do this; it does not want to move. This is inertia. (Law of inertia: an object at rest will tend to stay at rest, until an external force is applied to it.)*
MISUNDERSTANDING THE MEANING OF WORDS OR PHRASES

Sometimes a student can misunderstand what an item requires because words or phrases in the item mean something different to him or her than to a native speaker. This is actually an item writing issue, as compared with a scoring issue. However, if scorers identify that several students interpret the item in a particular way, attention should be paid to how scorers might interpret the text. For instance, a constructed response fourth grade item in California asked students how to use the “fewest” number of coins to purchase an object. A substantial number of ELLs interpreted “fewest” to mean smallest value or smallest size.

MISUNDERSTANDING BASED ON DIFFERENT VALUES AND EXPERIENCES

Students read an item based on their values and experiences, and the values and experiences prevalent in their culture. Since the values and experiences of the students’ native culture and the U.S. culture may be dissimilar, there is room for misinterpretation. Chapter 6 gives the example of an assessment item that asks students to create a fair race, which may elicit unexpected responses from some students. While item developers expect students to create a racecourse in which all of the contestants have to run equal distances, some students on a Michigan exam interpreted “fair” to mean that all contestants have an equal chance of winning. This may be especially true in cultures that do not place as much emphasis on competition. As a result, these students created a race course in which the slower contestants will run shorter distances. On the basis of their interpretation of a race and the notion of fairness, this is a valid response.
Also, in Chapter 6 (Figure 6.3) is an ELL’s response to a test item about how a community water distribution system works. This student, a recently arrived Laotian who was learning English in the home country, sketched a bamboo channeling system for accessing and distributing fresh water to the community. The English that is used is spelled correctly, although the picture is clearly reflective of the student’s home experience. As the figure shows, the rainwater was harnessed from the spring and transported via use of the bamboo channels to a large water tank which appears to be used for storage. Then, presumably, the water is retrieved from the large tank in the buckets for final use in the homes and/or for the terraced garden, via, perhaps, another smaller storage tank. This response was unexpected (scorers were looking for an urbanized system such as those used in the U.S.). However, the concept of the response clearly shows the student’s understanding of how a water system functions, and should be scored accordingly.

Finally, Figure 6.4 in Chapter 6 reflects how an intermediate ELL Punjabi student, whose experiences and cultural background prohibits the consumption of beef but allows milk drinking, interprets and responds to a request to show a food chain. Although misconceptions or lack of language can be found in the Figure, including the sun giving food to the grass, the student seems to have a rudimentary understanding of a food chain as a flow of energy. The student also understands how food sources from living organisms are used by higher life forms in the concept defined as a food chain, via energy exchanges such as oil which was made into gas for the truck, and the acquisition of money from other sources which was used to buy the milk. Interestingly, the student is attempting to adapt his prior experiences to his new home as evidenced by his depiction of the Quick Stop convenience store in the picture. Again, this response was unexpected and rubric examples did not cover such a possibility. However, given the student’s prior experiences and his home value systems, it is argued that responses such as these should be considered more carefully when a broad range of ELLs are included in the test administration.

Language Acquisition Issues
As discussed in several chapters, ELL students typically develop facility with social English well before they are proficient in academic English, the language of large-scale testing. Because they are able to interact with other children in English, they may be placed in classes with native English speakers. While this placement is appropriate for some reasons, the ELL students may be at a disadvantage if assessment (and instruction, for that matter) is carried out under the assumption that the ELL students had the same grasp of academic English as their native-speaking peers.

In order to separate what a student knows in a subject from how well he or she can articulate it in their responses, it is helpful to consider certain
standard patterns of misunderstandings that are part of the typical process of language acquisition. The instances which follow are covered in Chapter 6 as part of access-based item writing. The three points are illustrations of what scorers should to be cognizant of, and examples of what should be broached as well during training. They are included here, also, to emphasize why rubric notes and training should emphasize and discuss the target construct for each item which will be scored. By being aware of the target objectives, the scorers can make fine-tuned judgments about target-relevant versus irrelevant misunderstandings. The goal, of course, is to provide guidance to the scorers about how the students would not be penalized for target-irrelevant aspects of the responses (even though they may be inconsistent with native speaker mistakes) while, at the same time, expect them to hold the students accountable for knowledge and skills related to the same target objectives.

**CONFUSION RELATED TO WORDS WITH MULTIPLE MEANINGS**

A brief discussion about dual meanings in English is one central point that should be made during training. Once scorers are aware of the item’s target knowledge or skills, they can judge when it is appropriate to substitute informal language and when it is not; when to be sensitive to spelling which may be due to homophonic reasons; or how to score items which have been poorly written for certain groups of students. For instance, sometimes the technical language in a discipline uses “everyday” words in a distinct way. Earth (the planet) vs. earth (soil), plate (as a tectonic structure) vs. plate (dish); and fault (geologic feature) vs. fault (error) are just a few examples of this. Homophones, for example hole vs. whole, break vs. brake, and passed (read as past) vs. past (time) can cause confusion as well. Likewise, colloquialisms and idioms are replete with phrases and words which, taken as a whole, mean something different than the words would otherwise suggest. Sometimes it should have been the item writers’ responsibility to catch dual meanings which could cause confusion, such as when left (as opposed to right) vs. left (what is remaining) is used in an item. However, when these points have been missed during development, scorers must try to decide whether incorrect responses were substantially caused by an item which needed clarification. Since the interpretation is not the fault of the affected test takers, scoring supervisors need to be alerted so they can decide how to handle situations such as this on a case by case basis.

**NOVICE SENTENCE AND PARAGRAPH STRUCTURES**

As Chapter 6 describes, it is extremely common for ELLs to exhibit emerging syntax, that is, sentence and paragraph structures characterized by such features as lack of subject/verb agreement, limited use of
comparative forms of adjectives, run-on and fragmented sentences, and basic, repetitive sentence structures. Figure 11.8, a Vietnamese student’s attempt to describe how mountains are formed, provides an example of emergent syntax. It is the responsibility of the scorers to differentiate emerging language from targeted knowledge, albeit in an immature fashion.

**ALTERNATIVE RESPONSE FORMATS**

English language learners may elect to minimize their use of text in responses by relying heavily on numbers, charts, and pictures. This approach is also consistent with instructional strategies by which many ELLs are taught, which typically emphasize graphic organizers, brainstorming outlines, and other visual aids. Depending on what is being measured, these alternative response formats may or may not be considered acceptable. As noted in Chapter 6, including illustrations as acceptable representations of content in the scoring guide criteria is one way to separate what ELLs know in a subject area from how well they can write. Further, many English learners initially develop higher levels of English listening and speaking proficiency than English reading and writing proficiency, particularly if they are not literate in their primary language (LaCelle-Peterson and Rivera, 1994; Wong-Fillmore, 1994). These students may be better able to show what they know if they are allowed an oral response option, for instance into a tape recorder, or a voice-recognition program on a computer. These options and others are described in Chapter 9. As discussed earlier, utilizing computers as the way to present tests to students opens up the development of items that take advantage of additional response possibilities otherwise not terribly realistic if the test is delivered in a paper-and-pencil form.

**Figure 11.8 Vietnamese**

```
Pacific Plate

It is caused by heat which tries to push up the make more plates so they side each other. It was making a mountain.
```

“*It is caused by heat which tries to push up the make more plates so they slide each other (this o) It was making a mountain.*

Intended meaning: It is caused by heat, which pushes up to make more plates. So, when they (the plates) slide into each other, they make a mountain.”
Explain in detail the change in potential and kinetic energy that occurs to a rock when it is first sitting at the top of a mountain, then when it starts to roll down the mountain, and finally when it reaches ground level.

“on the to[p] it only has potential—In the Drop It has Both kinetic and potential—Before it hits the floor It is kinetic—Then Potential once again when on the Floor—(Hey Boy Come Down Here)”

(* = Rock  k = kinetic  P = potential) chart and arrow/sequential coding.

**Figure 11.9** Spanish

Figure 11.9 is a Spanish-speaking high school student’s visual representation of the concepts of kinetic and potential energy. This is an example of the type of response an earlier ELL would provide, although this particular response lacks certain key understandings of what the item is requiring. It is included here, in part, to illustrate why responses with so little language, with language that is informal, and one that contains an elaborate visual depiction, could mislead scorers who might equate formal language and presentation with target ability.

**Access-based Scorer Training**

It appears that only one study has been conducted to determine whether scorer training which touches some of the issues raised earlier might impact
how responses from ELLs are evaluated. Some results from the investigation (Kopriva and Lara, 1997) have been mentioned above. This study, involving a re-scoring of NAEP science items with NAEP contractor trained scorers, suggested scoring might be improved when monolingual English scorers are given a relatively short training on the basics discussed above, and when bilingual ELL educators were integrated into the scoring process. While most ELLs were waived from NAEP testing in 1992, which is where the items and responses used in the study were selected from, a subset of the more advanced English learners was assessed. Twelve NAEP items were subsequently selected, and all of the ELLs’ work from these items, as well as a random subset of responses from native English students were pulled. For this study, ELL responses were blindly seeded within the total set of responses to be scored for each item. Due to the matrix sampling approach used in NAEP, each student received only some of the 12 items. In all, a total of 4,100 responses from the 12 items were scored (including 929 booklets from ELL students). Additionally, scores on the 12 items from a random set of responses were selected to be used as a control. This set were scored earlier by the scoring subcontractor to NAEP.

Three tables of scorers (four monolingual English and two bilingual scorers at each table) plus three monolingual native English table leaders participated in the scoring. The twelve native speaker scorers, mostly current or former educators in the state where the scoring was occurring, were selected by scoring subcontractor staff from among those with whom they are currently contracting to score NAEP items. Table leaders were also selected from those currently under contract, and neither they nor the native English speaker scorers had originally scored this set of items for the subcontractor. The bilingual educators were selected from participating states, but, unlike the native English speaking scorers, none of these educators had participated in NAEP scoring prior to the study.

After the half day regular scoring training for all participants on the twelve items (conducted by subcontractor staff), participants were presented with one hour of introductory material identifying key linguistic features that affect ELL responses. This part of the training was followed by a short review of ELL training examples from three of the twelve items to be scored in order to familiarize participants with the types of issues they might face during their work. The study training was conducted by a state testing person, with expertise in ELL issues. She was also available during scoring for consultation upon request by the table. Scoring occurred over a three day period. Throughout the scoring, the bilingual educators participated fully in the scoring of the responses, and, along with English scorers, participated in the table discussions that arose during scoring. They were not permitted to score any ELL responses other than those which
were seeded randomly among the responses assigned to them. Interviews with each scorer were conducted at the completion of training. The scoring process was conducted by the subcontractor within the procedures and guidelines they used for all scoring of NAEP items. This included the average rate of scoring, double scoring of a percentage of items, and the use of oversight and auditing procedures to check for scorer drift and handle discrepancies.

Averaging over all responses for the twelve items, the score mean was highest for the bilingual educators (1.90), followed by the average score of the native English speaking scorers (1.85) and lowest for the control group of scores obtained from the initial scoring of the items in 1993. The mean differences for the ELL papers were more distinct between control and study participants (both native English and bilingual scorers) than were score differences between the groups for the responses of native English students. Most findings were significant, particularly for the ELL papers. In general, the results from the study participants were more similar to each other while the largest differences were between the control group and each study group. For instance, the overall difference between the control group and the native speaking English scorers ($t = 10.37$), and the control scorers and the bilingual educators ($t = 10.19$) were similar to each other and substantially larger than the difference between the bilingual educators and the participating native-speaking English scorers ($t = 4.2$). The training plus the presence of the bilingual educators at the tables appeared to raise the average level of scores overall, and more so for ELL responses. Differences between study participants may have been because of experience of the native speakers, and/or additional sensitivity or leniency of the bilingual educators.

Interview results indicated that most native English speaking scorers believed they were more conscious of differences of ELLs after training, but also responses which seemed to be from native speakers with “writing problems” and those responses which the native English scorers thought might have been from students with regional or rural idiosyncrasies. They appreciated a short introductory training followed by ongoing considerations at the tables during the training process. The primary limitations of the study were the heavy use of bilingual educators at each table and the restricted range of ELL responses. It is unknown how results for the native English speaking scorers would have differed if there had been fewer bilingual scorers at each table (which would be more realistic in most scoring situations). Further, this investigation only used papers from advanced English learners, and future studies should include responses from ELLs whose work would probably more definitively reflect the issues discussed earlier.
Scoring Process Considerations

Since it appears that just a single study has been completed, only tentative guidelines can be presented at this point. However, based on the investigations and other literature associated with item writing and use of text based accommodations, it seems probable that the large-scale scoring of ELL responses could be improved if these considerations are addressed. The three targeted areas discussed below include the training of scorers, the development and use of materials, and other selected aspects related to the scoring process.

Training Considerations

One procedure for conducting training of salient ELL issues was outlined above. As with training for other purposes, the intent is to initially introduce concepts and issues and provide examples throughout training. This is followed by interaction with the trainees at various intervals as they complete their work and attempt to implement the important aspects identified during training. Following are a few points to consider:

• Since these issues are new to many large-scale scorers, training needs to a) include time for the raters to hear the points; b) time to process the issues that come up in particular items for which they are responsible; c) have opportunities for clarification. This suggests the best approach may be to start with a short introduction to the issues as long as opportunities for discussion and clarification are built into the scoring procedures later on. In the study, interviewers reported that hearing the questions and discussion of others, as well as having their own questions discussed, was useful. Participation as one of a group appears to be a meaningful component when new concepts such as these are included in the scoring process. Recently, scoring contractors have been depending more on raters scoring via computer and sometimes doing so remotely. In some instances, this has minimized the group process nature of the scoring event or even eliminated it. If this is the case for the reader, it is important that other effective procedures be developed to not only introduce material but allow time for processing, include clarification discussions at specific intervals, and include oversight to make sure the raters are properly considering the points discussed here.

• Currently, no handbooks are available to help instructors design training in the ELL issues raised in the chapter. The booklets from CCSSO may be a useful tool for scorers and the study’s training followed the format in these documents during the introductory
hour. It appears the range of issues are important to touch upon but some larger organizational chunks are called for as well so scorers have an avenue for referring back to specific points within a larger contextual issue. Section I in this chapter divided the issues up into three components (linguistic, cultural, and language acquisition considerations). This might be a useful way to arrange the presentation of the information.

- The foundation of the training should be for scorers to expect and understand how to view and use item-specific target objectives. Around this foundation, discussion should occur about what else any given item is measuring and why those additional aspects are target-irrelevant. From there, the training can ask scorers to evaluate the target vs. the irrelevant parts of students’ responses. This process can help the scorers learn to identify target knowledge in unexpected contexts and otherwise differentiate immature communication strategies from the knowledge and skills in question.

- As the points are being introduced, using a large number of examples about these points is suggested. Clearly, this is because many of the scorers may not have experienced the types of issues raised here. They also have probably not had opportunities to see how responses which incorporate these issues are evaluated. Remember to include examples of work from some high ability ELLs so scorers can see that ELLs range in content achievement and can produce quality work albeit in an unexpected fashion.

- As noted from the study, scorers found the ELL training useful for scoring responses of other students, including poor readers/writers, students with disabilities in language, and students who write in dialects or use experiences or colloquialisms which are regional in nature. It may be possible to develop a training which highlights common issues of some of these student groups. It is important, however, to realize that each group, including ELLs, has unique issues and enough of these need to be covered too. As training modules are completed, it will be interesting to see how they are designed to handle multiple types of problems across groups.

**Preparing Materials**

Supplementary materials which address the issues raised above need to be developed for any successful training. For instance:

- As part of the training and implementation, a hard copy handbook of the issues, examples and explanations needs to be built and given to the scorers to refer to as needed.
Rubrics should include a clear and well differentiated explanation of the target objective for each item.

Besides examples during the training, ELL considerations need to be integrated into the item rubric notes. These are notes which usually accompany the rubrics for particular items and explain in more detail how to evaluate certain types of responses that scorers may encounter. The rubric notes should explain in more detail how irrelevant components can be demonstrated differently. Additionally, it is important to integrate ELL papers into the scoring practice sets. These are usually either “mini-tests” which scorers-in-training need to pass at a certain level, or “mini-tests” which may be used at intervals during scoring to re-calibrate scorers. A part of regular scoring training is to provide examples of what each score point possibility looks like. These are often called anchor papers. Besides providing examples of the access-based issues identified in Section I, examples of ELL work at most if not all levels should be routinely included. This is particularly important at the higher scoring levels as this is where scorers seem to have made the most assumptions of what “good” looks like. Routine integration of ELL responses sends the message to scorers about the diversity of acceptable responses and the range of strategies which can be used by students to demonstrate their knowledge.

Other Scoring Process Considerations

Besides training and the use of discussions about the ELL material during scoring, a couple of other points should be considered:

- Table leaders (or the person who serves that function in computer-based scoring) are those who traditionally are responsible for resolving scoring discrepancies. Kopriva and Lara found that it was important for table leaders to be especially well-trained in the issues that affect ELLs. When this is not possible, an ELL expert should be on-site. The expert should be fluent in the issues raised above, and also in how the considerations apply to scoring. A strong knowledgeable person should offset the need of having so many ELL educators at each table while still retaining the benefit of their expertise.

- It is important that at least one scorer at each table be an ELL educator who is fluent in English as well as knowledgeable about the response characteristics of emerging English learners. Fully bilingual educators who can fluently read and understand the common languages of ELL test takers as well as English should also be onsite in order to properly score L1 and code-switching responses.
Provisions should also be made ahead of time for scoring responses from students who speak less common languages. Remember, the ELL scorers should have expertise in the cultural systems of the test takers, not only their languages.

- All large-scale scoring operations assume that papers will be scored at an average rate of speed. The study suggests that it may be possible to utilize monolingual English scorers to score some of the responses from ELL students. However, it is important to keep in mind that these scorers are not fluent in the issues noted above. That is, they themselves are emerging learners in such issues as understanding characteristic features such as code-switching, phonetic spelling, and cultural conventions. Because of that fact, they will need more time per response to score some of the work accurately. This is most likely true especially as more low English proficient English learners are included in the testing system. As Chapter 6 explains, some studies have shown that these students prefer this type of item. In that case, if the items are constructed properly, it should be appropriate for a broad range of ELLs to participate in assessments with constructed response questions. As such, the number of ELL expert scorers should be increased accordingly to reflect the increased scope of responses expected from English learner students.
The test development process is iterative. As assessment systems are designed, items are written, reviewed, tried out and piloted, and as tests are field tested and finally implemented, information is collected regarding how items and forms appear to function. This information undergirds decisions about which items are completed, retained, changed, and which are weeded out altogether. It also informs which test materials are developed (and how), and what kinds of accommodations will be utilized to minimize barriers for some students. However, do the assessments and items tap the knowledge and skills intended? Can developers and clients be reasonably assured that students who have the requisite knowledge and skills will be able to effectively interact with the items and forms? Do scores on tests and forms within the assessment system appear to be comparable so that inferences can be defended across the entire range of test takers? Answers to these questions presume that adequate care has been taken to ensure that information about the quality of the assessments has been obtained for all test takers, including those with diverse needs.

This chapter will focus on particular technical considerations in validity for English language learners (ELLs). It will also introduce some of the challenges in demonstrating score comparability. A proper treatment of most of the topics raised here is outside the scope of this book. Instead, what the chapter is designed to accomplish is to introduce a few pressing issues and selected approaches researchers and developers have found to
be promising for inspecting how well today’s assessments might be addressing the challenges of English language learners. This includes some discussion about data design, data collection methods, and techniques about how the data might be analyzed. As readers will notice, within sections, some topics will be given more consideration than others. For the most part, decisions about topic inclusion were based on either devoting space to less developed notions that need more attention, or on attempting to correct some misunderstandings common in the literature.

Given the pressure that test developers are under to produce assessments for the breadth of students in U.S. schools, it seems to be imperative that appropriate steps are taken to improve the methods developers and others are using to demonstrate the quality of the materials and procedures which will be used to test English learners. Considerations associated with well-designed, large-sample quantitative investigations will be highlighted because the field seems to be having some trouble developing and implementing a rigorous research agenda for these students. Better inclusion of ELLs in samples during brief data collections, as well as increased emphasis on in-depth methods, is also important in order to examine how well ELLs are interpreting items and effectively demonstrating their skills. Further, those who are analyzing data from large and small sample collections need to properly sample students and disaggregate their investigations so technical rigor can be assured for smaller groups of test takers as well as majority populations. Otherwise, voices will be lost in large aggregated analyses and improper conclusions about the technical adequacy of the tests for some populations, including this one, may result. As Lachat (1999) admonishes:

All too often, states develop and field test new assessments for the general population, allowing the technical demands of test construction to postpone consideration of whether these new assessments are appropriate and fair for English language learners. Once developed, tests are then reviewed to determine whether a native-language version or some type of (administration) accommodation would facilitate the participation of English language learners. However, addressing the needs of as an afterthought makes it most difficult to develop assessments that are inclusive, valid, and reliable for this population. (p. 63)

The 1999 Standards for Educational and Psychological Testing (AERA/ APA/NCME, 1999) explains that streams of evidence should be used to evaluate the degree of valid inferences which developers and clients can reasonably expect. A validation design model should guide the types of
Selected Technical Considerations • 281

evidence which should be collected for any particular assessment, and this design is primarily defined by the purpose of the assessment system. Considerations of potential test takers should also be paramount although this is less well defined in these measurement standards. Instead of referring to different types of validity (such as content, criterion, or predictive validity), the APA/AERA/NMCE standards argue that score inferences suggest a level of construct validity confidence constrained within and defined by the evidence produced as a result of the design model. To date, technical confidence is mostly determined through careful test construction and implementation procedures, and through a typical set of post hoc data analyses. These are mostly data to inform the level of reliability in the assessment system, and documentation related to the allegiance of items to unidimensional content ability assumptions which undergird Item Response Theory methods. Certainly it is important to ensure careful development, implementation and analytic procedures, and procedures have been routinely adopted and refined. However, there has been little direction about which of these procedures is sufficient and/or necessary in order to ensure construct validity for different purposes or student groups. Lately, Mislevy and others (Mislevy, 1996, 2003; Mislevy et al., 2003a, 2003b) identified test design frameworks that, if properly constructed, should produce testing systems that can be more defensible. Still lacking is coordinated guidance about what types of validation approaches are essential and/or valuable for determining what kind of technical evidence is important to collect for English learners. Kopriva (1999, 2000) and this book are designed to provide some of this information.

The next section will outline key considerations for designing a model which guides the types of documentation required for demonstrating that inferences are reasonably valid for ELLs. The following two sections will review selected large and small sample methods and analytic approaches researchers are using to evaluate aspects of validity for this population.¹

Developing an Interpretable Research Agenda for Documenting Validity

One of the shortcomings of the current status of test research for English learners seems to be the lack of coherent research agendas. This is particularly true when an assessment system is being inspected, but it also may explain the fact that, to date, the studies are often not well-coordinated within researchers (or test companies) or between them. It appears that two components of a well-designed agenda would be, first, an articulation of the inferential arguments and the kinds of data to back up these claims or support others. Second, the agenda needs to clearly lay out some kind
of coherent structure within and across research studies and data analysis designs. In this structure, variables and their interrelationships could be consistently restrained in their definitions and methods of operationalizing, or they could be purposely allowed to vary for particular reasons. Considerations of both of these issues will be discussed next.

**Validation Design Model**

Just as test blueprints guide the design of assessment systems, the measurement standards (AERA/APA/NCME, 1999) advocate that a validation design model should guide and prioritize the empirical studies used to defend the score inferences. The models would articulate *a priori* the intended inferences that scores from the test are proposing, and the types of data that provide adequate evidence to support the intended score interpretations. It should also clearly identify what alternative inferences may be, that, if true, would weaken or refute the accuracy of the targeted inference the test is intended to sustain. Therefore, to be able to support the intended interpretation, there needs to be adequate evidence which strongly suggests that the alternative interpretations are not very viable. What the standards did not emphasize, however, is that some of the alternative explanations are subgroup-specific. This is particularly true for subgroups such as English language learners whose challenges are in direct opposition to a number of the methods used in today’s large-scale standardized academic testing. Several authors, for instance Kopriva (2000) and Abedi (2006b), have identified and explained some of the alternative interpretations that might be probable for English language learners.

There appear to be two stages to developing a satisfactory validation design model. The first stage is an explanation of the validation argument that will guide a research or data analysis design. The second stage is the design itself, which would support or refute the validation argument. To implement the model, studies and analyses need to be undertaken. Findings from the research will comprise the evidence that is needed to ensure that the validation research design model is acceptable. In cases where the findings do not support the initial models, the results should either drive constraints in the argument (for instance, the argument may be valid only for certain groups of students), or other research should be identified which satisfactorily addresses rival hypotheses so that the initial argument can be upheld. As necessary, investigations should be specified to study any new pieces of the models and produce the evidence to support the arguments. Chapter 6 in the National Research Council (NRC) document *Keeping Score for All* (Koenig and Bachman, 2004) provides a good example of what a validity argument might look like. This argument forms the beginnings of an appropriate validation design model for ELLs, and illustrations from
TABLE 12.1 Definitions

- The Targeted Inference is the general interpretation that test designers want to make on the basis of observed data about the level of the test takers’ knowledge and skills. The test score inference is a cumulative interpretation based on targeted knowledge and skills which are tested by different items.

- The data are evidence from development efforts and studies that support that the targeted knowledge and skills are being assessed.

- The warrant is the explanation of why the inference can be supported by the data.

- The warrant is based on backing, which is derived from theory, research and experience.

- Alternative hypothesis is one or more rival possibilities that might account for observed performance on the assessment where the measurement of the targeted ability is confounded with ancillary abilities.

- Each explanation provides alternative rationales about why the alternative hypotheses may be viable. To the extent that they are viable, they could seriously weaken or distort the intended inference.

- Alternative data suggests the type of evidence that could be collected to suitably address each alternative hypothesis. Adequate evidence that the alternative hypotheses have been satisfied would suggest that the target inference is defensible for students whose data would otherwise be subject to the alternative hypotheses and explanations.
their work will be summarized here. In general the NRC committee’s descriptions are based on concepts of Evidence Centered Design as conceptualized by Mislevy and others (for example see Mislevy, 1996, 2003; Mislevy et al., 2003a, 2003b). The basic model is adapted from Koenig and Bachman and shown below in Figure 12.1. The explanation of the terms in the figure immediately follows it (Table 12.1).

In the case of English language learners (as well as some other students), accessible testing materials, accommodations, or other types of testing options can be proposed to weaken the alternative hypotheses and lend credibility to supporting the test inference. Koenig and Bachman applied the validation argument to an informational reading passage from NAEP, and illustrated how a developer might address aspects of the passage and items for particular English learner students with particular challenges. The passage, items and more explanation about the process can be found in their volume. Figure 12.2 replicates what they conceived might be a set of alternative hypotheses, explanations, and alternative data for Tina, one student with a specific set of needs.

It should be noted that, in reviewing the NAEP example provided in Koenig and Bachman, this example of what needed to be addressed for Tina is probably incomplete. For instance, there are concerns that the selected passage contains language that is probably not meant to be targeted for fourth graders. Specifically, it contains particular colloquialisms, idioms and other language that students who lack sufficient experience in the U.S. may not have learned. But the example is still relevant for illustration purposes.

An important and necessary component of any effective validation design model for English language learners is that it reflect the fact that different alternative hypotheses and data are needed make the test inference relevant and defensible for different subgroups of ELL students. This is explained thoroughly in several of the earlier chapters and reflects the heterogeneity of the population and their diverse needs and strengths. Therefore, within the overall validation design model there would seem to need to be “submodels” for the different subgroups. These submodels use the same warrant and backing but focus on different alternative hypotheses for students with different sets of needs.

Of course, the question arises about who is going to select the primary submodels, prioritize the variables so that only key hypotheses and evidence are reflected, and ensure that particular kinds of evidence are satisfactory. It appears that this is really the responsibility of the test consumer, probably at the state level (as compared with the test publishers). The reason is that different states vary in the kinds of ELLs they educate, how they define who they are, and what progress looks like. The state is in the best position to know the demographics of their region, know how the state defines ELLs,
Data 2: The reading passage was appropriate for fourth grade level and informational, and the task requested Tina to draw reasonable conclusions from the text, recognize relationships such as cause and effect or similarities and differences, and connect ideas in the text with her background knowledge and experiences.

Targeted inference: Tina is a basic-level fourth grade reader.

Explanation: Tina does not have the background knowledge required to successfully complete the task.

Alternative Hypothesis 1: Tina does not have the background knowledge required to successfully complete the task.

Alternative Hypothesis 2: Tina does not have the writing ability required to successfully complete the task.

Alternative Hypothesis 3: The rater gave Tina’s response a low score because of grammatical errors and inaccurate vocabulary.

Explanation: Scorers are sometimes influenced by features of the writing, such as grammatical errors or inaccurate vocabulary when they rate written responses to open-ended NAEP assessment tasks.

Explanation: NAEP scorers are regularly trained to ignore errors in written language (e.g., grammar and vocabulary) and to focus on content when scoring students answers.
and how they track their academic progress. While areas or districts within the state may have unique situations, they can always augment the models the state develops to suit their needs. Once the submodels are completed, the data which refute or support the alternative hypotheses need to be collected. On a regular basis (every couple of years or so) the evidence that is collected to hopefully weaken the alternative explanations should be evaluated to ensure that the general test inferences continue to be robust for the range of ELLs in the state. Of note: the submodels can be applied to whatever are the current academic state assessments. This includes those systems whose test designs and contractors remain stable or ones that change over time.

Clearly, designing validation design models is a serious undertaking that takes time to properly develop. However, Mislevy et al. (2003) and Koenig and Bachman (2004) argue that unless this kind of deliberation is completed, research findings will not be able to suitably address the degree of effectiveness and validity of the score inferences, especially for populations where alternative hypotheses are relevant. It is recommended here that, initially, some of the evidence related to the alternative hypotheses for ELLs might tentatively come from viable research findings based on other tests and with out-of-state samples. It should also come from “face” validity or judgment (as evidenced by use of specific procedures and methods) about the level of accessibility obtained during item and test development of the test in question. It is essential, though, that there is some empirical data associated with the particular assessment system. Further, it is essential that as the assessment system matures the initial inferential conclusions are confirmed (or refuted) based on empirical data. If successful, this empirical evidence would support the weakening of the alternative hypotheses and directly link the improved inferences to the alternative data for particular students. Only in this way will there be adequate evidence that the intended test inferences are appropriate for those English learners for whom appropriate data are present.

The next section will summarize considerations related to completing research designs which support the validation arguments and produce the data for the alternative situations. The rest of the chapter will be devoted to summarizing select technical approaches which may be used to implement pieces of the research design for ELLs, and lend support for determining the level of confidence agencies can have in the results of their academic assessments.

**Research Design Considerations**

Once the validation design models have been constructed, consideration needs to be given to how the evidence will be collected. Coherent research designs should be formulated. The focus across studies seems to be to ensure a “diversified portfolio” of data for various arguments and alternative
explanations. That is, over time, evidence from three classes of source data—judgment evaluations, and small and large sample empirical sources—should be compiled to adequately defend the test inferences. Other work has explained how each of these sources provides unique information about how to properly determine the meaning of test scores. Because of their unique contribution, it is important to keep in mind that no one class of source data is adequate. Efforts must be made to vary the types of evidence within each argument if the interpretation is going to be reasonably supported.

If possible, it would be prudent also to include information about the students’ opportunity to learn since the impact of these data would also help interpret what the scores represent. As Chapter 3 explains, however, to date there are few models for how to collect these data on a large scale. Therefore, including this evidence on a routine basis is something that may need to occur after the field has completed more work in this area.

Both within and across studies, the conceptualization of the research work needs to include the identification of key variables. For academic test research, these seem to include the focal test scores, criterion information and salient ancillary factors. Criterion variables would be other sources of information about the students’ target ability which the studies argue are similar to the target. Ancillary characteristics are those which appear to be construct-irrelevant but which researchers believe may be impacting scores. Across studies there should be attention paid to how the factors are operationalized so that some kinds of convergent and discriminant validity inferences can be made. Further, across studies attention should also be paid to hypothesized interrelationships among variables and how these purported relationships might be confirmed.

Within well-conceptualized research studies, the design of the research studies need to be sound, and of high quality. Besides clearly specifying and operationalizing salient variables and interrelationships based on previous research and theoretical grounds, one or more appropriate research approaches need to be identified. A crucial step is to control data collections so adequate inferences are not threatened. This means limiting the scope of studies so study inferences will produce smaller but stronger inferences for a more specific practice situation, group, or use. One weakness in much of the current work is that so much of the work is conducted after tests administrations are conducted. Several compromises associated with either omitting certain variables because data are not available, or loosening the control of how the data were operationalized leads to muddy inferences of questionable and indeterminate accuracy. The What Works Clearinghouse at the U.S. Department of Education (2007) has published standards they use to review educational studies for inclusion in their data base. To meet the evidence standards (one of the three stages of review that are necessary for studies to pass), a study
must involve a randomized controlled trial or a quasi-experimental
approach that uses equating, a regression discontinuity design, or a single-
case design. These are explained more fully below, and in the U.S.
Department of Education documentation. The rest of this section will summarize key quantitative and qualitative
or informal research approaches associated with analyzing data which are
commonly used or may be promising.

Selected Quantitative Approaches

Many technical questions relevant to determining the validity of scores for
English language learners correctly require quantitative analyses. These
techniques are useful for identifying and confirming relationships and
treatment effectiveness. They are also useful for systematically exploring
how complex associations of student and testing factors influence test
performance. This section will address some of these types of questions
and approaches that seem to capture the methods most commonly used
and/or which appear to have potential for evaluating data associated with
ELLs. It should be noted that the two sub-sections which follow are divided
in a rather artificial manner. That is, these sub-sections (Comparisons of
Groups and Treatments and Modeling Relationships of Variables) differ
primarily in the focus of the study questions and hypotheses. In many cases
they share the same analytic techniques.

Some researchers might find it curious that several pages are allotted to
quantitative approaches to modeling relationships when recently there has
been a push to complete experimental studies in education. This author
agrees that the agencies which fund educational research are correct in
demanding rigorous studies, and especially ones which investigate differ-
ences using experimental methods. There is a problem with exclusively
conducting well-designed treatment and group comparison studies,
though. That is, the field does not currently know enough about the subtle
interrelationships among variables which influence test performance. This
dilemma is not unique within educational research. As investigations move
from more general to more contextual studies throughout education,
researchers are still learning about the nuanced variables that impact
student learning and performance. For ELL research work, the resulting
ignorance seems to have had a major impact on making good decisions
about causal study designs, judging by the inconsistent (although well-
meaning) sets of quantitative studies completed to date. The point is, it
appears that in order to improve our measurement of ELLs, there needs to
be a much better understanding of the phenomena so that measurement
models can use appropriate variables to properly infer their target abilities.
This state of affairs is contrasted to the movement in recent years to include ELLs in general academic testing in general and in more qualitative and informal research in particular (although ELL participants have almost always been incompletely sampled in these data collections). Therefore, it is argued that the focus of research should now be more often on quantitative studies. The work needs to be implemented within the framework of thoughtfully designed conceptualizations, however, which systematically and quantitatively (if possible) elucidate the complex relationships. Subsequently, research should rigorously examine the effect of findings from the modeling studies for contrasting groups from the larger population and/or under varying treatment conditions. Since experimental methods are more straightforward in implementation and are better known, much of this section will address how to encourage well-considered investigations of relationships so causal implications can then be studied in an effective manner.

Currently, those developing newer large-scale academic systems appear to conduct few if any quantitative studies geared to establishing validity for ELLs. If they do, the most common are analyses of differential item functioning which can be conceptualized as reliability or validity. In either case, this analysis is often problematic for ELLs. It is correct that some of the analytic approaches which will be discussed here are those which are more relevant for researchers and for states or publishers with an ongoing validation agenda of established assessments. However, it is also advised that some quantitative studies always be done in order to validate the inferences, and in order to determine how valid the test score interpretations are for English learners and for those who receive accommodations. In particular, it is proposed that analysis of the test’s dimensionality for these groups be considered among the first large sample studies systems should undertake. It will also be important to compare results of the validation studies which focus on English learners with those from the mainstream population to provide support for utilizing the same inferences.

Modeling Relationships of Variables: Selected Methods

In this section, selected types of modeling methodologies will be briefly explained. Usually, the research questions that are addressed by methods such as these focus on exploring or confirming why targeted students might perform as they do on the academic tests or items. Within the domain of possible kinds of methodologies, the purpose the particular approaches highlighted is that they seem well suited to addressing research questions that ask about how test scores, criterion measures, and ancillary variables associated with tests, ELL students, or their schooling environments interact.
with each other. Many methods, such as growth modeling, will not be addressed here. Further, hierarchical designs, while considered to be essential within the school climates will not be specifically discussed. Instead, readers are encouraged to consider the following arguments and approaches and apply them to hierarchical designs and other related types of nested and combined methods models.

For the most part the focus of the approaches is to estimate interactions. Interactions, statistically, refer to the phenomena where relationships among one set of variables differ depending on the values taken by the other variables. For example, if structural relationships among variables vary between two categories of persons, then there is an interaction between the variable defining the person categories and the other variables.

This section is included because of the lack of understanding about how various ELL student variables interrelate with each other and interact with aspects of tests. Lawmakers are moving ahead with implementing legislation that demands more inclusive tests, and researchers recognize that cognitive and socio-cultural models related to test performance are less well understood for some students (including ELLs) than for others. To explore these questions, the section will be divided into two general, arbitrary artificial, parts. The first will review some typical modeling methodologies which answer more restrained questions and where the independent and dependent variables are usually some combination of continuous and categorical. The second will summarize a couple of methods which analyze complex sets of categorical factors variables.

**MODELING DATA FROM CONTINUOUS AND CATEGORIES VARIABLES**

Many types of modeling procedures have been developed over more than 100 years. Crocker and Algina (1986), among others, provide a basic explanation of this type of modeling, and Stapleton (1995) and Hair et al. (2006) outline many of the procedures.

**Classical Linear Models** Classical observed score approaches include correlations (including simple and multiple correlations) and regressions. A seminal reference for traditional or classical linear modeling is Rao (1965). Regression analyses can involve simple linear relationships or they can utilize polynomials and model curvilinear relationships. In many cases the dependent factor is continuous (or treated as such) as are the independent variables even though they may be basically categorical and even dichotomous. This class of models includes both simple (utilizing one dependent variable) and multivariate regression. Multivariate regression is not to be confused with multiple regression; rather, the methods handle
more than one dependent variable. When the independent factors are categorical, classical regression is actually consistent with the Analysis of Variance (ANOVA) which has been commonly adopted as a method of group comparison.

Logistic regression has been developed to handle dichotomous dependent variables by transforming the probability of correct response into the logit metric. This method is often used to explain factors which might influence dichotomous item scores. It is sometimes also used with constructed response items by decomposing results into dichotomous contrasts. For instance, the regression would be computed for those with a score of “1” vs. all others or another particular level. Generally, because categorical dependent variables are treated as such, statisticians consider that the results from logistic regression are more stable than if traditional regression procedures which assume continuous variables were used.

Some studies that include English language learners or students whose first language is not English have been done utilizing these methods. Initially the field investigated very simple models that had only limited usefulness. With the technological advances and increased interest in inspecting construct irrelevant influences, more complex models can be fit which may yield more informative findings. For instance, Finch et al. (2006), Emick and Kopriva (2007) and Kopriva et al. (2007d) have used both test and item score regression methods to investigate targeted and non-targeted effects for these populations.

Structural Equation Models A substantial class of analyses has become popular over the last 30 years or so to model and estimate relationships between latent variables that underlie manifest performance. For the most part these have generally fallen under the rubric known as structural equation models (SEM), and many volumes, including Bollen (1989), explain the basic methodology and many of the analytic procedures associated with this approach. Essentially, the methods specify latent variables and model their relationships in latent space. Research questions are usually similar to those addressed by the types of models just explained, and they are answered by investigating the relationships between observed phenomena associated with the latent variables. However, the structure of the analyses differ from the class of models just explained in that each latent variable is not directly estimated, but is estimated by using observed data. In the classic regression models observed relationships are modeled and modeling error is a function of unexplained variance that has not been accounted in the dependent variable. In this approach, besides documenting modeling error, errors of measurement associated with estimating each latent variable with the observed data are also part of the
models. The advantage of the SEM models is that the results are considered to be more precise because they recognize and attempt to account for errors in the independent variables as well as error in the dependent variable.

For modeling questions associated with achievement of English language learners, one set of structural models, latent class analysis, have been used to investigate the number of achievement levels a test might warrant for ELLs (see Kopriva et al., 2007c, for a brief explanation), and these models has been used to interpret differential item functioning relationships in English language proficiency data (Samuelsen, 2005). Latent class analysis estimates the number and quality of groupings that appear to underlie the observed data. Latent trait analyses on the other hand, assumes the underlying distributions are continuous whereas latent class analysis assumes the distinctions among latent variables are discrete. The most common example of a type of latent trait analysis is the set of analytic procedures associated with Item Response Theory (IRT). Item Response Theory, at its most basic, assumes the underlying latent construct of student mastery of a particular content is continuous and usually unidimensional. Students’ performance on tests and items can be modeled in terms of this underlying construct of achievement. Most current testing systems use IRT to guide development and interpret student scores. Kopriva et al. (2004a) used structural equation modeling to operationalize the Multi-trait Multimethod (MTMM) technique. In this study they investigated the relationships among a new English language proficiency test and older tests using multiple manifest indicators of latent English language proficiency traits. Sireci and Khaliq (2002b) utilized structural equation modeling as well as other methods to examine the structure of a dual language form as compared with the standard English form.

**Factor Analysis**

Today, the procedures of factor analysis are often operationalized as one class of structural equation modeling procedures. The genesis of factor analysis is somewhat distinct, however, and so it is addressed separately. Factor analytic theory and procedures associated with it were first conceptualized by Louis Thurstone in the 1930s (cited in Thurstone, 1947). Essentially, factor analysis analyzes the covariation among observed variables in order to detect underlying structures which account for the observed variation and covariation. It was originally developed to analyze the relationships among observed scores from multiple tests that measure the same and different phenomena. It was assumed that, while all tests to some degree measured different things, there were some common elements that could be identified among them. Lawley (1940) applied maximum likelihood methods to factor analysis concepts and at that time first
statistically differentiated latent underlying constructs from manifest test score data. From that point on, the procedures associated with factor analysis have generally differed from Thurstone’s initial developments but the conceptual framework remains intact.

One key constraint of many of the factor analytic procedures is that they assume the distributions of the phenomena being studied are continuous. That is why the focus of the analyses was initially on test scores rather than on item level data. About thirty-five years ago researchers began to consider the application of factor analytic techniques to the binary level data of dichotomous items. In 1978, Muthen published a paper which explained how dichotomized variables might be investigated in factor analysis. As methodology was developed to handle these data (e.g., see Bock and Aiken, 1981; Bock et al., 1988; and Muthen and Muthen, 2000) many researchers have used this approach to investigate the dimensionality of tests. In other words, factor analytic methods have been used to take item level scores in tests and sort them according to underlying structures that may be observed in the data. Like many of the techniques summarized in this section, there continues to be debate about the best way to handle these types of binary data (for instance, Mislevy, 1986).

Most academic assessments today assume that the tests are measuring a unidimensional trait, for instance degree of science mastery. A few studies have used factor analytic methodologies to investigate if this assumption holds in academic tests for students who receive and do not receive accommodations, or for students with disabilities or English language learners as compared with those without disabilities or native English speakers (respectively). For instance, Huynh and others (Huynh et al., 2004; Huynh and Barton, 2006) focused on students with disabilities and looked at both of these questions. Tippets and Michaels (1997) investigated the accommodations question for students with disabilities while Cook et al. (2006) did the same for ELLs, and Fitzpatrick et al. (2006) studied the structures of English language proficiency tests for ELLs. While these studies essentially found a similar-enough structure of the scores for the focal group(s) (e.g., English learners with and without accommodations, students with disabilities, or students who had been accommodated) as compared with the reference group (usually native English speakers, not disabled, or those who didn’t need accommodations) it was clear that items were performing somewhat differently between groups. Other factors that were identified in these studies were generally not interpretable, however, at least in terms of other targeted constructs. Little has been reported to determine the extent to which the factors might be identifying construct-irrelevant influences.
Other Methodologies for Modeling Dimensionality

Verifying that the same dimensions or underlying targeted constructs are being measured for different students is important because comparability of large-scale test results needs to be established for score interpretations to extend beyond the mainstream population to diverse types of students. This is particularly the case when varying testing methods are being used, and as such, comparability will be discussed further at the end of this chapter. Further, it is important to be able to detect other salient dimensions if indeed they do exist, especially dimensions that reflect systematic construct-irrelevant constraints of some groups. However, other methodologies besides those discussed above may need to be used. As the studies just mentioned have suggested, other latent dimensions appear to be a possibility, and assuming the same construct dimension without investigating if this is appropriate seems to be irresponsible. In this section, a few additional methods will be summarized.

Hattie (1985) identified several techniques for identifying dimensional structure in the test data. In linear factor analysis, a linear relationship between factor loadings and factors is assumed. In order to detect dimensions that these techniques may not identify well, some researchers have suggested binary data SEM methods which are sensitive to non-linearity and multidimensional space. Some of these are non-linear procedures, operate directly on item scores rather than correlation or covariance matrices, and are akin to multidimensional Item Response Theory techniques (e.g., Bock and Aiken, 1981; Bock et al., 1988; Fraser and McDonald, 1988). TESTFACT and NOHARM, respectively, are two programs which utilize the procedures of these authors and others. Conceptually, these techniques derive from well-known factor analytic models and assume an underlying parametric IRT structure of the test data (Gorsuch, 1983). However, they use different parameter estimation methods.

Barton and Finch (2004) utilized a program called DETECT (Zhang and Stout, 1999a, 1999b) to investigate if and how non-targeted dimensions appeared to impact the scores of students without disabilities and with disabilities (both with and without accommodations). This method does not assume any particular model as having generated the test scores; rather, it is built as a type of cluster analysis technique estimating latent variables from the data. Covariances are estimated using a contingency table approach that assumes no particular parametric of item response function and condition estimates on the student’s total score. The method reports the number of clusters and provides results indicating the degree of approximate simple structure for the entire exam. Simple structure is the degree that a typical item is influenced by only a small number of latent traits. Sireci (2005, 2007) has used multidimensional scaling techniques to evaluate language
variations of tests. The methodology he used in both investigations studies test dimensions using a distance metric. Like Barton and Finch (2004), this method examined the relationships among items and how they cluster together for different groups of students. Robin et al. (2003) used DIF and multidimensional scaling to investigate the item and structural consistency across translated and English forms for small samples of examinees ($n = 100, 173, 180$). The researchers found that these procedures appear to be suitable even for samples of this size.

**Modeling Frequency Data from Categorical Variables**

This section extends the type of modeling procedures discussed above, modeling frequency data from categorical variables or variables that are treated as categorical. Two approaches will be mentioned here.

*Loglinear Models* Traditionally, contingency tables were analyzed by calculating chi-square statistics. This has been largely replaced by loglinear techniques which model the relationships among multiple categorical variables. Essentially, loglinear techniques are omnibus methods that model cell probabilities over variable levels. Variables are not usually defined as independent or dependent and the questions are focused on how levels of variables interact with each other to explain the distribution of persons over cells. In the case where variables are explicitly defined as dependent and independent, logistic regression, explained above, is a better technique. Bishop et al. (1975) is an older resource volume that builds on the work of L.A. Goodman and thoroughly explains a wide range of methods, including loglinear. More contemporary texts, such as Agresti (2002), also explain many of the procedures.

Loglinear methods model observed probabilities. While statisticians use different methods, essentially they are modeling the logs of the cell frequencies in terms of the contrast between the marginals for each variable. The models also allow one to investigate whether the cell probabilities vary non-uniformly across the variables. Significance tests using likelihood ratio procedures estimate the probability of a significant difference for hypothesized contrasts, and full or restricted models can be fitted. For example, in investigating how student variables affect correct item responses of ELL students, a $3 \times 2$ contingency table may be constructed for each item which indicates the numbers of students in each cell. The two variables could be student factors: home language literacy and a judgment about testwiseness skills. Three levels of home language literacy could be identified (low, medium and high) and two levels of testwiseness skills (they have them or they don’t). In this example a contrast between marginals would
answer the question “Over home literacy levels, do students with low testwiseness skills get items correct at about the same rate as students who have testwiseness skills?” The interaction data answers questions such as “In investigating the impact of testwiseness skills, does percentage correct vary or not vary in the same way for students with low language literacy as it does for those with higher levels of home language literacy skills?” Kopriva et al. (2007d) used loglinear analyses to investigate how individual access-based changes in items impacted ELLs, poor readers, exited ELLs, and native speakers. Analysis at this grain size is one of the greatest advantages of these methods.

Bayesian Networks  The set of methods known as Bayes networks or nets is one application of a large class of Bayes procedures which applies to multivariable and usually large probabilistic structures. Each structural element (analogous to a cell on a contingency table) is composed of the frequency of response data that occur for a specific combination of variable levels. Like loglinear models, levels of categorical variables interact with each other to form the structural elements in a Bayesian network. Results are conditional probabilities of data occurring in a particular cell. Observed data are conditioned by a prior probability distribution estimate based on other research or theory in order to provide the best fitting posterior distributions for targeted probability parameters. This method answers such queries as “If students have a high reading proficiency and have been judged by their teachers as having a high level of knowledge and skills about an item’s targeted content, what is the distribution of the probability of these students responding correctly to a particular item?” If each of these variables (reading level and content knowledge level) has three levels (a 3 x 3 model), the probability results will be specified for each of the nine conditions. Gelman et al. (2003) is a good reference which explains Bayes nets and the statistics behind them.

The Bayes approach relies on conditional estimates of the probability distribution occurring for each element in the structure. This is because it is believed that these data can increase the precision of the probability estimates over and above what might be calculated using only the observed data. Prior distribution estimates influence posterior parameter results more precisely if they are similar to the observed data. Therefore, inappropriate priors would result in less focused posterior parameter distributions than would priors which are more proximate. Clearly, as the final parameter estimates are always informed to some extent by the prior distribution information, if prior data are misleading, this would lead to skewed results.

A model such as this could be used to analyze data from a complex research study which investigates if students with different needs who receive different accommodations, tailored and not tailored to the needs
of individual students, respond on items as criterion information about the students’ abilities suggests they should. For this question, each parameter estimate can be aggregated up to determine effects of particular student or accommodation variables or more and less potent combination of different variable levels. For instance, a parameter estimate at one point may find that a correct item response has a high probability of occurring given that the criterion suggests these students have the targeted knowledge and skills and the accommodation package matches the students’ needs. Over additional variables (e.g., other items and other accommodation choices), other aggregated estimates may suggest that the same accommodation package works best over many of the items but not for all students with the same set of needs.

Recently, two studies were completed which utilized Bayes nets methodology in modeling the impact of complex sets of variables for English learner students. Koran (2005) examined the potential for using this type of probabilistic structure in matching test accommodations to individual ELL students based on their specific needs and strengths. Using two data sets related to this work, she illustrated methods for establishing and updating conditional probability values within this approach and discussed how researchers and developers might apply the work as additional accommodation research is completed. Wei et al. (2005) utilized this methodology in investigating the interaction between ELL student and item characteristics, using extensive cognitive lab data from eighty-four English learners. While this work needs to be extended to large-scale data, it provides an example of how this approach might address the analyses of these issues.

Comparison of Groups and Treatments: Selected Methods

This section will be a brief overview of some approaches which could be proposed in examining the impact of accommodations or other testing methods on the test performance of English learners. Differential item functioning (DIF), a popular item level analytic technique designed to flag performance differences between groups once student ability is held constant, will be discussed in more detail. The section will be divided into four sets of data comparisons which could occur between groups or between treatment methods: evaluation of differences in test score data, item level data (including both DIF and other item level methods), and evaluations of differences in model structures. Considerations of effect sizes and causality would often be the outcome of the work when the research has been well designed and executed.

Comparison of Test Score Level Data

In the English learner testing literature, the most common comparisons focus on evaluating test score data for different groups (for instance, ELLs,
non-ELLs) or under different treatment conditions (most often different accommodations). Pennock-Roman and Rivera’s (2006) meta-analysis of experimental studies from 1997 to 2004 provide cites of some of this work. The research designs sometimes involve repeated measures, equivalent groups, or comparisons of self-existing groups. Indices of central tendency (e.g., means), standard errors, and sometimes other distributional characteristics are most often compared and techniques such as t-tests, Chi-Square tests and ANOVAs (including use of covariates and multiple dependent variables) are usually employed. These methods are all widely known and won’t be discussed here.

Two often heard criticisms related to the work that has been done to date center around (1) the lack of well-controlled studies, and (2) the inability to easily generalize findings to the population. Problems tend to be because often self-existing groups are not equivalent, and because the interactional quality between student factors and testing conditions are usually not considered. Much of the literature is composed of quasi-experimental studies or experimental studies which do not take into account key interactional characteristics. As such, mixed results could reflect faulty research designs as much as they reflect genuine findings about student performance. As the other chapters repeatedly encourage, proper comparisons are key to making inferences and to obtaining effect sizes which are trustworthy and generalizable. This suggests the need for well designed studies which thoughtfully employ well known design techniques that produce equivalent groups through student randomization or other means. It also suggests that student and testing characteristics, including interactional effects, should be controlled, constrained, minimized, or measured, and that outcomes should clearly and properly interpret findings within the constraints of uncontrolled influences.

DIFFERENTIAL ITEM FUNCTIONING

The most common method for determining if a test’s items (and therefore the test) are behaving in a similar fashion for different groups of students are techniques associated with differential item functioning (DIF). Typically this method has been used to evaluate item viability for groups sorted by gender and racial-ethnic membership, and lately, developers and researchers have been using it as one way to evaluate item quality for ELLs. They have also used this technique in some investigations of accommodation effectiveness. The big advantage of DIF is that this approach looks at functioning at the item level rather than at an aggregate test score level where significant differential data may be masked. Addressing validity at the item level is an important contribution to demonstrating the validity of test score inferences. However, the method has some serious problems
for reliably assessing the validity of items for ELLs and this will be discussed below. First, an explanation of DIF will be summarized so readers can understand how this methodology works.

**What is DIF?** In 1984, the court settlement between the Golden Rule Insurance Company and the Illinois Department of Insurance/Educational Testing Service specified that a raw difference of 0.15 or more in an item’s $p$-values (level of difficulty of an item), favoring white over African American applicants, was evidence that the item is biased. The case focused on inclusion of items in the Illinois insurance licensing examinations. The agreement determined that biased items, as defined by the above criteria, should not normally be included in the test.

While this judgment has some initial appeal, there was concern that the level of academic mastery of students or populations, and a test’s ability to accurately measure this mastery, were confounded. In other words, what if the test was measuring ability accurately, but overall mastery by group was indeed different? Those two conclusions need to be differentiated. A sounder distinction lies in the identification of items where population subgroups respond differently from each other within comparable ability levels. Differential item functioning methods essentially seek to hold the academic ability constant while estimating if students from various groups are responding differently to particular items.

Clearly, differential item functioning may or may not occur because of characteristics associated with a particular group. It may also highlight an unequal opportunity to learn the particular concept being evaluated, in other words, an unequal educational access issue. If differences in opportunity can be ruled out or otherwise addressed in the inferences, differences may still be due to non-target characteristics in items that are differentially impacting students. However, differences may also signal multidimensionality in the construct wherein groups behave differently on different construct-central dimensions (a valid explanation, but one which needs to be understood and clarified by test developers once items are flagged). For instance Shepard (1982) emphasized that “the assumption of unidimensionality underlying all of the (DIF) methods is not merely a statistical prerequisite, but it is central to the way in which item bias is defined” (p. 25).

There are two steps in completing differential item functioning analyses. The first step is to apply a statistical procedure to the data. The second step is judgmental and involves determining why statistical flags may have occurred with some items. Statistical methodologies have been identified which can handle both forced-choice and constructed response data although DIF techniques are most often used with forced-choice response
data. In each case the results are an estimate of differential item functioning between a focal group (e.g., ELLs or women), and a reference group (e.g., non-ELLs or men). In general, the higher the DIF result is per item, the more unequal is the functioning between groups for that item. Using what have become rather standardized methods of flags, items are identified as having low, medium or large amounts of DIF. Most of the methods assume unidimensional ability or similar distributions, most often normal, within the populations being analyzed. Some recent procedures loosen these constraints.

The most commonly used statistical approaches are the Mantel Haenzel (Holland and Thayer, 1988), Rasch or other Item Response Theory methodologies (see Holland and Wainer, 1993, for a summary), and logistic regression (Swaminathan and Rogers, 1990). Most require rather large sample sizes and most use observed total scores as the conditioning variable. SIBTEST (Simultaneous Item Bias TEST, Shealy and Stout, 1993) is a multidimensional approach which detects crossing of item response functions, and uses an estimate of latent true scores as the conditioning variable. There is some evidence that it can handle smaller group sizes although power is influenced by unequal sample sizes, whether these sample sizes tend to be large or small (Narayanan and Swaminathan, 1994; Barton and Finch, 2004). Roussos and Stout (1996) discuss DBF—differential bundle functioning—where corresponding bundles of items detect more modest amounts of DIF that, together, would build upon one another in most test score aggregation schemes. Besides Holland and Wainer (1993), Zumbo (1999) is another book which explains many of the differential item functioning methodologies.

The judgmental step seeks to rule out differences due to multidimensional influences that are target-central (if the method does not handle more than one dimension up front). Target-central is defined as influences which are part of the targeted knowledge or skills and which are, therefore, part of what an item should be measuring. Once this is completed the remaining items are examined in order to determine what might be the source of the construct-irrelevant differences. Linn (1993) suggests that the judgmental step is where much of the uncertainty lies about how well DIF findings actually determine how to identify problems and make substantive changes to offending items. He reflected that:

Far fewer general principles about test construction have been derived as the result of DIF analyses than most researchers expected. The majority of items with large DIF values seem to defy explanation of the kind that can lead to sounder test development.
practice. More often than not, judges have a rather poor record of predicting which items will or will not be flagged. (p. 358)

In practice, problematic items are sometimes simply omitted but there is routinely little notion as to why. Too often, even when problematic items are sent to review groups to consider, the knowledge of issues associated with ELLs is sufficiently lacking because the membership in many of these groups is too constrained or members are not trained to detect this type of issue (see Chapter 7 for more information). In any case, even though there is broad use of this procedure it has not brought about the kinds of understanding about what DIF is actually flagging, or how to better address item construction or otherwise “fix” problematic items.

In reviewing a large set of science NAEP items, Kopriva and Lara (1997) noted that, even after DIF analyses were conducted, there were a significant number of the items that appeared to have problems for English learners. This led to a series of discussions about how and why this might occur. The next section will summarize the primary issue and consider if and under what circumstances the DIF procedure might be useful for evaluating validity for this population.

The Issue of Pervasive Influences Because of how the DIF statistics are currently computed, an essential issue for English learners (and many students with disabilities) is that all of the methods inaccurately assume that items not flagged do not contain bias. Rather, approaches assume that the scores from the test (and its administration and response conditions) from which the questionable items are flagged are generally valid and fair for both the reference and the focal groups. This assumption allows the DIF methodology to use the students’ total score as the conditioning variable. For example, if a subset of ELL students can’t read a science test in English (or have little oral English ability if it is presented in this format), then their total score would be close to 0 or would tend to reflect chance as they randomly responded over multiple choice items. This would result in two conclusions, both of which are faulty. The first is that the scores would suggest the students have low ability in the tested subject area. Items then would be flagged only if they were unusual for low ability or unmotivated non-ELLs (or whoever the reference group might be). The second conclusion is that only the flagged items are problematic for the ELLs. In truth, however, most or all of the items are inappropriate as they are not measuring the targeted content. The bottom line is that this method can be compared to the wrong reference group, and it does not presently detect pervasive problems in items—invalidity which occurs over many
items. The flags only pick up certain types of problems unique to individual items and relative to the improper reference group. This is helpful but, meanwhile, the much larger issues are not being identified by the methodology.

For many ELLs, it is primarily their English language limitations that impact their ability to demonstrate what they know throughout the test. For low English proficient ELLs, oral and literacy requirements are also often confounded with cultural considerations and home language conventions. Further, as Abedi (2004) reported in his review of several mathematics and science NAEP exams, this problem persists for even more accomplished English speakers. His findings showed that language minority students who, as a group, had a relatively broad literacy base before they took the tests (they were advanced ELLs, exited, or were bilingual and never needed special ELL services offered as part of USED Title III funding) still seemed to differentially respond to more items because of the language rather than because of what was being measured (as compared with the non language minority reference group). As previous chapters have explained, issues such as standard U.S. conventions of testing, amount of text, complex sentence and language structures, format complexity or standard administration constraints often impact many of the items. Students may also have limited access to visuals, appropriate language aids, or other compensatory tools which could offset these problems. Finally, because the test may be inaccessible, it is not unusual for some students to simply shut down because of fatigue, disgust, and/or confusion (any of which would have a persistent impact on the items as well).

Several authors (for instance, see Holland and Wainer, 1993) explain the pros and cons of using the students’ test score as the conditioning variable to detect differential functioning of items. Unfortunately, to date, the field has not found another criterion that might be viable. To the extent that the total score (observed or otherwise) is generally a useful criterion, then it is a reasonable estimate to use to evaluate if ELLs are responding similarly to non-ELLs, or if subgroups of these students are responding differentially. However, unless alleviation of the pervasive problems have been otherwise demonstrated for all kinds of students taking the test, given this methodology, there appears to be little way of knowing which items are still problematic. Some have suggested adding one or more additional reference groups, say of a comparable set of non-accommodated ELL students (if the focal group is accommodated in ways that meet their needs) or of ELL students who don’t need accommodations. However, these won’t help because pervasive issues could lead to masked detection for any of these additional groups as well. Further, reference
groups made up of students who don’t need the accommodations used by a focal group could differ in other important ways. As such, they are not a useful comparison.

Recently, several researchers have tried using DIF to estimate the effectiveness of particular accommodations (for instance, Finch et al., 2006; Laitusis et al., 2004; Cahalan-Laitusis et al., 2004). In reporting her DIF findings related to accommodated students with disabilities, Bolt (2004) reasoned that “if accommodations are serving their intended purpose, the measurement characteristics of accommodated test administrations for students with disabilities should be similar to those for non-accommodated students without disabilities.” This makes sense, but, for the most part, if there are still problems with any persistent influences, the same considerations apply. Likewise, if similar DIF levels apply across the focal accommodated ELL group and non-ELLS, this may be because the accommodation method is working or it could be because it isn’t. If more DIF items are identified in and for the focal group, this might signal that the accommodation may be favoring this group vs. the reference group. This may be useful evidence but it may also mean that, while it is favoring the focal group for a relatively small number of items, it may be operating as intended for the majority. If that is the case, this becomes a judgment call but not necessarily a clear signal that the accommodation is not appropriate.

On the other hand, if there were more differential functioning flagged for the reference group, this might mean the accommodation isn’t fully removing the construct-irrelevant influences in particular items. On the other hand, it may reflect that the accommodation is generally working for many items and so it is highlighting items that are not responding properly and are indeed problematic. Either way, it appears that the alternate interpretations make this methodology less than ideal for this purpose.

There just does not seem to be a good way to make use of this approach as it is currently conceived. The methods appear to work for some focal groups where the literature is clear that persistent test-related influences do not systematically tend to occur. For other groups, it seems that for DIF to be useful, methodologists need to address the points raised above. In particular, future improvements in DIF should probably try to improve and probably broaden the conditioning information. For English language learners it would seem that this should include more relevant information about the student, like other ability information about the content area being tested, and language proficiency data on a test that measures the range of English literacy and oral skills, including academic proficiency. For lower English proficient ELLs the language information should be combined with information about their culture relative to the way items and tests are compiled, and about their L1 skills. This student index may be somewhat
similar to the STELLA index discussed in Chapter 10. Likewise, more information about the assessment in question and different weighting of test data relative to thresholds in student information should help condition the criterion and make differential functioning more sensitive to a greater range of item issues. Knowing the dangers of using a total score as the conditioning variable, Thissen et al. (2003) suggested an item anchor technique when estimating and comparing distractor curves for items across groups. This technique, specified in the Multilog program manuals, uses a modified IRT approach. Maria Martiniello has also argued for using particular items with known characteristics among the focal and reference groups to anchor the DIF procedures (2006).

OTHER METHODS FOR ANALYZING ITEMS

Before selecting a final set of items for tests, developers have typically evaluated item performance by conducting point biserial correlations (or biserial correlations) on total population scores in order to determine if individual items correlate properly with the total scores. Fitting ICC curve parameters is completed by item as well and some researchers have completed either of these analyses by population subgroup, for instance ELLs. Cahalan-Laitusis et al. (2004) classified items by their aspects (e.g., difficult grammar, artwork necessary, extraneous information) and evaluated if a DIF statistic flagged items with certain characteristics for different disability groups. Similarly, Laitusis et al. (2004) placed items in different bundles by verbal load, complexity, item type and other characteristics, and evaluated the differential bundle functioning (DBF) using procedures that are a variation of those used in DIF. A recently funded project (Boals and Cranley, 2006) has proposed utilizing a technique for decomposing target and irrelevant influences in items and estimating revised target difficulty levels from both standard items and those built to minimize the language and cultural impact for low English proficient learners. If the method is found to be effective it may be a useful way of determining the target fit of items when variance due to irrelevant influences is controlled. Other work is continuing as well. For instance, Huff and Goren’s symposium at the 2006 annual meeting of the National Council of Measurement in Education focused on innovations in modeling item difficulty. While the focus of the papers was on using processing abilities of students to inform or explain item difficulty (e.g., see Embretson, 2006), these types of modeling techniques may be useful for disentangling when and to what extent ELL student performance is a function of non-targeted as well as targeted abilities.

In a project that has just been completed (Kopriva and Cameron, 2007), researchers conducted distractor analyses on multiple choice items for
different subgroups of students, including English learners and others with language difficulties (such as learning disabled students and those with hearing impairments). Abedi (2006c) also recently used distractor analyses to examine items of students with disabilities. This may be a promising approach as well.

To supplement the item analyses conducted on items being considered for a new English language proficiency test, a latent class analysis procedure estimated the proportion correct on each item in every latent class, hypothesizing that the proportion correct should increase with latent class. The ability of items to discriminate among the ordered classes was also calculated, with particular interest in differences of proportion correct between adjacent classes (Kopriva et al., 2004c). If applied to academic tests, analyses such as these could be performed for different subgroups of students.

Several researchers, including Cahalan-Laitusis et al. (2004), Laitusis et al. (2004), and Kopriva et al. (2007c) have evaluated items by inspecting the impact of their characteristics. The first two authors inspected non-targeted item aspects considered to influence the score performance of some students; the last authors evaluated the construct complexity of items and compared them with ordered achievement levels for ELLs. Once the field understands what statistical methods might be valuable for highlighting items for English learners, incorporation of these kinds of variables should help facilitate what parts of the items are troublesome or not.

**Comparison of Model Structures**

As the complexity of the target and non-target student/test interactions becomes known, comparing model structures would seem to be an important approach for determining the effectiveness of treatments or the impact of test condition phenomena. Section 1 above discusses the various ways relationships might be modeled. Most of these methods could be employed for equivalent groups and results could be compared at different points within the models.

Recently, some researchers have begun looking at conducting this type of analysis in conjunction with issues relevant to English language learners. Most of the work to date is focused on the item level, and investigates how items within tests interrelate with each other and possibly with other variables. For instance, Finch et al. (2006) completed logistic regressions on items flagged by DIF methods to study if any of the accommodations or total score of the non-DIF items were predictors for students with and without disabilities. Kopriva et al. (2007d) computed logistic regressions to measure the impact of item factors for native English speakers and ELLs after an estimate of their target abilities were regressed on test scores. Solano-Flores and Li (2006) used generalizability methods to investigate
the influence of language and dialect on test performance for students from
different languages, dialects, and speech communities.

To investigate whether convergent and discriminant validity differed for
students of different backgrounds and school experiences, researchers of
ELDA, the new language proficiency test, performed a number of distinct
multitrait-multimethod analyses for different subgroups of students. For
each subgroup, tests of significance determined whether the covariances of
sub-test scores on the latent trait were significantly greater than zero,
or whether the loadings were significantly different across groups (Kopriva
et al., 2004c). Cook et al. (2006) conducted a series of repeated factor
analyses for ELLs and non-ELLs to evaluate the levels of factors and how
the items loaded for each group. They used matrices of tetrachoric
correlations, and included a factor analysis to fit the asymptotic matrix for
the total group in order to confirm the equivalence of the groups. Finally,
as noted above, both Sireci (2005, 2007) and Barton and Finch (2004)
compared multidimensional scaling results across groups.

In each of these cases, the comparisons provide some information about
how students with certain characteristics perform on items or tests and why.
One main challenge in utilizing these methods is to be able to design the
research in order to target key influences and make generalized conclusions
about elements of the analyses. This type of design is clearly more complex,
and often involves the characteristics of tests and students which are
randomized in regular methods. Here, however, the factors and the
interactions between subjects and test characteristics are more often
measured, in such a way that the magnitude of these influences is quantified.
Another challenge for comparing ELLs and other small sample groups is
the issue of statistical power, of course. Some researchers are collecting
data over years while some are combining grades when that design is
viable for a particular purpose. Finally, issues of multi-collinearity and
associated concerns point to the effectiveness of using nested models, such
as hierarchical modeling, to extract problematic error and increase the
robustness of the findings.

Qualitative Approaches
The first section will quickly address some considerations about a couple
qualitative approaches which can yield informal, but useful, data to inform
how English learners interact with items or other aspects of tests. The
second section will outline some recommendations for ELLs around more
in-depth formal data collections such as cognitive labs. These issues will be
only superficially covered, even though they are a vital part of obtaining
data for English language learners. In both cases, the data which are
obtained may include some types of rudimentary quantitative data or information which can be quantitatively coded. Most often, the information is more “subjective” in nature, and analyses tend to be judgment-based.

**Tryouts and Pilots**

Tryouts are defined here as preliminary data collections of item responses and/or the impact of other testing components (such as directions, or accommodation options) where teachers (and their students) are convenience sampled, or the schools or teachers are purposefully targeted for some reason. Rarely, if ever, are these tryouts randomized in the population or considered as representative of the entire content domain. Tryouts could include an item writer testing students in her school on items she is working on. Pilot tests or other, more formalized, data collections (which are usually smaller and less structured than formal field-testing would typically be) could also be included here. In this case, participating schools would be chosen to focus on certain student demographic elements with a local or more widespread sampling scope. In each of these situations, results would be used to inform revisions to items or other testing components, and have the advantage of being relatively inexpensive and involving very little time. The drawbacks are they do not collect much thoughtful, in-depth information, and the results cannot be empirically defended as being broadly generalizable. Many developers and researchers conduct such evaluations. What is less frequent is that the evaluations investigate and record how English learners are responding to items.

During pilot testing in the mid-1990s, the state of California asked students to circle words and phrases which they did not understand (Kopriva and Lowrey, 1994). The sampling included a large number of ELLs at several levels of English proficiency, and from a number of schooling and cultural backgrounds. These data were collected from their teachers and recorded, and, later, the data were matched with students’ responses. The ELL responses (and circled words and phrases) were also compared with those from the other non-ELLs sampled in the statewide pilot test. Teachers were asked to circle words or phrases they believed would be troublesome for their students as well, and also to provide suggestions for improvement. State staff found both the student and teacher information helpful in identifying how items might be edited. While this procedure is usually not done, it is strongly recommended that the practice should be considered for this subgroup as well as some others, because it can highlight concerns not obvious to the item writers. This is particularly important when the item writers do not have the experience with these types of students or with their linguistic or cultural backgrounds (Durán, 1997).
If possible, it is suggested that some time be taken after the pilot test for teachers to ask their students (as a group) two questions about specific items selected by the developers ahead of time:

1. What was the item asking them to do?
2. Why did they answer as they did?

On multiple choice items, it is recommended that a third question be added:

3. Why were the distractors incorrect?

Teachers should record the student answers on a sheet which is part of the testing materials. When spiraled forms are used within a classroom, students could be grouped by form. All recording sheets should be identified by classroom, so responses can generally inform the student responses. This type of method was utilized during the development of a recent English language proficiency test (Winter et al., 2003). Teachers were asked to lead a short discussion with their students after a pilot test to determine what students thought selected items were asking them to do Protocols guided the data collection and teachers noted student responses. Again, the data were found to be helpful in making changes to the final items.

Cognitive Labs

In recent years formal and more in-depth qualitative approaches have been used more frequently than in the past to examine processes underlying student performance on items. The best known approach is the use of cognitive labs to collect information from students as they are completing items or shortly after they have finished. The National Assessment of Educational Progress initially called this approach “cognitive interviewing” and defined it as a form of interview used to uncover the mental processes (Levine et al., 2002). Paulsen and Levine (1999) asserted that, properly designed, cognitive labs provide developers with information about how students approach test items and whether they are using the cognitive strategies and processes assumed by the item writers. In their study with several hundred test takers, they reported that the cognitive labs, in conjunction with expert review, gave them the most information about problematic aspects of mathematics items. Usually, however, this work appears to be done with a small number of students from the entire population of test takers. Like the tryouts and pilots, the purpose of these small scale studies is to collect formative information midway through the item development process, which can help in the revision of the selected items.

In the type of cognitive lab study that used the think aloud approach, researchers observe students as they attempt to solve problems, asking them
to explain as they work. In a type of retrospective verbal probing or simulated recall, researchers would ask students the questions afterward. Paulsen et al. (1999) reported that it is most comprehensive and accurate to include both methods because the former offers insight into the student’s thought processes during the task, and the probing reveals information about the their recollection and understanding of the experience. Researchers may also interview teachers or collect additional information such as grades and student work samples, in order to gather more information about the students’ learning strategies, skills, learning opportunities, preferences, and achievement.

Few test developers appear to have focused on special populations in their cognitive labs. Stancavage (see Stancavage et al., 1996) of American Institutes for Research completed some cognitive labs with students with disabilities, and some researchers have likewise followed suit. A small number of test developers or researchers have used this technique to evaluate items for ELLs. For instance, developers with NECAP (the New English Common Assessment Program) utilized a modified cognitive lab technique when they were creating plain language items for ELLs and others. Similarly, Winter et al. (2006) and Kopriva et al. (2006) utilized cognitive labs as a preliminary part of their research to investigate the interactions between the needs of English learners and aspects of items. Investigating what students are thinking the items are asking them to do and why, as well as what strategies students use, is important for ELLs. This is because their experiences effect how they address items, and the literature has suggested that these experiences are not like those of native speakers in significant and fundamental ways (see Chapter 2). The limited work completed to date found that this appears to be true for multiple choice as well as constructed response items, where distractors make assumptions about incorrect strategies that may or may not be considered in the same way by ELLs as they are by their native speaking peers.

Kopriva (2000) and Winter et al. (2004b) outlined procedures for conducting cognitive labs with English language learners. They recommend that English learners should represent the primary language groups and all English proficiency levels. It seems to be particularly important to include students from cultures which are disparate from the U.S. culture. Every effort should also be made to identify students who span the range of academic achievement. Further, they suggest that exited and, if possible, a few native English speakers participate as well. Depending on the purpose and scope of the assessment, as a general guideline, it is recommended that there be at least five ELL students who fit each relevant set of criteria in order for any type of tentative interpretation to be made about the results.
They suggest that interviewers should be alert to student misunderstandings in the wording of the items or the answer choices; the meaning of the item, item stem, or answer choices; the item context; the format of the items or form; and if students are restricted from responding because of response constraints in the item. Probes would follow up on why students had these misunderstandings, and, as possible, whether students have learned the target skills. If the assessment is in the student’s home language, sensitivity to their proficiency in this language is important as well. Obviously, it is essential that interviewers not lead the students; rather, they should be cognizant of these elements in their responses.

As the study design dictates, the researchers suggest that related information about opportunity to learn the targeted information, the students’ language proficiency levels (English and L1), and target achievement data (in a form which is unclouded by literacy challenges) are key. The focus of the analyses would be to not only identify item problems, but also be able to differentiate patterns of item responses across students and groups. If possible, these patterns should identify misunderstandings that occur across items, recognize relationships between assessment items and across items and other factors, and identify irrelevant factors that otherwise contribute to accurate and inaccurate interpretations of student scores.

Comparability Considerations

Appropriate attention to the technical adequacy of validity, as discussed above, as well as to reliability, is essential for demonstrating comparability of forms for a given academic test or testing system. Researchers are just starting to consider when traditional methods of calculating comparability might be effective, and when other methods may need to be used. A central issue seems to be to perhaps define comparability more broadly to focus on not if a particular statistical method of putting tests on the same scale can be used, but, rather, if there is enough appropriate validity evidence to support that it should be used. As large-scale testing conditions are allowed to vary within a given purpose, and as forms include items or even collection methods (e.g., portfolios) which differ in ways not originally conceived, can inferences about common constructs still be maintained? In other words, do the scores imply construct equivalence across standard and non-standard methods so that they have the same meaning?

Defining Comparability

In recent years large-scale testing has usually relied on standardized test construction and administration procedures, the use of IRT or similar methodologies, content and bias reviews, and standard setting results in
order to reflect that items were measuring the same constructs within the same unidimensional domain (e.g., mathematics). Implicit in these procedures is the assumption that common inferences can be sustained over form scores. Based on these common and standard development processes, the field utilized identified sampling and/or test implementation designs to collect data about how items perform relative to one another so test scores could be placed on the same standardized score scale. Sampling and/or statistical techniques have been developed (for instance, see Kolen and Brennan, 1995) to handle various kinds of situations that arise so that test scores can be scaled together. Over time, the term “comparability” was often restricted to the statistical manipulations that scaled items and forms. The development processes weren’t explicitly considered to be part of comparability, as it was defined in this way. However, the statistical methodologies assumed standardized development and implementation, and that the processes could be technically defended.

The Standards for Educational and Psychological Testing (AERA/APA/NCME, 1999) explains that comparability seems to be reliant on these two elements, as well as on a clear understanding of the purposes for which the test is being used. That is, for a given purpose, comparability appears to be a function of both construct and statistical equivalence. And so, as non-standard conditions and data collection methods are allowed to vary, it seems prudent that developers and researchers focus on both types of equivalence under the umbrella of how comparability might be defined and defended. This is the approach taken by the multiple comparability studies within the Bazemore (2006) project and the investigations of computer-based item adaptations proposed by Boals and Cranley (2006). Given the recent literature, it also seems that researchers are increasingly thinking about what types of content or construct equivalence data would be defensible to make an argument for putting scores obtained under different methods on the same scale (for instance, the literature reviewed earlier where studies focusing on scaling and dimensionality asked these kinds of questions).

Further, an additional point is implied if and as both types of data are considered to be part of the comparability construct. That is, during their discussion of steps for achieving comparability when accommodations or non-standard forms are used by some students, Haertel and Wiley (2003a) argued that target or construct equivalence should be determined before statistical equivalence is considered; otherwise, the statistical manipulations have little meaning. Recently, this sequential approach was echoed by Winter and Rabinowitz (2006) and others (e.g., Sireci, 2007).

As noted above, the Testing Standards also considered the function of purpose related to how comparability is defined and executed. Examples of different purposes in educational testing include estimating the
comparability of the achievement standard a student’s performance is intended to reflect or the comparability of a student’s scaled score within a particular standard level. The meaning and utility of purpose in determining comparability seems to pivot on whether or not more or less precise differentiations need to be made. For instance, when students are classified according to four achievement standard levels (e.g., below basic, basic, proficient, and advanced), less precision appears to be required than when student scale scores are compared across students. Comparing scaled scores means that the meaning of a scale score of, say, “510” should be the same across students receiving the same scale score and this score should have a somewhat different meaning than the meaning for students who receive a “520.” Thus, scale scores require more precise distinctions than achievement standard levels. In this way, then, more precise evidence of comparability, both construct equivalence and statistical equivalence, needs to be documented for scaled scores than when students are classified by achievement standard. Haertel (2003) differentiated between comparability objectives for norm-referenced and criterion-referenced tests (like the standards-based assessments being used today) and the implications for these tests when standard and non-standard conditions were presented. He also specified comparability objectives under student-to-student comparisons, student-to-achievement level standards comparisons, and student-to-quantifiable criterion measures. While he suggested that judgments may be the primary avenue for making comparability decisions when student score inferences are at the level of achievement standards, he didn’t explain how comparability might be accomplished across non-identical forms or conditions when more precision within achievement standard levels is desired.

Mislevy (1993) differentiated three levels of linking academic forms—equating, calibration and social moderation. Feuer et al., (1999) extended the methods to four: equating, calibration, projection, and moderation. In both taxonomies, the methods are hierarchically arranged in terms of assumptions and precision of inferences arising from the results. That is, assumptions and precision are relaxed as approaches move from equating to moderation. Mislevy’s top level, equating, is the typical approach developers and researchers use to produce comparable forms. This level supports the finest distinctions in ability gradations. The methods evaluate test comparability through the use of statistical procedures where comparisons are made directly between performances across forms. Besides building forms from the same blueprints, the goal of content equivalence has typically been achieved by using identical development procedures, materials, and testing conditions. It is not clear whether or not this method of score consistency or equivalence is sufficient for producing forms with
comparable inferences when forms include both standard and non-standard versions. To date, it does not appear that other score equivalence methods have been considered to handle forms from the same blueprints but where presentation or testing conditions are not identical.

Calibration, Mislevy’s second level of linking, assumes that a well specified content domain is the common frame of reference, for instance content standards, and it evaluates the degree to which each form reflects that referent. The forms are only indirectly compared with one another. In development, calibration seems to assume that the forms do not use the same test specifications but substantively refer to the same referent throughout construction. As such, part of demonstrating adequate calibration will revolve around a quantified criterion estimate of the referent and/or detailed judgments from expert raters about the degree the alignment of the items on forms with the corresponding aspects of the target reference domain. Depending on the precision of analysis, comparisons may be made at the achievement standard levels, and possibly at some designations within the standards as well. Social moderation is the third level of linking where the referent is levels of performance (for example, the academic achievement standard levels). Here, the evaluation of students’ knowledge and skills is not designed to be parallel, and a looser form of expert judgment than calibration is utilized to evaluate how well the combined cognitive demand, or other aspects of the content domain on each form, supports comparability of performances. Empirical evaluations of linking in this case could compare the judgments about the methods or forms, the sub-score or total score performance of students, and perhaps some other independent judgments about the target abilities of the students. This level produces the least specific degree of comparability. As an example, comparability of portfolio scores with on-demand large-scale test scores for students could occur at either the calibration or social moderation type of linking, depending on the quality and types of the construct equivalence documentation.

Phoebe Winter (2007) recently illustrated some points along the range of both construct equivalence or comparability and score comparability (see Figure 12.3). Seeing the two continua in tandem, the reader can observe the level of score comparability that might be expected assuming the referents which form the basis of the test variations. In her model, appropriate technical documentation would be geared to the proper referents.

Analyzing Construct Equivalence

Adequate levels of construct equivalence seem to be a necessary prerequisite for producing construct equivalent scores. When non-standard forms,
methods, or test conditions are considered, it seems clear that sufficient evidence is required to support construct equivalence claims. For equating, both adequate judgments and sufficiently rigorous empirical validation of the target equivalence need to undergird claims of construct equivalence. Some elements of empirical support should supplement the judgments at the other linking levels as well.

For making judgments about construct equivalence for ELLs, development methods designed to promote correspondence across items are explained in detail in earlier chapters. Back translation and simultaneous test development methods (across languages) are examples of techniques which are used when the focus is construct equivalence between English and translated forms (for example, see Ercikan, Gierl et al., 2004). Alignment analysis and other types of independent expert evaluations are examples of judgments that are also needed. For instance, judgment review procedures of item variations targeted to the same test specifications include those used by Gierl and Khaliq (2001), and alignment reviews such as those utilized by Webb et al. (2005) could be used to evaluate forms. Some researchers have used judgment techniques to evaluate the content similarity and comparability of cognitive complexity levels in items across forms (for instance, O’Neil et al., 2003–04; Kopriva et al., 2007c). Sax et al. (2004) explored how and when comparability might be affected when open-ended responses were scored using human and automated graders. After analyzing the judges’ criteria for assigning scores and how the judges appeared to draw conclusions, they discussed how internal discrepancies might be handled to mitigate differences that arise.

![Figure 12.3 Continua of comparability expectations depending on the referent basis of test variations.](image)
For forms not built to be equivalent, content experts may review the bodies of knowledge and skills assessed across forms and determine if the same level of content complexity exists in both. Quality of judgments can be evaluated using quantitative methods such as the confidence interval approach proposed by Penfield and Miller (2004) or those used in standard-setting. Approaches defined in multidimensional scaling or other similar content validation methods may also be appropriate to use in some situations (e.g., Sireci, 1998; Haertel, 2003).

Several empirical methods could support the judgments of construct equivalence. Comparisons of interclass correlations across forms, and pairwise comparisons used by Webb et al. (2005) are examples of the types of descriptive summaries that can be used to support the judgment work. Although differential item functioning procedures are limited, they have been used for this purpose. For instance, Gierl and Khaliq (2001) used a DIF technique to identify items in English and translated versions of a test and identified four sources that help explain how items differ across translations. Similarly, Ercikan et al. (2004) found differences across an English and translated form designed to be equivalent when they looked at DIF patterns across versions. They reported that, while curricular differences explain some amount of DIF, they believe that differences in examinee culture and instruction may also be key variables.

Of particular relevance for comparability purposes are statistical methods which investigate the dimensionality and structure of item responses at the test level. Stone and Yeh (2006) reviewed three factor analytic methods. Tate (2003) briefly outlined several techniques which may be appropriate, such as methods that examine dimensionality, estimation of abilities contributing to test scores and methods for inspecting relationships, and approaches for determining other types of structural complexities inherent in the tests. Several of these methods and some relevant research are explained above.

One area that has received recent attention is the use of computerized administrations, used as an option to the administration of paper-and-pencil forms. As noted in earlier chapters, this test level presentation option may be particularly attractive for ELLs and other populations who are eligible for certain supplementary tools, and/or administration and response accommodations. To date it appears that virtually all of the analyses of construct equivalence have been conducted for the total population, as compared with evaluating effects for subgroups such as ELLs. In general, the work has focused on comparing means and most found small or no differences due to administration mode (e.g., Fitzpatrick and Triscari, 2005; Poggio et al., 2005; Pommerich, 2004). However, in a reanalysis of a state’s studies, Court (2005) found differences in score distributions and
performance levels classifications, even though mean differences had been slight. Further, Pommerich (2004) and Johnson and Green (2006) found differences at the item level for some items. It appears that research is needed to understand how to evaluate mode as well as work that identifies if mode effects are different for subgroups.

The coordinated series of four studies are investigating the construct consistency and score consistency of standard forms to presentation options salient to English language learners: translations, plain language forms, portfolios and a similar non-parallel alternative format, and computerized forms (Bazemore, 2006). To determine degree of construct consistency, three of the investigations will include some type of item review techniques to evaluate equivalence of items when variations are designed to include parallel items. The study with non-parallel methods will rely on alignment reviews and an evaluation of how these are similar across test formats, as well as an expert judgment review of item targets and estimates of cognitive demands for the portfolio entries. All studies will produce and interpret descriptive level data, and, at least two to three of these investigations will analyze the dimensionality and items structure of test variations and general test, using techniques such as a multi-group confirmatory factor analysis.

So far, this section has focused on form or method differences. In situations where forms are the same but the supplementary tools, administration, and/or response accommodations are not, other chapters have emphasized that construct equivalency work also needs to be completed. This includes not only materials and protocols that guide the development and implementation of these accommodations, but evidence that they are being assigned and used properly AND evidence that the item structure results (and other quantifiable and judgment indicators) are consistent across accommodated and non-accommodated students. The maturity of this work, as well, will help define the level of comparability which can reasonably be defended.

Analyzing Score Equivalence

As noted above, score equivalence methods are a statistical approach that essentially put forms or tests on a common scale so results from different forms or tests can be compared. It is important to remember that these procedures assume the instruments are measuring the same targets (for criterion-reference tests) or that they are validly ranking students with respect to ability in some content domain (as in norm-reference tests). That is why the prior step is so essential. Once construct equivalence has been adequately demonstrated, score equivalence techniques can produce scores on the same scale so that common inferences can be supported across forms or methods at the appropriate level of comparability.
When standard and non-standard forms are designed to be parallel, statistical equating is the preferred approach to obtaining score equivalence because of the precision with which the equated scores can differentiate performance. If possible, equating is also the preferred approach if it can be demonstrated that scores from students who used the same form with additional accommodations are reflecting the equivalent target performances as scores from students who did not use accommodations. On the other hand, Mislevy’s calibration or social moderation approaches involve technically rigorous methods that may be defensible if performance at a particular achievement standard is the level where comparisons between students or schools are being made. They may also be the level of comparability which can be defended when the data have not confirmed that accommodated results adequately capture the knowledge of the intended objectives measured by the non-accommodated instrument.

Explanations of equating methods are outside the scope of this book. Basically, texts such as Kolen and Brennan (1995) summarize a number of methods that collect test data from either equivalent or non-equivalent groups. When the distributions of groups are considered to be equivalent (for instance, through random selection), linear equating and equipercentile techniques have been derived, and similar techniques have been developed to handle non-equivalent groups as well. For most of these methods, data are collected on different forms or tests for the different groups. Today, most companies have moved to using item response theory (IRT) with non-equivalent groups to produce equated scores. This approach specifies that a sub-set of common items are given to the different groups as well as items which vary across groups. Item parameters on the common items are set across groups and maximum likelihood techniques are used to estimate the parameters on the rest of the items.

There are a number of different types of calibration and social moderation procedures which have been identified in the last few years. Most often these look like modified standard-setting procedures such as the Modified-Angoff and Bookmark methods (for instance, see Cizek and Berg, 2005; Brennan, 2006). Depending on the level of detail, some of the alignment procedures may serve this function as well. Bazemore (2006) has proposed using a criterion specifically tailored to the same standards-based indicators as a way to demonstrate the degree of calibration or social moderation.

Current procedures at these levels of comparability would appear to be flexible enough to handle non-standard forms, as long as experts are properly trained and any criterion instruments do not confound target ability with construct-irrelevant influences. The question here is whether the equating procedures need to be expanded in any way to properly utilize data from standard and non-standard forms designed to be parallel. For instance, Embretson and Reise (2000) summarize many of the typical
IRT procedures, but include some other variations that may be useful in this context. Since research on score equivalence methods for standard and non-standard forms is in its infancy, at least two of the projects which are addressing this challenge (e.g., Bazemore, 2006; Boals and Cranley, 2006) will also evaluate findings relative to other analyses.

The work needed to properly detect and quantify the degree of comparability in situations where forms or testing conditions are not identical is just beginning to emerge. As the field matures in its choice of validation questions and how they proceed to investigate them, score equivalence will take on increased importance. Currently, research results suggest that form and test condition accommodations are generally improving the validity of score inferences for students who need them. It is not yet clear, however, that validity is equivalent to that obtained for students who do not need accommodations. Comparability research is one of the key questions that future research will need to tackle. As this occurs, it is expected that innovations in how the field views and evaluates comparability will be instrumental in the how the field conceives and more fully operationalizes the measurement of learning for all students.

**Endnotes**

1. Many thanks to David Wiley for helping to research the literature, and organize and review this chapter.
2. This document was written by a convened committee of the NRC, comprising Lyle Bachman, Jonathan Dings, Judy Elliott, Margaret McLaughlin, Mark Reckase, Lourdes Rovira, Maria Medina Seidner, and Rebecca Zwick, with Judith Koenig, Alexandra Beatty and Michael DeCarmine acting as study director and staff, respectively.
3. Bias, as it is defined here, refers to a systematic error that is the result of characteristics of the test that confound group membership factors with student academic ability. The measurement field also uses the term “bias” to refer to any systematic error but the more limited definition is what will be assumed here.
References


References


References • 325


Kopriva, R.J. (1996b) Developing accessible tests from the ground up. Invited presentation to Council of Chief State School Officers, Washington, DC.


Kopriva, R.J. (2005b) Language, literacy, and psychosocial factors: Preliminary finding related to the interaction of ability, student variables and large-scale test characteristics. Paper presented at the annual meeting of the Council of Chief State School Officers Large-Scale Assessment, San Antonio, TX, June.


Kopriva, R.J. and Lara, J. (1997) Scoring English language learners’ papers more accurately. In Y.S. George and V.V. Van Horne (eds.), *Science education reform for all: Sustaining the...*
science, mathematics and technology education reform (pp. 77–82). Washington, DC: American Association for the Advancement of Science.


Solano-Flores, G. and Li, M. (in preparation) The interaction of student, item, and language of testing as a source of measurement error in the testing of English language learners whose first language is Spanish.


Wong-Fillmore, L. (2002) Teachers teaching or teachers testing? A question of priorities. Handout from the Plenary Session at the annual meeting of the National Association for Bilingual Education, Philadelphia, PA.


Index

academic achievement 8, 14, 17–20, 25, 30, 33–4, 86, 173, 186, 193–8, 313, 316
Access-Based Item Matrix 73, 163–4
access-based item templates 63, 92, 98, 101, 170
access-based scoring 275–81
access-based tests 1–11, 62–4, 84–8, 92–8, 106–9; access-based item matrix 6, 73–5, 163–4; context 109–17; and home language 139–40; issues 85–6; rubrics 140–4; structure 117–22; tools 138–9; visuals 125–38; vocabulary 123–5; writing 7, 98–104
Access Specifications Package 62–3, 72–5
Accommodation Station 241–2, 250–1, 258
Accommodations Summary 73–4
accountability 3–4, 17–19, 29–33, 42–4, 52, 55, 57, 69, 76–9, 159, 193, 257
administration accommodations 1, 86, 207–12, 284
administration options 5, 8, 73, 79, 206–12
African Americans 23–6, 34, 143, 303
Alaska 24, 26, 182
alignment studies 5, 78–81, 163
American Educational Research Association (AERA) 51, 68, 173, 284, 286, 314
American Psychological Association (APA) 51, 68, 166, 173, 284, 286, 314
Analysis of Variance (ANOVA) 45, 295, 302
ancillary skills 6, 115, 241–2, 256
anthropology 20–1; see also culture
Aschbacher, Pamela 41
Asian languages 16–17, 34
assessment systems 2, 29–37, 44, 54, 68, 77, 88, 98, 159, 174, 193, 196, 283, 286; access-based 1–11, 50, 62–4, 73–4, 84–6, 88, 92, 97–108, 111, 118, 145, 156, 275–81; accessibility 5, 58–66; see also access-based item development
August, Diane 16–17, 32, 35, 37, 137, 159, 161, 165, 195, 210, 227–8, 240
Improving America’s Schools Act (IASA) 52, 54, 87
improving testing, growth in importance 3
India 14
Individual Education Plan (IEP) 55–6, 233–4, 236, 242, 246, 258
Individuals with Disabilities Education Act (IDEA) 55, 133, 185, 233
informal assessment 76, 155, 238
informal language 188, 273, 275; acquisition 31
informal research 292–3, 310
invented spelling 264
item characteristic curve (ICC) 45–6, 308
item format 41, 110, 117, 132–3, 138, 157, 160
Item Response Theory (IRT) 44–5, 296, 298, 304
Item-specific subject matter (ISSM) 93–7, 102–3, 164
item structure 6, 26, 88–91, 98, 319
K-12 17, 31, 68, 170, 243, 251, 253
Katz, Anne 183, 185–6, 188–91, 193
Ketterlin-Geller, Leanne 87, 232, 241, 250, 258
Khaliq, Shameem 160, 202, 317–18
Koran, Jennifer 33, 193, 216, 222, 228, 235, 243–4, 253–4, 301
L1 see home language
LaCelle-Peterson, Mark 35, 173, 225–9, 231, 274
Lachat, Mary Ann 140, 143, 284
language aids 7, 145–7, 207–10, 218, 227, 232, 274, 306; see also tools
language dominance 176–7, 179
language liaison 74, 162, 207, 210, 231
Lara, Julia 17, 26, 35, 37, 84, 87, 90–1, 116, 137, 159–61, 165, 210, 213, 228, 269–60, 264, 276, 280, 305
large-scale testing 26–7, 29–31, 34–7, 72, 76, 142, 160–3, 205, 217–20, 238–40; computer-based 96, 158–9; criticised 17, 37, 99, 152, 278, 286, 293; defined 1; and difficulty parameters 156; and fluent students 2; and group size 208–11; history of 39–44; and social English 272–3; support 9, 67
LAS 29, 185–6
limited English proficient (LEP) students 11, 30, 54, 143, 180–1, 260
linear models 294–5
Linn, Robert 17, 43, 71, 155, 166, 304
literacy, history of 187
loglinear models 299–300
Lohman, David 59–60
Lord, Carol 87, 205, 246
Lowrey, Karen 26, 36, 70, 87, 100, 121, 137, 149, 209–10, 259, 308, 311
MacSwan, Jeff 178, 185–6
Malcolm, Shirley 99, 112, 146
Mandarin 15–16
measurement theory models 44–52
Messick, Samuel 42, 47, 51, 91, 173
meta-analysis 117, 206, 230, 301
Mexico 14, 16, 22, 27, 181
Minnesota Online Decision-Making Tool 236, 245–6
misunderstanding 139, 271–3, 313–14; see also grammar; home language
Monroe, Ryan 27, 110, 130, 135–6, 146, 215, 218–20, 228
multiple choice tests 35, 40–1, 45, 110, 137–8, 312; distracter choices 26, 118, 137, 152; ease of giving 8, 100; items 95, 97, 162, 209, 212–13, 217, 259, 305, 308, 311; SAT 62
National Assessment Governing Board 99, 167
National Center on Educational Outcomes (NCEO) 87, 101
National Center for Educational Statistics 34
National Clearinghouse for English Language Acquisition 15, 181, 190, 291
National Council on Measurement in Education (NCME) 51, 68, 166, 173, 284, 286, 314
National Research Council (NRC) 16, 286
Native Americans 23–7, 181
native speakers, ELLs outperform 34
New England Combined Assessment Program 87
No Child Left Behind (NCLB) acts 17–18, 30, 55, 77, 182, 184, 193, 199, 222, 315
NOHARM 298
numbers of ELLs 14

O’Brien, Gisela 178, 182, 184, 187, 189
Oakhill, Jane 127, 130, 134
Office of Civil Rights 166, 222
orientation 24, 132, 218
Pacific Islanders 16, 23, 26, 34
American Indian 27
peer reviews 54, 77
Pellegrino, James 46, 48, 59, 84, 89, 91–2
Pennock-Roman, Maria 206–7, 210, 230, 301
poetry 62, 116, 151–2
Pommerich, Mary 158, 319
Popham, W. James 51, 76, 92, 166, 173, 187
prerequisite knowledge 106, 112, 115–16
prerequisite skills 94, 115–16
pretest support 8–9, 85, 205–6, 215–16, 230–1
Principled Assessment Design and Inquiry (PADI) 90
Prins, Elise 99, 118, 122, 124, 135
pronunciation 122, 177, 191, 203
PsychCorp see Dynamic Assessment of Test Accommodations (DATA)
psychology 21–2, 48–9, 59, 84, 90, 166, 219, 284, 314
psychometrics 4, 43, 45, 50, 88–91, 197
Punjab 15, 113, 272
Rabinowitz, Stanley 100, 184, 315
race 17, 25, 52–3, 70, 165, 302
Rasch, George 46, 304, 321
Reckase, Mark 47, 63
Reise, Steven 59, 321
research design 146, 208–9, 286, 290, 301–2
Resnick, Lauren 41, 91
resource allocation 39, 222, 255
reviews; external 143, 145, 163; fairness 165–7, 169–70, 266; internal 163–5;
Romaine, Suzanne 175–8, 187
rubrics 34, 43, 68, 73, 75, 80, 112, 140–4, 163, 166–9, 215, 220, 267, 270, 272–3, 280, 295
Russian 16, 161
Saez, Sharon 25, 32, 260
Samuelsen, Karen 88, 96, 99, 213, 296
Saunders, William 178, 182, 184, 187, 189
Scholastic Aptitude Test (SAT) 40, 58, 117, 143
science 42–4, 47–8, 101, 110, 151–7, 191, 318; assessment 18, 26, 28, 33–4, 44, 90, 142, 151, 153; cognitive 41, 48, 61, 101; double meaning words 125;
measuring achievement 1, 6, 26, 62, 259, 276, 305; Science Nights 219;
spelling of 261
seasonal migration 114–15
Selection Taxonomy for English Language Learner Accommodations (STELLA) 33, 240, 243–4, 248, 251–5, 258, 308
Sexton, Ursula 26, 113–14, 260
Sharrocks-Taylor, Diane 99, 121, 127, 233
Shavelson, Richard 91, 197, 199
Shepard, Lorrie 40, 51, 91, 303
SIBTEST see Simultaneous Item Bias TEST
simplifying vocabulary 99, 156–7, 174
Simultaneous Item Bias TEST (SIBTEST)
Sireci, Stephen 117, 174, 196, 202, 222, 318
Siskind, Teri 8, 33, 136, 153, 157–8, 171, 206, 241, 250, 308, 318, 321
small sample pilots 5, 8, 10, 69–70, 284–5, 310, 318
Snow, Richard 59–60, 99, 120–1, 132, 137–8, 185, 189–90
Solomon, Charlotte 71, 221, 237, 257
Spanish 14–19, 26–7, 161, 186, 194, 200, 250, 261–3, 267–9, 274; dialects 16, 26; length of phrases 137, 227; prevalence of among ELLs 14, 137, 203; states, ELL populations in 15; STELLA see Selection Taxonomy for English Language Learner Accommodations; structural equation models 295–6, 318; Swierzbin, Bonnie 36, 245–6; Swisher, Karen 24, 26–7, 181; syntax 10, 139, 177, 188–9, 191, 261–2, 273–4; Taiwan 16–17; target-irrelevant knowledge 6, 92, 94–8, 156, 164, 273; Teachers of English to Speakers of Other Languages (TESOL) 190; TESTFACT 298; test materials 5–6, 8, 83–6, 91, 101, 117–18, 123, 127, 144, 208, 255, 283; test specification matrices 5–6; text amount 7, 107, 117, 136–8; textbooks 47, 88, 216; theory models; Third International Mathematics and Science Study (TIMSS) 29–30; time allocation 5; Tindal, Gerald 87, 91, 146, 209, 215–16, 222, 232, 235, 240–2, 250–1; Title I program 17, 52–4; Title III program 17–19; tools 6–7, 36, 47, 62, 72–4, 85–6, 106–8, 138–9, 145–7, 153, 231; computerized 145, 150, 171; flexibility 79; psychological 22; rubrics 43, 163; training material 3, 87–8, 169; Triscari, Robert 158, 319, 321; Trumbull, Elise 24, 27, 33, 35, 37, 99, 122–3, 136, 146, 180, 185, 193, 203, 210, 226–8, 231; Ulijn, Jan 99, 118, 122, 124, 135; Valdes, Guadalupe 176–7, 191; Vietnamese 15, 19, 273; visuals 7, 100, 102, 107, 110, 112, 117, 125–32, 138–9, 146, 168–9, 213, 306; vocabulary 7, 87, 92, 97, 112, 120, 122–5, 186–91; Vygotsky, Lev 21–2; Wan, Lam 215–16, 219–20; Ward, Martha 22–4, 26; Webb, Noreen 91, 197, 199, 317; Wei, Li 64, 176, 301; What Works Clearinghouse 291; Whitley, Susan 46, 63; Wiley, David 33, 37, 43, 51, 58, 63, 68, 90–1, 118, 206, 285, 309, 315; Winter, Phoebe 26, 33, 36–7, 62, 87, 90, 95–6, 100, 110, 118, 123, 127, 133, 136, 157, 206, 223, 295, 309, 312–15; Wong-Fillmore, Lily 17, 31–2, 81, 99, 120–1, 132, 137–8, 181, 189–90, 260–1, 274; Yuill, Nicola 127, 130, 134