CHAPTER 12
Selected Technical Considerations

The test development process is interactive. 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 be 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, and increased emphasis on in-depth methods are important as well 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 this population may result. As Lachat (1999) admonishes:

> "All too often, state 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." (pg. 63)

Validity

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 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 score inferences which are constrained within and defined by the evidence produced as a result of the design model suggest a level of construct validity confidence.

To date, technical confidence is mostly determined through careful test construction and implementation procedures, and through a typical set of *post hoc* data analyses. 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) have explained how one might consider such technical designs and what the components need to include. 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 evidence 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.

Developing an Interpretable Research Agenda

One of the shortcomings of the current status of test research for this population seems to be the lack of coherent research agendas. This is particularly true when a particular 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 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 where variables and interrelationships are consistently restrained in their definitions and methods of operationalizing, or otherwise purposefully 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 is essential. This model would articulate *a priori* the intended inferences that scores from the test are proposing and the types of data which 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 inference, 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, J. (2006, April), 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 implemented. Findings from the research will comprise the evidence that is needed to ensure that the validation research design model is acceptable. In cases where this evidence is not sufficient, other research or data analyses should be identified. In cases where the findings do not support the initial models, the initial results should drive refinements in the argument. 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[1]) provides a good example of what the validity argument might look like. This argument forms the beginnings of an appropriate validation design model for ELLs, and illustrations from 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 and the explanation which 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

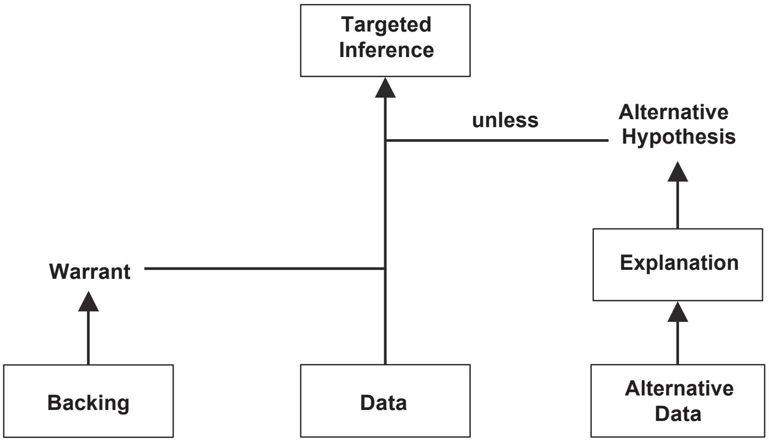


Figure 12.1 Diagram of the Basic Structure of a Validity Argument Adapted from Koenig and Bachman, 2004.

TABLE **12.1** Definitions

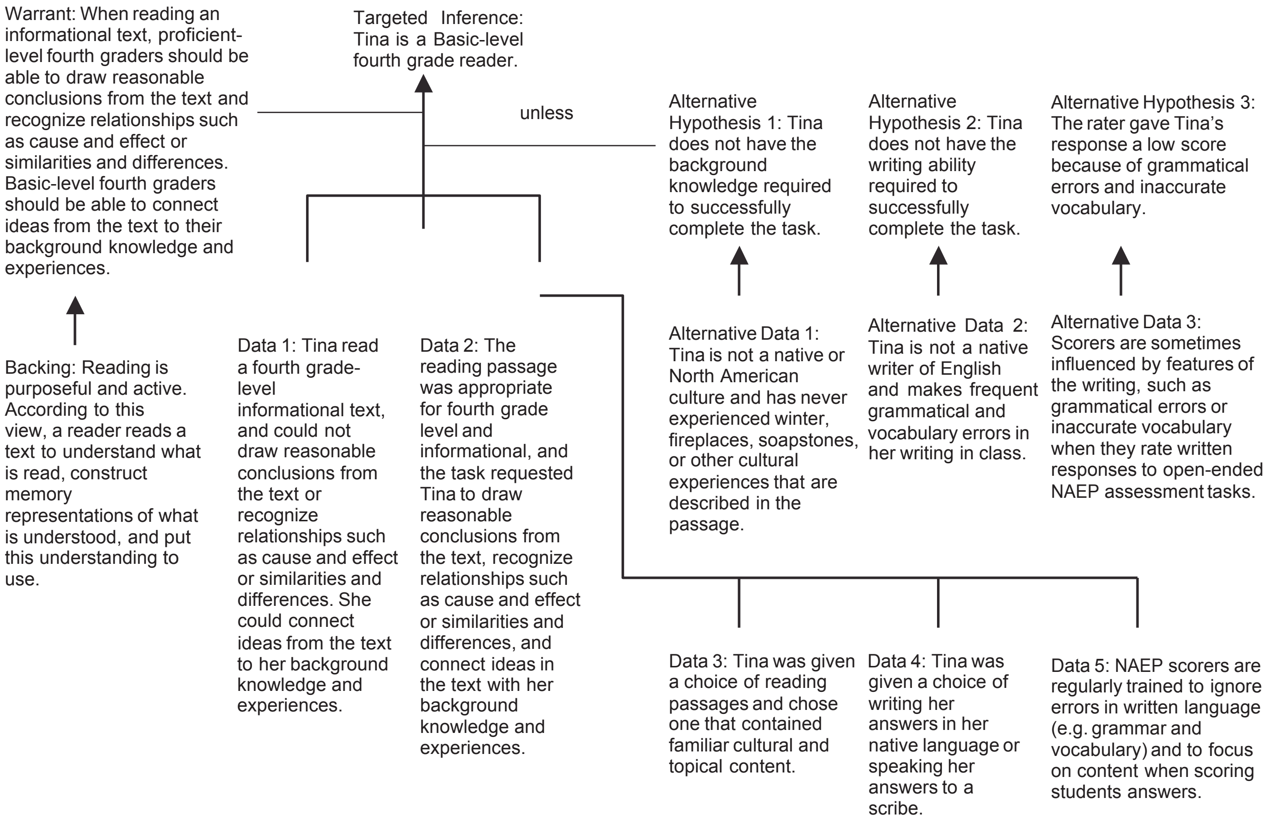
- The *Targeted Inference* is the general interpretation that test designers want to make on the basis of observed data about the level of the test takers' knowledge and skills. The test score inference is a cumulative inference based on targeted knowledge and skills which are tested by different items.
- 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 (but which are ancillary to the intended meaning of the target inference).
- 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.

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 a package of specific needs and challenges. The passage, items and more explanation about the process can be found in Chapter 6. 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 which need to be considered. Therefore, within the overall Validation Design Model there would seem to need to be "submodels" for the different subgroups. These submodels would be built based on different packages of needs that are exhibited by each identified subgroup.

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, 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 year 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



Warrant: When reading an informational text, proficient-level fourth graders should be able to draw reasonable conclusions from the text and recognize relationships such as cause and effect or similarities and differences. Basic-level fourth graders should be able to connect ideas from the text to their background knowledge and experiences.
Targeted Inference: Tina is a Basic-level fourth grade reader.
unless
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.
Backing: Reading is purposeful and active. According to this view, a reader reads a text to understand what is read, construct memory representations of what is understood, and put this understanding to use.
Data 1: Tina read a fourth grade-level informational text, and could not draw reasonable conclusions from the text or recognize relationships such as cause and effect or similarities and differences. She could connect ideas from the text to her background knowledge and experiences.
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.
Alternative Data 1: Tina is not a native or North American culture and has never experienced winter, fireplaces, soapstones, or other cultural experiences that are described in the passage.
Alternative Data 2: Tina is not a native writer of English and makes frequent grammatical and vocabulary errors in her writing in class.
Alternative Data 3: 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.
Data 3: Tina was given a choice of reading passages and chose one that contained familiar cultural and topical content.
Data 4: Tina was given a choice of writing her answers in her native language or speaking her answers to a scribe.
Data 5: 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.

Figure 12.2 Example of the Structure of a Validity Argument for Tina

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. Every few years or as needed they can be updated to reflect the makeup and policies of the state.

Clearly, designing Validation Design Models is a serious undertaking that takes time to properly develop. However, Mislevy, Steinberg, and Almond (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) regarding 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 has been constructed, consideration needs to be given to how the evidence will be collected. Coherent research designs across and within studies 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 information sources 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 data 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 design must be a randomized controlled trial or a quasi-experiment 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 sections which follow are divided in a rather artificial manner. That is, these sections, comparisons of groups and experimental conditions 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 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 differences using experimental methods. The problem with exclusively conducting well-designed treatment and group comparison studies, though, is that the field which investigates how to measure English language learner achievement doesn't 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, for the most part 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 properly focus on measuring the target abilities.

This state of affairs is contrasted to the movement in recent years to include ELLs in 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 in the chapter that the focus of research should now be on quantitative studies. However, the work needs to be implemented within the framework of thoughtfully designed conceptualizations which systematically and quantitatively investigate 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, a bulk 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 (DIF). However, as readers will note below, these procedures are not recommended for this population as there are serious problems with the present methodological approach which confounds the interpretation of results for ELLs. It is correct that many of the analytic approaches which will be discussed here are those which are more relevant for researchers and for states or publisher's with an ongoing validation agenda of established assessments. However, it is also advised that a limited number of quantitative studies always be done in order to validate the inferences in newer testing systems, 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 new 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 Analyses

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 analyses 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. As such, 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 models.

For the most part the focus of these 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 categories and the other variables.

As noted above, 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 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, if rather 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 variables.

Modeling Data from Continuous and Categorical Variables Many types of modeling procedures have been developed over more than 100 years. Crocker and Algina (1986) provide a basic explanation of this type of modeling, and Stapleton (1995) and Hair *et al*., (2006) outline many of the procedures. A seminal reference for traditional or classical linear modeling is Rao (1965).

CLASSICAL LINEAR MODELS Classical observed score approaches include correlations (including simple and multiple correlations) and regressions. 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 a 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.

Many studies which 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, Barton and Meyer (2006), Emick and Kopriva (2007, April) and Kopriva, Winter, Wiley, Emick, and Chen, (2007, Under revision) have used both test and item score 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 approach 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 have 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. It is similar to latent trait analysis except that the latter 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 uni-dimensional. 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.

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 analysis 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 underling 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 analysis 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 & Aiken, 1981; Bock *et al*., 1988; and Muthen & 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 binary data (for instance Mislevy, 1986).

As noted above, most academic assessments today assume that the tests are measuring a uni-dimensional 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 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 speakers, not disabled or those who didn't need accommodations) it was clear that items were performing somewhat differently for some of the focal groups. Other factors that were identified in these studies for the subgroups 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. Taking the results from these studies as a group, it appears that this class of methods (as compared to DIF procedures discussed below) may be able to identify if and to what extent the same dimension is occurring for the majority group and other subgroups. The approach may also help capture which other dimensions or factors are present in the data.

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, particularly 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.

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. An older resource volume that builds on the work of L.A. Goodman thoroughly explains a wide range of methods, including loglinear, is Bishop *et al*., (1975). More contemporary texts, such as Agresti (2002), also explain many of the procedures.

Loglinear methods model observed probabilities. Rather, they model the data appearing in contingency table format as a function of levels of categorical variables. 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 ξ 2 contingency table may be constructed for each item which indicates the numbers of students in each cell. The two variables are student factors, home language literacy and a judgment about testwiseness skills. Three levels of home language literacy are 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 "Does percentage correct in terms of testwiseness skills 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 impact 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 Bayes networks or nets methods 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 which occur for that 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 ξ 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 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 appropriate. 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 these needs, 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 (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 all for 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 Analyses

As noted above, 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, 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 three sets of comparisons which could occur between groups or between treatment methods: evaluation of differences in test score and item level data, 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 sometime 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.

For the amount of work that has been done to date, two criticisms seem to be centered around the lack of well-controlled studies, and 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. Rather, much of the literature is composed of quasi-experimental studies, or experimental studies such as Abedi *et al.*'s, 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 as they are limited by uncontrolled influences.

Item Level Analyses Considerations associated with Differential Item Functioning will be discussed next, followed by some notes about other item level methods. Just as with test scores, designs for comparisons of items based on different groups or treatments need to be thoughtfully considered and implemented.

Differential Item Functioning (DIF) 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. Recently, 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[2]. The case focused on considerations for inclusion of items in the Illinois insurance licensing examinations. This agreement determined that biased items, as defined by the above criteria, should not normally be included in the test (Council of Chief State School Officers, 2000).

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 test access to demonstrating 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, for instance (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 uni-dimensionality 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 close-ended 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 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 uni-dimensional 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 IRT (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 this 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 items are sent to bias 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 5 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 only be flagged only if they were unusual for low ability or unmotivated non-ELLs (or whoever the reference group might be). The second conclusion, and the one that makes the DIF approach problematic for this population, 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 does not presently detect pervasive problems in items—invalidity which occurs over many items. It only picks up certain types of problems unique to individual items. This is helpful but, meanwhile, the much larger issues are not being identified by this methodology.

For many ELLs, it is primarily their English language limitations that impact their ability to demonstrate what they know throughout the test. For early 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 (i.e., they were advanced ELLs, exited, or were bilingual and never needed English literacy services) 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 structure, 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. It would also seem to be difficult to interpret flagged items based on an ambiguously incorrect estimate of content ability. Some have suggesting 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 an unknown amount of 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. That is, DIF may or may not detect the item issues which impact many items. Some have suggested using a control group here as well—perhaps a reference group of like students who do not receive the accommodation or accommodation package under investigation. However, under-detection of any pervasive issues for the non-accommodated group could incorrectly confuse the issue. 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 earlier 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.

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. Undoubtedly 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 characteristicsfor 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. In all of these cases, however, the referents are total scores or thetas, and there are assumptions made about the validity of the total score or about the accuracy of the estimation of student abilities. 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 early 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.

In a project being completed in 2007 (Siskind *et al*., 2004) researchers are conducting distractor analyses on multiple choice items for different subgroups of students, including English learners and others with language difficulties (such as some learning disabled students and those with hearing impairments). Descriptive findings from Kopriva and Lowrey (1994) suggest that distributions of distractor choices are different for ELLs as compared with native speakers. 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, but investigate 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, Winter, Wiley, Emick and Chen (2007, under revision) computed logistic regressions to measure the impact of item factors for native 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.

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 robust for a particular purpose. Finally, issues of multi-colinearity 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.

In the mid-1990s, the state of California asked students to circle words and phrases during pilot testing 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 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 or "think-alouds" to collect information from students as they are completing items or shortly after they have finished. 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) assert 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 a cognitive lab study, researchers observe students as they attempt to solve problems, asking them to explain as they work. In a type of retrospective verbal probing, they may also ask students questions afterward. Paulsen *et al*., (1999) report 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 (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, the New England Compact (which produces a common test for some of the New England states) utilized a modified cognitive lab technique when they were creating plain language items for ELLs and others. Similarly, Winter *et al*., (2004) and Kopriva, Winter, *et al*., (in press) 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). This appears to be true for all item types, including multiple choice, 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) outline 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 speakers participate as well. As a general guideline, it is recommended that there be at least five ELL students who fit each criterion 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. In the latter case, it is clear that it will be beholden to the developers and researchers to document why these other methods appear to be technically defensible. Alone, simply putting forms on a common scale that have not been adequately made equivalent during development is not sufficient. Use of expert judgment only is also probably not defendable, particularly when inferences demand a level of precision beyond broad statements of performance. This section will first define comparability, and then briefly discuss two elements of comparability that appear to be essential if common score interpretations across standard and non-standard forms, and/or for accommodated and non-accommodated students, are considered to be meaningful.

Defining Comparability

Comparability of score inferences suggests that the meaning of the scores is the same, whether students take form A or form B. Comparability seems to include two steps: first, development methods and empirical evidence need to demonstrate that the forms are measuring equivalent knowledge and skills in the content domain of interest. Second, assuming content equivalence across forms, statistical methods place scores from the forms on a common scale so comparisons can be made across forms. The *Standards for educational and psychological testing* (AERA/APA/NCME, 1999) specifies that comparability is easiest to establish when procedures for test development and implementation are identical. In cases where formats or other conditions vary, other evidence may be required. In all cases, the requisite degree of comparability is defined in terms of context and purpose for which the tests (or forms) are being used. For example, to compare the performance of individual students over time or with other students a more precise definition of comparability would most likely be required than when equivalence is determined at the pre-specified achievement standards level (for example, basic, proficient, and advanced) with school-level data (i.e., for NCLB Adequate Yearly Progress).

In their discussion of steps for achieving comparability when accommodations or non-standard forms are used by some students, Haertel and Wiley (2003a) focused on the necessity of determining target equivalence before any consideration of statistical equivalence is considered. 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 forms are considered. He also specified comparability of test administrations 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 when student score inferences are at the level of achievement standards, he didn't explain how comparability might be accomplished across non-identical forms when more precision within achievement levels is desired. Winter and Rabinowitz (2006) defined two conditions, both of which they believe are necessary for evaluating comparability. The first of their two conditions is construct consistency—at the targeted level of comparison (i.e., to other students or to content standards), do the forms measure the same content information? The second condition is that of score consistency. At the appropriate level of comparison, do the same scores or same performance levels, across forms, reflect the same level of abilities? They emphasized that adequate evidence is essential to document the equivalence at each level.

Winter and Rabinowitz argue that, only after an adequate level of content equivalence has been established, should score equivalence methodologies be implemented. Mislevy (1993) differentiates three levels of linking academic forms—equating, calibration and social moderation. Feuer *et al.*, (1999) extend 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 level of achievement standards, 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 levels). Here, forms are 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 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.

Analyzing Construct Equivalence

Adequate levels of construct equivalence are a necessary prerequisite for producing construct equivalent scores. When non-standard forms 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 precise correspondence across items are explained in detail in earlier chapters. Back translation and simultaneous (across languages) test development methods are also important when the focus is target 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*., 2003–2004; Kopriva *et al*., 2007). 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.

For forms not built to be parallel, 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 statistics 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., Sirici, 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 are explained above.

To date, some relevant research has utilized these types of techniques. For instance, Barton and Finch (2004) examined the dimensionality of items flagged for DIF in both mathematics and ELA tests at different grades, for students with disabilities (with and without accommodations) and students without disabilities. They used a cluster analytic method called DETECT (Zhang and Stout, 1999a, 1999b), which also includes a procedure to determine the degree of consistency across clusters. Their investigation, which analyzed the clusters by group and by codings of item aspects, yielded similar numbers of clusters across several of the grade/group cells but failed largely to explain the dimensions with the item information. Cook *et al*., (2006) compared the dimensionality of a test for fourth grade ELLs (with and without accommodations) and non-ELLs using a factor analytic method under each condition, and following up with a combined group confirmatory analysis to determine if the results could be compared across groups. While other review procedures had suggested equivalence, findings indicated that there seems to be a reasonable amount of difference among groups. Estimating the particular amount of difference could not be completed, however, because the combined group analysis would not converge. Sireci and Khaliq (2002, April) utilized factor analysis, structural equation modeling, and multidimensional scaling to examine the structure of a dual language form as compared with the standard English form. Results also emphasized that empirical support, as well as judgment, needs to be produced to provide evidence of the construct equivalency of forms.

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. In a project finishing summer of 2007, researchers are conducting a series of item distractor analyses and multi-factorial comparisons across four types of forms (including the standard form and an “emerging language” form) where item variations were built to be equivalent (Siskind *et al*., 2004). Interestingly, these analyses will be conducted in four subject areas and six grades so comparisons will also occur across these variables. In a recently funded project, Boals and Cranley (2006) proposed the development of science items for early ELLs which will be built to be equivalent to standard items but which will not use the standard item types found in state academic tests. A series of reviews and pilot investigations will follow this work to determine if content integrity appears to be sustained over variations.

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.

A coordinated series of four studies have just been funded to investigate the construct consistency and score consistency of standard forms to presentation options salient to English language learners: translations, plain language forms, portfolios or 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 items will rely on alignment reviews and an evaluation of how these are similar across test formats. 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 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 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. Mislevy's calibration or social moderation approaches involve technically rigorous methods that may be defensible if performance at a particular proficiency level 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 (i.e., through random selection), linear equating and equi-percentile 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 techniques (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.

Researchers are currently completing goodness of fit analyses on IRT findings for various populations of students in South Carolina (Siskind *et al*., 2004). These analyses are being conducted on each of four forms, including two standard forms and an "emerging language" form. Item variations are being calibrated to parameters currently utilized by the state, using common anchor items which appear on each form. Fitzpatrick and Triscari (2005) post equated their online and paper-and-pencil tests by using the paper-and-pencil Rasch item parameter for a subset of online items and allowing the values of the other items to be calibrated freely. In the Bazemore project, equating studies will be conducted by comparing computer versions, translations, and plain language versions to their standard form counterparts. Part of their work will investigate selection methods for the common items so that mode effects are not masked. Boals and Cranley (2006) have proposed 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 early English learners. They are convening an expert panel to determine, among other things, what construct equivalence and equating technique might be applied that utilizes the target difficulty levels (as compared to the traditional "confounded difficulty levels" that incorporate both sets of influence) to estimate ability for participating students.

Since research on score equivalence methods for standard and non-standard forms is in its infancy, most of these projects will also evaluate their score equivalence results relative to other analyses. The Siskind and Bazemore projects will conduct additional construct equivalence analyses, and Boals and Cranley will also compare findings based on traditional methods with results from their experimental equating technique.

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. Thus, both construct equivalence and score equivalence studies are essential in addressing the comparability issue.

Endnotes

1. 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.
2. 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.