

## PARA Research Program Details Page

Para's research program is consists of studies that will occur both during the research phase in the second year of the project and during the field testing phase of the project. The research questions and studies are described here.

**Research questions.** The research questions that are addressed in Goal 2 are:

1. Which students need which kinds of assistance to access the types of reading assessments that states currently administer? (Sampling Issues)
2. How can a valid assessment be developed that will permit students with disabilities that affect reading to demonstrate their reading proficiency? (Validity of Accessible Assessments)
3. What common forms of assessment are appropriate for group testing that are accessible to students with disabilities that affect reading? (Universal Design and Individual Student Characteristics)
4. What opportunities do students with disabilities that affect reading have to access reading through alternatives to print in the classroom (Instructional Sensitivity/Opportunity for Use of Access Tools)

**Study 1. Sampling Issues.** Difficulty interacting with print cuts across disability categories. Yet, many studies have used disability category to exclude students. Mistaken assumptions that all students with certain disabilities have trouble reading print are just as inaccurate as assumptions that students without disabilities have no trouble reading print. A study on sampling issues will be conducted in two phases: 1) data collection and 2) field study.

**1. Data Collection.** During Year 1, data will be collected on how the study participants interact with print. Student statewide test results, information about accommodation use, data from student IEPs as appropriate will be gathered to inform the field study design and analyses.

**2. Field Study.** A group of students who have difficulty interacting with print will be selected (see process below) as the experimental group and another group of students with no difficulty

interacting with print will be selected as the control group. A more sensitive screening process will be used to identify participants in the experimental group to assure that those students actually have difficulty interacting with print regardless of disability category. Details of the difficulties experienced by students will be noted and catalogued. Analyses will be conducted of how different accommodation policies and different types of large-scale assessments impact the performance of students who have difficulty interacting with print. Students in the experimental group with difficulty interacting with print will be divided into two groups: 1) those with disabilities and 2) those with no apparent disabilities. A third group will be students with no difficulty interacting with print. We will assign students to the three experimental conditions randomly to the extent possible.

Individual students who participate in the study may be identified through teacher nomination, interviews (student, teacher, and parent), analysis of student IEPs, and analysis of previous large-scale assessment results to determine whether each potential participant has a disability that may affect reading. Only students who know how to use any needed accommodations will be included in the experimental group with disabilities (group 1).

Performance of students across the three groups will be compared on several measures of reading, including scores of standard achievement tests, grade points in reading, and reading testlets developed using definitions from Goal 1. These comparisons will be done using a multivariate analysis of variance (MANOVA) model. Following possible overall significant results, performance differences of subgroups of students can be compared using a series of *priori* or *posteriori* contrasts. Comparing reading performance of students with and without disabilities, both with difficulty interacting with print, will reveal possible sampling issues.

**Study 2. Validity of Accessible Assessments for Students with Disabilities.** Research has demonstrated that assessment results for students with disabilities that affect reading may be confounded with nuisance variables. These variables may be a source of construct-irrelevant variance and may negatively affect the reliability and validity of assessments for these students.

Problems affecting readability of assessment instruments, how assessments are “read,” issues surrounding test instructions, and inconsistencies in test administration and scoring are examples of nuisance variables. The effects of these variables on assessment outcomes may be more serious when they differentially impact students with different types of disabilities related to reading. Haladyna and Downing (2004) cited 21 variables that may be considered as nuisance variables; Thompson et al. (2004) noted that assessments that measure only a small range of concepts using print-only methods may also lead to construct-irrelevant.

In addition to identifying major sources of construct-irrelevant variance in both traditional tests and those designed using Goal 1 definitions, we will examine validity effectiveness of potential assessment strategies through think aloud/cognitive lab and experimental design (Ericsson & Simon, 1995). Differences among tests assessing a narrow range of concepts and versus a wide range of concepts will also be noted. We will focus on four aspects of assessment: 1) effectiveness, 2) validity, 3) feasibility, and 4) differential impact of assessment formats.

**Study 2A: The effectiveness of assessments** will be examined by comparing new testlets to existing large-scale reading measures. Assessments will be deemed effective if they can, in a way that is both accessible and understandable to students, encompass a wider range of reading concepts than currently existing assessments. Both a qualitative phase, where teachers and district-level personnel rate each of the three items using a 5-point Likert scale, and a field test phase. In the field test, students will be assigned to either an experimental group (those having difficulty interacting with print) or a control group (students with no difficulty interacting with print) and performance of the two groups will be compared on both traditional tests and the new tests using definitions constructed in Goal 1. The new test will be deemed effective if the performance gap between experimental and control groups on the test is reduced substantially.

**Study 2B: The validity of accessible assessments** will be tested by comparing the performance of students with and without print difficulties on differing types of tests (see above).

If providing accessible assessments increases the performance assessments of students without print difficulties for whom access tools are not intended, then the validity of access tools may be questionable. **(1) Qualitative Phase.** The accessible assessments will be presented to teachers and content specialists who will be asked to judge whether the accessibility factors impact the construct being measured. A 5-point Likert-scale rating scale will be used for this evaluation. Judges will use the rating system to identify any validity issue in the accessible assessment. **(2) Field Test Phase.** Skilled print readers within a classroom will be assigned randomly to an experimental group where they receive the same access conditions that are used for students with print reading difficulties, and a control group where they receive no access conditions. Individual students will be identified in similar ways to Study 1 with the goal being to ensure that each participant has a disability that may affect reading. A hierarchical linear model will be used to detect any possible class, teacher, and school effects that might remain after random assignment of conditions. For assessing the validity of access conditions, performance of the two groups will be compared using an analysis of covariance model in which some the important background variables such as SES will be used as a covariate. If students in the experimental group (skilled print readers receiving access conditions) perform significantly higher on reading than the control group (skilled print readers receiving NO access conditions), then the validity of access conditions would be questionable.

**Study 2C: Feasibility of accessible assessments** will be determined by interviews with state assessment directors and test vendors, as well as with district superintendents, principals, and teachers. They will be asked to rate the feasibility level of assessments using a 5-point Likert scale. Some types of assessments that may be highly effective and valid may not be feasible.

**Study 2D: Differential impact of assessments** will be studied by investigating the possibility of some forms of assessments that would have a broader application across both types of readers (print or alternative modalities) and concepts. For example, many assessments

only measure basic comprehension via reading print. In this study we will examine the possibility of a wider range of concepts by using a wider range of access tools (that will help a larger number of students with print reading difficulties). We will identify major variables that will have potential impact on accessible assessment. These variables will be identified by the Technical Advisory Group (TAC) and the General Advisory Group (GAC) during Year 1. Groups of students would be formed based on the categories of these variables and possible performance differences across the subgroups will be examined using a multiple discriminant analysis or logistic regression. Results of this study will identify major variables that could impact the “universal design” of the accessible assessments.

**Study 3. Universal Design and Individual Student Characteristics.** The purpose of this study is to establish a balance between individual students’ needs and group testing. We will explore the possibility of universally accepted access tools that would be relevant for different groups of students, depending on ability to access print. The specific design of the testlets used during this study will be created as a result of the TAC and GAC meetings, and will be created by staff during the last half of Year 1. Results of our study in the “differential impact of assessment” discussed above also will shed light on this aspect of assessments.

**Study 4. Instructional Sensitivity/Opportunity for Use of Access Tools.** Students with disabilities that affect print reading may not benefit from the same level of opportunity to learn (OTL) compared with their print reading peers due to print reading deficits. Thus, teachers have either to adjust their instruction according student needs or spend additional time with these students, both of which may be difficult (Deshler et al., 2001). We will interview teachers and students with disabilities and observe during regular classroom instruction, focusing on the reading concepts students with print difficulties are taught. We will also use an OTL questionnaire with teachers and students. OTL instruments developed by CRESST researchers to examine OTL for ELLs (Abedi, Herman, Courtney, Leon, & Kao, 2004) will be modified to fit OTL requirements for students with disabilities and used after field testing.

## **Development and Field-Test of Accessible Reading Assessments**

Based on input from the Definition Panel, research findings, and developed principles and guidelines, a testlet or set of testlets will be prepared for field testing in Year 4. A rigorous field test methodology will yield results appropriate for accountability assessments. Westat will manage the field testing. CRESST will conduct a validation study. NCEO will conduct a cost analysis study.

**Identify Districts and Schools for Sample.** Westat will design a sampling plan to compare two measures of reading proficiency for students with disabilities; one based on a developed testlet and one from another measure of reading proficiency. The objective of the sampling plan will be to select a representative sample of sufficient size from the disability groups of interest at the 4<sup>th</sup> and 8<sup>th</sup> grades to allow meaningful comparisons of the two measures of reading scores within each disability group and for students without disabilities. The precision of the estimated average difference in individual reading scores will depend on several factors such as the variation of the individual-level difference in scores, the sample design effect (i.e., effect of clustering the sampled students at the district or school levels), and the sample size. Based on a preliminary analysis of available data, we propose to select a sample of 200 students from each disability category and to select a sample of at least the same size of students without disabilities. The sample will be selected from the five participating states.

**Secure District and School Participation.** Westat expects that the participating states will secure an initial promise of cooperation from the sampled districts and schools. After we receive confirmation from the state contact we will send an introductory letter to the districts.

**Identify and Train Data Collection Staff.** Westat Field Liaisons will be responsible for hiring qualified and motivated individuals to administer the field assessments. They will work closely with our Field Supervisors who will manage the work of the Assessors and Assessment Assistants and maintain contact with districts and schools during the field assessment period. In

all hiring decisions, Westat will strictly apply our program of affirmative action to employ and advance the employment of qualified individuals with disabilities.

**Create Data System.** Westat will design and develop two related databases for this study using Microsoft SQL Server. One of those databases will contain the data needed for the management of study activities, while the other will contain the research (assessment data and other student background data). This architecture will enhance system availability and security while explicitly separating management data from study data, thus maintaining confidentiality. The management portion of the overall system includes the management database itself, a data entry system, the Field Supervisor's management system, and the Westat home office management system. The research portion of the data system supports data feeds from the data entry systems, and, in turn, provides data to the analysis and delivery systems.

**Conduct Field Assessments.** Assessments will be administered in 5 states, 8 districts within each state, and 10 schools within each district. Written parent consent materials will be sent to the School Coordinator at each school. For budgeting purposes, we have assumed one testlet that takes about one hour will be administered. As the design work for this study proceeds, it is likely this assumption and others will be adjusted.

**Clean Data and Create an Analysis File.** Westat will perform quality monitoring and control at every step of the data collection process to ensure accurate data entry. We will review the data using such procedures as verification, keying verification, and machine editing by specifically designed computer programs.

**Validity Studies (CRESST).** Validation studies will involve several major components: (1) item analyses, (2) estimating reliability, (3) estimating validity, (4) scaling issues, (5) issues concerning composite scores, (6) identifying test items with substantial cultural/linguistic biases, and (7) issues concerning standard setting.

**1. Item Analyses.** Classical item-analysis approaches will be used. Descriptive statistics such as mean (p-values in case of MC items), standard deviation, and frequencies of the

distractors would be obtained for the multiple-choice items. Item discrimination power would also be estimated. Item-total correlation will be obtained to help identify test items that are not consistent with the overall measure of the construct. These statistics will help improve the quality of test items by modifying them based on the results of items analyses. Similarly, for open-ended items, mean, standard deviation and item-total correlation would be obtained.

**2. Reliability.** For estimating reliability of the tests, different approaches could be applied. An internal consistency approach using Cronbach's alpha would be utilized to provide an estimate of internal consistency of the items. Additionally, principal components analysis would be conducted to examine the dimensionality of items. Since Cronbach's alpha is extremely sensitive to multi-dimensionality of test items (Cortina, 1993), the results of principal components analyses would shed light on the outcome of internal consistency analyses. Having a low alpha coefficient may be due to high level of measurement error, or multi-dimensionality of items or a combination of both of these factors. In the principal components analysis, factor correlations (correlation between the components) will be estimated to examine the possibility of a higher level of uni-dimensionality.

**3. Validity.** Multiple approaches for estimating the validity of the reading proficiency test could be utilized. A criterion-related validity approach (concurrent and predictive validity) may be conducted. Since it may be difficult to find a single valid criterion for validation, we propose multiple criteria, using a latent composite as a criterion for concurrent validation with confirmatory factor analytical techniques. A latent composite of the newly developed reading proficiency would be correlated with a latent composite of the external criteria. The external criteria for this study would include standardized achievement test scores in reading/language arts, student's GPA, and teacher's rating of student reading achievement. A multi-trait/multi-method (MTMM) approach may also provide useful validation data in this study.

**4. Scaling issues.** Issues concerning horizontal versus vertical scaling will be examined. Since reading proficiency is more developmental rather than grade-related, a vertical scaling

methodology would be more appropriate. The focus of this part of the study will be on the validity of horizontal versus vertical scaling. During the item development phase, we will prepare enough common sets of items across age/grade levels.

**5. Issues concerning composite scores.** Since the newly developed accessible reading assessment is not a unidimensional construct, it may be unrealistic to consider a simple composite of the subscale scores of this test. While there might be high correlations between subscales of the reading tests, there may be components that are specific to each subscale. If the purpose of this assessment is to provide a single measure of reading proficiency, then only the common share of the variance of the components must be considered. We will examine the validity of a simple composite versus a latent composite of the subscale scores. We will also compare the composite scores (simple and latent) with individual subscale scores to find the best approach in reporting the results of the reading proficiency tests.

**6. Identifying test items with substantial cultural/linguistic biases.** There might be interactions between nuisance or extraneous variables and some items. For example, test items high in cultural/linguistic biases may show differential level of performance between students with and without disabilities when adjusting for overall differences between the two groups. Different approaches have been suggested for examining the possibility of differential functioning of items (see Allen & Donoghue, 1996). Different approaches for examining DIF in such items include Quasi-chi-square (Scheuneman, 1975, 1979), log-linear (Alderman & Holland, 1981; Loyd, 1984), Mantel-Haenszel (MH) (Holland & Thayer, 1988), the standardization procedure (Dorans & Kulick, 1983), SIBTEST (Shealy & Stout, 1993); and the logistic regression approach (Spray and Carlson, 1988). NAEP has frequently used two different approaches, a graphical method which could be conducted by the modified version of BILOG program (Mislevy & Bock, 1984), and the MH procedure suggested by Holland and Thayer (1988). We will explore different possibilities and use different approaches for computing DIF statistics. Items that are identified with a large DIF would be marked for revisions.

**7. Issues concerning standard setting.** Standard setting is a major milestone in any test development, particularly in this project. It is essential to know how students with disabilities perform in this test. We will provide cut-scores based on different approaches and examine their consistencies. Some of these approaches are examinee-centered such as the Contrasting Groups approach and some are test-centered such as the Modified Angoff and Book-marking approach (see, for example, Kiplinger, 1997; Buckendahl, Smith, Impara, & Plake, 2001). Cut-scores obtained by standard setting would be validated through a concurrent validation approach.

Cost Analysis. Information is needed about how much accessible reading assessments would cost if widely adopted by states, so that good decisions can be made that are based on reliable cost data instead of on “guesswork or politics” (Levin & McEwen, 2001, p. 2; cf. Hoxby, 2002). We will collect cost data during the field study and make projections about how much it will cost for a state to develop and administer a more accessible reading assessment. By forecasting the costs associated with the widespread implementation of an accessible reading assessment, we will be able to provide information about the costs associated with various policy options. The analysis will project both first year costs (anticipated to be the most expensive due to development and implementation costs), and costs during the later years for typical small, medium, and large state. It is anticipated that projected costs may include test development, assessment review, reliability testing, Website development, production, and administration of the assessment.