# Ability or *Access*-ability: Test Item Functioning and Accommodations for Students with Visual Impairments on Pennsylvania's Alternate Assessment

by

**Kim T. Zebehazy**

Bachelor of Arts, Western Michigan University, 1997

Master of Arts, Western Michigan University, 1998

Submitted to the Graduate Faculty of

The School of Education in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2006

UNIVERSITY OF PITTSBURGH

SCHOOL OF EDUCATION

This dissertation was presented

by

Kim T. Zebehazy

It was defended on

November 21, 2006

and approved by

Naomi Zigmond, PhD, Professor
Department of Instruction and Learning, University of Pittsburgh

Jane Erin, PhD, Professor
Department of Special Education, University of Arizona

Louise A. Kaczmarek, PhD, Associate Professor
Department of Instruction and Learning, University of Pittsburgh

Audrey T. Kappel, PhD, Research Assistant Professor
Department of Instruction and Learning, University of Pittsburgh

Dissertation Advisor: George J. Zimmerman, PhD, Associate Professor
Department of Instruction and Learning, University of Pittsburgh

**Ability or *Access*-ability: Test Item Functioning and Accommodations for Students with Visual Impairments on Pennsylvania's Alternate Assessment**

Kim T. Zebehazy, PhD

University of Pittsburgh, 2006

This study explored issues surrounding the validity of Pennsylvania's Alternate System of Assessment (PASA) for students with visual impairments. The PASA is a performance-based assessment that assesses a sub-set of math and reading skills delineated by the State's alternate standards. Data from 286 students with visual impairments who took the 2005 Level A PASA at grades 3/4 or 7/8 were analyzed. Descriptive and statistical analyses compared achievement on the PASA between three groups of students with visual impairments at different levels of functional vision as well as to a matched group of peers without visual impairments. The latter comparison investigated differential item functioning (DIF) on each individual test item using the Wilcoxon Signed Ranks test. In addition, types of accommodations made for students with visual impairments to provide access to the assessment and potential factors contributing to test bias were documented. Overall, the study confirmed expected patterns of accommodation selection by functional vision level with layout/set-up accommodations being the most frequently used. It also revealed a set of test items flagged for DIF statistically that did not always coincide with the test items judgmental reviewers would expect to be problematic or different for students with visual impairments. Among the three functional levels and the students with visual impairments as a whole, 29 instances of DIF in which a test item may have been potentially harder were found. In addition, there were 12 instances where a test item may

have potentially been easier. A qualitative logical analysis highlighted a variety of variables that interact with the decision-making process to pinpoint potential reasons for the presence of DIF. Under-accommodation, the frequency of lucky guesses, score change patterns, and experience level with content were all factors suspected of contributing to performance on different types of test items. Discussion of these variables as well as interesting patterns in accommodation selection or the absence of accommodation selection is included. Challenges of and recommendations for adapting the PASA for students with visual impairments are provided as well as general discussion regarding aspects of assessing this population of students.

## TABLE OF CONTENTS

## LIST OF TABLES

## LIST OF FIGURES

## ACKNOWLEDGMENTS

*Twenty years from now you will be more disappointed by the things that you didn't do than by the ones you did do. So throw off the bowlines. Sail away from the safe harbor. Catch the trade winds in your sails.*
*Explore. Dream. Discover.*

*-Mark Twain-*

As my travel through the sea of doctoral study comes to an end, I am grateful for this opportunity to reminisce and appreciate those who in their own unique ways have supported, inspired, and challenged me. My decision to pursue a PhD has not only been an academic voyage but a journey of self-understanding. Each of the "22 Steps to the Doctorate"- wavy and unpredictable as they may have been at times- have resulted in new perspectives, educational growth, refined skills, and a greater appreciation for the effort it takes to create change.

I want to first acknowledge all of my family and friends across the United States and the globe who took the time to check in on my progress and to offer words of encouragement. The phone calls, instant messages, electronic cards, and the "How are you doing?" emails have meant a lot. In particular, I thank my mom who has vicariously experienced both the rough and calm waters of this voyage. Her love and support has always been unwavering- lending a listening ear even during those late hours when she was not accustomed to being awake! I also thank Penni,

my “Pittsburgh mom,” who offered me escape, a return to sanity, friendship, and good home-cooking.

I wholeheartedly acknowledge all my fellow travelers who sailed right along with me. Without them, this voyage would never have been the same. In particular, I send my deepest appreciation to Amanda, Karen, Lynn and Wendy whose support and friendship have been invaluable. They were always there to struggle over the hurdles, to celebrate the milestones, to find humor in the absurd, to discuss the future and the issues, and to just plain “goof-off” sometimes. Their presence along this journey and the lasting friendships that remain have made the stormy times easier and the triumphs sweeter.

I thank all of the PASA crew whose time and attention made it feasible for me to analyze my data according to my set course. They, along with the hard work of dedicated teachers of the visually impaired, data enterers, and the support of other colleagues in the field made conquering the logistics of this dissertation possible. Also, I extend a sincere thank you to everyone in the Department of Instruction and Learning who took an interest in my progress and lent innumerable support both logistical and otherwise. Their stories, laughter and friendship often brightened my day. In addition, I want to acknowledge Dr. Elaine Rubenstein whose statistical expertise and affirming demeanor helped calm the seas and bring the horizon into view as we problem-solved my first challenge in analysis.

Last, but certainly not least, I would like to express my heartfelt appreciation to my dissertation committee and to my advisor. I send my gratitude to Dr. Jane Erin for taking the time to be my outside committee member. Her expertise, support of my work, and thought-provoking input have encouraged me and broadened my thinking on the topic. I thank Dr. Louise Kaczmarek for her careful review of my document, her insights, and for her

encouragement. I can now say that I am “más que menos” rather than “menos que más!” I thank Dr. Audrey Kappel for her statistical expertise, willingness to answer my questions, and her orchestration in pulling together the data base for my analyses. I also express a sincere thank you to Dr. Naomi Zigmond who has supported me in numerous ways including the opportunity to focus on my dissertation these past few months so that I could finally sail into port and bring closure to my voyage. Her knowledge, dedication and generosity have inspired me. Finally, I wish to extend my genuine gratitude to my advisor Dr. George Zimmerman who has afforded me many professional opportunities and experiences throughout my doctoral study. His words of encouragement and friendship will always be cherished.

To all of you: friends, family, colleagues, and mentors
thank you for being a part of my journey.

# 1.0 INTRODUCTION

PL 94-142, the Education of All Handicapped Children Act (EHA) of 1975, opened a new era for the education of students with disabilities, acknowledging that these students deserved a free and appropriate public education just like any other student. Since the enactment of this federal legislation, reauthorizations of the EHA under its new name, the Individuals with Disabilities Education Act (IDEA), have continued to acknowledge the rights of students with disabilities to individualized education. However, increased scrutiny of special education for lack of outcome data on student learning and achievement that would justify specialized instruction has lead to philosophical changes within IDEA.

Influenced by the No Child Left Behind (NCLB) Act of 2001, the Individuals with Disabilities Education Improvement Act of 2004 (IDEA) not only has a stronger focus on outcome measures for Individual Education Program (IEP) goals, but also stronger requirements for students with disabilities to participate in large-scale, high-stakes state accountability assessments. It states that "all children with disabilities are included in all general state and district wide assessment programs, including assessments described under section 1111 of the Elementary and Secondary Education Act of 1965." (section 612A (16)). Under this requirement, students with disabilities are expected to show progress on state standard content in math and reading just as any other student. At least 95% of students with disabilities must be tested and scores on the standard or alternate assessment are to be included in the measure of whether or not

a school is making adequate yearly progress (AYP). In addition, the scores of students with disabilities must also be disaggregated and reported as a category of their own. Failure of schools to consistently make AYP can ultimately result in school restructuring and loss of jobs (U.S. Department of Education, 2002).

Since student advancement, graduation, and teacher and school reputations can all be on the line with the results of yearly state accountability tests, a heightened focus has emerged on the adequacy of assessments to accurately measure progress. In particular, for students with disabilities, ensuring that they have every opportunity to demonstrate their skill levels within the state standards is essential if scores are to be considered comparable to the scores of students without disabilities and are to contribute to AYP in an accurate manner. Beyond the high-stakes implications of the assessments, it is useful if the assessments serve to monitor progress and inform about instructional needs. This can only be done if the assessment results are interpreted appropriately (Linn, 2002).

Of particular concern regarding assessments and students with disabilities are the effects that accommodations have on the results. Whether students with disabilities are taking the regular state assessment, a modified assessment now allowed for 2% of the population of students with disabilities (U.S. Department of Education, April 7, 2005), or the state alternate assessment appropriate for about 1% of students, they are allowed reasonable and appropriate accommodations (IDEA, 2004; U.S. Department of Education, 2002). Since state accountability tests are to have high technical quality including “validity, reliability, accessibility, objectivity, and consistency with nationally recognized and technical standards” (200.2 (b) and 200.3 (a) 1 as cited in Title I Regulations, 2003), efforts both to evaluate how accommodated conditions affect technical adequacy and to evaluate the types and appropriateness of accommodations selected for

students are essential if test scores are to be used for accurate decision-making regarding student ability and high-stakes determinations about how well schools are educating these students.

Overall, the underlying intent of inclusion of students with disabilities in these large-scale, high-stakes assessments is that students with disabilities should not be forgotten or disregarded but should be afforded the opportunity for instruction. The focus now is on the core content considered important for all students as dictated by state content standards. The requirements to focus on core content are meant to heighten a school's responsibility for educating students with disabilities and for working to help those students meet the state standards (Thurlow, Elliot, & Ysseldyke, 2003). This philosophy includes students with the most severe cognitive disabilities. These students, too, are to be taught math and reading content with the measurement of progress being through state alternate assessments that test progression on alternate content standards.

The philosophical drive to acknowledge that the education of all students is important is probably strongest at the alternate assessment level. Students with the most severe cognitive disabilities are the ones who typically would have been exempted from state testing in the past. Now, with new legislation, not only are they assessed, but they are assessed on alternate state standards that are supposed to focus on reading- and math-related skills (U.S. Department of Education, 2002). Beyond serving as a measure of educational progress for AYP, alternate assessments, since they theoretically focus on skills that may not typically have been on IEPs for this population of students in the past, serve as a catalyst for change towards this new philosophical viewpoint of NCLB. If consequential validity, or a change in what teachers are attending to in the education of their students with the most severe disabilities is to occur, alternate assessments need to be viewed as worthwhile and informative by those teachers.

First and foremost then, alternate assessments should accurately measure the intended skill base while still being flexible enough to allow for a wide variety of accommodations or modalities of communication (Gong & Marion, 2006). There are a variety of different types of alternate assessments employed by states including portfolio assessments, checklists, and performance-based assessments. Of the range of assessments, performance-based assessments most closely link to the traditional manner of testing. They contain test items that are to be administered on-demand to the student and then scored either externally or by the test administrator (Thompson & Thurlow, 2003). Because test items are prescribed (in contrast to portfolios that typically allow teachers to choose the types of tasks that will be highlighted) performance-based assessments, like standard assessments, should be screened for test items that function differently for students in different disability categories or under different accommodated conditions. In other words, there is an obligation to answer the question: do the test items for different groups of students accurately assess the intended constructs?

Students with visual impairments are a group in particular need of research regarding test item functioning, reliability, and validity on assessments including performance-based alternate assessments. While tests that incorporate principles of universal design can alleviate some of the need for accommodated conditions, these principles do not eliminate the need for accommodations within a test for all students on all test items (Thompson, Johnstone, Anderson, & Miller, 2005). Traditionally, students with visual impairments are a group who may require additional adaptations to an assessment. Since alternate assessments based on alternate standards generally test lower level reading and math skills, or precursory skills to reading and numerical literacy, many items on a performance-based assessment will contain pictures as the focus of the skill. Adaptations to these types of performance items can be particularly problematic for

students without usable vision. Furthermore, students who take alternate assessments often have multiple impairments that can limit their range of experiences. The addition of a sensory impairment such as diminished or complete loss of vision can create added limitations on the extent of experiences the student brings to learning situations (Barraga & Erin, 2001; Warren 1994). Test items, therefore, have the potential of being biased, or differentially difficult, when assessed through a different sensory channel.

To date, little research has been conducted specifically on students with visual impairments on state assessments in general, and more specifically on large-scale, high-stakes alternate assessments. Due to its low incidence nature, visual impairment as a disability category may be included in research studies within the global category of students with disabilities, but is rarely disaggregated as a category of its own. In order to assure that performance-based assessments accurately inform about the skills of students with visual impairments, both for AYP purposes and more importantly for instructional and consequential purposes, research needs to be conducted more regularly on this population of students despite the challenges of small sample sizes.

## 1.1 REASEACH QUESTIONS

This study began exploring the issues of performance-based alternate assessments for students who have visual impairments by examining use of accommodations and test item bias on the 2005 administration of the Pennsylvania Alternate System of Assessment (PASA). A descriptive analysis of accommodations made by test administrators was followed by statistical and

judgmental testing for differential item functioning (DIF) and a logical analysis of possible reasons for DIF. In particular, the following questions were explored:

Q1. Were there significant differences in the scores of students with visual impairments at different functional vision levels on the 2005 grade 3/4 or 7/8 A level PASA?

Q2. What accommodations did teachers make to adapt the 2005 PASA for students with visual impairments?

a. Are there relationships between the types of accommodations made and level of functional vision or type of test item?

b. Were there accommodations that seemed to change the intent of the skills being tested?

Q3. Were there significant differences on individual 2005 level A PASA math and reading test items at the 3/4 and 7/8 grade levels of students with visual impairments as compared to students without visual impairments who had similar ability profiles on the constructs of interest?

Q4. Considering the accommodations made and student performance on different types of test items, what are the potential reasons that some items (here denoted as "flagged" items) functioned differently?

## 1.2 DEFINITION OF TERMS

### 1.2.1 Terms

The following terms, presented throughout this dissertation, are important to understanding the issues related to alternate assessments for students with visual impairments in general, and on the PASA in particular.

1. *Accommodations:* "Changes in testing materials or procedures that enable the student with disabilities to participate in an assessment in a way that allows abilities to be addressed rather than disabilities" (Thurlow et al., 2003, p. 28).
2. *Accountability System:* Large-scale system to measure and report student and program progress. The system involves development of standards, assessment of those standards, and public reporting of results (Linn, 2002).
3. *Alternate Assessments:* Large scale assessments based on alternate content standards developed for students with the most severe cognitive disabilities who cannot participate in the standard state accountability assessment even with accommodations.
4. *Consequential Validity:* A measure of value implications and social consequences regarding whether an assessment refines the teaching and learning process, whether teachers perceive the assessment information as helpful, and whether there are unintended side effects of use of the assessment (Gersten, Keating, and Irvin, 1995 as cited in Gersten and Baker, 2002).
5. *Construct Irrelevant Variance:* Scores that result from "sources other than students' knowledge, conceptual understanding, and skill, or their ability to apply knowledge,

concepts and skills in some performance." (Taylor, 2002). The test item is measuring something other than the intended construct.

6. *Differential Item Functioning (DIF):* A procedure for determining relative difficulty of test items for different groups of individuals as compared to a reference group. DIF procedures do not interpret the reasons for differences in difficulty. Paired with logical analysis, however, DIF procedures can help identify items that contain test bias (Camilli & Shepard, 1994).
7. *Individuals with Disabilities Education Improvement Act of 2004 (IDEA):* Federal legislation that protects the rights of students with disabilities to a free and appropriate public education. IDEA as reauthorized requires that students with disabilities participate in large-scale state accountability assessments.
8. *Individual Education Program (IEP):* A yearly educational plan containing individualized goals and objectives for students in special education. IEPs are a requirement of IDEA.
9. *Internal Consistency:* A measure of reliability that evaluates to what extent test items within a scale measure the same underlying attribute or construct (U.S. Department of Education, n.d.).
10. *Modifications:* A change in test setting, timing, scheduling, presentation, or response that causes the construct of the skill being measured to become different than the intended construct during standard administration.
11. *No Child Left Behind (NCLB):* Federal legislation that requires that all states make annual yearly progress (AYP) towards all students achieving math and reading state standards by 2012. Accountability is determined through annual large-scale testing

which includes students with disabilities who may take the standard assessment, a modified assessment or an alternate assessment depending upon eligibility.

12. *Pennsylvania Alternate System of Assessment (PASA):* Performance-based alternate assessment used to assess students with the most severe cognitive disabilities for AYP purposes.
13. *Reliability:* The consistency or repeatability of a measurement. *Inter-rater reliability* refers to the consistency of observations made of the same situation between more than one observer (U.S. Department of Education, n.d.).
14. *Scoring Rubric:* Method for evaluating performance-based assessment that involves assigning a score to test items based on the level of independence of the student's response according to a specified set of decision-making rules.
15. *Validity:* The degree with which the interpretation of test scores are meaningful, useful, and appropriate (Zumbo, 1999). Validating a test refers to accumulating evidence that shows that the inferences made from test scores for different populations are appropriate (Linn, 2002).

### 1.2.2 Study Specific Acronyms

16. *Levels of Functional Vision (V, CV, NV):* Students with visual impairments were classified under one of three groupings of functional vision:

    *V:* Primarily uses vision for most tasks

    *CV:* Uses a combination of vision and other senses for most tasks

    *NV:* Uses other senses in place of vision for most tasks

# 2.0 LITERATURE REVIEW

Adapting mainstream assessments for students with visual impairments has always been a challenge. Just as IDEA requires that no single measure or assessment be used to determine a student's eligibility for special education services (614 B (2)), general recommendations have been the same for assessing students with visual impairments for any reason (Bowen & Ferrell, 2003; Bradley-Johnson, 1994) including educational performance. Since the presence of a sensory impairment such as vision loss calls into question the validity of standardized tests, educators have interpreted assessment results with caution. However, with NCLB, state accountability is based upon one annual assessment. Therefore, it has become more crucial for educators to analyze the issues surrounding the assessment of students with visual impairment in order to promote the best possible assessment situation and interpretation of results. Considerations include necessary adaptations to the test itself and adequate selection and use of accommodations. This analysis should not ignore students with visual impairments who also have additional disabilities such as students who take state alternate assessments for accountability.

## 2.1 STUDENTS WITH VISUAL IMPAIRMENTS

Students with visual impairments are a heterogeneous group. From an educational perspective, this disability category includes any student with diminished vision that adversely affects educational progress (Huebner, 2000). Students may range from being totally blind to having varying degrees of low vision. Students may be print or braille readers, congenitally (prior to two years old) or adventitiously visually impaired (after two years old), and may have stable or fluctuating/progressive eye conditions. The Office of Special Education Programs (OSEP) in its 26th Annual Report to Congress estimated that in Fall 2002, 28,598 3-21 year olds receiving special education services were identified as being visually impaired and 1,788 students were identified as deaf-blind (U.S. Department of Education, 2004).

This estimation is probably low, however, due to the fact that OSEP reporting is by primary disability only (Huebner, 2000). Students with additional disabilities may go unreported as also having a visual impairment. In fact, OSEP in its 25$^{th}$ Annual Report to Congress does report that 15% of students with disabilities age 6-12 have three or more co-concurring disabilities and 30% have two or more disabilities. Twenty-eight percent of students in the 13-17 year old age category have three or more co-concurring disabilities, and 19% have two or more disabilities (U.S. Department of Education, 2003). In addition, the 2004 American Printing House Federal Quota Census, which uses a stricter definition of visual impairment, registered 49,270 children, infant through school-age as visually impaired- approximately 37% more students than reported by OSEP. This number excludes the adult category, but includes the "other registrants" category in the total (which may not coincide exactly with the 6-21 year old category in OSEP). Many of these additional students may be students with multiple impairments whose primary disability was designated as something other than visual

impairment. In fact, Kirchner and Diament (1999) estimated that 53% of students with visual impairments had additional disabilities. It is probable that a good portion of these students with additional disabilities are the students with visual impairments who take state alternate assessments whether or not their primary disability is registered as visual impairment.

## 2.2 PERFORMANCE-BASED ALTERNATE ASSESSMENTS

The manner in which these students are assessed will depend upon the state and its alternate assessment policy since states differ on the type of alternate assessments they have developed (Thompson & Thurlow, 2003). The most common types of alternate assessments that states have used to comply with NCLB requirements include: body of evidence or portfolios, checklists or rating scales, performance-based assessments, or in some cases, a combination of strategies (Thompson & Thurlow, 2003; Quenemoen, Thompson, & Thurlow, 2003). In Pennsylvania, students take a performance-based assessment called the Pennsylvania Alternate System of Assessment (PASA).

Performance-based assessments are less common than other alternate assessment methods. Overall, approximately four states (8%) reported using this type of alternate assessment (Quenemoen et al., 2003). They are generally not paper and pencil tests but instead involve a student and teacher working through direct measures of skills with manipulatives. They can be time-intensive, so they generally survey a smaller range of skills than a portfolio that is collected over time would (Roeber, 2002). Performance-based assessments probably provide more accurate results than checklists because the student is actually performing the skill. In some states, the teacher rates the student's performance after each skill is executed based on accuracy

and level of independence (Quenemoen et al., 2003). In other states, including Pennsylvania, outside scorers make that determination.

### 2.2.1 Scoring and Focus of Alternate Assessments

The criteria for evaluation of student performance can also vary between states even if they have the same type of alternate assessment. With NCLB, academic content standards should have become the focus for most states in their alternate assessments (Quenemoen, Rigney, & Thurlow, 2002; Thompson & Thurlow, 2003). By 2003, 80% of states were found to have aligned their alternate assessments with state content standards by grade-level or through expanded standards. Some of these states continue to incorporate functional skills into the assessment as well (Thompson & Thurlow, 2003).

In order to score achievement, the majority of states in 2003 (40 states) used some sort of rubric, particularly those states using portfolio or performance-based assessment. The PASA falls into this category. The contents and focus of the rubrics can vary. For example, alternate assessments may focus on student performance (student criteria), on program opportunities (system criteria), or both (Quenemoen et al., 2003; Roeber, 2002). Student criteria measure actual student performance but can range from just measuring accuracy (right or wrong) to judging student level of independence, level of progress, and/or ability to generalize the skills. System criteria evaluate the quality of the system including whether students were provided opportunities to perform certain skills such as evaluating their own work, whether students were provided instruction in multiple settings, and whether they were provided with appropriate supports (Quenemoen et al., 2003).

In order to gain better insight into the interpretability of results on an alternate assessment for a particular group of students, such as students with visual impairments, details about the assessment process within that state must be gathered, some of which can be found within states' technical manuals. Technical manuals will report on validity and reliability processes as well as how cut-off scores for performance levels were determined.

### 2.2.2 Specifics of the PASA

#### 2.2.2.1 Structure

As just mentioned, Pennsylvania has chosen to assess students with the most severe cognitive disabilities in compliance with IDEA and NCLB using an alternate performance-based assessment. The Pennsylvania Alternate System of Assessment (PASA) consists of approximately 20 reading and 20 math performance items that are videotaped or recorded as narrative notes and sent in for scoring. Up to an additional five assessment items may be included in each assessment as test items that are not factored into the students' reported score but that inform on how new items function. Students are assessed in grades 3-8 and in grade 11. Grades 3 and 4 take the same assessment with different achievement criteria as do grades 5/6 and 7/8. The philosophy of the PASA is that:

> Student participation…underscores school district accountability for holding high expectations for all students; establishes school district accountability for teaching, challenging and supporting every student's accomplishment of maximum potential knowledge and skills; provides information to assist teachers, parents and students in evaluating student progress and performance relative to

the general education standards and curricula; and provides information to assist teachers in curriculum and instruction decision-making (www.pasaassessment.org).

As with most assessments, it is meant to be a snapshot of student performance on reading and math related skills based on alternate standards derived from the state content standards. The assessment contains three testing levels: A, B and C. The assessment is multiple-choice in nature for the majority of test items, providing an array of choices from which the student selects a response. Level A contains the least complex skills, level B contains intermediate skills, and level C contains the most complex skills associated with the alternate standards at each grade level. As one moves up in grade and/or level, materials move from objects to pictures to words, and the range of answer choices increases or become open-ended. For example, in reading at level A, a third grade student may be asked to find an object named out of three choices with two choices being very different from the target (e.g. “find the glove” with choices of teddy bear, folder, and glove). At level B, a student may be asked to select a picture named from four choices. Some pictures, depending on the grade level, may be similar in appearance (e.g. “find the plate” with picture choices of plate, clock, cookie and pizza). At level C, a student may be asked to select a word named from five choices (e.g. “where is the word math?” on a school schedule with words: science, math, reading, lunch, writing). Depending on the grade level, words may all have the same beginning and ending letter. At higher levels, one type of open-ended response question might ask the student to recall what was just read (e.g. “What are two things you just read about turtles?”).

In the 2005 assessment, teachers selected the assessment level most appropriate for the student and also filled out a skills checklist on each student that links directly to the content of

the PASA. Appendix A provides a specific list of the tasks and skills assessed on the 2005 administration of the PASA for grades 3/4 and 7/8 (PASA State Report, 2005). Appendix B contains one example of the skills checklist at both of these grade levels. Interestingly, teachers selected level A most frequently for students with visual impairments. Appendix C gives examples of a reading and math test item at level A for grades 3/4 and 7/8.

The 2005 assessment was sent to school districts and special schools with a testing window of approximately six weeks. The amount of time individual test administrators had to prepare and administer the assessment varied among districts depending on resources and how quickly assessment packets were distributed. A test administrator's manual was sent with the materials that provided guidance on administering the assessment, optimally videotaping the assessment, and selecting accommodations for different disabilities.

#### 2.2.2.2 Scoring

The PASA assessments are scored by teams of two teachers who have been trained and screened for reliability. Scoring is based on a 0-5 scale and reflects the accuracy of student responses as well as the level of independence. Scores on a skill are lowered when multiple prompts are needed or when modifications must be made to make the skill easier. Performance on the PASA is based on a general scoring rubric (Table 1) that indicates the level of independence at which the student was able to perform the skills being assessed in each test item.

**Table 1: 2005 PASA Scoring Rubric**

| 5 | 4 | 3 | 2 | 1 | 0 |
|---|---|---|---|---|---|
| Performed **correctly** and **independently** with initial instruction only **and** demonstrated target skill | Performed **correctly** with 1 or more additional **prompts**, redirections or corrections **and** demonstrated targeted skill | Performed **correctly**, but on an easier (**modified**) version of the targeted skill | Performed **incorrectly**, or Demonstrated **skill different** from the targeted skill **or** Performed skill when the **correct response** was **ensured** | **Passively participated**; did not demonstrate targeted skill **and** Assessor ensured correct response **or** **Component not completed** by student or assessor | **Not observed**: item **omitted** or item **not recorded** |

(www.pasaassessment.org)

To illustrate the scoring process, consider a math test item asking the student to select the biggest circle from a choice of three. The student would receive a score of a five if the correct circle was selected after the first time of being asked to do so. The student would receive a score of a four if the teacher had to repeat the prompt of, “find the biggest circle”, or if the student selected wrong and the teacher said, “try again” after which the student selected correctly with this additional prompt or multiple versions of the prompt that did not change the skill level. A score of a three could be earned in various ways, but might involve the teacher making the task easier by removing one of the three circles to choose from, leaving only two. If the student selected the wrong circle and the teacher then showed the student which was the biggest, the student would receive a score of a two. A score of one would occur on this item if after asking the student to, “find the biggest circle”, the student shows no engagement in the task and the teacher takes the student’s hand and puts it on the largest circle. A zero would be given to this item if the teacher skipped the item completely. After scoring all the items in this fashion, final scores are then weighted depending on the difficulty level of the task (A, B, or C) and turned into scaled scores and proficiency levels for reporting.

#### 2.2.2.3 Technical Adequacy

In order to understand how accurately the PASA informs about a student's performance, reliability in scoring and validity of the test items themselves are important. Technical adequacy specifics of the 2005 PASA are reported in Chapter 3. In general, however, PASA technical supplements report that judgmental item reviews were conducted by special education and technical experts to screen items for obvious bias against certain sub-groups. In particular, screening is conducted for gender and setting (rural, urban, suburban) bias. Items were also reviewed to determine if they contained contexts with which students with the most severe cognitive disabilities as a whole might have had direct experience. In addition, items were piloted with at least one appropriately matched student at each grade and test level to review for problems in administration. No specific item review was conducted for students with visual impairments. In order to gain a better understanding of the aspects of technical adequacy that may affect the interpretation and use of assessment results for students with visual impairments, it is first important to review the concepts of validity, reliability, and test bias.

## 2.3 VALIDITY AND RELIABILITY

### 2.3.1 Validity

Validity is a central issue in determining the appropriateness of an assessment including alternate assessments (Ryan & DeMark, 2002). It refers to whether the test actually measures what it claims to measure (Bradley-Johnson, 1994; Geisinger, 1994; Mcloughlin & Lewis, 1994;) and, more specifically for state accountability tests, refers to whether the assessment appropriately

measures proficiency on each of the established standards. Traditionally, validity was separated into different types: content related, criterion-related, and construct related (Linn, 2002). It was common to select one of the methods to report an assessment's validity (Zumbo, 1999). However, the contemporary focus on validity is no longer on the measure itself, but on how valid the inferences are that are made from the outcomes of the measure. It is the consequences of test decisions and the use of those decisions that has become the focus of validation (Zumbo, 1999; Messick, 1995 as cited in Linn, 2002). This conceptual framework of validity is particularly important for high-stakes, large-scale accountability assessments where decisions about student progress and program effectiveness are being made from annual test scores.

### 2.3.2 Integrated View of Validity

The process of establishing validity of a test involves both procedural and empirical documentation of evidence that a test measures what it claims. The quality of this evidence helps to determine how a student's score on the test can be interpreted (US Department of Education, n.d.). Based on the 1999 *Test Standards*, Linn (2002) states:

> The trend is toward an integrated view of validity as a unitary concept that incorporates the use of a variety of types of evidence and logical analyses to make an evaluation of the degree to which a specific use or interpretation of assessment results is justified. Evidence would include a consideration of content relevance and representativeness as well as correlations of scores with other variables. This information might include judgments about the degree of alignment of the test with content standards. Correlations of student characteristics and instructional experiences (e.g. measures of opportunity to learn the material assessed) as well

> as potential criterion measures such as teacher grades would be considered relevant as well. In addition, the relevant evidence would be expected to include information that in the past would have been associated with construct validation, such as information about the internal structure of the test and the cognitive processes used by students responding to test items (p. 32).

With this integrated view, an array of evidence that comes out of the traditional categories of content, criterion, and construct-related validation informs about the decisions that can be made from assessment measures (Linn, 2002). For example, content investigated by reviewing documentation that the assessment was developed using appropriate testing standards including having experts who have verified that the test items match to specific standard objectives as intended would serve as one piece of evidence of an assessment's validity. Validity may be further investigated by considering whether the cognitive processes in which the students are expected to engage are truly the ones that are being measured. This type of validity can be determined through content experts as well, but is more strongly supported when students are asked to engage in *think aloud* procedures to better understand actual cognitive processing that takes place during testing (US Department of Education, n.d.). Content validity may also be strengthened when constructs being measured on the test are shown to have a relationship with other measures that assess the same constructs (US Department of Education, n.d.).

### 2.3.3 Reliability

Reliability goes hand in hand with validity. Evidence that is gathered to validate the specific use or interpretation of assessment results and the assessment results themselves must be reliable. Zumbo (1999) views reliability within the new conceptual framework of validity as an issue of

measurement precision. As little measurement error as possible is important when using measures to make inferences. Reliability is measured in different ways, but the general concept is if a student were tested and then re-tested after a short period of time, the results would be the same with little measurement error (Mcloughlin & Lewis, 1994). Reliability can be affected by both internal and external variables and needs to be established for diverse groups of individuals taking the test such as for students with disabilities. Internal variables can include student background characteristics and motivation. External variables can include variations in test administration and evaluator bias (US Department of Education, n.d.). The types of reliability that are important depend upon the type of assessment being given. For the PASA, inter-rater reliability is of importance since subjectivity or individual judgment in scoring can affect the variability of scores (US Department of Education, n.d.).

## 2.4 CONSEQUENCES OF ASSESSMENT

As just discussed, the contemporary viewpoint of validity is to inform on the appropriateness of the decisions made based on test results, and reliability informs on measurement precision when making those decisions. Since the ultimate concern is regarding how test results are used for decision-making, consequences regarding the use of the test must be investigated. Validity of an assessment, like reliability, depends upon the level of standard administration that was conducted. The more standard the administration, the easier it is to attribute student scores to actual differences in knowledge or ability. Deviation from standard procedures, such as when accommodations or alternate test formats are used, can affect validity and the comparability of scores and must be investigated carefully (Geisinger, 1994).

This is an important consideration for students taking alternate assessments. Students with the most severe cognitive disabilities are a heterogeneous group of students who often require a wide variety of accommodations to access materials and learning. Additionally, students, such as students with visual impairments, taking alternate performance-based assessments may require adaptations to the test questions themselves. The assessments, then, should report on how the assessment functions for different sub-groups of students as part of the body of evidence that validates the assessment and the appropriateness of its use for different decision-making. For students with disabilities, among other investigations, consideration of accommodation selection and its impact on validity and consideration of test bias are all important.

## 2.5 ACCOMMODATIONS AND VALIDITY

Since many students with disabilities use accommodations during high-stakes testing, states must consider the effects of those accommodations on the validity and comparability of results to students who do not use accommodations or who use different accommodations. As defined by Thurlow et al. (2003), accommodations are "changes in testing materials or procedures that enable students with disabilities to participate in an assessment in a way that allows abilities to be addressed rather than disabilities" (p. 28). Compared to a modification which is considered a change that causes the construct of the skill being measured to become different than the intended construct during standard administration (ASES, n.d.), allowable accommodations should positively impact the performance of the student by compensating for the disability without giving an unfair advantage over students without disabilities. That is, the

accommodation should have little-to-no impact on the performance of students without the disability. Finally, the accommodation should not alter the psychometric properties of the assessment (the construct being tested) (Thurlow, McGrew, et al., 2000; Tindal & Fuchs, 2000). States vary on the types of accommodations that they allow and in the way they report tests taken under accommodated conditions (Thurlow, House, Boys, Scott, & Ysseldyke, 2000). This is an indication that consensus has not been reached on the effects of different types of accommodations for individual disability groups (Thurlow & Bolt, 2001). While some research exists regarding the effects of accommodations, the research base is still insufficient to draw any conclusions. The quality and type of research also varies, which limits the ability to generalize the results (Tindal & Fuchs, 2000).

Thurlow, McGrew, et al. (2000) identify imperative questions that future accommodation research needs to address. These questions include determining if items under standard and accommodated conditions are comparable (differential item functioning), whether scores obtained under standard and accommodated conditions measure the same abilities or constructs, whether scores under the two different conditions correlate similarly to outcome criteria (criterion-related validity), and whether the cut-off score used to make decisions about students should be the same under each condition.

### 2.5.1 Types of Accommodations

Thurlow et al. (2003) describe six categories for accommodations: setting, timing, scheduling, presentation, response and other. Setting refers to changes that are made to the place where the assessment is given. Some examples of accommodations under this category would include administering the test in a separate room, providing special lighting, or seating the student close

to the test administrator. Timing refers to changes in the "duration or organization of time during testing" (p.54). Accommodations under this category could include extended time, additional breaks and multiple sessions. Scheduling refers to changes in the time or order of test administration. This category includes rearranging subtests, and giving the test at a different time of day. Presentation consists of format alterations, procedure changes, and use of assistive devices. Format alterations include providing the test in braille or large print, increasing the spacing, and providing instructional picture cues in the test booklet. Changes to procedure include simplifying the directions, giving extra examples, using a reader, or answering questions about items to clarify. Use of assistive devices includes use of low vision magnification equipment or a computer, use of audio taped directions, or use of a template to reduce the amount of visible print. Other accommodations include encouragement during testing and instruction in test-taking skills.

### 2.5.2 State Policies on Accommodations

All states have accommodations policies, but the extent of those policies varies greatly from simple lists of acceptable accommodations to documents that provide guidance on selection procedures. States also vary on what accommodations are allowed with and without restrictions and with and without consequences to the scoring and reporting of the assessment results (Clapper, Morse, Lazarus, Thompson, & Thurlow, 2005). Of the six classifications of accommodations, those falling under the category of presentation are generally the ones that are the most controversial (Clapper et al., 2005; Thurlow et al., 2003). Overall, an analysis of state policies on accommodations reveals read aloud, calculator, spellchecker, and proctor/scribe to be the most controversial accommodations (Clapper et al., 2005).

### 2.5.3 Selection of Accommodations

Because some accommodations are identified as controversial by certain states, the accommodations needed by each individual student have to be appropriately identified. Thurlow et al. (2003) emphasize the need for accommodations used in testing to be based on those accommodations used in instruction. New accommodations should not be introduced at the time of the assessment, and care must be given not to over-accommodate. A majority of states seem to agree with this process whether or not they provide direct supports to teachers to ensure it happens. Forty-five states include used for classroom instruction as a variable to be used in decision-making (Clapper et al., 2005). The process should not start with a review of what accommodations are accepted without restriction by states, but be informed by IEP team decisions about what is needed by the individual student based on the intent, type, and content of the assessment (Thurlow et al., 2003). Fuchs et al. (2000) further advocate for the use of data collection by teachers to confirm their beliefs about the benefits of chosen accommodations. Once accommodations are selected, then they should be compared with the list of acceptable accommodations by the state and proper measures taken according to state policy if non-standard accommodations are necessary (Thurlow et al., 2003). In fact, 35 states include maintains validity of test and resulting scores as a variable for decision-making (Clapper et al., 2005). This requirement highlights the question about whether teachers and IEP teams have the support and means to make a designation about validity. It was already noted that state policies vary in the level of detail provided about accommodation selection (Clapper et al., 2005). In addition, study findings indicate that the quality of teacher decision-making regarding accommodations may vary widely. Destefano, Shriner, & Lloyd (2001) found that teacher decision-making about accommodations improved after training. If teachers are not receiving such training, the accuracy

and carefulness of accommodation selection is called into question. It is important to understand how much knowledge teachers have about assessment and underlying assessment issues and concepts (Tindal & Fuchs, 2000). Wise, Lukin and Roos (1991 as cited in Tindal & Fuchs, 2000) suggest that most teachers learn about testing from trial and error in the classroom, and that many institutions offering initial teacher certification do not require an assessment course. Helwig and Tindal (2003) conducted a study on teacher decision-making in appropriately selecting the read-aloud accommodation for students on mathematics tests. They found that teachers were only as accurate as chance in determining if a student would benefit from the accommodation. This finding again focuses attention on the need for teachers to have data collection skills and training in assessment issues to improve decision-making skills about accommodations. Thompson, Lazarus, Clapper, and Thurlow (2004) document essential knowledge and skills that teachers need to support achievement of students with disabilities, and highlight national and state standard models that are being incorporated into some teacher preparation programs to support new teachers in a high-stakes accountability teaching culture. These standards include ensuring the ability of a teacher to make appropriate participation and accommodation decisions for students with disabilities.

No direct studies about accommodation selection for students with visual impairments are available; however, it can probably be assumed that the same issues of teacher under-preparedness to make quality selection decisions based on data of classroom accommodations applies to teachers of the visually impaired (TVIs) and the respective IEP teams, at least in some cases. Literature has noted, for example, the tendency for TVIs to select large print for students without substantial reasons (Lussenhop & Corn, 2002) when use of optical devices may actually improve silent reading speed and silent comprehension (Corn et al., 2002). Considering that

some teachers may be over-selecting accommodations, inaccurately choosing accommodations, or inaccurately judging the affect of accommodations on validity, it is even more important to review what research indicates so far regarding accommodations for students with visual impairments and their affects on assessment.

### 2.5.4 Accommodations for Students with Visual Impairments

Seemingly, less concern is projected in literature about selection of accommodations for students with sensory impairments because the need is obvious to the public as compared to accommodations being selected for a student with a learning disability where the disability, such as difficulty in reading, is closely related to the construct being tested (Elliott, McKevitt, & Kettler, 2002; Phillips, 1994; Thompson, Blount, & Thurlow, 2002). This lack of concern about accommodations for students with visual impairment is probably based on people thinking mostly about large print or braille versions of a test (presentation). Whereas these particular accommodations seem more pure for students with visual impairments, there is still a need for research in the area. Thompson, Blount, et al. (2002) state the need precisely when they say:

Although it is important to focus research on the largest number of students affected by accommodations use, additional research is needed on accommodations by students with visual, hearing, and physical disabilities. These students are smaller in number than those with learning disabilities, but often have very complex accommodation needs, including braille, sign language interpretation, and assistive technology" (p. 17).

Considering that teachers and IEP teams may not be well-trained enough in their decision-making, and that, while necessary to allow access, accommodations to compensate for a visual impairment, including large print and braille, can still have an effect on test validity and

reliability, additional research is essential to overall understanding of the challenges in assessing students with visual impairments. Furthermore, while there are some accommodations that are vision specific, a student with a visual impairment may be utilizing many of the same accommodations as other students, including read aloud accommodations that are controversial. Yet, research on these controversial accommodations generally do not include a group of students with visual impairments. In addition, although literature downplays concern about accommodations for visual impairment, state policies do vary on accommodation decisions, including those that are vision specific.

The American Printing House for the Blind (APH) Test Access guide (Allman, 2004) acknowledges that, "braille, large print, and audio (taped or a reader) are accommodations that some students with visual impairments will use interchangeably" (p. 25). In addition, use of the abacus, braillewriter, other assistive technology devices to read or produce written work (e.g. screen readers, braille note takers), magnification systems and optical devices, and special lighting are listed as some of the additional vision specific accommodations. While not vision specific, extended time is also a typical accommodation selected for students with visual impairments based on studies that suggest that alternate test formats of braille, large print or audio could require anywhere from 50% to 200% more time (Lowenfeld, Abel & Hatlen, 1969; Nolan, 1966; Wetzel & Knowlton, 2000 as cited in Allman, 2004). In addition, some students may use a computer, require extra breaks, need a separate testing location, mark answers in the test booklet, or give answers orally to the test administrator.

As for other disability groups, certain accommodations may be considered controversial for students with visual impairments under certain circumstances. Extended time, for example, while generally accepted more often for students with visual impairments than for other

disabilities, can still be controversial or affect test validity. Allman (2002) conducted a survey of 33 states about their inclusion of students with visual impairments in state accountability testing. She asked states to indicate from a list of accommodations that might be selected for students with visual impairments, which would be allowed. Using the random criterion of less than half of the states allowing the accommodation to analyze the findings, Table 2 lists the accommodations that emerged as the most controversial:

**Table 2: Accommodations Allowed by Less Than Half of 33 States Reporting**

| Accommodation | Number of States Allowing | Number of States Not Allowing |
|---|---|---|
| Enlarging at the local level | 11 | 16 |
| Brailling at the local level | 8 | 21 |
| Use of computer presentation of the test | 12 | 18 |
| Use of scanner or screen reader/voice output on a reading comprehension section of the test | 13 | 17 |
| Use of paraphrasing or simplification of stimulus material or test questions | 14 | 18 |
| Use of oral reader or tape-recorded test from math test | 13 | 17 |
| Use of dictionary for writing test | 13 | 17 |
| Use of spell check on tests where spelling and writing will be scored | 8 | 23 |

(Allman, 2002)

A pattern of 13 states, perhaps the same states, allow the use of read-aloud accommodations for comprehension tests for students with visual impairments. This seems like a particularly controversial accommodation when considering test validity and comparability of scores. The field of visual impairment is divided in its opinion of whether access to print via audio accommodations can be interpreted as literacy (Hatlen, 2003; Koenig & Holbrook, 1995).

Since it is possible some teachers and IEP teams will select an audio format for reading comprehension tests, research is particularly needed in this area. Allman (2002) also notes that from states that gave two different perspectives on allowable accommodations, teacher responses sometimes differed from state level responses, again, highlighting possible barriers to appropriate selection of accommodations by teachers.

In addition to Allman's 2002 findings, Clapper et al. (2005) in the area of material accommodations record only 17 states allowing the use of the abacus without restrictions- generally a vision specific accommodation- and five states allowing it in some circumstances. Manipulatives, which may be a selection for some students, are listed as allowed in only 11 states, allowed under certain circumstances in two states, allowed under certain circumstances with implications for scoring in two states and prohibited in one state. Some states, obviously, did not have a policy on all the accommodations investigated. Even braille and large print were not unanimously reported as allowed without restrictions by all states. Large print was allowed by 47 states, but one state each either allowed it under certain circumstances or allowed it with implications for scoring. Braille was listed as allowed in 38 states with five states allowing it with implications for scoring and six states allowing it in certain circumstances either with or without scoring implications (three each). These findings show that some states may be more aware of the possibility that large print and braille versions of a test may change certain constructs and should be interpreted cautiously.

### 2.5.5 Accommodations and Alternate Assessments

#### 2.5.5.1 General Issues

Issues of accommodations will differ slightly for alternate assessments. Many accommodation studies focus on comparison between accommodated conditions and standard conditions of students without disabilities. For alternate assessments for students with the most severe cognitive disabilities, there will be no students without disabilities taking the assessment with whom to compare. It is still important, however, to understand how various accommodations affect the results and interpretability of the assessment with a focus on how accommodations affect the construct being assessed. In many ways, investigations of accommodations on alternate assessments will be more complex than for standard state accountability assessments. For performance-based assessments, standard administration may vary more widely because of the severity of the disabilities of the students who take alternate assessments as well as the likelihood that students will use more than one accommodation, making the effects harder to research. Also, for students with visual impairments, particularly those students without usable vision, test items on alternate performance-based assessments may be more challenging to adapt or accommodate without changing the intent of the skill being measured because there may be many picture-based items representing pre-literacy skills. The format of test items may also contain bias for this population of students.

#### 2.5.5.2 Accommodations and the PASA

Accommodations are allowed on the PASA without penalty as long as they do not change the intent of the skill. Since students taking the alternate assessment are a heterogeneous population, flexibility is given for adaptations to the test. It is acceptable for teachers, for

example, to replace suggested objects with other objects if they better fit a student's range of experience as long as the substitutions do not change the level of difficulty of the skill. It is also acceptable for teachers to make modifications (for a score of no higher than 3) if those modifications are truly what a student needs to demonstrate some level of the skill. While the flexibility is important to best accommodate student needs, it does add another element to the assessment: teacher accuracy in execution and judgment in adapting and accommodating.

A teacher, when adapting the PASA to suit her student, ideally, must go through a thought process such as that illustrated in Figure 1.

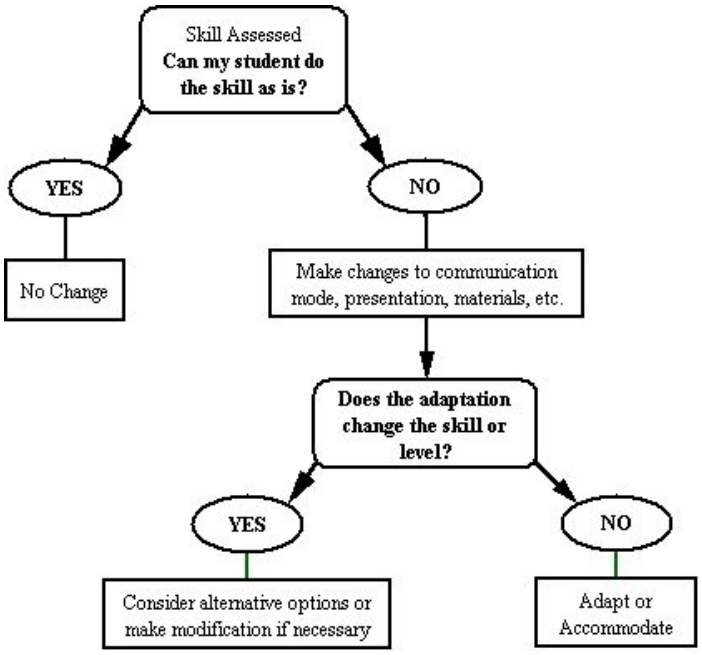



**Figure 1: Decision-making for adapting PASA skills**

The ability of teachers to make accurate decisions in order to accommodate as necessary depends on how well the teacher understands the construct of the skill, and to an extent, how invested the teacher is in the process. For teachers of students with visual impairments, more

adaptations will be needed particularly at the B level where the skill assessments are very picture based. The amount and type of adaptation will depend on the level of the student's vision, level of experiential background, as well as additional physical and communication needs. For skills teachers determine are biased, adaptation may be more extensive to support the student's needs. However, there is potential for teacher error by making changes that change the intent of the skill. Some skills, even, may be too hard to adapt without changing the intent.

## 2.6 BIAS AND VALIDITY

For diverse populations such as students with visual impairments, validity cannot be assessed without considering bias. Accommodations play a direct role in gauging the impact of potential bias. Tests should be produced to be as culture-fair as possible. This means that the items are written to minimize factors that would lower the performance of individuals from diverse groups (Gonzales, 1982 as cited in Mcloughlin & Lewis, 1994). "When an assessment does not reflect the experiences, linguistic, cultural or cognitive styles of the examinee, validity can be compromised" (Geisinger, 1994, p. 62). A culture-free test, theoretically, would allow results to be interpreted equally across diverse populations (Mcloughlin & Lewis, 1994). The question of what is meant by nondiscriminatory assessment must be explored (Mcloughlin & Lewis, 1994). True bias, also known as systematic error or construct irrelevant variance, introduces a factor unrelated to the intended construct that unfairly affects the score either positively or negatively (US Department of Education, n.d.). The Standards for Psychological Testing (1999) acknowledge different levels of testing fairness, some more unanimously agreed-upon than others. Considerations of fairness relate to the test itself and the outside conditions. Opportunity

to learn is one condition for fairness that is harder to gauge. If a student did not have the opportunity to learn material being tested (as compared to having had the opportunity but failing to learn), application of the test score could be unfair for certain purposes. It is important to consider whether certain items on a test are differentially more difficult for certain groups of students with disabilities, even under accommodated conditions, and whether they are testing the same constructs.

Use of accommodations or the nature of a disability itself could mean that different cognitive processes are being utilized on test items than what was intended, resulting in a different construct being assessed. This is of particular concern for students with visual impairments when test items are being accessed through a different sensory channel from the standard administration.

In fact, limited data suggests that students with visual impairments tend to score lower on assessments than individuals without disabilities. Jackson (2003) reported that a group of students with visual impairments taking the Arizona state accountability assessment on-level scored one to two stanines lower than their peers. Small sample size could affect this comparison, but the trend of lower assessment results is also indicated in data gathered on standard state accountability tests by the National Center on Low Incidence Disabilities (NCLID) and in the Special Education Elementary Longitudinal Study (SEELS) data. The limited NCLID data from 10 states indicated an average of 15% fewer students proficient in reading as compared to students without disabilities and 20% fewer in math. When compared to non-disabled peers based on Woodcock Johnson norms, students with visual impairments were found to be about 0.4 to 0.8 years behind (Wagner & Blackorby, 2004, p. 17).

Knowing that the appropriate use and selection of accommodations can affect assessment outcomes and that accessing test items through a different sensory or limited sensory channel can change the intended cognitive processes being engaged during a test item, it is very important to investigate test bias for students with visual impairments on standard and alternate assessments before interpreting the results. Are the differences in performance actual differences in knowledge or ability, or are the differences due to construct irrelevant variance?

### 2.6.1 Investigating Test Bias

There are different manners by which items can be analyzed for test bias- judgmental and statistical (Camilli & Shepard, 1994). Judgmental methods are often used during test development. A panel of reviewers with expertise in the areas of interest will review test items for those that might be more difficult for the sub-group of students of interest. As part of this process, it is recommended that test developers consider the use of principles of universal design to not only screen for bias, but also to reduce the amount of accommodations that may need to be utilized during testing (Thompson, Johnstone, Anderson, & Miller, 2005).

### 2.6.2 Judgmental Methods

#### 2.6.2.1 Universal Design

The National Center on Educational Outcomes (NCEO) advocates for large-scale assessments to be produced using principles of Universal Design (Thompson & Thurlow, 2002). Universal Design, as it relates to assessment, is based on the belief that tests should be accessible in their regular administration to the widest range of students possible (Thompson et al., 2005;

Thompson, Johnstone, & Thurlow, 2002). With good preparation during development, aspects of accessibility in the presentation can be present without the need for separate accommodations. Of course, some students will still require additional accommodations (such as braille), but with good design up front, fewer accommodations may be needed and/or adaptations in medium will be easier.

A study conducted by Johnstone (2003) of 231 sixth graders from under-performing schools or populations found that students scored significantly higher on a test designed with Universal Design principles than on the original test without the principles. Students reported that the universally designed test was more readable and understandable, that they recognized material they had learned in class more easily, and that unlimited time was helpful. Johnstone's (2003) study incorporated elements identified by Thompson and Thurlow (2002) as being directly applicable to assessments. These elements are described in Table 3 and include ensuring that constructs are well defined, items are non-biased, and the test is legible, clear, and concise.

**Table 3: Elements of Universally Designed Assessments**

| Element | Components |
|---|---|
| Inclusive Assessment Population | Tests designed include every student in its design and field testing |
| Clearly Defined Constructs | Clear definitions of constructs to facilitate removal of construct irrelevant cognitive, sensory, emotional and physical barriers |
| Accessible, Non-Biased Items | Accessibility built into the items from the beginning. Bias review procedures used to ensure quality of items- developed by individuals who understand student characteristics |
| Amenable to Accommodations | Facilitates use of needed accommodations (e.g. all items can be brailled) and is compatible with a variety of accommodations |
| Simple, Clear, and Intuitive Instructions and Procedures | Instructions should be easy to understand |
| Maximum Readability and Comprehensibility | Features of maximum readability are considered including student background, sentence difficulty, and organization of text. |
| Maximum Legibility | Characteristics are applied to text that supports decipherability including contrast, type size, spacing , leading, typeface, justification, line length, blank space, graphs and tables, illustrations, and response formats |

(Thompson & Thurlow, 2002)

#### 2.6.2.2 Universal Design and Visual Impairment

In addition to the principles of Universal Design described by Thompson and Thurlow (2002), The American Printing House for the Blind's Accessible Test Department has developed a guide for test publishers, developers, and state assessment personnel that delineates guidelines for creating accessible tests in large print, braille, and audio format (Allman, 2004). The guide includes suggestions about elements that should be included in a contract with test developers to ensure that large print and braille versions of the assessment (including practice tests) are accurate and available at the same time as standard versions. It also includes suggestions for the test development teams, and specific guidelines for creating the test in alternate media. Suggestions are based on research, best-practice, and principles of Universal Design.

In terms of Universal Design, APH does state that some of its guidelines for constructing a test for braille, large print, or audio format may not adhere to the specific recommendations of universal design or test publisher policies for a larger population but are specific to the needs of students accessing a particular adapted medium. Thompson et al. (2005) also notes that the process of Universal Design is challenging. Sometimes changes to a test that benefit one group of students may make it less accessible to another group of students. Theoretically, however, tests that were originally developed under Universal Design principles should have fewer items that need to be deleted or substituted because it has already been screened for *non-bias* and for *amenability for accommodations* in which case it would take into account APH test access principles. Johnstone (2003) mentions some features that can be built into a test that would facilitate the correspondence of the braille and standard test. These features include "avoiding the use of construct irrelevant graphs or pictures, avoiding vertical or diagonal text, not placing keys or legends in locations such as at the bottom where they are more difficult to locate in braille,

avoiding items that depend on reading graphic representations without accompanying verbal or text descriptions, and removing distracting pictures that are not needed to accomplish the task" (Johnstone, 2003, p. 7).

Among other suggestions in the APH test access guide, test developers are encouraged to only delete or substitute items if they cannot be provided in braille, tactile graphics, large print, or audio format without significantly changing the items and their intent. In order to facilitate development of non-biased items from the beginning, APH, just as Thompson and Thurlow (2002) mention, encourages that a person with specialization in visual impairments be involved in the item development. For universally accessible tests to be accomplished, the development team needs to contain a variety of people with different expertise. Ideally, representatives that understand cultural and ethnic differences, and those who understand the nature of different disabilities that students being tested may have all should be part of the test development in the early stages (Thompson et al., 2005). In addition, representatives who understand measurement and the content and constructs being tested must be present in order to help determine if decisions regarding universal design and bias still maintain the intended construct.

In addition to having a diverse group of experts involved with test item decision making, Thompson et al. (2005) propose a set of considerations based on the applicable elements of Universal Design for assessments. The considerations are meant to be a more comprehensive checklist to support test development teams in determining if items are universally designed. The checklist was reviewed through a Delphi study by a group of experts representing content areas and the fields of assessment, assistive technology, computer-based testing, second language acquisition and special education. The revised checklist contains questions to consider for each test item with sub-categories under each major topic area and notes that help clarify sub-areas

and/or indicate potential challenges that may arise when using the considerations. For example, one question area involves whether the test has clear visuals when essential to the item. This section of the checklist includes determining if item visuals are needed to answer the question, if the visuals have clearly defined features with minimal use of grayscale or shading, if there is sufficient contrast between colors, and whether visuals are labeled. Notes include reminding the reviewers that labels on pictures are helpful even if the picture seems obvious, and that there may be instances when grayscale and shading are appropriate to provide relevant information (Thompson et al., 2005). When using the considerations during the review process, the authors recommend that in addition to including disability, technology, and language acquisition experts in test item reviews, professional development for both item developers and reviewers regarding the considerations for universal design is necessary. They also recommend that items should be tried out with students and field tested in accommodated formats. They further acknowledge that the considerations checklist can help support open discussion about test design throughout the whole development process, but by no means does a universally designed test guarantee accessibility to all students. However, states that incorporate an item review process such as that indicated by NCEO and APH, are more likely to create assessments that produce results that are meaningful for a wider range of students.

#### 2.6.2.3 Test Item Review Committees

Thompson and Thurlow (2003), in a report on special education outcomes, surveyed all 50 states on their approaches to achieving universally designed assessments. Of the 22 states that reported having a disability representative on the assessment bias review committee, only 10 had a representative for visual impairment. In addition, only 14 states reported training test developers and only 17 reported training test item reviewers. The relatively small percentage of

states providing training or coordinating a team that contains a representative who understands the different disabilities resurfaces the question about how prepared individuals are who make decisions about test items.

Of the states that reported having a disability representative in visual impairment, only one state, Minnesota, has reported on its bias review process for students with visual impairments in particular (Knowlton, Seeling, Martin, & Archer, 2003). Minnesota's review committee consists of five members whose combined experience includes knowledge about testing laws, test design, test construction and testing timelines; experience teaching students who are blind or visually impaired; certification in braille transcribing including Nemeth code (math braille) and tactile graphics; and knowledge of state resources for converting tests to braille or large print. The five individuals conduct reviews of each item, of alternate formats (braille, large print, or audio) including administration instructions, of response formats, and of accommodations. During the item review, the committee considers the text content, accompanying graphics, and procedures for student response when deciding if an item is potentially biased both in its original format and alternate formats. The committee also makes decisions about the types of accommodations allowed (amount of extra time, adaptive equipment, etc.). Minnesota's experience in reviewing tests, while not reporting specific data of how well the process has worked, yielded similar recommendations to those already presented in this review for success in creating un-biased or universally designed tests for students with visual impairments. Insights included recognition that test item order in alternate format may need to be rearranged since some items (graphically based) considered easier for the general population of students may be more difficult for students with visual impairments. They also recognized that test administrators are the final control for the adequacy of the test and need a way to report

discrepancies in alternate formats that occur, and that test item review needs to happen well in advance to allow for proofreading of alternate formats.

### 2.6.3 Statistical Methods for Test Bias

Judgmental methods alone, however, are often not sufficient to screen test items for construct irrelevant variance affecting certain groups of students. Several studies in the 1980's found that expert judgments about items that may be more difficult for some groups of students was no better than chance in most cases (Jensen 1980; Reynolds, 1982, Plake 1980; Mille (1980) as cited in Camilli & Shepard, 1994). Statistical methods, both descriptive and inferential, also exist to test for differential difficulty and are recommended as a way to "flag" items that have potential for being biased. These differential item functioning (DIF) methods have been evolving. Classical methods of testing for DIF included use of average p-value differences or ANOVAs. However, these methods have been questioned for their appropriateness of detecting differential difficulty (Camilli & Shepard, 1994). A variety of newer methods considered to be better measures of differential difficulty, have emerged. These methods include Item Response Theory (IRT), use of contingency tables, logistical regression, standardized mean differences (SMD), and SIBTESTs (Camilli & Shepard, 1994; Zumbo, 1999; Zwick, Thayer & Mazzeo, 1997).

#### 2.6.3.1 General Characteristics of DIF

While the best method suited for different data sets and situations may differ, the general intent of differential item functioning (DIF) analysis is to investigate variations in test item difficulty between two groups of equal ability (Standards of Testing, 1999). An internal variable,

such as total test score or score on selected items including the item of interest, is generally used to match students on ability. The underlying theory is that when students are matched on overall ability, one would expect that different test items would be equally difficult and that relatively the same percentage of individuals in each ability grouping will respond correctly to the item. When this does not occur, then the item is said to be functioning differentially for a particular group. Individual test items that function differently for a target group can then be further investigated for bias. DIF research can be conducted at the pre-assessment or post-hoc level. At the pre-assessment level, a pilot version of the assessment can be used to analyze how different test items function between a target group of interest and a referent group. At the post-hoc level, DIF informs about the functioning of items and is one step to help determine the interpretability of the results.

#### 2.6.3.2 Selection of DIF Procedures

When deciding on a DIF procedure, it is important to weigh the advantages and disadvantages of each method to select one that best fits the research situation and intent of the study. Considerations include available sample size, uniformity of the sample, and type of test data (i.e. dichotomous vs. polytomous or ordinal). For smaller samples which often occur when investigating students with disabilities and in particular students with low incidence disabilities like visual impairment, nonparametric procedures may be needed (US Department of Education, n.d.). Descriptive measures can also be used with smaller samples to investigate DIF. For example, the frequency distribution of scores (partial credit or scores based on level of independence) around a certain item can be compared for different groups or accommodated conditions (US Department of Education, n.d.). The type of DIF to be detected factors in as well. For example, some test items represent uniform DIF; that is, the item functions differently

across the whole target group regardless of ability level. Non-uniform DIF represents an interaction with ability level; the item is easier or harder for a particular ability grouping (Zumbo, 1999). The selection of the method to analyze item difficulty, then, can affect the results. Johnstone, Thompson, Moen, Bolt, and Kato (2005) found that different items from the same test were identified as potentially problematic when different statistical methods were used and different groupings made (e.g. disability groupings versus accommodation groupings). However, they did acknowledge that the statistical tests combined with pragmatic rules "such as finding patterns across disability groups and across analysis techniques aid in reducing the complexity of items found to have universal design issues." (p. 18).

For performance-based assessments, like the PASA, a procedure that can handle polytomous scoring is needed. In other words, since the PASA, which uses a scoring rubric that encompasses more than just indicating whether a student is correct or incorrect on an item, procedures that can account for a range of scores on one test item is essential to really examine for DIF. In addition, if students with visual impairments are the group to be evaluated non-parametric polytomous procedures may show more promise because unlike parametric procedures, fewer assumptions are made (Penfield & Lam, 2000).

#### 2.6.3.3 Limitations of DIF Procedures

For students with visual impairments, DIF procedures will inevitably contain limitations because of sample sizes and variability in levels of visual functioning. In fact, rarely have common DIF procedures, such as IRT or logistical regressions methods, been carried out for DIF on just students with visual impairments. It may be necessary to use a non-traditional procedure to explore DIF while trying to take into account the components, such as ability level matching, that comprise all DIF studies.

DIF procedures for alternate performance-based assessments will be somewhat limited as well due to the wide range of extraneous variables that could confound the results when working with students who have multiple disabilities and wide ranges of accommodations. Performance-based assessments may also contain multi-dimensionality in the test items, either planned or unplanned, that introduces additional factors, such as teacher interaction, that affect scores on a test item beyond ability alone (Tate, 2002). Furthermore, as discussed earlier, predominantly picture based alternate performance-based assessments may be particularly difficult to adapt for students without usable vision. If DIF procedures match on an internal variable alone, like total test score, the procedure may be less successful in detecting pervasive DIF, or differentially difficult items that span across the whole assessment (Camilli & Shepard, 1994). Regardless, the identification of items that function differently due to potential bias or construct irrelevant variance is just as important for these groups of students who are traditionally left out of analyses due to sample sizes.

Since DIF procedures may inform less accurately, or be inconclusive, for students with visual impairments and in particular for those students with additional disabilities, information gained from the methods should be paired with additional pieces of information to support the interpretation of the results. In this manner, DIF methods may still provide some useful comparative information.

## 2.7 RESEARCH ON STUDENTS WITH VISUAL IMPAIRMENTS

### 2.7.1 DIF and Accommodations on Large-Scale Assessments

As just alluded to, few studies on accommodations and DIF have addressed students with visual impairments or alternate assessments. This is particularly true for large-scale, high-stakes school-age assessments. Koretz and Hamilton (1999) analyzed state assessment data in Kentucky, and Koretz and Hamilton (2001) analyzed a pilot administration of the New York Regents exam. In these studies, the researchers discussed differential item functioning under different accommodated conditions. Students with visual impairments were part of the samples in each study; however, due to sample sizes, Koretz & Hamilton (1999) only disaggregated results for students with learning disabilities, mild mental retardation and emotional/behavioral disorders, and Koretz and Hamilton (2001) did not disaggregate by disability category at all. Unfortunately, this means that issues in DIF and the use of different accommodations specific to students with visual impairments were hidden within the broad category of students with disabilities using accommodations or students with disabilities without accommodations.

In addition to the Kentucky and New York studies, Barton and Huynh (2003) analyzed data from the South Carolina High School Exit Examination. This study analyzed the types of errors made by students using oral reading accommodations on a multiple choice reading test. Three types of oral reading accommodations were analyzed: use of a reader, use of an audiotape, and use of a videotape (for students with hearing impairments). The results were analyzed by disability group with the physical disability group containing students with speech, hearing, visual or orthopedic impairments. This collective group was the smallest (91 individuals) with the majority of the group having an orthopedic impairment. Again, while overall the researchers

found a weak association between type of disability and the type and number of errors made on the reading test, no specific information regarding students with visual impairments can be gathered from this study.

Jackson (2003) was the only accommodations study found that used state test data and looked specifically at groupings of students with visual impairments. Jackson collected accommodation information on students with visual impairments who took the regular SAT-9 either on-level or out-of-level. Seventy-one students were classified by ethnicity (white (55%)/non-white (45%)), home language (English (89%)/non-English (11%)), disability (visually impaired (76%), visually impaired plus additional disabilities (24%)), and reading medium (print (34%), large print (45%), and braille (21%)). Other accommodations or modifications were controlled for in part of the analysis. Of the additional accommodations used beyond reading medium selection, extended time, allowing answers to be marked in the test booklet, frequent breaks, and use of a flexible schedule were the most frequently cited.

Compared to peers without disabilities, the students with visual impairments taking the SAT-9 at grade level (33% of the 71 students) scored one stanine lower in reading and language and one to two stanines lower in math. Compared to each other, however, no statistically significant differences solely by reading medium were evident in reading, math or language (after post hoc analysis). Jackson suggests, however, that practical significance should still be considered. She found that controlling for test modifications did not significantly change the total math scores of the students, and interactions in math and language between reading medium and other variables such as ethnicity, home language, and disability category (additional disabilities or not) existed. No tests for differential difficulty were conducted in this study.

### 2.7.2 Other DIF Studies

Bennett, Rock, and Jirele (1987) investigated score levels, test completion rates, and reliability for examinees with disabilities as compared to examinees without disabilities. Only students who indicated that English was their best language were used in the study. For students with visual impairments, two groups were analyzed: students taking the standard administration under timed conditions and students taking the large-type edition under extended time. Students taking the braille version of the exam were not included because fewer then 100 students were in that group. The test, overall, was found to be reliable for all groups analyzed. The performance of students with visual impairments taking the standard administration was not found to differ at a practical level defined as scoring 0.2 standard deviations different from the comparison group of students without disabilities. However, this group was slightly less likely to complete test sections.

For the group of students taking the large-type extended time version of the exam, the mean analytical score was .28 standard deviations higher then the comparison group. This is considered an indication that the analytical section may not be testing the same attributes in the group of students with visual impairments. Through a factor analysis, Rock, Bennett, and Jirele (1988) found that the analytical section did not function effectively as a single factor for students taking the large-type version of the test, indicating that analytical scores and total scores (in which analytical was a part) had different meanings for this group. Typically, higher quantitative and analytical scores are achieved by students whose majors are in the sciences (Bennett, Rock, & Jirele, 1987). The two groups of students with visual impairments were disproportionately underrepresented in the science areas, which is contradictory to the better performance of students taking the large-type edition with extended time. The researchers hypothesize that the

use of extended time could be the contributing factor to the difference; extended time may give an unfair advantage. It should be noted, too, that the overall results of this study differed from the findings of Braun, Ragosta, and Kaplan (1986 as cited in Bennett, Rock, & Jirele, 1987) that showed students with visual impairment scoring *slightly lower* on the verbal scale of the GRE and *substantially lower* on the quantitative scale. Braun et al. (1986), however, noted a discrepancy between the performance of students who are blind compared to other students with visual impairments which might also account for the difference between the two studies; Bennett, Rock, & Jirele (1987) did not look at students who were blind, (e.g. the group of students who would use a braille administration of the test).

Braille administration was more closely analyzed in the two studies investigating the SAT for different disability groups. Bennett, Rock, and Kaplan (1987) analyzed differences in difficulty level of test items for nine disability-accommodations groupings that had a sufficient number of students to analyze (98 or more). Extended time administration was a constant for each grouping. Of the nine groupings, three groups were comprised of students with visual impairments: students taking the exam in braille, students taking the exam in large print, and students taking the regular exam. Test items were analyzed as clusters since the number of forms and test items used made individual test item analysis too difficult. Clusters were based on logical groupings of items that might be problematic for different groups of students. Clusters that resulted in differentiation of 0.2 standard deviations or more on both forms were then analyzed by individual items for statistical and practical importance. No clusters surfaced as differentially significant at the 0.2 level for students with visual impairments taking the standard or large-type administrations. For students with visual impairments taking the braille version of the test, however, two math clusters fit the 0.2 designation: multiple choice items with graphics,

and miscellaneous multiple choice items. A closer review of the tables also shows two additional math clusters for braille users that came close to meeting the 0.2 standard deviation criterion: items involving graphic comparisons and geometry multiple choice items. Overall, within the two identified clusters meeting the 0.2 criterion, 10 differentially difficult items were identified as too hard and three items identified as too easy. The researchers discussed two examples to highlight possible causes for increased difficulty. One example dealt with the use of the tally system with the braille version being changed to clusters of five tallies instead of the common visual tally system of four tallies with the fifth diagonally crossing the four. The question was changed to attend to this difference with the speculation that perhaps the newly worded question was more complex. The second example used a special symbol to denote an operation with the potential that, tactually, the special symbol was confusing.

Based on the results of this study, Bennett, Rock, & Novatkoski (1987) examined the differentially difficult items on the braille version of the math section of the SAT more closely. Three years of SAT scores within two states were analyzed for students who used the braille version of the test and had English marked as their dominant language. The researchers investigated potential reasons for differential item functioning by searching the ERIC and Psychological Abstract databases for literature under *mathematics* and *cognitive processing in blind students* and by consulting with professionals knowledgeable about students who are blind. Three main factors were identified: 1) Congenitally blind students may have less developed spatial abilities due to lack of visual experience; 2) Visual input makes some processes easier than tactual input including being able to eliminate certain options in a test based on visual examination of size and the ability to compare more than one item at a time; and, 3) Braille takes up more space and includes additional symbols to represent visual characteristics. Potential

barriers include the need to break up items which can create a higher demand for decoding and longer processing time.

As with Bennett, Rock, and Kaplan (1987), this study created clusters that represented items that could be more difficult for students using the braille version of the SAT in math. The same criterion of 0.2 standard deviations or more across all forms was used to identify clusters that had meaningful differences between the braille version and standard print version of the test. Clusters meeting this criterion were then analyzed by item for statistical and practical significance. Differential difficulty was discovered, particularly in items that included figures in the stimulus, items where spatial estimation was helpful in determining the answer, and items that presented small or medium sized figures. As is common with differential item research using test data (quasi-experimental), the underlying reasons for the difficulty can only be hypothesized and might be due to other reasons such as a true difference in mathematical skills that could have been learned by a student with visual impairment but were not (Bennett et al., 1989). Bennett et al.'s (1989) findings regarding spatial components in test items are supported for young children and school-aged children in part by studies discussed regarding the cognition of students with visual impairments.

### 2.7.3 Logical Analysis

Bennett, Rock, and Novatkoski's (1987) second study that attempted to pinpoint the reasons that certain items emerged during the DIF study, highlights the need to conduct a logical analysis of the items that emerge as differentially difficult in order to interpret or determine the reasons for the difference in difficulty level. DIF studies alone do not inform on test bias. Differential item functioning serves only as a "flag" for items that show multidimensionality which might be bias

if through logical analysis of the item, the difference does not appear to be relevant to the test construct (Camillli & Shepard, 1994). It could be that a type I or type II error occurred, that the difference is pertinent to the test construct, or that no pattern emerges on similar items. Further investigation is needed to analyze the reasons why an item may have functioned differently for a particular group of students. As noted earlier, for students with disabilities possible reasons for DIF can be due to a change in the cognitive processes being engaged because of an accommodation or adaptation made to an item, lack of opportunity to learn the material for certain sub-groups of students, or content within an item that is outside the realm of experience of the sub-group of students (Gersten & Baker, 2002).

### 2.7.4 Cognitive Development of Students with Visual Impairments

Since students with visual impairments must gather information non-visually or with diminished visual input, it is possible that experience with the world through different modalities and a lack of incidental learning that occurs through the use of vision could result in different development cognitively, which may contribute to factors that explain instances of DIF. Studies in the areas of language, concept development, classification, and intelligence tests have provided starting points, sometimes contradictory to each other, to understanding factors associated with the cognitive development of children with visual impairments and barriers that surface in testing procedures. While conclusions are very often difficult to make due to the complexity of sorting out factors that contribute to the achievement of students with visual impairments, who ultimately represent a heterogeneous group of students, studies in the area of cognition can help provide starting points that can be applied to a logical analysis of test items that emerge as differential different for students with visual impairments in a DIF study. Literature reviews

conducted by Warren (1994) and Barraga and Erin (2001) in the area of cognition of students with visual impairments seem to suggest that cognitive abilities, fundamentally, do not differ from sighted peers, but that delay and/or cognitive limitations can occur due to lack of experiential background or additional disabilities.

#### 2.7.4.1 Word Meaning and Concepts

The types of words that children who are blind first acquire have been found in some studies to differ from sighted children's beginning vocabulary. Mulford (1988) synthesized information from a variety of studies on the first 50 words of children who were blind and compared the synthesis to results found by Nelson (1973 as cited in Mulford, 1988) about sighted children. A few differences emerged. First, specific nominals (specific names of items or people used to refer to only one instance of a category) were more prevalent than general nominals (names of classes of objects) in the vocabularies of children who were blind as compared to sighted children. Also, action words were more frequently used to relate to oneself (egocentric) for a longer period of time as compared to sighted children who moved on to applying action words to refer to other people or things more quickly. Similarly, Dimcovic and Tobin (1995) found some children who were blind (age 6-11 years old) in their study to be able to pick the item in a verbal task that does not belong, but be unable to assign the name of the super-ordinate class for the items that did go together. Warren (1994) also found in his review support from studies that an understanding of referential words such as pronouns and spatial terms like *here* and *there* can be delayed in children with visual impairments. These potential differences are not surprising considering that initial understanding of these words would most likely be supported by seeing actions attached to the word use.

Researchers have differed in their interpretation of why these differences occur. In terms of specific versus general nominals, the appearance of more specific nominals may be due to restricted experience with a word, leading the child to only have a single instance for applying that word (Bigelow 1987 as cited in Warren 1994). Another theory is that it has more to do with "limited ability to form concepts to which words are attached as labels" (Anderson et al., 1984 as cited in Warren, 1994 p. 138). That is, due to overall experience, the child who is blind might lack general concepts which lead to the use of less general nominals. In both cases, experience plays a role, and either explanation, if accurate, can affect the cognition of children with visual impairments. If concepts are under-developed, cognitive processing could be affected.

At the 100 word stage, word types appear to be similar for blind and sighted children (Anderson et al., 1984; Landau, 1983 as cited in Warren 1994), but children who are blind invented words less frequently and over-extended words less frequently (e.g. applying the word dog to all four-legged animals) (Warren, 1994). Anderson et al. (1984 as cited in Warren 1994) attribute the difference to the possibility that the children who were blind more readily accepted limited meanings to words and that this reflects differences in the richness of the underlying concepts associated with those words. Again, experience seems to be the underlying cause for these differences. Children who are sighted often extend word use by what they are observing visually and taking in incidentally, whereas a child who is blind receives less information incidentally and is limited either to the direct experiences provided him or her or to experiences that are within arms reach or contain auditory quality. Additional factors would include the amount of mobility a child has to explore his/her environment and the amount of encouragement the child is given to do so (Barraga & Erin, 2001).

Millar (1983 as cited in Warren 1994) evaluated children who were congenitally blind aged 8 to 13 years on their ability to recognize mismatched adjective-noun pairs (e.g. meowing dog). She found that the younger children in the study had difficulties with adjectives that had spatial or visual meanings. Anderson (1979) and Anderson & Olsen (1981 as cited in Warren 1994) found that children who were congenitally blind age 3 to 9 years old were asked to define and describe tangible and less tangible items. Responses were compared to sighted children of similar ages. They found that sighted children were more responsive to the less tangible items and were less likely to refer to tactual qualities of items. Also, the children who were blind were more likely to give more concrete and less abstract functional attributes as compared to sighted children's responses. However, despite these differences, there did not appear to be a difference in meaning attached to the objects. Warren notes that the collective work of Anderson et al. in this area led them to the conclusion that, "the response patterns of blind children demonstrate that their language does not simply reflect the usage of the surrounding language environment [empty language]. Instead, it appropriately reflects the experience-specific conceptualizations of objects the children obtain via touch and other non-visual senses" (p. 144). Similarly, Higgins (1973 as cited in Warren 1994) found that children who were blind developed classification skills in the same way that Piaget observed for sighted children, but that there was a difference between abstract and concrete concepts. He, too, concluded that experience was the driving force, not an actual difference in cognitive ability saying that it:

> reflected a child's previous activity with the elements about which he had to reason. The likelihood of a correct response was significantly greater if the child had performed perceptual or motor actions in relation to the elements specified in the class inclusion questions (p. 33 as cited in Warren, 1994 p. 149).

Overall, Warren (1994) summarized his interpretation of the contradictions and interpretations of cognitive development in the literature by saying:

> On the one hand, several studies of the meanings of individual words (DeMott, 1972; Dershowitz, 1975) and word usage (Millar, 1983) show the word meanings of children with visual impairments to be very similar to those of children with vision. The differences that do occur (e.g. Anderson, 1979) appear to be linked to the child's perceptual experiences, and specifically the role of visual experience, but there is little evidence from these studies that the underlying concepts that words represent are impaired in any significant way. On the other hand, the evidence of Anderson and her colleagues suggests more fundamental differences when we look at the interrelationships among words and their underlying concepts: these may be less elaborated for children with visual impairments (pg. 146).

While the general consensus seems to be that the cognitive development of children with visual impairments is not impaired due to the mere fact of the existence of a visual impairment, differences in the experience-base of children with visual impairment can create holes in conceptual understanding. This seems to be particularly possible, as noted in the Millar (1983 as cited in Warren 1994) study, when meaning is attached to visual or spatial components. The possibility that children with visual impairments may exhibit differences in cognitive performance on visually dependent or spatially based tasks is also supported by several other studies dealing with the measurement of cognition or concept development of students of varying ages.

#### 2.7.4.2 Cognition and Spatial or Visual Components

Brambring and Foster (1994) looked at cognitive development of children who were blind compared to sighted children using the Bielefeld Developmental Test for Blind Infants and Preschoolers. This test tried to create blind-neutral tasks so that cognitive development could be more accurately measured for students without vision who were not yet verbal. In order to examine whether the test achieved its goal of a blind-neutral test, three groups of children who were blind ages 36-, 42-, and 48-months and three groups of sighted children ages 32-, 28-, and 44-months were tested and compared. The sighted students took the assessment twice: once with vision and once without. Testing situations were counterbalanced with a week in-between each testing session. The assessment contained 26 items requiring tactile materials and 12 items that used auditory or verbal tasks. A test box that presented defined space in which to work was used. It consisted of three compartments that had tactual and auditory cues to help with localization. Overall, children who were blind displayed a 16 month delay from sighted children, and a 10 month delay from sighted children taking the test without vision. In particular, it was found that the assessment did not result in a fair comparison between sighted children and children who are blind, but that it could be used to compare skills and cognition among children who are blind (within-group). The largest differences between sighted and blind children were found in tasks where children, "must point in a certain direction, remove something from a specific location, or spatially arrange objects according to specific principles." (p. 6).

Hartlage (1969 as cited in Warren, 1994) also found that young children who were congenitally blind had more difficulty dealing with spatial over non-spatial concepts. For example, a question dealing with the concept of *in front of* was more difficult than *smarter than* (Warren, 1994). Warren (1994) indicates that Hartlage's findings suggest that by the beginning

of the fifth grade, children who are blind are equally able to deal with spatial and non-spatial concepts, but younger children show a difference. In comparison, Caton (1977), using a translated format of the Boehm Test of Basic Concepts (BTBC) called the Tactile Test of Basic Concepts (TTBC), tested 25 students in each of the grades of kindergarten through second. Results were compared to Boehm standardization norms of sighted children. Caton (1977) found no significant differences in overall performance between the two groups based on visual status; however, a significant interaction by grade did occur. Kindergarteners who were blind performed slightly better than sighted kindergarteners, but first and second graders who were blind performed worse than their sighted peers. Based on an item by item analysis of percent passing scores some items emerged as different in difficulty between children who were blind and sighted children. At the kindergarten level, four items were less difficult for children who were blind and one item was more difficult (see Table 4). At the first grade level, three items were easier, and 11 items were more difficult. At the second grade level, no items were less difficult, and twelve items emerged as more difficult. The more difficult items included those that required comparative judgments (half, middle, medium-sized, third, in order, pair, matches, and least) versus items where the child was able to use himself as the reference point, an indication, also, that items with more spatial complexity were harder.

**Table 4: Items Different in Difficulty for Children who are Blind**

| | Kindergarten | First Grade | Second Grade |
|---|---|---|---|
| **Less Difficult** | Other<br>Never<br>Always<br>Equal | Always<br>Forward<br>Equal | |
| **More Difficult** | Middle | Through<br>Some, Not, Many<br>Middle<br>Farthest<br>Around<br>Over<br>Widest<br>Most<br>Between<br>Different<br>Half | Top<br>Through<br>Inside<br>Some, Not, Many<br>Middle<br>Farthest<br>Over<br>Most<br>Corner<br>Half<br>Alike<br>Matches |

(Caton,1977)

As seen in Table 4, some of the concepts would also be supported by visual observation. Unless children who are blind were given direct experience with these concepts, they would be harder to obtain an understanding of without vision. In addition, Caton (1977) mentioned that comparative items were more difficult because of the modality used. Unlike sighted children who can observe all the figures simultaneously, children who were blind accessing the figures through tactual means had to "observe" each figure separately.

Dimcovic and Tobin (1995) confirmed this speculation about tactual exploration when testing 30 blind children aged 6-11 years compared to sighted children under blindfold of the same age. Children who were blind performed classification tasks that had a figure base (tactual shapes) less successfully than the sighted children. This was especially true when more than two elements had to be compared. Sighted children, despite being under blindfold during the task, still had visual experience with shapes and the underlying concepts for the classifications. It was

also noted that some children, even when they seemed to understand the figure tasks, could still not perform it. The potential for items to be differentially more difficulty due to the sensory mode used to access was present. Dimcovic and Tobin (1995), based on a qualitative analysis in their study of the children's language, behavior and interaction during testing, identified potential contributing factors for the poorer performance overall of children who were blind. These factors included passivity of some children, unfamiliarity with the tactual task, a sequential way of gathering information on the 4 item figure tasks, difficulty with generalization, and overall experience level affecting the depth of concept development. Dimcovic and Tobin (1995) also note that children who were blind in their study age 8 ½ to 11 years old showed dramatic improvement on the verbal aspects of the study as compared to the younger children, and 11 year olds were almost comparable on all tasks to the sighted children. The authors suggest that pace of development may be different due to differences in experience in encountering multiple referents for words. This study used the WISC-R vocabulary subtests to estimate general verbal competence prior to doing the classification tasks, but the authors caution to consider the validity.

Wyver and Markham (1999) examined items for visual interaction on the WISC verbal subtests of comprehension and similarities. Within these two subtests, they categorized items as being dependent upon visual experience, influenced by sensory non-visual experience, or abstract. Fifteen children with congenital visual impairment (visual acuities of less than 20/60) ages 6-12 years old and 15 sighted children approximately the same ages were tested. Within the comprehension subtest, significant effects of visual status were found with no interaction. The same was not found for the similarities subtest, however. Only an effect for age on all three types of items was found. The researchers extended a couple suggestions as to why visually dependent

material was more problematic on the comprehension subtest. Children may understand visual words but have difficulty applying them in everyday situations, or children with visual impairments may have had fewer opportunities to deal with the type of situations presented in visually dependent comprehension passage.

#### 2.7.4.3 A Note about Low Vision

Most of the studies just discussed focused on children who were congenitally blind, indicating that more is probably known about the cognitive development of children who do not have usable vision. What about children with low vision? This category, of course, is extremely heterogeneous in terms of the amount of visual loss, which can be one contributing factor to why studies in the area of visual impairment may vary in their results. Warren (1994) notes a lack of sufficient research that takes into account differences in visual status along with analysis of individual cases and experiential background. In his analysis of cognitive development, he indicates in his summary that some classification studies that did factor in level of vision, do reveal variations in performance-based on visual experience (age of visual loss and partial vision) among other factors like range of experience with the materials used to assess. Similarly, Dekker, Drenth, Zaall, and Koole (1990) assessed children with visual impairments ages 6-15 years old using an intelligence test designed to create a more comprehensive measurement that tapped into a wider range of cognitive processes than the WISC verbal subtests alone could do (typically what is used with children with visual impairments). They found that students with partial vision had an advantage on haptic-spatial tasks measuring spatial ability. Partial vision was not the only factor, however. Other factors for children without usable vision included educational placement (regular school or special institution).

At the same time, Groenveld and Jan (1992) included children with low vision in their analysis of the Weschler Intelligence Scale for Children- Revised (WISC-R) and the Weschler Preschool and Primary Scale of Intelligence (WPPSI). They found that while children with low vision did well producing geometric designs from an example, they had more difficulty with picture completion items and other items that involved reconstructing from memory, which is more closely related to visual experiences. They also found that some children made errors in sequencing pictures not because of faulty logic but from errors in how the pictures were perceived. Generally the effects were greater as acuity level was lower. In addition, Coleman (1990) conducted an investigation of how students with visual impairments understand length. Twenty-four children participated. Seven children (ages 7-18) took a braille version of the assessment, seven (ages 8-12 years old) took a large-print version, and 10 (two randomly selected from each age group) took a regular print version. Coleman found that braille readers had more difficulty measuring length than students who read regular print, but large-print readers were found to have the most difficulty with measurement of length out of the three groups. Coleman hypothesized that the severity of the vision problems of students accessing the test via visual input was the reason as opposed to the large-print accommodation itself. At least in Coleman's small sample of residential school children, students with low vision experienced similar or more difficulty than students without usable vision in demonstrating the concepts of conservation of length, a concept supported by vision.

Although conceptual development was not the main focus of her study, Milian (1996) observed higher mean scores by her monolingual (English) study participants taking the Tactile Test of Basic Concepts than students (legally blind and low vision) taking the Boehm Test of Basic Concepts. Milian speculates that, "greater emphasis [may be] placed on teaching basic

concepts to students who are functionally blind than to students who have low vision or are legally blind. If that assumption is true, then is it also true that educators assume that children who have low vision will incidentally learn concepts more readily than will children who are functionally blind?" (p. 395). Milian's suggestion is supported by Corn and Bishop's (1984 as cited in Barraga and Erin, 2001) finding when they tested 116 adolescents using the Test of Practical Knowledge. They found that adolescents with low vision (legally blind but not totally blind) had more difficulty than adolescents who were totally blind on the test, again indicating that students with low vision may no be developing concepts incidentally through visual observation as readily as educators may assume. Furthermore, Hull and Mason (1995) found children who were congenitally blind to perform better on a digit span memory test than sighted children, but children who had some usable vision performed similarly to sighted children. While this is only one area of compensatory skills (auditory memory), the study raises the question about whether children with low vision are developing sufficient compensatory skills to compensate for what vision is not providing.

#### 2.7.4.4 Possible Effects on Testing

In summary, synthesis of the studies regarding the cognitive development of children with visual impairments suggests that concept development may be less well developed in some areas if experiential background is not sufficient to provide enough direct experience with the concepts. The studies also suggest that spatial or visually dependent concepts may be particularly challenging for children who are blind. The impact on students with varying levels of usable vision is less well understood, but limited information indicates a need to be aware of the possibility that some concepts may not be fully developed. In addition, cognitive development of children with visual impairments in some areas may not be so much deficient as different due to

the manner in which the concepts were experienced. Also, accessing tasks tactually as compared to visually may make the task itself more difficult because of lack of experience, training (Warren, 1994) or fundamental differences in the sensory modality being used. It is important, then, to consider that accommodations, including a switch in the type of medium used, could affect the difficulty or intent of a task.

#### 2.7.4.5 Possible Effects on Alternate Assessments

The possible effects on testing just discussed would also apply to students with visual impairments who have additional disabilities and take alternate assessments, including the PASA. How well an alternate assessment will reflect the actual skill level of students with visual impairments and how much construct irrelevant variance will be present, will depend on the type of assessment, the content of the assessment, and the process by which the assessment was created.

In general, performance-based assessments like the PASA are the closest to standard large-scale tests because all students are performing the same set of skills in the same way. Theoretically they would also contain many of the same considerations for students with visual impairments as standard large-scale assessments would. Research on these types of alternate assessments is warranted to explore the extent to which the skills of students with visual impairments are being measured accurately, and to determine, ultimately, how the scores of students with visual impairments might be interpreted.

Performance assessments should be reviewed for how well tasks are universally designed, for bias toward students with visual impairments, and for how conducive skills are to being adapted into alternate formats. Since alternate assessments are for students with the most severe disabilities, it is likely that some of the skills being assessed are pre-cursors to the actual

target skill such as reading. Because of this possibility, performance assessments will also need to be reviewed for skills or constructs that are irrelevant to some students with visual impairments. For example, picture identification is often considered a pre-reading skill. However, for a student who is totally blind, picture identification would not be a relevant pre-reading activity. If this skill is assessed in a performance assessment or shows up on a checklist, the student may be unfairly evaluated, or the teacher would have had to adapt the skill or question. Additional challenges may arise on the scoring end of the process. If the scorers are not familiar with acceptable accommodations for students with visual impairments or typical skill progressions for adaptive skills (such as beginning braille literacy skills), student performance scores may not accurately reflect the student's abilities. Similarly, scores could be elevated if students are given the benefit of the doubt when scorers are unsure of the adaptive methods being used.

Finally, since students with visual impairments who would be taking alternate assessments will also have additional disabilities, there is an added challenge in determining whether construct irrelevant variance exists in the assessment due to visual impairment. This was already a possibility for students without additional disabilities. A question that is flagged as problematic on a differential difficulty study, for example, does not indicate the reason for the difference. It is always possible that the item was not actually biased, but that the student legitimately did not know the skill. Separating out the possible reasons for differences in performance will be confounded by additional variables, but it does not discount the importance of doing research in the area of alternate assessment for students with visual impairments. Use of various strategies including checklists for item analysis during test production, research methods

for analyzing differential difficulty, and the impact of accommodations will help support the process of separating out the pieces.

## 2.8 STATEMENT OF THE PROBLEM

To understand whether assessment results are being appropriately interpreted for students with visual impairments, more research needs to be done on the impact of accommodations and adaptations to test items, particularly when a change in sensory mode occurs by which the item is being accessed. This includes test items on performance-based alternate assessments. Items may no longer be testing the same skill or may be differentially more difficult for students with visual impairments. Teacher selection of accommodations for students may help alleviate some of the bias; however, since no state alternate assessments currently report on differential functioning or on specific accommodations being selected by teachers of students with visual impairments, it is unknown how well these assessments measure student ability. In addition, certain accommodations utilized by students may help them access the item but also change the intent of the skill being assessed (Tindal & Fuchs, 2000).

With state accountability systems under NCLB, decisions are currently being made about student and program progress towards the alternate content standards with little to no research about the effects of adaptations and accommodations on what the scores actually mean for sub-groups of students with disabilities such as students with visual impairments. In addition, if the consequential effect of alternate assessments is to transform teacher perceptions about math and reading content appropriate for students with the most severe cognitive disabilities based on alternate standards, the test items must model content that is flexible enough to be appropriately

adapted or accommodated for the student's primary sensory mode if teachers are going to be convinced of the content's viability. Test item and accommodation analysis are two ways to begin to better understand the validity of alternate performance-based assessments, such as the PASA, for students with visual impairments.

## 2.9 RELEVENCE OF THE STUDY

This study attempts to investigate differential item functioning on the Pennsylvania Alternate System of Assessment (PASA) as well as to describe what accommodations are being used. This study serves as a first step to recognizing potential problems in testing for students with visual impairments, particularly on the PASA, that may later lead to resolutions. While exploratory in nature, this study will provide initial information on how test items on the PASA might be better adapted for students with visual impairments, particularly those students without usable vision for purposes of taking the assessment. Outcomes of the study can lead to direct application when reviewing future test items for their appropriateness in measuring intended constructs for this sub-population of students with the most severe disabilities who also have a sensory disability. It can also promote conversations within the field of visual impairment as well as future research regarding skill progression in math and reading and the appropriate selection of accommodations and adaptations, particularly on performance-based alternate assessments.

# 3.0 METHODOLOGY

## 3.1 INTRODUCTION

The intent of this exploratory study was to begin to understand the components that may affect the score interpretation of students with visual impairments on the Pennsylvania Alternate System of Assessment (PASA). PASA is a performance-based assessment that is multi-dimensional in nature. That is, performance items contain both a student and a teacher factor that cannot always be separated. Due to the wide range and severity of disabilities of the students who take alternate assessments such as the PASA, teachers or educational teams are responsible for selecting appropriate accommodations for their students that do not change the intent of the skill being assessed. A student's score may be affected positively or sometimes negatively by the type of accommodations/modifications utilized. For students with visual impairments, particularly students who cannot access pictures or print visually, adapting the PASA may be particularly challenging. Test items themselves that focus on precursory reading and math skills such as selecting pictures named and matching sizes may no longer be testing the same skill when conducted through a different sensory mode. In order to better understand how the PASA functions for students with visual impairments, this study proposed to examine test item functioning through the use of a post-hoc, mixed methods model.

## 3.2 RESEARCH QUESTIONS

As mentioned earlier, the following research questions were explored in this study:

Q1. Were there significant differences in the scores of students with visual impairments at different functional vision levels on the 2005 grade 3/4 or 7/8 A level PASA?

Q2. What accommodations did teachers make to adapt the 2005 PASA for students with visual impairments?

a. Are there relationships between the types of accommodations made and level of functional vision or type of test item?

b. Were there accommodations that seemed to change the intent of the skills being tested?

Q3. Were there significant differences on individual 2005 level A PASA math and reading test items at the 3/4 and 7/8 grade levels of students with visual impairments as compared to students without visual impairments who had similar ability profiles on the constructs of interest?

Q4. Considering the accommodations made and student performance on different types of test items, what are the potential reasons that "flagged" test items functioned differently?

Based on the format of the assessment, it was expected that PASA would be most difficult to adapt for students with visual impairments whose functional vision is unusable for purposes of taking the assessment. At the A level of the PASA assessment for example, substitution of objects in place of the pictures at the upper grade levels would potentially make test items equivalent to items at lower grade levels since a movement from objects to pictures

was a built-in change in complexity across grades. Students without usable vision, then, would probably exhibit the most elaborate accommodations and adaptations to the assessment. It was expected that students with visual impairments who primarily use vision for most tasks would exhibit fewer instances of differential functioning or cases where elaborate adaptations were made.

## 3.3 PASA ASSESSMENTS USED IN THE STUDY

The 2005 PASA is the measure explored in this study. PASA was designed to meet No Child Left Behind (NCLB) federal regulations, particularly the requirement which allows students with the most severe cognitive disabilities to take an assessment based on alternate standards set by the state for purposes of determining adequate yearly progress (AYP). As explained more thoroughly in Chapter 2, the PASA is a performance-based measure consisting of multiple choice and short answer items in math and reading. Each content area contains between 20 and 25 items at three distinct levels of difficulty. Students take either the A, B, or C level of the PASA. Teachers supply a videotape of the performance or detailed narrative notes of the student's responses during the assessment, which are then scored by teachers across the state based on a scoring rubric of 0-5.

### 3.3.1 Level and Grade Selection

Since the majority of students with visual impairments (76%) took the A level assessment of the PASA, it was chosen as the focus of this study at grades 3/4 and 7/8. The A level version of the

assessment contains the least complex skills such as matching and selecting requested objects, pictures, or amounts. Appendix A summarizes the skills assessed at the A level for grades 3/4 and 7/8 in math and reading.

These two grade groupings were specifically selected for analysis because of the parallel comparison they provide. Grade 7/8 contains similar skills to grade 3/4 but marks the progression in task difficulty by moving from objects in 3/4 to pictures and by requiring closer discriminations between target answers and distracters. The grades are grouped for analysis because third and fourth graders take the same assessment, with the same being true for seventh and eighth graders. Different cut-points for proficiency levels are established between the different grades in a grouping to make the distinction between them, but this aspect of the PASA does not directly affect the intent of this study. Therefore, analyses were done by maintaining the groupings in order to benefit from larger sample sizes.

### 3.3.2 Technical Adequacy of Assessment Selections

According to the 2005 PASA technical supplement, the A level of the PASA consistently provides a greater challenge to the students taking it as compared to students taking the B and C levels of the assessment. The percentage of students scoring a 5 (highest score) on individual test items ranged from 36-58% with mean performance ranging from 2.9 to 4.6 on a scale of 0-5. Because it is the least complex level, A level raw scores received no additional weighting before being transformed into performance levels.

#### 3.3.2.1 Internal Consistency

Internal consistency, or the degree to which the test items measure the same underlying constructs, was high at the A level with Cronbach's Alpha statistics of .95 and .97 for third and eighth graders respectively in mathematics and .96 and .97 in reading. The average correlation of A level individual reading and math test items to performance levels and scaled scores was .64 for third grade and .73 for eighth grade.

#### 3.3.2.2 Inter-Rater Reliability

Since scoring of the PASA involves assigning a score from 0-5 to a test item based on the correctness of response and the degree of independence the student used to respond, it is important to report inter-rater reliability in assigning those scores. Teams of two individuals scored each tape. Reliability was assessed using a team-to-team approach by having each team score four tapes that were also scored by other teams. Thirteen percent of the third grade tapes were double scored, and approximately 17% of the eighth grade tapes were double scored. These tapes were randomly selected and included students taking the PASA at any of the three levels of difficulty (A, B, or C). Reliability percentages are reported for the third and eighth grades, but not specifically reported for the fourth and seventh grades in the technical supplement.

The average percentage of exact agreement of the scores assigned to each test item was 88.7% in reading and 85.4 % in math at the third grade. At the eighth grade it was 83.5% in reading and 83.4 % in math. In addition, an average of 8-12% of the scores were within one point of each other in grades three and eight. Average percentages of scores differing by two or more points from each other ranged from 3-4.8% in the two grade levels.

Since differences in score assignments to test items could result in a student being classified at a different performance level, consistency of performance level classification was

also calculated. In the third grade, 89.7% of the double scored tapes in reading and math would have received an identical classification. In the eighth grade, 79.0% of the double-scored reading assessments and 84.0% of the math would have received identical performance classifications.

#### 3.3.2.3 Threats to Validity

The 2005 technical supplement identifies three major threats to validity of the PASA. The first is the use of a single method to measure the construct of reading and mathematics achievement. The test structure, the student's ability to attend to one-to-one testing, the quality of the taped performance, and the recording of scores are all examples of ways that student performance may be artificially lowered. The PASA is meant to be merely a snapshot of a subset of math and reading skills. The length of the test was also identified as a factor potentially affecting validity. Each content area is kept to 20-25 items so as not to overly burden the student or the test administrator. Shorter assessments, however, can affect the ability to draw inferences from the test. Finally, it is recognized that some contexts and materials may be overused in the assessment due to an effort to present materials, as much as possible, that would be familiar to the population of students being tested.

## 3.4 PARTICIPANTS AND DATA

This study used existing data for all 286 students with visual impairments and an equivalent matched group of students without visual impairments who took either the grade 3/4 or grade 7/8 A level PASA assessment. Most of the identified 286 students with visual impairments (92%) took both the A level math and reading assessments. One student took the A level reading

assessment and a different level math assessment (B or C), and another 22 students took the A level math assessment and a different level reading assessment. Data from only the A level assessment of these 23 students were used for this study. Matches were created independently for the math and reading assessments. That is, the same student with visual impairments may have a different match for math than for reading. Available data for the study included test item scores and ability checklist profiles for both groups of students. In addition, supplemental data gathered from assessment videotapes of the majority of students identified as having a visual impairment (n=257) were available for analysis.

### 3.4.1 Identification of Students with Visual Impairments

The students with visual impairments were identified through a documentation form filled out by the test administrator and sent in with the student's assessment. The documentation form included four questions specific to visual impairment:

1. Is a visual impairment documented in this student's IEP? YES NO
2. What is this student's diagnosed pathology (e.g. cortical visual impairment, retinopathy of prematurity, etc.)?
3. Does this student receive service for visual impairment in his/her IEP? Yes No
4. To what extent does this student use vision to perform tasks (*please circle one)*?
    a. Primarily uses vision
    b. Uses a combination of vision and other senses (e.g. tactile, auditory)
    c. Uses other senses in place of vision (e.g., tactile, auditory)

Based on the supplied information to these questions and the student's reported primary disability, the following decision rules were then used to determine the population of 286 students with visual impairments who took the grade 3/4 (n=143) or 7/8 (n= 143) A level PASA:

1. All students whose primary disability was indicated as visual impairment when enrolled for the assessment were retained.
2. All students who were indicated as receiving services for visual impairment were retained.
3. Students who were indicated as not receiving services for visual impairment were reviewed:
    a. All students in this category whose level of vision was indicated as "uses other senses in place of vision" were retained.
    b. All students in this category whose level of vision was indicated "uses a combination of vision and other senses" were retained except one because the diagnosed eye pathology was marked as "N/A".
    c. All Students whose level of vision was indicated as "primarily uses vision" were retained only if they had a probable eye pathology indicated. The following cases were dropped:
        i. Students whose visual impairment only affected one eye (monocular);
        ii. Students whose visual condition only indicated use of glasses or a refractive error;
        iii. Students whose diagnosed eye pathology or visual condition was left blank, stated as not applicable, or unrelated to vision;
        iv. Students whose visual diagnosis was indicated as visual perception.

Of the 286 identified students, 29 % (n = 82) were indicated as primarily using vision (V) for most tasks, 52% (n = 148) were indicated as using a combination of vision and other senses (CV) for most tasks, and 16% (n = 46) were indicated as using other senses in place of vision (NV) for most tasks. An additional 4% (n = 10) of the students did not have a functional vision indication on their documentation forms. Table 5 shows the breakdown of the number of students in each functional vision category by content area and grade.

**Table 5: Number of Students Taking Each A level Assessment**

| Functional Vision for Most Tasks | 3/4 A Level Math | 3/4 A Level Reading | 7/8 A Level Math | 7/8 A Level Reading |
|---|---|---|---|---|
| **V** | 35 | 30 | 46 | 39 |
| **CV** | 79 | 75 | 69 | 66 |
| **NV** | 23 | 23 | 23 | 22 |
| **No functional vision category given** | 5 | 5 | 5 | 5 |
| **Totals** | 142 | 133 | 143 | 132 |

In total, the 286 students with visual impairments taking the PASA in grades 3/4 and 7/8 comprised approximately 5% of the students taking those assessments. The majority of the students with visual impairments had a primary disability designation other than visual impairment with the most common classification (57%) being multiple disabilities. The next two most common primary disability categories were mental retardation (17.1%) and visual impairment (16.8%, n = 48). There was a strong significant correlation between the use of visual impairment as the primary disability category and students receiving services at schools for the blind and visually impaired (Pearson correlation of .926, significant at p=.01). In fact, only two students of the 48 with a primary disability designation of visual impairment did not attend a

school for the blind or visually impaired. Thirteen percent (n = 37) of the 286 identified students were not receiving services for visual impairment. Of those 37 students, 35% (n = 13) were designated in the functional vision category of "primarily uses vision", 30% (n = 11) were designated as using "a combination of vision and other senses", and the remaining 35% (n = 13) were designated as "uses other senses in place of vision".

Approximately 16% (n = 46) of the students did not have an eye pathology or visual condition indication on their documentation forms, and an additional 5 % (n = 15) had an indication of "unknown". For the remaining students, cortical visual impairment was the most prominent visual condition indicated. Table 6 shows the breakdown of different visual conditions by grade level for those students with reported pathologies (n = 225). Please note that some students had more than one eye pathology or visual condition (e.g. cortical visual impairment and optic nerve hypoplasia).

**Table 6: Percent of Students with Specific Visual Conditions by Grade Level**

| **Visual Conditions** | **Grade 3/4 (n = 114)** | **Grade 7/8 (n= 111)** |
|---|---|---|
| Cortical Visual Impairment | 54% | 41% |
| Optic Nerve Hypoplasia or Atrophy | 10% | 13% |
| Retinopathy of Prematurity or Retinal Detachments | 9% | 8% |
| Visual Impairment due to a Syndrome | 6% | 3% |
| Field Loss from Hemianopsia or Retinitis Pigmentosa | 5% | 3% |
| Cataracts or Aphakia | 3% | 8% |
| Nystagmus | 3% | 2% |
| Glaucoma | 1% | 1% |
| Albinism | 1% | 0% |
| Other conditions | 12% | 26% |

### 3.4.2 Creating Student Matches

Students without visual impairments who took the 3/4 or 7/8 A level PASA were selected as matches for each student with visual impairments for each content assessment (math and reading) taken. Due to the heterogeneous nature of the population of students who take the PASA, matching was primarily done by mean score but supplemented with a skills checklist (see Appendix B). The addition of the skills checklist as a matching variable was used as an attempt to get the closest match possible given the possibility of confounding variables.

Skills checklists were filled out by the students' teachers and sent in with enrollment information as part of the assessment process. Each checklist, one for reading and one for math, contained 19 questions for which teachers indicated a student's ability to perform a specified type of skill using a three point Likert scale (always, sometimes, never). All possible students without visual impairments who had the same mean score on the PASA as a student with visual impairments and were at the same grade level were considered as a possible match. The checklist profiles of these possible matches were compared to the student with visual impairments by calculating the total number of differences for each item on the ability checklist. For example, if a possible match student was given a score of 1 (never) for a skill, and the student with visual impairments received a score of 3 (always), the difference for that checklist item would be 2. The closer the checklist profiles, the smaller the sum of the absolute value differences of the checklist items. The selected match student was the one who had the same mean score and the closest ability checklist profile (the smallest difference score). If a student with visual impairments did not have any possible matches based on exact mean score, the next closest mean score was used supplemented by the closest checklist profile. The highest summed difference score possible for a checklist would be 38 (all items opposite of each other: never vs. always).

Overall, at the third grade level 87% of the checklist differences of matched students were 10 or smaller and 35% were perfect matches or only had a difference of one. At the seventh grade level 95% of the matched checklists had differences of 10 or smaller and 43% were perfect matches or only had a difference of one. Approximately 4% of the students with visual impairments at each grade level did not have a checklist on which to match.

### 3.4.3 Additional Data on Students with Visual Impairments

In addition to assessment scores and ability checklists, additional data collected from the assessment videotapes were available for approximately 90% (n= 257) of the 286 students with visual impairments. An additional 7% (n = 21) of the assessments were not reviewed because they were submitted as narrative notes, and the remaining 3% of tapes (n = 8) were not available for review (i.e. missing or not duplicated). Data from the assessment videotapes were collected by a total of 22 tape reviewers who were all certified teachers of the visually impaired (TVIs) from across the State of Pennsylvania. The teachers came from a variety of service delivery settings, as delineated in Table 7, and participated in one or more tape review weekends.

**Table 7: Number of TVIs Reviewing Tapes by Service Delivery Model**

| Service Delivery Model | Number |
|---|---|
| Itinerant teachers of the visually impaired | 8 |
| Classroom teachers at schools for the blind and visually impaired | 8 |
| University instructors in programs related to visual impairment | 2 |
| Classroom teachers in special education | 2 |
| Resource room teacher | 1 |
| Itinerant adult service provider in orientation and mobility/rehabilitation | 1 |

Twelve of the 22 teachers (54.5%) had administered the PASA to a student with visual impairments, at least in part. Four teachers (5.5%) had scored the PASA assessment at a scoring conference. Twenty of the 22 (91%) currently work directly with students who have visual impairments and all teachers have had at least some experience working with students who have multiple disabilities.

Prior to participating in a tape review weekend, teachers were sent pre-training documents giving them an overview of how to score the PASA and some practice scenarios. Teachers who had never scored the PASA before were also sent a CD with short video clips to give them some additional scoring practice. During the weekend, teachers were provided with a 3.5 hour training where they learned more about scoring the PASA and learned the specific coding procedure for collecting the additional data (see Appendix D for the code sheets). Tape reviewers were given the scores that each student received on the assessment and then watched each test item on the performance assessments and recorded a series of observations that included:

1. *The reason the student received the assigned score:* Based on the professional judgment of the tape reviewers and specified guidelines on which they were trained, they indicated for each test item whether the score reflected the student's perceived ability with the skill, reflected a lucky guess, or reflected an error in administration. There was also a place for reviewers to indicate if they did not agree with a score that was assigned to a test item and the reason.

2. *The accommodations used to provide access to the test items*: For each test item, the

tape reviewers used a coding sheet to decide upon and record any accommodations used during the test to accommodate vision and/or other disabilities. The codes were divided into five main categories: substitution accommodations (e.g. replacing objects for pictures), picture or object enhancement accommodations (e.g. making a picture tactual), layout or set-up accommodations (e.g. creating a defined space to find choices), instruction accommodations (e.g. alternate wording), and response accommodations (e.g. use of an augmentative communication device). For certain codes within each main category, tape reviewers were asked to specify the accommodation. For example, if a reviewer coded that a low vision device was used for picture enhancement, s/he was also asked to specify the particular type of device such as "magnifier" or "Closed Circuit Television". If no accommodations for vision were present, the tape reviewers indicated, based on their professional opinion and their observation of how the test item was performed, whether the item was "okay as is" or "needed accommodations". If they selected "needed", then they also specified the reason.

3. *A judgment on how the accommodations affected the skill intent:* For both accommodations for vision and other accommodations, tape reviewers were asked to indicate whether they felt that the accommodations (or one in particular) changed the intent of the skill by circling "N" for no or "Y" for yes, and specifying the reason if they indicated yes.

Inter-rater reliability was conducted on these data by comparing the codes to a standard for a percentage of the total assessments. The researcher re-coded one subject test (either math or reading) of approximately 20% of the students. This amounted to a reliability check on approximately 10% of the assessments (math and reading together) distributed across the two

grade levels and the two content areas. Tapes were randomly selected with attention given to re-coding at least one subject assessment for each teacher of the visually impaired who participated in the data collection. Reliability was calculated as percent agreement for each sub-component of the data: reason for score, score changes, accommodations, and change in skill intent. Percentages were computed globally and by test item for each subject test at each grade level. In order to balance the varying numbers of accommodations that each assessment tape might contain, percent agreement for the accommodations section was conducted by first calculating the percent agreement for each individual tape (corresponding to one tape reviewer) on each item, then summing the percentages of all the tapes for that item and dividing by the total number of tapes. Reliability obtained from this procedure is reported in the results section.

## 3.5 DATA ANALYSIS PROCEDURES

In order to explore the research questions as fully as possible, a mixed methods approach was used. Information from student documentation forms, aspects of the test items, test item scores, and data from the tape reviews were all reviewed and analyzed.

### 3.5.1 Procedures for Research Question One

A descriptive and inferential approach was taken to answer research question one: Were there significant differences in the scores of students with visual impairments at different functional vision levels on the 2005 grade 3/4 or 7/8 A level PASA?

Percentages of students with visual impairments that were classified as proficient, novice, and emerging on the assessment were calculated. Within the categories of proficient and novice, percentages of students at each level of functional vision achieving these classifications were also calculated and discussed. In addition, the Kruskal-Wallis test, a non-parametric version of a one-way ANOVA, was conducted for grades 3/4 and 7/8 in math and reading to compare mean score differences between students with visual impairments grouped at the three different levels of functional vision: primarily uses vision (V) for most tasks, uses a combination of vision and other senses (CV) for most tasks, and uses other senses in place of vision (NV) for most tasks. Significance was tested at the p=.01 level and a Mann Whitney U test was conducted as a post-hoc analysis to confirm between which groups significance was found for total score means.

### 3.5.2 Procedures for Research Question Two

Accommodation patterns were analyzed to answer question number two: What accommodations did teachers make to adapt the 2005 PASA for students with visual impairments? Particular focus was given to discovering the relationships between accommodations and the identified variables in the following sub-questions:

1. Are there relationships between the types of accommodations made and level of functional vision or type of test item?
2. Were there accommodations that seemed to change the intent of the skills being tested?

In order to explore the relationships between accommodations and these other variables, several approaches were taken. For accommodations made by level of functional vision, descriptive and Chi-Square statistics were calculated and any emerging patterns were noted. For relationships with test-item types, the patterns of accommodations provided to students at the three levels of

functional vision across test items in an assessment were reviewed for any changes in patterns that were noteworthy. For example, if an accommodation that was being provided to about 10 students consistently across test items suddenly drops to only 3 students on an item, that item was reviewed to gauge if a drop in the accommodation would be expected. Noteworthy changes in pattern were indicated and discussed. Finally, to explore accommodations that changed the skill intent, the accommodations and opinions of the tape reviewers were reviewed for items they flagged as having a change in skill intent. Common accommodations that were consistently flagged as changing the skill intent were noted.

### 3.5.3 Procedures for Research Question Three

A statistical and judgmental approach was taken to explore question number three: Were there significant differences on individual 2005 level A PASA math and reading test items at the 3/4 and 7/8 grade levels of students with visual impairments as compared to students without visual impairments who had similar ability profiles on the constructs of interest?

Differential item functioning (DIF) analysis was used to explore this question statistically. Traditional or popular methods of exploring DIF were not used due to the large sample size requirements of most of these tests. Typically with item response theory (IRT) analyses, focus and reference group members are stratified into ability categories with the need for a sufficient amount of subjects to be within each grouping. The typical negatively skewed distribution of students taking the A level PASA combined with smaller sample sizes, particularly when students were divided into the three functional vision levels, prohibit the effective use of these techniques.

Instead, the Wilcoxon Signed Ranks Matched Sample test was used. In place of stratification, students with visual impairments were matched with a student without visual impairments as described in the participants section of this chapter. Matching was done on total test score (mean score), with a secondary matching variable using the checklist profiles. While this nonparametric test will not identify at what level of the matching variable differential item functioning is suspected (e.g. for students with high total mean scores versus low total mean scores) since students are all pooled, matching on total test score should provide a test of overall DIF. Another benefit of the Wilcoxon Matched Sample Signed Rank test as compared to conducting a series of matched sample t-tests is that the Wilcoxon takes into account the direction and magnitude of the differences of each test item score and more easily allows for closer inspection of the distributions of scores by providing the number of negative and positive ranks, and ties between the matched sample. It also corresponds with the ranked nature of the data itself. In addition, being a non-parametric test, fewer assumptions are potentially violated due to the skewed nature of the distribution of students taking the PASA. All items in math and reading at grades 3/4 and 7/8 were tested at the p=.05 level. This level was chosen over .10, which was originally considered to more thoroughly explore possible areas of DIF, because Type I error rate is already inflated when conducting multiple comparisons on the same sample.

The judgmental review was conducted to gather information about the factors that experts in the area of visual impairment consider when making a decision about whether a test item is in need of accommodations and whether the item is measuring the same construct after adaptation. The flagged set of items that reviewers felt may be problematic or different for students with visual impairments were also used as a comparison to the statistical method used to answer this question. Two experts in the field of visual impairment reviewed each actual test item in the 3/4

and 7/8 A level assessments. Both experts who conducted the review have significant experience with children with visual impairments. One expert has a specialty in literacy for students with visual impairments and the other has considerable training in working with students who have multiple disabilities including deaf-blindness. Both test item reviewers evaluated the PASA test items independently using a set of instructions provided to them (see Appendix E). Using these instructions, the reviewers considered each test item "as is" for each level of functional vision (V, CV, NV) and marked those items that they felt might function differently for those students. They then reconsidered the items when accommodated at each level of functional vision and considered whether the item functioned differently after the accommodations were made (e.g. easier, harder, changed intent, etc.). When doing this second portion of the review, the reviewers specified the type of accommodations they were considering. The independent review served as a catalyst for two group conferences consisting of the reviewers and the researcher. The purposes of the conferences were to come to consensus and to discuss the particular issues surrounding accessibility of the test items for students with visual impairments. The researcher guided the conferences, posed questions, and recorded the comments and decisions of the test item reviewers.

### 3.5.4 Procedures for Research Question Four

Descriptive and qualitative analyses were conducted to explore question four: Considering the accommodations made and student performance on different types of test items, what are the potential reasons that "flagged" test items functioned differently?

Since the DIF results from the statistical test in question three do not provide insight into the reasons items were significant, a logical analysis of the factors that impact the test questions

was necessary. In order to try and tease out the potential reasons for DIF, the number of lucky guesses, occurrences of administration errors, scores of zeros, and suggested increases and decreases in score changes were extracted from the videotape data for these items. These frequencies were compared to the frequencies of an item measuring the same construct that was not flagged- if available. Any apparent patterns were noted. Changes in the number of particular accommodations between significant and non-significant items of the same skill were also considered as well as the judgmental item reviewers' comments and tape reviewers comments. In particular, significant items that were flagged were reviewed for materials or contexts that may be outside of the experience of the students with visual impairments or patterns in suggested accommodations that were needed. Again, noteworthy patterns were noted and discussed.

# 4.0 RESULTS

The purpose of this research study was to explore the performance of 286 students with visual impairments who took the 2005 A level PASA math and reading assessments at grades 3/4 or 7/8. In particular, inferential, descriptive and qualitative analyses were conducted on individual item and mean scores of the students and their matched peers and on supplemental accommodations data of students with visual impairments. The overall goal of the study was to discover the potential factors that may affect the performance of students with visual impairments on the PASA and to better understand sources of test bias and challenges in making accommodations. Results for all four research questions are presented in this section.

## 4.1 QUESTION ONE: PASA ACHIEVEMENT

### 4.1.1 Achievement Categories

Since the PASA is used as part of AYP calculations for schools, scores are transformed into proficiency levels that can be reported with the standard state accountability assessment. At the A level, the following cut scores, displayed in Table 8, were used to classify student achievement on the PASA for AYP purposes in 2005 (PASA technical supplement, 2005):

**Table 8: Cut-Scores for A Level Proficiency Categories**

| Proficiency Category: | Cut Score Range (mean raw score): |
|---|---|
| **Advanced** | Not possible at A level |
| **Proficient** | 5.0 (perfect) |
| **Novice** | 3.75-4.99 |
| **Emerging** | 0.0-3.74 |

Of the 286 students with visual impairments taking the A level PASA, no students achieved proficiency on mathematics, and six students (2%), all at the 7/8 grade level, achieved proficiency on reading. Because the A level proficiency category is very stringent, the novice category was also explored. On the mathematics assessment, 77 students (27%) achieved novice status, and 71 students (27%) achieved the novice status on the reading assessment. Of the total number of students achieving proficient status, 83% of them were students who primarily used vision (V) for most tasks. Only one student (17%) in the CV group, uses a combination of vision and other senses, also achieved proficiency in reading. The same pattern is evident at the novice level. Of the total number of students achieving novice status on the PASA, 51% and 58% in math and reading respectively were classified in the V functional vision group. An additional 43 and 34% of the novice students in math and reading fell into the middle functional vision category- the CV group. Table 9 shows the breakdown of students with visual impairments in the three functional vision categories who achieved proficient, novice, or emerging status compared to the total number of students within their own level of functional vision grouping. It should be noted that the total number of students in this analyses is reduced slightly since students with missing scores on any item could not be included in proficiency classifications.

**Table 9: Percent Proficiency Classifications within each Functional Vision Level**

| | Reading | | | Math | | |
|---|---|---|---|---|---|---|
| | **Proficient** | **Novice** | **Emerging** | **Proficient** | **Novice** | **Emerging** |
| **V** | 9% (n=5) | 51% (n=35) | 42% (n=29) | 0% (n=0) | 48% (n=38) | 52% (n=41) |
| **CV** | 1% (n=1) | 16% (n=23) | 83% (n=116) | 0% (n=0) | 22% (n=32) | 78% (n=115) |
| **NV** | 0% (n=0) | 11% (n=5) | 89% (n=40) | 0% (n=0) | 11% (n=5) | 89% (n=41) |

### 4.1.2 Mean Score Comparisons

In addition to the above comparisons of students with visual impairments in different functional vision categories, the non-parametric Kruskal-Wallis test was also conducted to compare the mean scores and individual test item scores of students by functional vision level at each grade and subject. Table 10 shows the mean scores and standard deviations for each functional vision grouping by grade and subject assessment.

**Table 10: Mean total Score out of 5.0 by Functional Vision Level**

| | A Level PASA Mean (Standard Deviation) | | | |
|---|---|---|---|---|
| **Functional Vision Level** | **3/4 Math** | **3/4 Reading** | **7/8 Math** | **7/8 Reading** |
| **V** | 3.36 (.908) | 3.50 (.936) | 3.71 (1.14) | 3.73 (1.36) |
| **CV** | 2.54 (1.16) | 2.68 (1.17) | 2.76 (1.20) | 2.74 (1.09) |
| **NV** | 2.33 (1.16) | 2.47 (1.29) | 2.23 (1.20) | 2.34 (.985) |

Based on overall mean score, there was a significant difference in total score between students with visual impairments by different functional vision classification at the $p \leq .01$ level for grades 3/4 and 7/8 in both math and reading. The .01 level was used to help account for the increase in Type I error rate from conducting multiple comparisons. Even if a Bonferroni adjustment were used in place of an alpha =.05, the majority of items reported below would still have been significant.

A review of the mean ranks, as shown in Table 11, reveals that students with visual impairments who primarily use vision (V) scored better overall on the A level PASA than students who use a combination of vision and other senses (CV) and students who use other senses in place of vision (NV). The mean ranks listed in the table represent the sum of the rank orders of student raw scores. Ranks were assigned to scores as a whole group with a rank of one corresponding to the lowest score. These ranked numbers were then redistributed into the functional vision categories for comparison. Therefore higher mean ranks correspond to higher scores on the PASA.

**Table 11:Mean Ranks and Significance of Total Test Score by Functional Vision Level**

| | Mean Ranks by Vision Level | Chi- Square Statistic (df=2) | Significance |
|---|---|---|---|
| **3/4 Math** | V: 92.32<br>CV: 63.36<br>NV: 54.13 | 16.590 | .000 |
| **7/8 Math** | V: 90.63<br>CV: 60.50<br>NV: 45.57 | 25.108 | .000 |
| **3/4 Reading** | V: 84.95<br>CV: 58.78<br>NV: 53.46 | 13.100 | .001 |
| **7/8 Reading** | V: 87.09<br>CV: 56.92<br>NV: 44.32 | 24.103 | .000 |

These results were confirmed by conducting a post-hoc analysis using the non-parametric Mann-Whitney U test to compare performance between the V and CV functional vision groups, the V and NV groups, and the CV and NV groups. As suspected by analysis of the mean ranks, V group students scored significantly better overall on the PASA at grades 3/4 and 7/8 in math and reading than both the CV and NV groups. This finding was significant at $p \leq .01$. There was no significant difference at either grade level on either subject test between the CV and NV functional vision groups.

A closer analysis by item of the Kruskal-Wallis revealed that this pattern of significance held true for most test items. Table 12 shows the number of skills in each grade and subject test that were *not* significant at $p \leq .01$ level.

**Table 12: Skills on which the V group Did *Not* Perform Significantly Better**

| 3/4 Math (7 out of 22 skills) | 7/8 Math (1 out of 24 skills) | 3/4 Reading (7 out of 23 skills) | 7/8 Reading (2 out of 20 skills) |
|---|---|---|---|
| ✓ Selects a lot (p=.078)<br>✓ Matches sets of 1 (p=.183)<br>✓ Selects most or least (p=.031, p=.011)<br>✓ Matches length-shortest and longest (p=.017, p=.056)<br>✓ Selects biggest-area and volume (p=.049, p=.011)<br>✓ Matches smallest-volume (p=.018)<br>✓ Selects heaviest or lightest (p=.016, p=.149) | ✓ Clock: Which one shows time? (p= .015) | ✓ Orients (p=.030)<br>✓ Matches identical objects-pencils, baggies, tissues (p=.080, p=.116, p=.066)<br>✓ Select object by function-pitcher (p=.382)<br>✓ Demonstrates function of object-toothbrush (p=.019)<br>✓ Selects similar objects-notebook/coloring book, book/magazine, crayon/marker (p=.174, p=.029, p=.133)<br>✓ Selects related objects-cup/straw, scissors/paper, soap/washcloth (p=.248, p=.618, p=.046)<br>✓ Answer literal what or who question- hat, boy (p=.032, p=.014) | ✓ Selects related picture-washcloth/soap (p= .020)<br>✓ Selects picture by function- cup (p=.013) |

In summary, the students who had more functional vision (V group) performed better on the PASA overall and on at least half of the individual test items at each grade level. There was

no significant difference in performance between the other two functional vision categories. Additional analyses of accommodations and other factors are needed to understand why this is.

## 4.2 INTER-RATER RELIABILITY

### 4.2.1 Accommodation and Change in Skill Intent Codes

Inter-rater reliability of the accommodations data available from student assessment videotapes was conducted by calculating percent agreement between tape reviewer data and a standard of re-coded tapes by the researcher. The varying numbers of accommodations per student were balanced by calculating percent agreement for each individual tape reviewer for an item, then averaging those percentages across all tape reviewers for that item. Overall, percent agreement for recorded accommodations observed on the videotapes was 88.8%. More specifically at the grade 3/4 assessment level, percent agreement on accommodations was 88.3% in math and 91.9% in reading. At the 7/8 assessment level, there was 85.1% agreement in math and 90.5% agreement in reading. Table 13 summarizes this information and shows the percentage ranges across individual test items. In addition it lists the percent agreement as to whether or not tape reviewers felt that accommodations changed the intent of the skill.

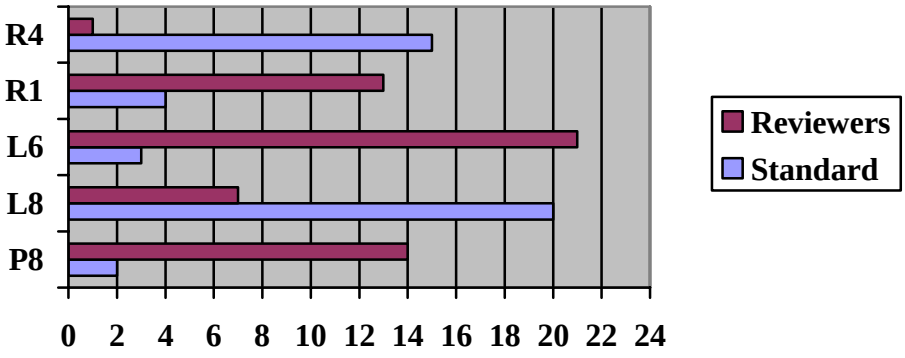
**Table 13: Percent Agreement in Math and Reading by Grade Level**

| Grade and Subject | Percent agreement: Accommodations | Ranges across test items | Percent Agreement: Change in skill intent |
|---|---|---|---|
| **Overall** | 88.8% | 72.1-100 | 98.7% |
| **3/4 Math** | 88.3% | 75.8-97.7 | 100% |
| **3/4 Reading** | 91.9% | 75.0-100 | 99.6% |
| **7/8 Math** | 85.1% | 72.1-97.7 | 98.1% |
| **7/8 Reading** | 90.5% | 81.1-97.7 | 96.8% |

A closer analysis of the reliability helps to put the percent agreement into perspective. Tape reviewers were documenting accommodations across a whole set of test items: 22 items in 3/4 math, 23 items in 3/4 reading, 24 items in 7/8 math, and 20 items in 7/8 reading. Observations were not mutually exclusive because the majority of students received similar accommodations on most test items. Some disagreements occurred across multiple test items for the same accommodation. Specifically, 63% of the disagreements in grade 3/4 and 59% in grade 7/8 were codes that occurred across several test items for one individual. There are two probable reasons for this to occur: either there was an accommodation that either the reviewer or standard felt was noteworthy but the other consistently did not, or there were test items where the code was consistently missed. Given the multiple test items reviewers watched containing the same types of accommodations for a student, it is likely they could have missed putting down a code for some items. In either instance, the high percentages of disagreements of the same code across test items suggests that reliability, in terms of the number of disagreements for unique codes within a student's assessment, would be higher.

In addition, Figure 2 illustrates the accommodation codes that had unequal patterns of disagreements- being either over or under reported by tape reviewers as compared to the

standard. The “standard” columns on the bar graph indicate the number of times the researcher used a code and the reviewer did not, and the “reviewers” column represents the number of times that the reviewers used a code and the researcher or “standard” did not. While many of these code discrepancies were repeated by the same few tape reviewers, they may still indicate that the reported results may be slightly high for eye gaze (R1), anchoring of objects (L6), and use of high contrast backgrounds (P8), and slightly low for reorientation of students to the location of answer choices (L8) and the use of other response modes not specifically on the coding sheet (R4).




**R4= other response modes; R1=eye gaze; L6=anchoring items; L8= reorientation; P8=high contrast background**

**Figure 2: Accommodation codes with unequal disagreement patterns**

A note should also be made about the percent agreements for “change in skill intent.” These reliability percentages were high, but seem to disguise the true pattern that emerged on the reliability tapes. The majority of accommodations on most test items were marked as not changing the intent of the skill by the reviewers and the standard, leading to high percent agreements. However, when looking at just those items where a change in skill intent was

suspected, only 20% were agreements of the reviewers to the standard. Sixty-seven percent were disagreements in which the standard suspected a change in skill intent, but the reviewers did not. The remaining 13% of the cases reflected the reviewers suspecting a change in intent, but the standard not. In other words, while overall agreement about intent was high, deciding whether or not accommodations changed the intent of the skill appears to have been a difficult task for everyone.

### 4.2.2 Reason for Score Codes and Score Changes

Reliability was also calculated for the "reason for score" and "score change" data collected from the videotape assessments of students with visual impairments. "Reason for score" refers to the tape reviewers' judgments about why a student received a particular score on a test item. Was it reflective of ability, a lucky guess, or some error in administration of the test item? "Score change" data refers to instances when tape reviewers indicated that they felt the given score assigned to the test item was incorrect. These data in combination with the accommodations data were used to logically analyze for potential reasons why certain test items were flagged as significant for DIF. As with the other portion of the reliability check, percent agreement was calculated. Overall, reliability was 85.2% for the reason for score data and 88.7% for score change data. Table 14 shows the reliability percentages by grade and subject along with the percent agreement ranges across test items.

**Table 14: Percent Agreement for Reason and Score Change Data**

| Grade and Subject | Percent agreement: Reason for score | Ranges across test items | Percent agreement: Score changes | Ranges across test items |
|---|---|---|---|---|
| **Overall** | 85.2% | 63.6-100 | 88.7% | 63.6-100 |
| **3/4 Math** | 86.4% | 72.7-100 | 89.6% | 72.7-100 |
| **3/4 Reading** | 90.5% | 72.7-100 | 92.9% | 72.7-100 |
| **7/8 Math** | 77.3% | 63.6-100 | 89.0% | 72.7-100 |
| **7/8 Reading** | 88.2% | 72.7-100 | 82.7% | 63.6-100 |

Again, a closer analysis adds some perspective to these results. It appears that tape reviewers had a greater tendency than the standard to assign a reason code of "ability" to test items. At grade 3/4, 42% of the disagreements in reason codes was due to scorers assigning an "A" for ability when the standard gave the same test item an "L" for lucky guess. At grade 7/8, 23% of the disagreements fit this distinction. This could be due to tape reviewers not wanting to misjudge a student since indications were being made based on a subjective sense of the child's response on each item. For some students it was difficult to make this determination.

Reviewers also had a tendency to assign "A" over "E" or "error in administration", with this distinction varying consistently between tape reviewers. In particular, this tendency was apparent for students who received scores of all ones or twos and were not making direct selections of choices. One of the decision-making rules given to tape reviewers in training was to consider an error in administration for low scoring or passive students when the student was not given a chance to respond. Some tape reviewers followed this rule more diligently than others. For example, when a student was very unresponsive on the assessment tape and the teacher was just going through the motions, some reviewers felt that an attempt to accommodate should have been given regardless and assigned an "E" for reason. Other reviewers, however, felt that the

teacher must know that the student could not perform on the test and therefore assigned an “A” for ability to those test items. As with the accommodations codes, inter-dependency also affected the reliability of the reason for score codes. If a reviewer assigned “A” for the situation just discussed, it was likely that that same code was assigned to all test items and vice versa if “E” was selected. At grade 3/4, 51% of the disagreements were from the standard assigning “E” for error and reviewers assigning “A” for ability. At grade 7/8 this percentage was 68%. As with accommodations, reliability percentages reported reflect the repeated disagreement regarding a reason code. Percentages would be higher when considering just the “unique” disagreements for an individual student.

Score change disagreements also followed an interesting pattern. The majority of disagreements, 63% in grade 3/4 and 79% in grade 7/8, were due to the standard indicating a score change and not the reviewers. An additional 11% of disagreements in 3/4 and 6% in 7/8 were due to both the standard and reviewer indicating a score change from the original assigned score, but assigning a different replacement. Score change disagreements were not as affected by inter-dependency between items. The only time this was apparent was when a reviewer felt that a student who received all ones for passive participation actually deserved twos for active participation. However, 64% of this particular score change was in accordance with the standard.

## 4.3 QUESTION TWO: ACCOMMODATIONS

### 4.3.1 No Vision Accommodations by Functional Vision

When a student did not appear to have any accommodations made for visual impairment on a test item, tape reviewers were asked to indicate whether, in their professional opinion and based on what they observed, the student needed an accommodation. Differences in the percentage of students not receiving vision accommodations and those indicated as needing them varied slightly between grade levels, content areas, and test items. These ranges are specified in Tables 15 and 16. Calculations for students not receiving vision accommodations are based on the minimum and maximum number of students not receiving accommodations for any given test item divided by the number of students within that functional vision grouping taking the assessment. Percentage ranges of students who might have benefited from accommodations were calculated by dividing by the average number of students from the functional vision grouping range of students *not* receiving accommodations.

**Table 15: Ranges of Students Not Receiving Vision Accommodations in Math by Item**

| | **3/4 Math Functional Vision Levels:** | | | **7/8 Math Functional Vision Levels:** | | |
|---|---|---|---|---|---|---|
| **Percentage ranges of:** (number of students) | **V** | **CV** | **NV** | **V** | **CV** | **NV** |
| **Students *not* receiving vision accommodations per test item** | 28-50% (9-16) | 21-43% (16-33) | 15-35% (3-7) | 33-58% (13-23) | 17-37% (10-22) | 9-23% (2-5) |
| **Students who might have benefited from accommodations*** | 9-37% (1-4) | 23-47% (6-12) | 24-71% (1-3) | 6-17% (1-3) | 13-39% (2-6) | 34-100% (1-5) |

*percent ranges based on average number of students in that grade and subject not receiving vision accommodations

**Table 16: Ranges of Students Not Receiving Vision Accommodations in Reading by Item**

| | 3/4 Reading Functional Vision Levels: | | | 7/8 Reading Functional Vision Levels: | | |
|---|---|---|---|---|---|---|
| **Average Percentage of:** (number of students) | **V** | **CV** | **NV** | **V** | **CV** | **NV** |
| **Students *not* receiving vision accommodations per test item** | 30-63% (8-17) | 31-46% (22-33) | 20-35% (4-7) | 31-46% (11-16) | 12-18% (7-10) | 5-14% (1-3) |
| **Students who might have benefited from accommodations*** | 0-27% (0-3) | 24-47% (6-12) | 18-70% (1-4) | 7-22% (1-3) | 23-58% (2-5) | 61-100% (1-3) |

*percent ranges based on average number of students in that grade and subject not receiving vision accommodations

For the most part, these results are not surprising given the definitions of the different functional vision groups and the changes that occur in the test format between grades 3/4 and 7/8. One would expect that more students in the CV and NV categories would need accommodations for vision and thus the average percentages of students not receiving accommodations for a test item would be lower as usable vision decreased and as one compares grade 3/4 to 7/8. For the most part, this is what occurred. It also makes sense that percentages in general regarding the tape reviewers' opinion of the need for accommodations increased as usable vision decreased and increased as one moved from grade 3/4 to 7/8. Chi-Square tests confirm that there was a significant difference at the $p \leq .01$ level among the three levels of functional vision at all grade and subject levels in terms of the number of students not receiving accommodations. Post-hoc analyses at $p \leq .05$ indicate that the V group was significantly different from the CV and NV groups, except for 3/4 Math for the CV group ($p=.858$). The CV and NV groups were also not significantly different from each other for grade 3/4 Math ($p=.637$).

In terms of the number of students who the tape reviewers felt might benefit from accommodations, significant differences were present among the three levels of functional vision for all assessments except grade 7/8 reading. A post-hoc analysis revealed that the V group was

significantly different from both the CV and NV groups on the remaining assessments, but no significant differences were present between the CV and NV groups. Table 17 summarizes the Chi-Square test statistics among the three functional vision groupings.

**Table 17: Chi-Square Statistics among Functional Vision Groups**

| | No Accommodations | | Accommodations Needed | |
|---|---|---|---|---|
| **Grade and subject:** | **Chi Square (df=2)** | **Significance** | **Chi Square (df=2)** | **Significance** |
| **3/4 Math** | 28.128 | .000 | 13.520 | .001 |
| **7/8 Math** | 186.983 | .000 | 19.833 | .000 |
| **3/4 Reading** | 16.706 | .000 | 33.030 | .000 |
| **7/8 Reading** | 198.694 | .000 | 5.964 | .051 |

One interesting aspect of these results is that an average of four students in the NV group of "uses other senses in place of vision" did not receive accommodations on some test items. This is particularly surprising on the grade 7/8 reading assessment that is predominantly picture based. Given the definition of the NV functional vision category, one would expect at least the accommodation of L1-allowing the student to feel each object/picture- to be noted for all students. It is possible that tape reviewers missed making this notation. However, comments from tape reviewers such as, "Test administrator stated that the student cannot see the pictures but there's no way to adapt this test", "Teacher made no appropriate accommodations, seemed to be protesting the test", and "Teacher does not orient, no time to explore" suggest that for some students at least, the test administrators appear to have just gone through the motions of the test, felt that the student could not do the skills on the test, and/or were not aware that accommodations can be made to the materials. Considering the percentage ranges for needed accommodations, it is possible that this may be the case for some students at the CV level of

functional vision as well. In addition for the NV category, some of the suggestions by the tape reviewers for accommodations included using high contrast and placing materials in the student's field of view. It is possible that some students categorized as using other senses in place of vision were still encouraged to use their vision or were incorrectly categorized by the test-administrator.

### 4.3.2 Frequency of Accommodations by Functional vision

Initially, tape reviewers were asked to separate the vision accommodations from accommodations for other reasons when recording their observations of test items. However, many tape reviewers noted difficulty in making this distinction since some accommodations potentially served dual purposes for this population of students. For this reason, all accommodations recorded for students with visual impairments were collapsed together for analysis. The number of accommodations that teachers provided to an individual student (for vision or other reasons) on any one test item ranged from 0-8. For most students, the same type and numbers of accommodations were generally used across all test items, as is evident by a consistent pattern of the number of students receiving a particular accommodation within an assessment on each item. For example, if a student used a slant board and colored pictures on the 7/8 reading assessment, the student was likely to use these two accommodations on most of the test items. Therefore, if 10 students at the V level of vision used colored pictures in one item, the pattern of 10 students was generally seen across most test items within a couple of students. When it was not, it was analyzed to see if the difference was noteworthy. Of course, it should be noted that the students receiving a particular accommodation on one item may not always be the same students receiving it on another item. This may be the case in some instances, but notes

taken during data entry suggest that the majority of the patterns observed are due to the same students receiving the same accommodations across items.

Overall, layout/set-up accommodations were used most frequently with students across grade level and subject assessments. Out of a total of 29 accommodations for which the reviewers coded including "other" categories (see Appendix D), only 12 different accommodations among the grade and subject assessments were used by 10 or more students on a test item at once. Table 18 shows the range and percentages of students receiving a particular accommodation on the same test item by functional vision grouping.

**Table 18: Ranges and Percents of Students with the Same Accommodation by Test Item**

| | V | CV | NV |
|---|---|---|---|
| **3/4 Math** | 0%-56%<br>(n = 0-18) | 0%-45%<br>(n= 0-34) | 0%-60%<br>(n =0-12) |
| **7/8 Math** | 0%-18%<br>(n = 0-7) | 0%-42%<br>(n= 0-25) | 0%-73%<br>(n= 0-16) |
| **3/4 Reading** | 0%-30%<br>(n = 0-8) | 0%-34%<br>(n = 0-24) | 0%-65%<br>(n = 0-13) |
| **7/8 Reading** | 0%-29%<br>(n = 0-10) | 0%-49%<br>(n = 0-28) | 0%-67%<br>(n = 0-14) |

As with the percentage ranges when no accommodations were given, the results for percentage ranges in Table 18 are also not surprising, except for the ranges in grade 3/4 math for the V group of functional vision. In general, a higher percentage of students were given accommodations on a test item as level of functional vision decreased. The jump in the range for the V group in grade 3/4 math compared to other grade levels is actually due to an anomaly on a couple of test items: "selects a lot or few", and "selects lightest or heaviest." For "selects a lot or few" test administrators were to lay out three pieces of paper forks or spoons in different

groupings. The color paper chosen by the teachers may or may not have been selected for the student's vision accounting for the sudden increase in the number of students receiving the accommodation "uses a high contrast background" (P8). Also, for the "selects heaviest or lightest" test items, the regular instructions ask the test administrator to have the student feel the weight of each item. The sudden increase in the number of V group students receiving the accommodation of "allows student to feel each item" (L1), would most likely be due to this fact and is an occurrence of over-reporting actual accommodations. The range without these test items for the V group at grade 3/4 in math would be 0-9 students.

Using the adjusted range, Chi-Squares confirm that there was a significant difference at $p \leq .05$ level between the high ranges for the groups on all assessments except grade 3/4 Math. Post-hoc analyses revealed that the V group was significantly different from the NV group on all significant assessments, but only different from the CV group in 7/8 Math. The NV and CV groups were only significantly different from each other in 3/4 Reading. Table 19 summarizes the statistics between all three groups.

**Table 19: Chi-Square Statistics among Functional Vision Groups**

| **Maximum number of students receiving an accommodation on the same test item between functional vision levels** | | |
|---|---|---|
| **Grade and subject:** | **Chi Square (df=2)** | **Significance** |
| **3/4 Math** | 5.341 | .069 |
| **7/8 Math** | 11.706 | .003 |
| **3/4 Reading** | 7.511 | .023 |
| **7/8 Reading** | 7.412 | .025 |

### 4.3.3 Most Popular Accommodations by Functional Vision

As suggested by the ranges of students receiving the same accommodation on the same test item just discussed, there were some accommodations used more frequently with students than others. Tables 20 and 21 summarize the top four accommodations for math and reading by functional vision groupings (V, CV, NV). The selections were determined by averaging the number of students within a functional vision group across all test items who received each accommodation and selecting the accommodations with the highest averages. The "top four" are listed in order for each grouping with the most frequent first.

**Table 20: Four Most Frequently Used Math Accommodations by Functional Vision Level**

| | Functional Vision Grouping: | | |
|---|---|---|---|
| **Grade and Subject:** | **V** | **CV** | **NV** |
| **3/4 Math** | Specifically reoriented the student to the location of each of the choices<br>Used a high contrast background<br>Held each object/picture in student's field of view<br>Anchored pictures/objects | Held each object/picture in student's field of view<br>Used a high contrast background<br>Used eye gaze<br>Allowed student to feel each object/picture | Allowed student to feel each object/picture<br>Used eye gaze<br>Created a defined space to find objects/pictures<br>Specifically reoriented the student to the location of each of the choices |
| **7/8 Math** | Used a high contrast background<br>Held each object/picture in student's field of view<br>Used a slant board<br>Anchored pictures/objects | Held each object/picture in student's field of view<br>Used a high contrast background<br>Allowed student to feel each object/picture<br>Anchored pictures/objects | Allowed student to feel each object/picture<br>Used objects in place of pictures<br>Specifically reoriented the student to the location of each of the choices<br>Created a defined space to find objects/pictures |

**Table 21: Four Most Frequently Used Reading Accommodations by Functional Vision Level**

| | Functional Vision Grouping: | | |
|---|---|---|---|
| **Grade and Subject:** | **V** | **CV** | **NV** |
| **3/4 Reading** | Held each object/picture in student's field of view<br>Used a high contrast background<br>Used a slant board<br>Used eye gaze | Held each object/picture in student's field of view<br>Allowed student to feel each object/picture<br>Used eye gaze<br>Used a high contrast background | Allowed student to feel each object/picture<br>Created a defined space to find objects/pictures<br>Specifically reoriented the student to the location of each of the choices<br>Used eye gaze |
| **7/8 Reading** | Held each object/picture in student's field of view<br>Colored the pictures<br>Used a slant board<br>Used objects in place of pictures | Held each object/picture in student's field of view<br>Used objects in place of pictures<br>Allowed student to feel each object/picture<br>Used a high contrast background | Allowed student to feel each object/picture<br>Used objects in place of pictures<br>Created a defined space to find objects/pictures<br>Held each object/picture in student's field of view |

The majority of the top four accommodations for each functional vision level coincide with that grouping's definition of functional vision. For example, "held each object/picture in the student's field of view" would be expected for students with low vision. Tape reviewers generally recorded this accommodation not only when the teacher held the object or picture up for the student or brought the materials closer on the table, but also when the student initiated doing so.

Although most of the "top four" were expected, there were also some surprises. First was the use of objects in place of pictures in the V vision group for six students on the 7/8 grade reading. If a student is labeled as primarily using vision for most tasks, the substitution of objects for the pictures seems as if it would be an unnecessary accommodation, at least from a vision standpoint. Instead one might expect more accommodations within the area of picture/object enhancement such as coloring pictures, which was the only one in this category that emerged in the top four for V level vision. A closer look at those V students who received object substitution

as an accommodation revealed that they were the same six students across all test items and that four of the six came from the same school for the blind. Among the other two, one was served in a different school for the blind and the other served by an Intermediate Unit (IU).

Another interesting "top four" accommodation was "used eye gaze" in grade 3/4 math and reading for four students in the NV vision group- uses other senses in place of vision. Intuitively this communication accommodation does not make sense for a student who is truly using other senses in place of vision. Again using the test items that were significant for DIF, a closer analysis of these students revealed that the students were the same four across most test items. All four students had cortical visual impairment. They all came from different schools, two attended special schools that were not schools for the blind and two received IU services. Interestingly, two of the students received some scores of twos, but also fours and fives on certain items. When receiving these higher scores, one student was marked by the tape reviewer to be answering based on ability and the other was marked as receiving the higher scores based on lucky guesses. In these cases, it could be that these students are utilizing vision for educational purposes more than their functional vision designation would suggest. This situation was also suspected in the results of those few students in the NV category not receiving vision accommodations on some test items. At least two of the four students analyzed were also among those students not receiving vision accommodations- for one of whom tape reviewers indicated that the test items were "okay as is."

Finally, what is really more noteworthy than the accommodations most frequently used are the accommodations that were rarely used if at all. The lack of use of certain accommodations that would be expected of the different functional vision groups is probably a direct reflection on the test materials and on the population of students to whom the A level test

is geared. The use of object substitution instead of tactual adaptations to pictures is one of the most prominent examples. For the NV grouping of students and possibly some of the CV students, tactual representations of the pictures provided in the 7/8 reading assessment would be difficult unless tactual symbols were used. As indicated by the judgmental item reviewers, for these tactual symbols to be useful, they would most likely need to be familiar to the student. If the student uses a tactual symbol system, it may be challenging for the teacher to figure out how to utilize it to fit the PASA test items. In fact, except for one occurrence in reading, the accommodation "making pictures/objects tactual" was only used in math for items that did not involve the need to actually identify the tactual representation. This seems like an appropriate way to use tactual adaptations for the PASA that is not specific to a student's communication system. These items included matching sets, matching length, identifying the result of addition or subtraction, and matching time and numbers. These tactual adaptations were used with 2-6 students in the CV and NV functional vision groupings combined.

The fact that other picture/object enhancement accommodations did not make the "top four" list may be due to the fact that pictures within the PASA assessment are already fairly large and bold. The lack of use of low vision devices and augmentative communication devices highlight an important area for discussion. Are these accommodations not typically used in the classrooms of this population of students, does the PASA not lend itself well to the use of these accommodations, or is the PASA accessible enough without them? Further investigation would be needed to determine whether low vision devices or augmentative communication devices are being routinely used in the classroom.

#### 4.3.3.1 Accommodation Patterns by Test Item

As discussed earlier, whether accommodations were frequent or not, the majority of accommodations were used consistently for students across all test items. When the accommodation distribution pattern changed (for example, 10 students in the CV category were using colored pictures on most test items, but one item showed a drop to 3 students), the majority of the occurrences were expected, particularly in math. In other words, coloring pictures would be expected for only those items in math involving pictures. Another expected occurrence was alluded to earlier in the increase in the coding of "allows student to feel each object" for heaviest and lightest test items.

There were some noteworthy changes, however. One of the most interesting accommodation pattern changes occurred in the 7/8 math assessment where object replacement was used for a range of 28-34 students on each test item across the three functional vision categories except for the test item, "comprehension- where question" in which object substitution dropped to being used for only 11 students. This is the only A-level item that contains complex pictures. The pictures- a playground, grocery store and farm- are named and the student is asked to select the one where swings can be found. Object replacement in this case would require using objects symbolically (e.g. use of a ball to represent the playground scene represented in the picture). This type of abstract substitution teachers may find problematic to make accessible to students who cannot functionally use the pictures provided. Paralleling the decrease in object substitution on this test item, there was a slight increase in the use of the accommodation: "made test item auditory in place of pictures/objects." However, the increase did not account for the full drop in the use of objects. A closer inspection of this test item for the CV and NV functional vision groups reveals that the pictures were used for some of the students, particularly for CV

students, with an increase of the accommodation of coloring pictures apparent. A few CV students were presented with the pictures, but the pictures were described.

In general, it is interesting to also note the type of test items where a pattern emerged even if just for a few students for making the test item auditory. In the 7/8 math assessment, for example, auditory substitution was employed for a few students for the skills of matching time and numbers, and for the test item, "selects daytime/nighttime activity" in which a student is provided three pictures- a bed, sunglasses, and a rake- and asked to select the one people usually use at night. Beyond the complex picture "where" question already discussed in 7/8 reading, auditory substitution seemed to be used across the board for one or two students, but in particular for one of the "selects category" questions that involved selecting the category of "things you ride in" and for a similar function question where the student was given three picture choices- cat, tree, and chair- and asked to select the one used like the target picture- a couch. Again, the increase in the use was very slight (by one or two students), but given the type of items these are, object substitution could be more difficult.

#### 4.3.3.2 Changes in Skill Intent

Investigation of changes in skill intent proved to be difficult, which was not surprising given the results of the inter-rater reliability check. Tape reviewers unpredictably flagged instances where they felt the accommodations may have changed the intent of the skill. Rarely was an item type flagged consistently for the same accommodation; however, a couple of patterns were noted. First, there were more instances of reviewers flagging an item for change in skill intent in math and in 7/8 reading. In 3/4 math the most frequent indication of change in skill intent was when the target item was a more attractive color than the distracter choices. In 7/8 math and reading the most frequent occurrence of indicating a change in skill intent was when

the test administrator deliberately named the pictures or objects when the instructions indicated otherwise. Rarely were the items flagged for change in skill intent DIF items. In addition, there was also a tendency for item reviewers to flag an item as a change in skill intent when other modifications were made such as a reduction in the number of choices that the student was given, or use of wording that helps the student such as “here’s a nice big one”. Because of the inconsistency and difficulty test reviewers had with deciding if a skill was changed, results are more likely an indication of teacher awareness about the effects of accommodations and the aspects of test items they attend to rather than a confirmation of the types of accommodations that change the intent of a skill.

In summary, the following items and accommodations were occasionally flagged, some of which also corresponded to a score reduction to a 3 according the PASA rubric on modifications (e.g. naming items):

- Matching numbers and time: The original test item involves the student matching a print number or print digital time with two pictures being the distracter choices. Substitution of objects for the picture distracters may decrease the likelihood of actually matching numbers or digital time. Paper may be what is being matched instead since only the target and answer choice remain in this medium. It should be noted that this example was used in the training of the tape reviewers, which may have increased the frequency that this item type was indicated as potentially changing the intent of the skill.
- Objects in place of pictures may change the level of the skill
- Category items: naming items instead of the category may change the skill; object substitutions feel the same and offer no help about the category label

- Naming items changes intent (e.g. for test items like “selects object named” where the test administrator names all the choices as an accommodation for visual impairment)
- Gave functions of items or description of items (particularly for items such as “selects object/picture by function”)

Further analysis would be needed to gain more insight into whether these examples or other instances of particular accommodation use actually change the construct being measured.

## 4.4 QUESTION THREE: DIF ANALYSIS

### 4.4.1 Inferential Outcomes

In order to explore test items that might be functioning differently for students with visual impairments under the various accommodation conditions just discussed as compared to their matched peers without visual impairments, the Matched Samples Wilcoxon Signed Ranks test was used. Comparisons were run for the students with visual impairments as a whole group, as well as within the three different levels of functional vision as compared to their matched peers without visual impairments. Although increased Type I error rate was not of primary concern for this exploratory study using smaller sample sizes than typical DIF procedures, the decision was made to “flag” items at the $p \leq .05$ level instead of a more liberal .10 due to the fact that multiple comparisons will already increase the chance of over-identifying.

In grade level 3/4, six test items in math and eight test items in reading were flagged for DIF at the $p \leq .05$ level for one or more vision groupings: V, CV, or NV. In grade 7/8, nine test

items in math and five test items in reading were flagged. When considering those items flagged for more than one vision group, 11 total instances in math and 13 instances in reading were flagged for DIF at grade 3/4, and 11 instances in math and 6 instances in reading were flagged for grade 7/8.

#### 4.4.1.1 Significant Items in Math

In math, 55% (6 out of 11) of the instances on the grade 3/4 assessment and 64% (7 out of 11) of the instances in grade 7/8 were items that were differentially *more difficult* for students with visual impairments within the vision groupings flagged. That means that 45% of the instances at grade 3/4 and 36% of the instances at grade 7/8 were indications of items being differentially *easier* for students with visual impairments in the flagged groupings as compared to their matched counterparts without visual impairments. Tables 22 and 23 show the breakdown of harder and easier items that were significant at $p \leq .05$ by vision level groupings.

**Table 22: DIF Items by Functional Vision Groupings on Grade 3/4 Math Assessment**

| | 3/4 Math: More Difficult | 3/4 Math: Easier |
|---|---|---|
| **All students with visual impairments** | ✓ Matches area- smallest<br>✓ Selects "money"- quarter<br>✓ Selects "money"- penny<br>✓ Finds smallest- square | ✓ Finds biggest- square<br>✓ Finds heaviest item |
| **V group** | ✓ Finds smallest- square | ✓ Finds biggest- square |
| **CV group** | ✓ Selects "money"- quarter | ✓ Finds heaviest item |
| **NV group** | None at $p \leq .05$ | ✓ Finds biggest- square |

**Table 23: DIF Items by Functional Vision Groupings on Grade 7/8 Math Assessment**

| | 7/8 Math: More Difficult | 7/8 Math: Easier |
|---|---|---|
| **All students with visual impairments** | ✓ Selects "shows time"- clock<br>✓ Finds group with least | ✓ Finds biggest- square |
| **V group** | ✓ Selects "shows time"- clock<br>✓ Matches length- longest strip of paper | ✓ Matches numbers (picture distracters) |
| **CV group** | ✓ Finds "money"- dollar bill<br>✓ Finds group with least | None at $p \leq .05$ |
| **NV group** | ✓ Matches area- smallest item | ✓ Matches set of 2- strips of pennies (pictures)<br>✓ Finds heaviest item |

Overall in math there seemed to be an emerging pattern of the diminutive version of a skill being more difficult (smallest vs. biggest) as well as money questions being more difficult. Conversely, items dealing with a more tactual based skill- finding heaviest or lightest- emerged as easier as well as some of the superlative versions of matching and selecting skills (biggest vs. smallest).

#### 4.4.1.2 Significant Items in Reading

In reading, 92% (12 out of 13) of the flagged instances on the grade 3/4 assessment and 67% (4 out of 6) of the flagged instances in grade 7/8 were indications of items being differentially *more difficult* for students with visual impairments within the vision groups flagged. That means that 8% (n=1) of the instances at grade 3/4 and 33% (n=2) of the instances at grade 7/8 were indications of items being differentially *easier* for students with visual impairments in the flagged groupings as compared to their matched counterparts without visual impairments. Tables 24 and 25 show the breakdown of harder and easier items by vision level, significant at $p \leq .05$.

**Table 24: DIF Items by Functional Vision Groupings on Grade 3/4 Reading Assessment**

| | 3/4 Reading: More Difficult | 3/4 Reading: Easier |
|---|---|---|
| **All students with visual impairments** | ✓ Matches objects- pencils<br>✓ Answers literal what question- hat<br>✓ Selects object named- glove<br>✓ Matches objects- baggies | None at p≤ .05 or .10 |
| **V level functional vision** | ✓ Matches objects- pencils<br>✓ Answers literal who question- boy<br>✓ Selects related items- cup/straw<br>✓ Selects related items- paper/scissors<br>✓ Selects similar function- crayon/marker | None at p≤ .05 or .10 |
| **CV level functional vision** | ✓ Selects object named- glove<br>✓ Matches objects- baggies<br>✓ Selects similar function- crayon/marker | ✓ Selects related items- paper/scissors |
| **NV level functional vision** | None at p≤ .05 or .10 | None at p≤ .05 or .10 |

**Table 25: DIF Items by Functional Vision Groupings on Grade 7/8 Reading Assessment**

| | 7/8 Reading: More Difficult | 7/8 Reading: Easier |
|---|---|---|
| **All students with visual impairments** | ✓ Selects category- "things you ride in" | None at p≤ .05 |
| **V level functional vision** | ✓ Scans items<br>✓ Selects related items- soap/washcloth | None at p≤ .05 |
| **CV level functional vision** | None at p≤ .05 or .10 | ✓ Selects picture named- backpack |
| **NV level functional vision** | ✓ Scans items | ✓ Selects related items- sock/shoe |

Overall, the number of items flagged as more difficult for the V group of students: "primarily uses vision for most tasks" was surprising. In particular, this group seemed to have difficulty with the "selects similar" test items that involve being shown a target item and asked to select the other item from three that would be used the same way. It was expected that this grouping of students would need the least accommodations to have access to the assessment and was also the group out of the three functional vision levels that performed best overall. The other

emerging pattern at grade 3/4 was difficulty with “matches objects” test items where students are given a target item and asked to “find the one that is the same.” At grade 7/8 less of a pattern is noticeable and fewer items were flagged overall.

### 4.4.2 Judgmental Item Review

Along with a statistical exploration of DIF on the PASA, two experts in the field of visual impairments conducted a judgmental item review by considering each individual test item in the grade 3/4 and 7/8 level A assessments. After an independent review using a set of guidelines (see Appendix E), two conferencing sessions took place to discuss the items each reviewer flagged and the factors affecting their decisions. Flagging a single set of items that had the potential for DIF was challenging. Both reviewers felt that whether a test item was accessible and measured the intended skill depended on a wide variety of circumstances including the student’s level of experience and the types of accommodations made.

#### 4.4.2.1 Initial Independent Review

Despite working from the same set of instructions, the two reviewers took very different approaches to the independent review. For example, when reviewing the items “as is” (i.e. before accommodations), one reviewer took a literal perspective and flagged any items that did not explicitly state in the instructions a non-visual way of accessing the materials as negatively biased for the CV and NV levels of functional vision. This resulted in her flagging every item except the weight items on the 3/4 and 7/8 math test since the directions for this item type explicitly stated to have the student pick up each object. In contrast, the other reviewer flagged fewer “as is” items because she assumed the teacher would allow the student to touch the

materials if needed even if it was not explicitly stated. Both reviewers did agree that the assessment at both grade levels and in both subjects should be accessible "as is" to the students at the V level of functional vision. They also agreed that all items in the 7/8 reading test, which contains all pictures, were inaccessible to the NV level of functional vision without accommodations.

The two reviewers also took a slightly different approach when considering DIF for test items under accommodated conditions. This was most apparent within their reviews of the 7/8 A level reading test. One reviewer assumed that teachers would automatically substitute real objects for pictures, while the other reviewer considered using tactile symbols for many of the items. Consequently, the reviewer who assumed the use of real objects did not flag any items with accommodations for CV or NV level functional vision while the other reviewer flagged all the items as either containing DIF (n=12) or possibly 'containing DIF' depending on the circumstances (n=8). In addition, this reviewer also flagged seven items for the CV level of functional vision as possibly containing DIF, again depending on the circumstances.

#### 4.4.2.2 Results from Conferencing

Based on the results of the initial independent review, the reviewers were asked to reconsider the test items for the CV level and NV level of functional vision using a re-established framework. During two conferences, they considered the items under a variety of accommodations except context substitution. In other words, they discussed the different accommodation possibilities such as substitution of objects that represent the same thing as in the pictures, use of tactile symbols, use of simplified pictures, etc., but initially restricted themselves from considering substitutions of the type of items or pictures (e.g. replacing a couch and chair picture in a "selecting similar function" item with a brush and comb picture). Table 26

summarizes the items in math that the reviewers agreed to flag as having potential for being differentially harder for students with visual impairments at the CV or NV level of functional vision.

**Table 26: Items Flagged by Reviewers in Math**

| Grade Assessment: | Items Flagged: | Reasons and Comments: |
|---|---|---|
| **3/4 A Math** | ✓ Finds clock | A wall clock may be outside of the experience of B and C level functional vision. Replace with talking watch or alarm clock. |
| **7/8 A Math** | ✓ Match digital time- 2 occurrences in test (1:00 and 2:00) | If really trying to get at matching time for a student without usable vision, needs to be done orally. This one could be easier or harder. If objects are substituted may be easier by just matching two things that are not objects, but it could be harder if the objects are more distracting. |
| | ✓ Identify result of addition | Too many materials. Spatial layout is confusing, and a student doing this tactually cannot take in everything at once |
| | ✓ Find "money"- dollar bill | One distracter is too similar tactually and also potentially for reduced vision |
| | ✓ Identifies result of subtraction | Too many materials. Spatial layout is confusing, and a student doing this tactually cannot take in everything at once |
| | ✓ Selects "shows time"- clock | A wall clock may be outside of the experience of B and C level functional vision. Replace with talking watch or alarm clock. |

Of the items flagged in math, two of the items in grade 7/8 did come up as significant in the statistical test: find "money"- dollar bill for the NV level functional vision grouping, and selects "shows time"- clock for all students with visual impairments and the V level grouping.

Reviewers flagged more items in reading than they did in math. Based on the conversations during the conferencing, this may be predominantly because substitution of pictures or use of miniatures seemed more problematic for the reviewers and the focus of more of the discussion than making tactual comparisons on math skills. Table 27 summarizes the flagged items and comments for reading.

**Table 27: Items Flagged by Reviewers in Reading**

| Grade Assessment: | Items Flagged: | Reasons and Comments: |
|---|---|---|
| **3/4 A Reading** | ✓ Selects Category- "things you ride in" | Memory load issue with so many items, miniatures make it harder-"You don't ride in a toy car" |
| | ✓ Selects object by function- pitcher for "holds water" | Distracter is white out- bottle with liquid can be a hard distracter especially when explored tactually |
| | ✓ Selects similar function- 4 occurrences- coloring book/notebook, paintbrush/sponge, book/magazine, crayon/marker | Coloring and painting may be outside of the student's experience. Crayon/marker- replace with something student is familiar with to "make marks." Book/magazine- Item requires connecting two things that are very visual if not in student's own media. Tough call for CV level vision if it would be harder, but most likely for NV |
| | ✓ Literal who question- 2 occurrences-boy and girl | Miniatures are harder the worse vision is. An open-ended question might be easier for some students. |
| | ✓ Selects related object- scissors/paper | Students are less likely to have experience with scissors |
| **7/8 A Reading** | ✓ Selects picture named-3 occurrences- book, towel, backpack | Harder if tactual symbols used unless the student is using them. It is rare that tactual symbols stand for an object, but they stand for something more abstract like "bath time". Potentially easier if objects are used |
| | ✓ Selects picture by function- 2 occurrences- CD, Cup | Harder if tactual symbols used unless the student is using them. Potentially easier if objects are used, but might not help memory load, but complicate it. Perhaps oral is the way to go if the selection is not integral to the task |
| | ✓ Selects similar function- 2 occurrences- couch/chair, crayon/marker | Difficult one to represent tactually, it is unlikely that a child has a tactual symbol for both chair and couch. Use of miniatures makes this harder too. Coloring may be outside the experience of the student |
| | ✓ Selects category-2 occurrences- "things you eat with", "things you ride in" | Too many materials, possible miniature problem with animals and things you ride in (you do not ride in toys). Doing this item orally would depend on student's understanding of the category labels- the visual reminder of what the category means is not there |
| | ✓ Answers literal where question- complex pictures- playground/swings | Potentially harder done orally since there are no clues from the pictures to help (swings are pictured on the playground), tactual symbols would be harder if the student is not used to using them |

Of all the reading items that were flagged, three were also flagged as significant in grade 3/4 on the statistical test: answers literal who- boy, selects related- scissors and paper, and selects similar function- crayon/marker. What is interesting is that the "answers literal who" was significant for the V level of functional vision, not the CV and NV level that the reviewers expected. As mentioned earlier, it is also interesting that "selects related- scissors and paper" was significant for the CV level of vision because it was an *easier* item as compared to the matched group. In contrast, it was significant for being harder for the V level of functional vision which was the group the judgmental item reviewers felt would be most likely to have experience with scissors out of three functional vision groupings.

On the 7/8 grade assessment, two of the items flagged by the reviewers were also significant in the statistical test at $p \leq .05$: "selects category- things you ride in", and "selects picture named- backpack." However, the selects picture named item was significant for the CV level functional vision because it again was *easier* than its matched group counterpart, not because it was potentially harder for tactual students (NV level and possibly CV level) if symbols were used. Interestingly, the "things you ride in" category item was one of the items where a slight increase in the use of auditory accommodations was discussed earlier. The reviewers' concerns about memory load and the difficulty with using miniatures might have been on the mark for this test item.

#### 4.4.2.3 Reviewers' Considerations and Questions

It should be noted that the reviewers flagged the presented set of items with hesitancy, because again, both reviewers during this process felt DIF ultimately depended on a variety of factors. Table 28 lists some of the common factors and questions that emerged during the

discussions about whether items may exhibit DIF for the CV or NV functional vision groups of students with visual impairments.

**Table 28: Factors and Questions Impacting DIF Decisions**

| Factors Affecting Judgmental DIF Decisions | Additional Questions Posed |
|---|---|
| ✓ Student's level of experience and exposure to the contexts<br>✓ The amount of vision students in the B level functional vision category have and - grey area group<br>✓ The personal nature of tactual symbol systems. Does the student have symbols for the contexts presented? Is use of a symbol meaningful for those contexts?<br>✓ Use of miniatures, particularly for students who are tactual learners may make it harder<br>✓ Amount of materials - memory load issue<br>✓ Presentation of item- memory load issue<br>✓ Teacher decision on accommodations<br>✓ Material selection- E.g. use of sturdy, distinctive materials for size comparisons<br>✓ Tactual similarity of distracters with target item | ✓ Does familiarity with distracter choices affect the level of difficulty?<br>✓ For comprehension questions, is selection of the correct choice integral to the skill being assessed, or is answering orally equivalent? Does use of objects or symbols in place of pictures meant to help memory load help or hinder?<br>✓ Is using an object as a symbol more difficult than using a picture as a symbol because of the direct link of being able to physically use the object for what it actually is?<br>✓ Is progression from objects to abstract symbols to braille really the progression for this population of students considering how symbols must be presented? Is functional use more important to learning a system than whether it is concrete or abstract? |

## 4.5 QUESTION FOUR: DIF LOGICAL ANALYSIS

The comments and review provided by the judgmental item reviewers began to give some insight into possible reasons that certain types of test items might have been flagged as functioning differently for students with visual impairments. However, as discussed previously, there were surprises that emerged in the DIF statistical analysis that were not anticipated by the judgmental item reviewers which warranted specific analysis of the flagged items. Being a performance-based assessment and given the complexity of the student population who takes alternate assessments, there are a multitude of variables to sift through to begin to understand the reasons items may be functioning differently for students with visual impairments. Several possible explanations to consider for why items might exhibit DIF on the PASA include:

1. There is a fundamental difference in knowledge on a particular skill type for students with visual impairments.
2. The types of accommodations made changed the intent or level of the skill.
3. Insufficient accommodations were made to provide access to the test item.
4. The test item contains visually biased materials or contexts.
5. Students happened to guess accurately more frequently on a test item.
6. Errors in the administration of the item affected scores.
7. The heterogeneous nature of the population makes it difficult to predict.

In combination with the test item reviewers' global comments and the tape reviewers' comments about needed accommodations, data from the videotape assessments on the reason for scores, suggested score changes, and frequency of accommodations were analyzed for any emerging patterns. Flagged items were compared to non-significant counterparts if available. However, in 7/8 math in particular several of the flagged items were one of a kind in the assessment. The most noteworthy items to start with were the skills that were flagged more than one time (if more than one occurrence was in the assessment). These would be expected to be the items that might fundamentally be different for students with visual impairments compared to test items that surface one time due to a flaw in the testing materials or some other reason. The logical analysis began with these items that also had a non-significant comparison and then moved into considering the other grouping and other items that were single occurrences.

### 4.5.1 Skills Flagged Multiple Times with a Comparison

In grade 3/4 "selects related items" and "matches objects" were flagged more than once for some type of DIF for at least one vision grouping: all students with visual impairments, the V group,

the CV group, and/or the NV group. These constructs also had a non-flagged item on which to compare. Table 29, summarizes the results of the logical analysis for "selects related items". The 3/4 reading items listed were compared to "selects related- soap/washcloth" and the 7/8 reading items listed were compared to "selects related- crayon/marker." As a reminder, objects are the standard presentation materials for 3/4 reading, and pictures are the standard presentation format for grade 7/8 reading.

**Table 29: Logical Analysis Patterns of "Selects Related" to Item Comparison**

| **Comparisons:** | **Significant "Selects Related" items:** | | | |
|---|---|---|---|---|
| | **3/4 Reading Cup/Straw** | **3/4 Reading Scissors/paper** | **7/8 Reading Sock/shoe** | **7/8 Reading Soap/washcloth** |
| DIF Direction | V group- harder | V group-harder<br>CV group-easier | NV group-easier | V group- harder |
| Reason for Score Patterns | V- none | V- none<br>CV-higher instances of lucky guesses, less errors in administration, fewer zeros | NV- more instances of luck (3) | V- counter-intuitive: slight increase in lucky guesses and slight decrease in administration errors |
| Score increase/decrease patterns | V-none | V-none<br>CV- lower suggest score decreases | NV- none | V-none |
| Accommodations patterns | V-Slight increase (2 students) in students allowed to feel the object on non-significant item | V-Slight increase (3 students) in students allowed to feel the object on non-significant item | NV- none | V- non-significant item had a more frequent occurrence of reorienting the student to the choices |
| Needed accommodations patterns | V- allow to feel items, orient and reorient student to items, contrast needed with clear straw | V-few suggestions-no repeats<br>CV- answer choice was biggest/brightest | NV- few suggestions-not repeats | V- no repeat suggestions- one person: "use actual objects for a possible 5" |
| Item Reviewers comments | Felt this item was okay | CV and NV students may have less experience with scissors | This item should be okay and may be easier when objects are used | Item seemed okay |
| Probable Reason: | V- possibly needed more direct orientation to item and choices and better contrast | V- possibly needed more direct orientation to item and choices<br>CV- Function of reason for score | NV- probably due to lucky guesses; may also have been due to familiarity with content plus use of objects | V- possibly students needed orientation/reorientation or other accommodations |

From the analysis it appears that for the V group of students, there may be a need for more or different accommodations in order to ensure visual access to this particular type of test item. Perhaps test administrators over assume the students know where the choices are located since the item type involves making a comparison between a target and another set of items. The pattern is not overly strong, but the only pattern that emerged as an explanation for the V students. Of course, there is also the possibility that the V group has a weaker understanding of the concept of "goes with" (e.g. "which one goes with the straw") than their matched counterparts. The clear straw could also be considered a flaw in materials in terms of visual access for the V group. The easier "selects related" item for the CV group seems to be strongly explained by the combination of lucky guesses and lower instances of administration error. The pattern for the easier NV "selects related" item is less strong but probably explained by the increase in lucky guesses. However, since object replacement is what gave access to the majority of NV students, it is possible that the test item is made easier.

Table 30 summarizes the analysis for the item "matches objects" at the 3/4 reading level. The flagged items were compared to the non-significant item "matches objects-tissue boxes."

**Table 30: Logical Analysis Patterns of "Matches Objects" to Item Comparison**

| | Significant "Matches Objects" Items | |
|---|---|---|
| | **3/4 Reading Matches pencils** | **3/4 Reading Matches plastic baggies** |
| DIF Direction | All group- harder<br>V group- harder | All group-harder<br>CV group-harder |
| Reason for Score Patterns | All-none<br>A- counter-intuitive increase in lucky guesses for significant item | All- none<br>CV- none |
| Score increase/decrease patterns | All- higher level of suggested score increases (7 more)<br>V- slightly higher suggestion for score increases (2 more) | All- slightly higher level of suggested score increases (3 more)<br>CV- slightly higher level of suggested score increases (3 more) |
| Accommodations patterns | All, V- none | All, CV- increase in object substitution some were to replace clear baggies (5 instances) |
| Needed accommodations patterns | All- need for orientation/reorientation to objects; objects to match should be the same color<br>V- few suggestions, no repeats | All, CV-better layout and orientation to objects, don't use clear baggies, contrast |
| Item Reviewers comments | On the basis of matching should be okay. May be slightly harder if objects are novel, for NV spatial layout and multiple comparisons may be a problem | On basis of matching should be okay, clear baggies may be a problem if not replaced, for NV spatial layout and multiple comparisons may be a problem |
| Probable Reason: | All- Could be an instance of scoring inaccuracy, or possibly a need for better orientation to item choices<br>V- inconclusive | All, CV- could be an issue of materials or possible a scoring discrepancy issue |

While the pattern would have to be confirmed by checking the accuracy of the suggested score changes, the greater indication for score increases by the tape reviewers, particularly for "matches pencils", may be the prominent factor for this set of items. If accurate, five of the suggested increases for "matches pencils' were jumps of greater than one score category (e.g. from a score of a 1 to a score of 5). There also seemed to be a material factor for "matches baggies" similar to the one noted with the clear straw.

### 4.5.2 Other Skill Patterns of Items Flagged Multiple Times

Other skill groupings that were interesting in 3/4 and 7/8 Math from the DIF analysis were the multiple occurrences of "find money" being more *difficult* and the multiple occurrences of matching area-smallest, or selecting smallest being more *difficult*. Of particular interest is that the superlative "biggest" was often flagged as *easier* for students with visual impairments. Some of these items had a comparison test item that was not significant, others did not. The "money" items were analyzed together for emerging patterns as were the biggest/smallest occurrences of DIF.

Since all money items were flagged, comparisons focused on the materials, level of accommodations and recommendations by the tape reviewers for needed accommodations. From the analysis the most probable cause of the difference, particularly at grade 3/4 that used coins, was size of the target items. Use of accommodations of "bringing items into student's field of view" and " reorients student to location of answer choices" varied between the two "finds money" items. Tape reviewers stated the need for more contrast and more orientation/reorientation to the items as well. They also felt that the bigger and brighter distracters might have kept students from noticing or picking the correct coin choice. Fewer comments were made regarding the money item using a dollar bill by the tape reviewers, but the use of "reorienting student to the location of the items" was low for this item as well, and the judgmental item reviewers were concerned about the tactual similarity of the dollar bill to the receipt. It appears, too, that at least in some instances, fake money was used which would cause the money to look even more similar to a receipt for a student with diminished vision or tactual strategies.

For the biggest/smallest test items, there was an evident pattern of greater occurrences of lucky guesses for the biggest test items, but no differences in the accommodation patterns were evident. However, tape reviewer comments about needed accommodations centered on the small item getting lost among the big items, students being distracted by the big items so as not to choose the small item, and the need for orientation to the choices and better spacing of the materials. Also, the match volume item at grade 3/4 for smallest was not flagged for DIF like the "match area" item was. This most likely was due to the greater ease of experiencing the volume items tactually as well as visually. In fact, there was a slightly greater occurrence of allowing the student to feel the items for "match volume." In addition, the heaviest/lightest weight items showed a similar pattern of DIF where heaviest was flagged as *easier* for all students with visual impairments and the CV group at grade 3/4, and *easier* for the NV group at grade 7/8. In these cases as well there were slightly greater instances of lucky guesses marked. However, there was also a greater tendency in the accommodations to reorient the students to the item choices, to allow the students to feel the items, and to bring the items into the student's field of view on the "heaviest" test item in grade 3/4 as compared to the "lightest" item.

### 4.5.3 Other Supporting Patterns

For the remaining test items that were flagged for DIF, there were some reoccurring patterns that have already been discussed. For most of the remaining reading items that were flagged, there was a greater tendency (by at least 3 students) to allow students to feel the items (L1) or to reorient the students to the items (L8) on the non-significant test items used as comparisons. A difference of three students or more was noted since this accommodation also repeatedly surfaced in the tape reviewers' comments. A difference in the frequency of L1 or L8

accommodations for the CV group occurred between the comparison item and "selects similar function-crayon/marker" item. It was also evident for all students with visual impairments to the "What question-hat item." At grade 7/8 reading, this same pattern was noticed in the significant "Category-"things you ride in" compared to "things you eat with." There was a greater tendency to allow students to feel the items on the non-significant test item for those students using object substitution.

In addition to the slight variations in the use of accommodations that would ensure the student experienced all the answer choices, the judgmental item reviewers comments regarding experiential base with test materials may also come into play here. When object substitution is used for "things you ride in," use of miniatures may make this item more difficult than when you replace items for "things you eat with." In the words of one of the judgmental reviewers, "You don't ride in miniature vehicles, you play with them. They are toys. The category might be better represented as things you play with." Finally, in reading, the item "selects picture named-backpack" in 7/8 reading flagged as *easier* for the CV group seems to be due to a greater number of lucky guesses (13 versus 2 for a comparison item).

For some of the remaining items that did not have any skills with which to compare the results, fewer patterns were able to be explored. For example, the significant item of "selects clock" when given the prompt of "which one shows you the time?" (which was *harder* for all students with visual impairments and the V group) was the only of one of its kind in 7/8 math. However, the judgmental item reviewers felt that a wall clock might be outside of the experience of students with decreased distance vision. In addition, this item in its standard form used objects all of similar round appearance (clock, plate, CD in case). Tape reviewer comments seemed to indicate that the test administrator's choice in the type of time piece is the important component

here. Tape reviewers noted a few instances where a talking watch or a tactile version of a clock would be preferred, and they noted some occurrences of a clock being drawn on a paper plate which they felt was confusing especially since one of the distracters was a plate. Another item, "match length-longest," which was significant as harder for the V group, also did not have any comparison items, but tape reviewers noted several instances of more contrast being necessary.

### 4.5.4 Summary of Logical Analysis

In summary, there appear to be a variety of reasons why certain items were flagged as significant for DIF. However, the emerging patterns suggest a few themes to keep in mind:

1. The manner of presentation and the amount of reorientation support provided to students with visual impairments appears to be an important factor.
2. For some items, lucky guesses appear to play a role in the reason for DIF, especially for items that are the answer and also bigger and/or brighter than the other items.
3. There was some indication that scoring discrepancies could affect the outcome of the DIF analysis. These instances should be further investigated for accuracy in the patterns observed.
4. The large number of items flagged as more difficult for the V group could be due to a tendency to under-accommodate for visual impairment.

Finally, what was also discussed in the literature and not discussed in this section is the possibility that some of the items flagged repeatedly, although they seem to be explained by lucky guesses, may also be occurrences where an item is truly easier or harder for students with visual impairments. Reasons could be due to certain concepts that may be more likely taught

(such as descriptive words like “biggest”) or less likely taught (“money” which may be experienced incidentally less often by students with visual impairments) to students with visual impairments. Another alternative is that the spatial layout or the need for multiple comparisons when items cannot be experienced all at once with reduced fields or no vision could make some test items more complex.

# 5.0 DISCUSSION

Students who take the PASA are a heterogeneous group whose varied levels of experiences, cognitive challenges, and multiple disabilities impact the focus and goals of their educational programs and their performance on standardized performance-based alternate assessments for accountability. The distinctive mix of qualities that contribute to who these students are as learners, in combination with the fundamental nature of performance-based assessments provide multiple challenges in research when attempting to pinpoint reasons for differences in performance on the PASA. The challenge is increased with the addition of a sensory impairment. Considering that students with visual impairments without additional disabilities are under-represented as a disaggregated group in assessment research, those students who do have additional disabilities are at risk of being easily forgotten since they are one of the lowest low-incidence groups of an already low-incidence population: students with the most severe cognitive disabilities. However if in addition to accountability requirements, the practical goal of alternate assessment is to promote meaningful instruction in content areas and to promote positive outcomes of consequential validity, then the same question needs to be addressed for students with visual impairments who take alternate assessments as it does for other students: Who are these unique learners and how well does assessment inform about their abilities and needs?

This exploratory study using a mixed methods approach attempted to gain insight into these two questions. It attempted to paint a preliminary picture of who the students with visual

impairments are who took the 2005 Level A PASA at the 3/4 and 7/8 grade levels with the hope of also gaining some insight in general about the abilities and needs of students with visual impairments who take alternate assessments. The specific purpose of this study was to investigate achievement on the PASA for these students. In particular, the study focused on documenting the accommodations that are selected for students with visual impairments, discovering instances of differential item functioning (DIF) at three levels of functional vision, and logically analyzing the contributing variables to identify potential reasons for DIF occurrences. This analysis was conducted using data from 286 students with visual impairments with focus on the following research questions:

Q1. Were there significant differences in the scores of students with visual impairments at different functional vision levels on the 2005 grade 3/4 or 7/8 A level PASA?

Q2. What accommodations did teachers make to adapt the 2005 PASA for students with visual impairments?

a. Are there relationships between the types of accommodations made and level of functional vision or type of test item?

b. Were there accommodations that seemed to change the intent of the skills being tested?

Q3. Were there significant differences on individual 2005 level A PASA math and reading test items at the 3/4 and 7/8 grade levels of students with visual impairments as compared to students without visual impairments who had similar ability profiles on the constructs of interest?

Q4. Considering the accommodations made and student performance on different types of test items, what are the potential reasons that "flagged" test items functioned differently?

The findings related to these questions were presented in the previous chapter and are further elaborated upon in this chapter with a focus on the connections between the results of each research question. In addition, limitations of the study, implications of the research, and suggestions for future research are discussed.

## 5.1 SYNTHESIS OF RESULTS

### 5.1.1 General Achievement

Students with more functional vision (the V group) performed significantly better on the A Level PASA assessment than students with less functional vision (both the CV and NV groups). Consequently, the V group also had a higher rate of classification as proficient or novice. While not a surprising finding, Question One, which compared the general achievement of students with visual impairments on the PASA warrants some discussion about what the finding might mean. Actually, this finding is the fundamental basis for this exploratory study and a good way to begin discussion about the factors affecting assessment for these students.

This finding could lead to misinterpretation when combined with the facts that 73% of the students with visual impairments who took the PASA at grades 3/4 or 7/8 took the easiest test level, and that 70% of those taking harder levels were in the V vision group. Is it really the case that students with more severe visual impairments or blindness do not have the ability to take the

higher level PASA assessments? At face value, it would be easy to assume that students with more severe visual impairments also have a tendency to have more severe disabilities, legitimately affecting their performance on assessments. Or it could be assumed that these students because of visual impairment simply do not achieve as many academic based skills.

For some students, of course, this assumption may be true. For students with multiple impairments, visual impairment is often a secondary disability as a result of extreme pre-maturity and complications during development (Gates, 1985 as cited in Barraga & Erin, 2001). Intuitively, it makes sense that the more severe the multiple disabilities are, the greater the potential for severe visual impairment. However, this may not always be the case, and the assumption should not be made based on a single assessment score. There are other explanations that need to be considered because the assessment itself could be affecting the outcomes. Research conducted on different types of students with visual impairments (young children, college-age students, etc.) suggests that assessments may contain elements that change the intent of the skills being tested under accommodated conditions, particularly when there is a change in sensory mode such as the need to access items tactually (Bennett, Rock, & Kaplan, 1987; Brambring & Troster, 1994; Caton, 1977; Wyver & Markham, 1999). Bradley-Johnson (1994) and Bowen and Ferrell (2003) support the importance of a multi-faceted view of a student's abilities for this very reason. For any assessment, including the PASA which was analyzed in this study, alternative possibilities to the equation that achievement scores = ability levels must at least be considered:

1. How universally designed is the assessment?
2. Do the skills upon which the assessment focuses easily adapt into alternate versions of the same construct?

3. Are there aspects of the higher levels of the performance-based assessment that make access more difficult?
4. Are there contexts that are potentially biased?
5. Are there skills that do not make sense for students with visual impairments at different levels of functional vision?
6. Were test items appropriately accommodated?

Sifting through these possibilities as well as others was the challenge of this study. Overall, the mixed findings of this study suggest that a combination of factors contribute to a score's interpretation- some that appear connected to vision, and some that may be more connected to the heterogeneous nature of students with the most severe cognitive disabilities. The very fact that items were flagged as having DIF- some easier and some harder- indicates that there are aspects of the assessment process that may prohibit interpretation of the "scores" of the students into the "abilities" of the students on some of these flagged skills.

### 5.1.2 Accommodations

#### 5.1.2.1 Frequency of accommodations

The study's description of the frequency and the types of accommodations made on the PASA provided some insight into the investigation of separating out ability from accessibility for students with visual impairments. Teachers focused mostly on layout/set-up accommodations which centered on the use of a slant board and holding an object/picture within the student's field of view for students in the V and CV groups. Layout accommodations mostly consisted of allowing the student to feel the objects for the NV group. Use of a high contrast background was also a frequently used accommodation. Reorientation to the choices occurred but was

surprisingly low in frequency per test item. Interestingly, some of the most frequently used accommodations were also the accommodations that tape reviewers felt were needed on different test items, particularly contrast and deliberate orientation to test items (whether by allowing the student to feel and explore items or by showing the student each item at an optimal distance and location). In addition, tape reviewers frequently indicated a need for deliberate re-orientation to the options; several noted test items in which they believed the student was not aware of the location of the choices, or some choices that appeared to be outside of the student's visual viewing distance. Logical analysis of the DIF results moderately supported these suggestions; on some items there was a mild to strong pattern in the differences of accommodation frequencies for layout between significant and non-significant items of the same skill type.

In summary, it appears at least in part, that access to the test items relied on the quality of the layout. More deliberate and systematic layout with enough time for the student to explore objects or look at pictures at a proper viewing distance could have positively affected scores on some items. In fact, in several instances tape reviewers noted that for students receiving scores of one's and two's, not enough wait time and no chance for exploration was given. Granted, some tape reviewers also supplemented these observations with comments along the lines of, "for this student I do not know if the accommodations would have helped."

#### 5.1.2.2 Under-Accommodated Students

These patterns of comments made by tape reviewers reveal another interesting factor related to the concept of consequential validity, or in other words, "the buy-in factor." There were some cases where tape reviewers became very "fired up" in their comments about the lack of opportunity or accommodations given to a student. One case in particular was of a student who was totally blind whom the test administrator said could not access the pictures on the test;

the administrator than went through the whole test using the pictures. In another instance, the tape reviewer wrote in the comments sections “vision consult needed” because, in the reviewer’s opinion, no appropriate accommodations for vision were provided. Tape reviewers varied in the way their comments were worded for students who were receiving scores of all ones’ or two’s because no selections were being made by the student. Some reviewers felt that accommodations should still be attempted (at a minimum orienting the students to the items either tactually or by presenting the items in the student’s field of view); they highly praised those teachers who made a gallant effort to provide access to the materials. Other tape reviewers, however, questioned the value of putting in so much effort. For students who are not responsive or for whom the skills are too difficult, they wondered why a teacher should bother. These few examples confirm what has always been suspected of administration of the alternate assessment: that the quality of accommodations provided- for vision or otherwise- is partially dependent upon the teacher’s perceptions of the value of doing the assessment. The teacher needs to see the connections of the skills to their student’s goals or programming.

This component of consequential validity also surfaces when considering the accommodations that were rarely used on the PASA. There were very low occurrences among all three functional vision groups of the use of low vision devices and the use of augmentative communication devices. Particularly for augmentative communication devices, one has to question: is it the case that this type of technology, typically a component of programming for students with multiple disabilities, is not being utilized with students with visual impairments? For low vision devices, it might be the case (only light boxes and one use of a Closed Circuit Television being available or used were noted), but it is harder to believe it is the case for augmentative communication devices. When making accommodations for assessments, literature

encourages the use of accommodations that are typically used with the student for everyday instruction (Thurlow et al., 2003). The alternative explanation to the question just posed is that use of some types of typical classroom and instructional accommodations are not being used on the PASA for students with visual impairments, and one would assume for other students as well. It may be that test administrators have trouble figuring out how to best incorporate a typical accommodation into the structure of the assessment, or may be restricted on time to make adaptations. They may also fear deviating from the assessment's general instructions and layout. In fact, there were some instances where test administrators went through all the alternate response options with the student as prompts saying, "look at the…, touch the…, point to the..." when it was clear that the student's physical limitations prohibited touching and pointing, and the student's visual impairment (blindness) prohibited looking. Occasional tape reviewer comments also supported this notion of typical accommodations not being used. One reviewer noted that, "supporting documentation says that the student uses eye gaze, but test administrator did not acknowledge during the test."

#### 5.1.2.3 Object Substitutions

Another accommodation infrequently used was "making pictures or objects tactual." Instead, replacement of objects for pictures was the accommodation of choice and was used frequently at grade 7/8 reading for the NV group and quite often for the CV group. As noted in the results, the use of objects with six V group students in 7/8 reading was surprising. The majority of these students attended the same school for the blind. It is possible that the school created an adapted test item kit that was used for all students, regardless of the level of vision. Other patterns of object use by teacher were evident as well. The pattern suggests that some schools and/or teachers may have gathered materials to use with several students, which is

understandable considering the multitude of items the test required and the limited time many teachers may have for making adaptations. The 2005 assessment year of the PASA was the last year in which no test kit was provided, so teachers had to gather their own materials. Theoretically, this would open up the opportunity to individualize object selections to the student's experiential base and/or to objects that are familiar to the student. Patterns seem to suggest, however, that this was not the norm. Some object substitutions were merely what the test administrator could find and were not substitutions because it was an item with which the student was more familiar. However, this is not to say that individualization never occurred. Comments by the tape reviewers cited some instances where the test administrator made an extra effort to use familiar objects. An example would be the test administrator who used familiar items and red bins turned on their sides to block out conflicting background for a student with Cortical Visual Impairment (CVI). Another test administrator used the student's favorite plate and other items in the student's favorite color.

The fact that object substitutions were used over tactile substitutions for students unable to access the pictures highlights a very important aspect regarding the interpretation and validity of the assessment for these students. Technically, for many items in reading, object substitution at the 7/8 level makes the test identical to the grade 3/4 level. As discussed in the results section, few conclusions could be made about accommodations that changed the intent of a skill because tape reviewers were inconsistent in marking changes- most likely an indication that it is a difficult determination to make. In fact, except for two instances (one for the CV group and one for the NV group) where items using object substitution were flagged as easier, no items where object substitution was used emerged in the DIF analysis.

The fact that tactile enhancement or use of textures or tactile symbols was rare could be interpreted in several ways:

1. Students in the CV and NV groups are not using tactile symbols in their regular programming.
2. As suggested by the judgmental item reviewers, tactile symbol systems are generally very individualized and therefore do not lend themselves well to fitting the context of the PASA.
3. Substitution of objects is an easier/quicker way to accommodate more students needing to take the same assessment by the same teacher or in the same school building.

If number one is possibly the norm for most students, then the question regarding whether functional or pre-literacy instruction is being incorporated into student programming should be posed. It also raises the question about what teachers consider as literacy instruction for students with visual impairments with additional disabilities in the CV and NV groupings. While rarely seen in the tapes, an interesting auditory accommodation was noted in which the test administrator provided a target sound then a loop tape of sound choices. The student was asked to match the same sound by hitting his head switch when it was presented on the loop tape. Intuitively, one would consider use of a tactile substitution as a more directly linked adaptation to matching pictures, but for students with visual impairments and additional disabilities this accommodation example does raise the question of auditory comprehension as literacy- a continuing debate within the field of visual impairments. For some students auditory skills may be a better, more practical, instructional option.

In addition, the judgmental test item reviewers also raised concerns regarding the use of tactile symbols or objects as symbols. Is the progression from objects representing themselves to tactile symbols or objects as symbols (in which case the student may have more trouble separating from its tangible use) more of an abstract leap than moving from objects to pictures? And, is concrete versus abstract as important as whether the student is familiar with the system? These questions posed by the test item reviewers are important considerations to investigate further if a better understanding is to occur of instances when accommodated versions of assessment items change the intent of the skill or construct ultimately being assessed, particularly for alternate assessments. A direct example of this issue was seen in the test item "answers where question" at grade 7/8 reading. This item was the only one that contained complex pictures which resulted in a large drop in the use of the object substitution accommodation. Those that did use the substitution attempted to use objects as symbols (e.g. ball for playground, plastic bag for grocery store, and stuffed cow for farm). The fact that fewer test administrators chose this route for this item has implications at the 11$^{th}$ grade A level assessment where complex pictures are used more frequently and at the B level assessment where pictures and complex pictures become the norm at certain grade levels. This cycles back around to the question posed at the beginning of the discussion section: do so few CV and NV students take the B and C level assessments because cognitively they do not have the skill level to do so, or is it at least in part due to an artifact of the test item structure or content of the PASA? Finally, given the unanswered questions just posed, what really would constitute a comparable skill progression in reading for students without enough vision to progress toward print literacy? Since for various reasons some students may be unable to progress toward functional braille literacy (e.g. due to physical limitations to access materials tactually), both a comparable tactual and auditory skill

progression may need to be considered for assessment. However, for students who do use tactile symbol systems, the individualization of these systems makes standardized assessment of tactual literacy skills challenging.

### 5.1.3 DIF Analysis

The DIF analysis revealed 29 instances of items being potentially harder for at least one vision group (V, CV, NV, or all students with visual impairments) and 12 instances where items may have been potentially easier as compared to matched peers. Several aspects of this outcome of the DIF analysis have already been discussed, but a closer look at the V group patterns in particular (where more items were flagged as harder than would be expected) and some additional discussion about judgmental item review is warranted given the results of this study.

#### 5.1.3.1 V Group Patterns

There were a surprising number of items flagged for DIF for this group of students given that both the judgmental test item reviewers and the researcher assumed that the test would be fairly accessible "as is" to a population of students classified as primarily using vision for most tasks. After all, in terms of universal design, the PASA has fairly large and bold pictures for most items. However, logical analyses of the DIF items for the V group were often inconclusive (could not be explained by lucky guess, score change suggestions, or administration error patterns) or showed a weak pattern of fewer layout or picture/object enhancement accommodations between significant and non-significant items. It is possible that some students in this group were under-accommodated on some test items. Test administrators could have assumed, as the judgmental item reviewers did, that the materials provided appeared to be

appropriately sized. Another possibility, however, is that vision could be affecting the experience level students have with the chosen materials on certain test items. As mentioned in the literature review, students with low vision may be given more credit for being able to interpret pictures visually (Groenveld & Jan, 1992) and for learning incidentally (Milian, 1996) than they really are able to do. For example, several test items of "selected related items" were flagged for the V group. One would assume that multiple occurrences of DIF on the same item type are an indication of something more significant than just a poor choice in test materials. It would suggest that either the item type was problematic because of its layout, or that the construct itself really is harder for this group of students with low vision. Further investigation would be needed to be certain, but the former explanation should be ruled out first given that there were instances where more reorientation to the choices on these items was sometimes indicated.

#### 5.1.3.2 Miscellaneous DIF Items

In addition to the DIF patterns already discussed, logical analysis of the DIF items as a whole revealed some competing explanations for the same type of test item. This is not surprising due to the wide range of variables that can affect a score on the PASA. However, some general patterns did emerge. For items that were easier for the CV group, there seemed to be more incidences of lucky guesses and fewer administrator errors noted. Another possibility of the use of object substitution was already noted. While math items of "find money" and "find biggest vs. smallest" might be explained away by the size of the answer target which potentially made it more difficult for students with visual impairments to notice or find the item within the array, the fact that they were flagged more than once could also indicate a different level of understanding of these concepts. In fact, when a dollar bill was used instead of coins at grade 7/8, the item was still flagged as more difficult. On the PASA, "money" seems to be one of those

items that many students know right away, so it was a surprise when all instances of money were flagged. While there were a couple instances of fake money (especially the dollar bill) being used that may have made the item harder, the majority of the students were presented with real dollar bills and real coins. It may be that students experience money by watching other people buy things. It could be an "incidental learning opportunity" missed out by students whose vision inhibits passive observation of events from afar. Whether this is the case is speculation, but it does merit a closer look. Perhaps money is a functional skill that needs to be more directly taught to students with visual impairments and additional disabilities.

The items involving the concept of "biggest" are also interesting to note. Is it possible that superlative comparisons are used more frequently with students with visual impairments and additional disabilities than with students who do not have additional disabilities? This possibility would coincide with Milian's (1996) theory that the higher achievement of students who are blind on some comparative concepts could be due to the greater attention given to direct instruction on these words in order to provide explicit language for understanding the environment without vision or very limited vision.

#### 5.1.3.3 Judgmental Item Review

The judgmental item review in this study proved to be a very interesting process which posed additional questions to consider when thinking about adaptations to the PASA and interpretation of assessment results. Helwig and Tindal (2003) have noted that a teacher's ability to decide whether an accommodation would benefit a student was no greater than chance. In a similar view, the judgmental review seemed to be no better than chance as compared to the results of the DIF analysis. However, that does not mean that judgmental item reviews do not have a place in the assessment process. Many potential reasons for DIF emerged during the

review process. In some respects the judgmental item review can be more informative or more accurate than the statistical DIF analysis particularly when small sample sizes are involved. In addition, it should be noted that item reviews are more challenging for performance-based alternate assessments like the PASA that are more flexible in the range of accommodations that are allowed. The item reviewers considered many items from the tactile symbol perspective which turned out to be used very little. For assessments with such variation in accommodations, a logical analysis after the assessment (if accommodation information is collected) might be a better strategy to use to inform accommodation choices and test item content for the following year's assessment.

## 5.2 LIMITATIONS OF THE STUDY

The results of this study and subsequent discussion of emergent patterns and factors contributing to the interpretation of PASA results for students with visual impairments should be considered with some limitations in mind. While this study had a larger sample size than is often typical of studies on students with visual impairments, sample sizes were still relatively small, especially when the sample was divided into three levels of functional vision. Sample size could certainly have affected the DIF analysis. Smaller sample sizes could increase the impact of an outlier on the inferential test. Conversely, fewer items might be flagged statistically because of lack of power. For the NV group in particular, power to detect significance was low. However, the NV group was also the lowest achieving group and was therefore matched with lower achieving students, reducing the variability in scores, making it harder to detect differences in general. This study attempted to balance the statistical tests with supporting information from descriptive and

qualitative approaches in order to paint a more balanced picture and derive richer information. Nevertheless, as is typical for studies of students with visual impairments, caution must be given to the interpretation of statistical outcomes.

Another potential limitation deals with the matching process used for the DIF analysis. Great care was taken to match on ability as closely as possible. Skills checklists were used as a secondary variable to help protect against possible cases where pervasive DIF would make matching on ability by total score alone inappropriate. In other words, if a student's access to the assessment prohibited answering skills correctly, that student may have low scores across the board, but scores might actually be higher if the test had been more accessible. By matching on the skills checklist, a closer match was hopefully found for actual ability, not cases where pervasive DIF interfered. It needs to be noted, however, that the skills checklists used (existing data) contain visual language in some of their descriptions. Test administrators or teachers filling out the checklists for a student with little functional vision may have placed the student low on those skills because they did not re-interpret the skill into use of a different sensory mode. Furthermore, the skills checklist was directly tied to the test, which means it was not a separate independent measure of a student's skill base. These limitations in combination with the heterogeneity of the population being studied could have introduced confounding variables to the DIF statistical findings.

Finally, several limitations in the logical analysis need to be noted. First, it was conducted using available videotape data from the assessments of the students with visual impairments. This same videotape data was not available for the matched group of students. Had it been available, a more in-depth comparison and analysis of the reasons for DIF specifically relating to visual impairment could have been conducted and perhaps stronger patterns and

conclusions might have emerged. Second, since the PASA necessarily limits the number of tests items that are assessed each year, there were several flagged test items that appeared only once within the test. This was especially true at grade 7/8 math. It is possible that some of the single items might have been flagged a second or third time, resulting in a better understanding about what currently seem to be “fluke” items.

## 5.3 IMPLICATIONS FOR PRACTICE

The results, discussion, and additional information from tape reviewer and judgmental test item reviewer comments highlight some important considerations and recommendations regarding assessment of students with visual impairments on the PASA and on alternate performance-based assessments in general.

### 5.3.1 Utility of the Standards

There are indications from the results of this study that teachers who have students with visual impairments may find it difficult to see the benefit of the alternate standards being assessed on the PASA. This may be principally true for students who cannot access pictures efficiently. For positive consequential outcomes to emerge for these students, teachers need to have a sense that the assessment is meaningful. While it is probably the case that a group of teachers who have students without visual impairments may feel similarly skeptical about the value of the alternate standards on which the PASA is based, the standards appear to apply more directly to their students. Whether teachers are fond of the standards or not, they provide guidance for the content

material on which students with the most significant cognitive disabilities are expected to advance. Based on the discussion regarding the PASA reading assessment and what constitutes a skill progression in literacy for students with visual impairments and additional disabilities, there are skills within the standards that may need to be further reinterpreted for students at the CV and NV functional vision levels in particular.

### 5.3.2 Practical Implications

Based on the study results, a set of recommendations can be made when considering how to accommodate students with visual impairments on the PASA- at the test administrator level and/or at the test development level. In many respects, these recommendations may be useful when considering accommodations for other assessments as well.

1. Consider the experiential base of the items being used. Is there a potential "incidental learning" or "visual bias" variable attached to contexts (e.g., wall clocks that may never be noticed by students with visual impairments).
2. Be deliberate and systematic when orienting students to the test item choices.
3. Allow enough time for students to explore and experience the objects or pictures before asking the question.
4. Consider the order of wording and its impact on memory load when pictures in the assessment meant to support memory are not a usable tool.
5. Reconsider the use of transparent items in the assessment (e.g. sandwich baggies, clear cups, clear straw, etc.)
6. Use accommodations that are familiar to the student- those that are used daily in regular instruction.

7. Even if skills are beyond the student's ability level, accommodate for the student's needs and consider ways in which the test can also be an opportunity for students to practice other instructional goals (e.g. switch use to communicate).
8. When adapting items requiring tactual size comparisons that are two-dimensional, use thicker material to make comparisons easier.
9. Evaluate the testing environment for lighting, glare and contrast to eliminate some unexpected assessment glitches. For example, tape reviewers noted several instances where a student with CVI who engaged in light gazing was tested facing a window or placed with an angle of view in the wheelchair up towards the ceiling lights.

## 5.4 FUTURE RESEARCH

Throughout this discussion, several questions emerged that could not be directly answered by the current study. The area of assessment of students with visual impairments and specifically the area of alternate assessment is in need of more frequent research in general for these students. For the PASA, as for other alternate assessments, a longitudinal study that evaluates for DIF would be important. Since skills are infrequently repeated within one assessment, patterns where a particular type of test item contains a fundamental component that functions differently will only surface when analyzing over several years worth of data. However, to do this, secondary disability status would need to be recorded yearly in order to identify students with visual

impairments beyond the students who attend schools for the blind (the students most likely to carry a primary disability status of visual impairment).

In addition, it is necessary to conduct an analysis of the B and C level versions of the PASA for students with visual impairments as well as an analysis at the other grade levels. A mixed methods model such as the one used in this study would be appropriate. However, statistical analyses would be more limited as sample sizes are extremely small at these levels. The B level in particular contains many item types that pose challenges in adapting the item into another sensory mode. To begin this study, it would be interesting to just record the types of adaptations that test administrators attempted. Further investigation would then be necessary to begin to tease out the types of accommodations that change the test item construct. Since this is often accomplished in other accommodation studies through the use of a "think aloud" procedure where the student taking the assessment verbalizes his or her thinking, coming to conclusions about changes in skill intent for this population of students will require some creativity in research design.

Finally, a third study that is needed would be to begin to gather information about the functional literacy programs in which students with visual impairments and multiple disabilities are actually engaged. Are there commonalities? If so, do those commonalities lend themselves well to alternate state standards in the content area of reading? Trial items incorporating a literacy progression could be tested with students at different functional vision levels.

## 5.5 CONCLUDING THOUGHTS

PASA is meant to be a snapshot of a student's skill base on math and reading related content. For students with visual impairments whose test scores have always traditionally been interpreted with caution, this is a wise concept for educators and state administrators to keep in mind. In the process of painting a picture of who students with visual impairments are who take the PASA and how their performance on the alternate assessment relates to their ability, this study revealed fundamental questions that deserve further investigation in order to create the best possible assessment situation that will yield meaningful results and consequentially result in meaningful instruction. The painting is still missing many of its details before the landscape is complete. One of the biggest challenges for this population of students with the most severe cognitive disabilities is to better understand when and how accommodations and combinations of accommodations change the intent of the skills being measured. For any large-scale assessment, the unique characteristics of these students make the challenge even greater. Additionally important for students with visual impairments taking alternate assessments is to become clearer on when the skills being measured are not appropriate for different levels of functional vision. Judgmental reviews and logical analyses are our best tools so far, but efforts to test what intuitively makes sense need to be made. After all, in the words of one of the judgmental test item reviewers: "It's not always a matter of being easier or harder, but different. Skills for students with visual impairments are sometimes just different."

# APPENDIX A

## 2005 PASA TASK AND SKILLS LIST

This appendix contains tables that list the skills that were tested at grade 3/4 and 7/8 on the 2005 PASA at all testing levels.

## 2005 TASKS AND SKILLS ASSESSED BY PASA READING TASKS

| Grade 3 | | |
|---|---|---|
| Level of Complexity of Task | | |
| A | B | C |
| • Orients toward set of objects<br>• Matches identical objects – objects are from different categories in appearance<br>• Selects object named – distracters are objects from different categories in appearance<br>• Selects similar objects – distracters are objects from different conceptual categories<br>• Selects related objects – distracters are objects from different conceptual categories<br>• Answers literal 'what' question by selecting object – distracters are objects from different conceptual categories<br>• Answers literal 'who' question by selecting object – distracters are objects from different conceptual categories | • Selects picture named – distracters are pictures from same category in appearance<br>• Selects picture with beginning sound named<br>• Locates picture named in 4-item display – distracters are pictures from same category in appearance<br>• Selects similar pictures – distracters are pictures from same conceptual category<br>• Selects related pictures – distracters are pictures from same conceptual category<br>• Identifies category of picture – distracters are pictures from different conceptual categories<br>• Selects picture by function – distracters are pictures from same conceptual category<br>• Selects picture by feature<br>• Demonstrates understanding of 2-10 word oral command<br>• Answers literal 'who' question by selecting picture - distracters are pictures from same conceptual category<br>• Answers literal 'what' question by selecting picture - distracters are pictures from same conceptual category | • Selects 1 word with beginning sound named in array of 5<br>• Selects 1 word named in array of 5 with all choices having same beginning letter<br>• Reads 1 word in isolation<br>• Reads 1 word in context<br>• Locates 1-2 words named in 5-6 item display with text or in real materials<br>• Selects word within same conceptual category<br>• Selects 1 word by function in array of 5 words<br>• Demonstrates understanding of 2-10 word 2-step oral command<br>• Selects picture representing 1 word read silently from array of 5 pictures<br>• Orders 3 pictures based on text<br>• Selects picture to identify main event from narrative text<br>• Predicts topic of story from picture by selecting from array of 5 words |

## 2005 TASKS AND SKILLS ASSESSED BY PASA READING TASKS

| Grade 8 | | |
|---|---|---|
| Level of Complexity of Task | | |
| A | B | C |
| • Scans set of materials<br>• Matches identical pictures – pictures are from different categories in appearance<br>• Selects picture named – distracters are pictures from different categories in appearance<br>• Locates identical picture in 3-item display – distracters are pictures from different categories in appearance<br>• Locates picture named in 3-item display – distracters are pictures from different categories in appearance<br>• Selects similar pictures – distracters are pictures from different conceptual categories<br>• Selects related pictures – distracters are pictures from different conceptual categories<br>• Identifies category of picture – distracters are pictures from different conceptual categories<br>• Selects picture by function – distracters are pictures from different conceptual categories<br>• Demonstrates function of item in picture<br>• Answers literal 'who' question by selecting picture – distracters are pictures from different conceptual categories<br>• Answers literal 'what' question by selecting picture – distracters are pictures from different conceptual categories<br>• Answers literal 'where' question by selecting picture – distracters are pictures from different conceptual categories | • Selects rhyming words<br>• Selects word with same beginning sound as target picture named<br>• Selects picture showing 2 features named<br>• Answers literal ‘who’ question – open-ended<br>• Answers literal ‘what’ question – open-ended<br>• Answers literal ‘where’ question – open-ended<br>• Answers literal ‘when’ question – open-ended<br>• Selects last word missing in sentence using clue from picture<br>• Names 2 details in picture | • Reads 11-29 words<br>• Demonstrates understanding of 11-29 word written command<br>• Answers literal ‘who’ question<br>• Answers inferential ‘who’ question<br>• Answers literal ‘what’ question<br>• Answers inferential ‘what’ question<br>• Answers literal ‘where’ question<br>• Answers inferential ‘where’ question<br>• Answers literal ‘when’ question<br>• Answers inferential ‘when’ question<br>• Answers literal ‘why’ question<br>• Answers literal ‘how’ question<br>• Orders three 1-5 word phrases representing main ideas from story<br>• Describes 3 events from narrative text<br>• Names 3 facts from expository text |

## 2005 TASKS AND SKILLS ASSESSED BY PASA MATHEMATICS TASKS

| Grade 3 | | |
|---|---|---|
| Level of Complexity of Task | | |
| A | B | C |
| • Orients toward materials<br>• Selects set with a lot/a few – smallest difference is 4x<br>• Matches 2 sets of items with 1 item each – smallest difference is 4x<br>• Selects coins – all items are dissimilar in appearance<br>• Selects set with most/least using items arranged in a pattern – smallest difference is 4x<br>• Matches items of same length – smallest difference is 4x<br>• Matches items of same size – smallest difference is 4x<br>• Selects biggest smallest item – smallest difference is 4x<br>• Identifies biggest/smallest item based on volume – smallest difference is 4x<br>• Matches items with same volume – smallest difference is 4x<br>• Identifies heaviest or lightest item – size and weight vary directly | • Counts items up to 5 in ordered array with the teacher pointing to each item<br>• Counts one-dollar bills or pennies up to 5 with the teacher pointing to each item<br>• Selects quantity named (highest or lowest) and shown from 1-5 from array of 4 fixed, ordered sets arranged in a pattern – all dissimilar sets<br>• Selects number named up to 5 from array of 4<br>• Reads number from 1-5<br>• Reads whole number price up to $5 or 5 cents<br>• Selects largest or smallest value from graph without numbers – ordered display<br>• Locates number named up to 5 in 4-item display – one variable only<br>• Selects longest/shortest item 1-5 inches in length from array of 4<br>• Measures item by counting units from 1-5 with teacher pointing to each item (area)<br>• Selects heaviest/lightest item from array of 4 – weight and size vary directly | • Adds 2 prices with sums < $9 or < 9¢ by counting sets of one-dollar bills or pennies<br>• Subtracts 2 prices < $9 or < 9¢ by counting dollars or pennies and using subtraction to take away<br>• Counts items in ordered array up to 9<br>• Counts out items up to 9 from larger set<br>• Counts one-dollar bills or pennies up to 9<br>• Counts out one-dollar bills or pennies up to 9 from larger set<br>• Selects quantity named and shown from array of 5 ordered sets arranged in a pattern with 1-9 items - 2 similar sets<br>• Selects one- or five-dollar bill from array of 5 - all choices look different<br>• Selects largest or smallest value from 1-9 in array of 5 ordered numbers<br>• Selects largest or smallest value from graph of ordered numbers from 1-9<br>• Measures item using fixed ruler<br>• Identifies shortest/longest straight line path starting from different locations from an array of 5<br>• Sorts 8 items into 4 groups - all groups are distinct but resemble each other - no model |

## 2005 TASKS AND SKILLS ASSESSED BY PASA MATHEMATICS TASKS

| Grade 8 | | |
|---|---|---|
| Level of Complexity of Task | | |
| A | B | C |
| • Scans materials<br>• Selects set with 1 - smallest difference is 2x<br>• Matches identical numbers from 1-2 – distracters are pictures<br>• Matches 2 sets of items with 1-2 items each – difference is 2x<br>• Selects dollar bills - one distracter is similar and one dissimilar in appearance<br>• Selects set with most/least using items arranged in a pattern - smallest difference is 2x<br>• Matches items of same length – smallest difference is 2x<br>• Identifies longest/shortest item - smallest difference is 2x<br>• Matches items of same size - smallest difference is 2x<br>• Selects biggest/smallest item - smallest difference is 2x<br>• Identifies biggest/smallest item based on volume – smallest difference is 2x<br>• Matches items with same volume - smallest difference is 2x<br>• Selects full/empty item<br>• Matches item to space - distracters are dissimilar<br>• Selects clock - both distracters | • Adds 2 prices with sums < $19 or < 19¢ by counting sets of one-dollar bills or pennies<br>• Subtracts 2 prices < $19 or < 19¢ by counting dollars or pennies and using subtraction to take away<br>• Counts items in ordered array up to 19 starting at 2 or more with bridge<br>• Counts out items to 19 from a larger set<br>• Counts one-dollar bills or pennies up to 19 starting at 2 or more with bridge<br>• Counts out one-dollar bills or pennies up to 19 from a larger set<br>• Selects quantity named and shown from array of 4 ordered sets arranged in pattern and containing 1-19 items - 3 similar sets<br>• Selects largest or smallest value from 1-19 in array of 4 unordered numbers with varying 10s place<br>• Selects longest/shortest flexible item 1-19 inches in length from array of 4<br>• Measures item by counting units from 1-19 - one dimension is constant (area)<br>• Measures item by counting 1-19 units (area) | • Adds 3 numbers with sums < 99 using a calculator, action, and a word problem<br>• Adds 2 prices with sums < $99 or < 99¢ using a calculator, action, and a word problem<br>• Subtracts 2 numbers < 99 using a calculator, a word problem, and using subtraction to take away<br>• Subtracts 2 prices < $99 or < 99¢ using a calculator, a word problem, and using subtraction to take away<br>• Counts fixed items in unordered array to 99<br>• Counts items using combination of 1s and 5s or 1s and 10s to 99<br>• Counts using a combination of one- and five- or one- and ten-dollar bills to $99<br>• Counts out money from a combination of one- and five-dollar bills, or one- and ten-dollar bills from a larger set up to $99<br>• Counts quarters, dimes, or nickels to $1<br>• Selects closest amount from array of 5 sets with 10-99 using model and 3 similar sets<br>• Identifies item that can be purchased given money available - from $1-$99<br>• Selects division of item from array of 5 continuous figures - 2 choices are unequally spaced<br>• Selects result of sum of 2 moveable figures from array of 5 continuous figures - all choices < 1<br>• Determines least or most likely outcome, given characteristics of population of items up to 99 - extreme proportions<br>• Measures item to within 1/2 inch |

| | | |
|---|---|---|
| are similar<br>• Matches digital time 1:00 or 2:00 from array of 3 - distracters are pictures<br>• Identifies heavy or light item - items differ in appearance and size and weight vary directly | • Identifies measured amount (1/2 cup, 1/4 cup) from array of cups and spoons<br>• Matches digital time to half hour from array of 4<br>• Selects unit of time associated with activity from array of 4 numbers with 3 time labels<br>• Selects activity associated with time from array of 4 | • Measures area using non-standard unit with enough items to measure the area<br>• Measures volume using non-standard unit with enough items to measure the volume (<10)<br>• Matches analog time to half hour with digital time in array of 5<br>• Identifies heaviest/lightest set of 2-3 items from an array of 5<br>• Sorts 9 items into 4 groups – 2 pairs vary on 1 dimension - no model |

# APPENDIX B

## SKILLS CHECKLISTS EXAMPLES

This appendix contains one math skills checklist for grade 3/4 and one reading skills checklist for grade 7/8. Skill checklists were completed by teachers on a student and sent in during enrollment. The checklist data were used as a secondary matching variable for the DIF analysis.

**Grade 3/4**

**Mathematics Skills Checklist**

**Please rate how often your student independently completes the activities described below by putting a check mark in one of the three rating boxes for each item. Please mark only <u>one</u> rating for each item. Specifically, you should mark 'Always or Almost Always' demonstrates the skill independently if the student always or almost always completes the activity correctly after receiving initial instructions from you. You should mark 'Sometimes' if your student completes the task independently on some occasions, or if he/she requires some support to complete the activity. The 'Never or Almost Never' rating should be used if the student is unable to demonstrate the skill at all or if he/she requires a lot of help to complete the skill successfully. An example of each skill appears in italics immediately below the item. Please review the example before you rate your student to ensure that your rating is based on the skill intended. If your student has never attempted a particular skill, you can try the example before assigning the rating.**

| My student can perform the following skills independently…… | **Always/ Almost Always** | **Sometimes** | **Never/Almost Never** |
|---|---|---|---|
| **1) Orients toward the materials**<br>*The student is given 3 objects (dollar bill, baseball, spoon). He/she looks at or touches the objects.* | | | |
| **2) Counts one-dollar bills or pennies up to 5**<br>*The student is given one-dollar bills (4). He/she counts the money aloud as the teacher points (1, 2, 3, 4).* | | | |
| **3) Counts one-dollar bills or pennies up to 9**<br>*The student is given pennies (7). He/she counts the money aloud (1, 2, 3 …7).* | | | |
| **4) Selects a set with a lot or a few**<br>*The student is shown 3 sets of items (2 spoons, 15 spoons, 10 spoons). He/she selects the set with a few (2).* | | | |

| | | | |
|---|---|---|---|
| **5) Selects a quantity named up to 9**<br>*The student is given 5 pictures of sets of items (8 pencils, 10 pencils, 19 pencils, 14 pencils, 17 pencils). He/she is then shown a number card (8). The student selects the amount on the card (8 pencils).* | | | |
| **6) Matches 2 sets with 1 item**<br>*The student is given 3 sets of items (1 penny, 13 pennies, 14 pennies). He/she is then given a target set with 1 penny. He/she matches the sets with the same number of items (1).* | | | |
| **7) Selects a number**<br>*The student is given 4 numbers (1, 5, 3, 4). He/she selects the number named (5).* | | | |
| **8) Orders 4 numbers**<br>*The student is given 4 numbers (1, 5, 2, 4). He/she puts the numbers in order (1, 2, 4, 5).* | | | |
| **9) Selects largest (or smallest) value from a graph without numbers**<br>*The student is given a bar chart with 4 bars. He/she selects the highest bar.* | | | |
| **10) Selects largest (or smallest) value from a graph with numbers**<br>*The student is given a bar chart with 6 bars. He/she selects the largest number.* | | | |
| **11) Matches items of the same length**<br>*The student is given 3 straws (13 inches, 16 inches, 3 inches). He/she is given another straw (3 inches). He/she matches the sets of items that are the same length (3 inches).* | | | |
| **12) Measures an item using a ruler and counting units up to 5**<br>*The student is given a ruler with blocks rather than numbers. The ruler is affixed to an item (VHS tape). He/she counts the blocks as the teacher points (4).* | | | |
| **13) Measures length of item using fixed ruler**<br>*The student is given a ruler with 1-inch markings. The ruler is affixed to the long side of an item (3 x 5 index card). He/she measures the length of the long side (5 inches).* | | | |

| | | | |
|---|---|---|---|
| **14) Selects heaviest (or lightest) item – one item is very heavy**<br>*The student is given 3 items (5 lb. bag of sugar, folded plastic bag, folded washcloth). He/she selects the one which is the heaviest (5 lb. bag of sugar).* | | | |
| **15) Selects heaviest (or lightest) item – size and weight are related**<br>*The student is given 4 items (small notepad, granola bar, 5 lb. bag of sugar, washcloth). He/she feels the weight of each item and selects the one which is the heaviest (5 lb. bag of sugar).* | | | |
| **16) Selects heaviest (or lightest) item – container size is the same**<br>*The student is given 5 items in sandwich bags (cotton balls, paper towels, marshmallows, sugar, cereal). He/she feels the weight of each item and selects the one which is the heaviest (sugar).* | | | |
| **17) Sorts 1 object into 1 of 3 groups**<br>*The student is given 3 groups of items that look very different (2 mittens, 2 forks, 2 toothbrushes) in separate bins. He/she is given a target item (3rd fork). The student puts the target item in the appropriate group (forks).* | | | |
| **18) Sorts 4 items into 4 groups**<br>*The student is given 4 groups of items (2 socks, 2 washcloths, 2 large towels, 2 undershirts) in separate bins. He/she is then given 4 target items (3rd undershirt, 3rd sock, 3rd washcloth, 3rd large towel). He/she puts the target items in the appropriate groups.* | | | |
| **19) Sorts 8 items into 4 groups**<br>*The student is given an unsorted set of 8 items (2 socks, 2 washcloths, 2 large towels, 2 undershirts) and 4 boxes. He/she puts the items that are the same into the boxes.* | | | |

**Grade 7/8**

**Reading Skills Checklist**

**Please rate how often your student independently completes the activities described below by putting a check mark in one of the three rating boxes for each item. Please mark only <u>one</u> rating for each item. Specifically, you should mark 'Always or Almost Always' demonstrates the skill independently if the student always or almost always completes the activity correctly after receiving initial instructions from you. You should mark 'Sometimes' if your student completes the task independently on some occasions, or if he/she requires some support to complete the activity. The 'Never or Almost Never' rating should be used if the student is unable to demonstrate the skill at all or if he/she requires a lot of help to complete the skill successfully. An example of each skill appears in italics immediately below the item. Please review the example before you rate your student to ensure that your rating is based on the skill intended. If your student has never attempted a particular skill, you can try the example before assigning the rating.**

| My student can perform the following skills independently……. | **Always/ Almost Always** | **Sometimes** | **Never/ Almost Never** |
|---|---|---|---|
| **1) Scans a set of materials**<br>*The student is given 3 pictures (toothbrush, light bulb, chair). He/she looks at or touches each of the pictures.* | | | |
| **2) Matches identical pictures**<br>*The student is given 3 pictures (door, dog, man). He/she is then given a matching picture (door). The student matches the items that are the same (door).* | | | |
| **3) Matches an identical word – words look similar**<br>*The student is given 4 word with the same beginning and ending letters (school, shell, sell, steal). He/she is then given a matching word (2nd word 'school'). He/she matches the words that are the same (school).* | | | |

| | | | |
|---|---|---|---|
| **4) Selects the picture named**<br>*The student is given 3 pictures (pencil, key, ball). He/she finds the picture named (pencil).* | | | |
| **5) Selects the word named – words look different**<br>*The student is given 4 words with the different beginning and ending letters (tree, hat, dog, sink). He/she finds the word named (dog).* | | | |
| **6) Reads a complex picture with 1 word**<br>*The student is given a picture (park). He/she is the given a word (girl). The student uses the picture to help decode the word (girl).* | | | |
| **7) Reads 11-29 words**<br>*The student reads 11-29 words (The pool opens in May. The hours change when school is out. It will be open all day until dark).* | | | |
| **8) Identifies category of a picture**<br>*The student is given 3 pictures of items (animals, clothes, school supplies). He/she is given 1 target picture (snake). He/she puts the item in the appropriate group (animals).* | | | |
| **9) Identifies the category of a word**<br>*This student is given 4 pictures (clothing store, garden, park, school). He/she is then given 1 word (hat). He/she puts the word with the appropriate group (clothing store).* | | | |
| **10) Demonstrates function of an item in a picture**<br>*The student is given a picture (sandals). He/she demonstrates what is done with the item (points to feet).* | | | |
| **11) Completes a cloze passage**<br>*The student is given a card with text (The boy wanted to know how long it would be until his mother picked him up. He looked at the _____.). (clock)* | | | |

| | | | |
|---|---|---|---|
| **12) Gives an example to complete a definition**<br>*The student is given a word (shoe). He/she completes a definition (a shoe protects your foot).* | | | |
| **13) Defines the function of a word read silently**<br>*The student is reads a word silently (watch). He/she tells what you do with the item (tells time).* | | | |
| **14) Answer a literal what question by selecting a picture**<br>*The student hears a sentence. (The girl bought jeans.). He/she is then given 3 pictures of items with very different uses (airplane, jeans, tree). The student selects the item that was bought (jeans).* | | | |
| **15) Answers a literal what question**<br>*The student hears a sentence (The girl bought jeans at the mall.). He/she names what the girl bought (jeans).* | | | |
| **16) Answers a literal what question**<br>*The student reads a paragraph (The girl likes to go clothes shopping on the weekend. Last Saturday she went to the mall with her friends. She bought a pair of jeans.). He/she names what she bought (jeans).* | | | |
| **17) Answers an inferential when question**<br>*The student reads a paragraph (The boy went to the restaurant. He got an apple, a juice box and some soup. Then he sat with his friends.) He/she tells what time of day the boy went to the restaurant (noon, lunch time).* | | | |
| **18) Selects a picture to identify main event from narrative text**<br>*The student hears 4 sentences (School is fun. I talk to my friends. I do crafts. I read books.). He/she is given 4 pictures (playground, school, store, park). He/she finds what the story was about (school).* | | | |

| | | | |
|---|---|---|---|
| **19) Describes 3 events from narrative text**<br>*The student reads a paragraph (My sister likes to go shopping on the weekend. Last Saturday she went to the mall. She looked at three pairs of shoes before she picked out a pair of sandals.) He/she then names 3 things that happened in the story (went to mall, looked at shoes, picked out sandals).* | | | |

# APPENDIX C

## TEST ITEM EXAMPLES

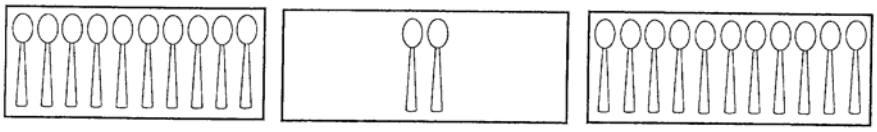
This appendix contains two grade 3/4 and two grade 7/8 example test items from the 2005 PASA A Level Assessment.

**Point to each of the materials as you introduce them, if applicable.*

# 18

**Materials:**

- objects: 23 spoons, 3 pieces of paper



**Test Administrator's Actions*:**

✓ Present 3 pieces of paper. Place sets with different numbers of spoons on each paper (10, 2, 11). *Do not name the amounts.*

| **Say:** | **Skill Assessed** |
|---|---|
| **Which group has a few?** | Selects set with a lot/a few - smallest difference is 4x<br>***Response: points to set with 2*** |
| **Or, instead, say:** | **Alternate Responses** |
| • Find the smallest group.<br>• Point to the smallest group.<br>• Show me the smallest group. | • Says "two"<br>• Picks up set with 2<br>• Touches set with 2 |

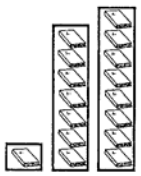
***Remove all materials for the next step.***

** Point to each of the materials as you introduce them, if applicable.*

# 23

**Materials:**

- pictures: 3 sets of books (1 book, 7 books, 8 books)



**Test Administrator's Actions*:**

✓ Present 3 picture cards with sets of different numbers of books (1, 7, 8). *Do not name the amounts.*

| Say: | Skill Assessed |
|---|---|
| **Which group has the least?** | Selects set with most/least using items arranged in a pattern - smallest difference is 2x<br>***Response: points to set with 1*** |
| **Or, instead, say:** | **Alternate Responses** |
| • Find the smallest group.<br>• Point to the smallest group.<br>• Show me the smallest group.<br>• Find the smallest amount. | • Says "one"<br>• Picks up set with 1<br>• Touches set with 1 |

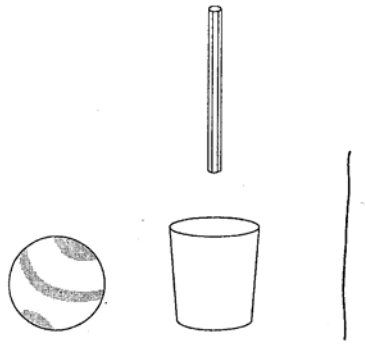
***Remove all materials for the next step.***

**Point to each of the materials as you introduce them, if applicable.*

# 15

**Materials:**

- objects: ball, cup, string, straw



**Test Administrator's Actions*:**

- ✓ Present 3 objects (ball, cup, string).
- ✓ **Say: This is a ball. This is a cup. This is a piece of string.**
- ✓ Present a straw.
- ✓ **Say: This is a straw.**

| **Say:** | **Skill Assessed** |
|---|---|
| **Which one goes with the straw?** | Selects related objects - distracters are objects from different conceptual categories<br>***Response: points to cup*** |
| **Or, <u>instead</u>, say:** | **Alternate Responses** |
| • What goes with the straw?<br>• Find something that would go with a straw.<br>• Show me something that would go with a straw. | • Says "cup"<br>• Puts straw in cup<br>• Picks up cup<br>• Touches cup |

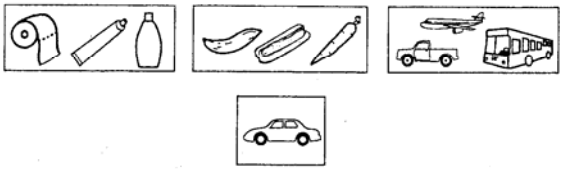
***Remove all materials for the next step.***

**Point to each of the materials as you introduce them, if applicable.*

# 14

**Materials:**

- pictures: roll of toilet paper, toothpaste, bottle of shampoo; banana, hot dog, carrot; truck, plane, bus
- picture: car



**Test Administrator's Actions:**

- ✓ Present 3 picture cards (roll of toilet paper, toothpaste, bottle of shampoo; banana, hot dog, carrot; truck, plane, bus).
- ✓ **Say: These are things you use in the bathroom. These are things you eat. These are things you ride in.**
- ✓ Present another picture card (car).
- ✓ **Say: This is a car.**

| **Say:** | **Skill Assessed** |
|---|---|
| **Which group does the car go in?** | Identifies category of picture - distracters are pictures from different conceptual categories<br>***Response: points to vehicles*** |
| **Or, instead, say:** | **Alternate Responses** |
| • Where does the car go?<br>• Where does the car belong?<br>• What group does the car go with? | • Says "vehicles" or "truck, plane, bus"<br>• Places car on top of vehicles<br>• Picks up vehicles<br>• Touches vehicles |

***Remove all materials for the next step.***

# APPENDIX D

## ACCOMMODATIONS AND REASON FOR SCORE CODES

This appendix contains the coding sheets that tape reviewers used when extracting data on reasons for score, accommodations and changes in skill intent.

## Accommodations/Adaptations Code Sheet

| Substitutions | Picture/Object Enhancement | Layout/Set-Up | Instructions/Directions | Response |
|---|---|---|---|---|
| **S1** Used different objects than the objects indicated- **please specify**<br><br>**S2** Used different pictures than the pictures indicated- **please specify**<br><br>**S3** Used objects in place of pictures- **please specify, underline if they are not the same thing as in pictures**<br><br>**S4** Made test item auditory in place of pictures/objects- **please specify**<br><br>**S5** Other- **please specify** | **P1** Made pictures/objects tactual (or used textures as the pictures)- **please specify**<br><br>**P2** Colored the pictures<br><br>**P3** Enlarged the pictures<br><br>**P4** Bolded the pictures<br><br>**P5** Simplified the pictures- **please specify**<br><br>**P6** Used a low vision device- **please specify**<br><br>**P7** Used supplemental lighting<br><br>**P8** Used a high contrast background<br><br>**P9** Other- **please specify** | **L1** Allowed student to feel each object/picture- **specify if it was done H-O-H**<br><br>**L2** Created a defined space to find objects/pictures (e.g. trays)- **please specify**<br><br>**L3** Held each object/picture in student's field of view<br><br>**L4** Used an alternate placement of pictures/objects<br><br>**L5** Used a slant board<br><br>**L6** Anchored the pictures/objects<br><br>**L7**- increased spacing between items<br><br>**L8** Specifically reoriented the student to the location of each of the choices (to be able to find them)<br><br>**L9** Other- **please specify** | **D1** Used alternate wording- **please specify**<br><br>**D2** Other- **please specify** | **R1** Used eye gaze<br><br>**R2** Used an augmentative communication device (switch, auditory scanning, etc.) **–please specify**<br><br>**R3** Used a yes/no indication<br><br>**R4** Other- **please specify** |

**REMEMBER:**

- ✓ **If you have comments about the quality of an accommodation or think an additional accommodation was needed, specify it in the comments box for that item.**

## Reason for Score Grid

| Code | Description | Examples |
|---|---|---|
| **A** | Score accurately reflects student ability on the test item (given the observed accommodations) | **Score 5:** Student knew the answer right away<br>**Score 4:** Student seems to know the answer after one or more additional prompts or redirection<br>**Score 3:** Teacher intentionally modified the item for the student's ability level and student was able to perform the item<br>**Score 2:** Despite given multiple tries by the teacher, the student just didn't know the answer.<br>**Score 1:** Student passively participated and the teacher made all efforts to involve the student (item was presented, accessible, prompts given, wait time given, etc), but student does not participate actively and does not give an answer. |
| **L** | Score reflects a lucky guess | **Score 5:** Student didn't seem to know the answer but hand or eye gaze happened to land on the correct answer.<br>**Score 4:** Student went through a process of elimination of all the answer choices to come to the correct answer or seemed to randomly pick until s/he got the answer.<br>**Score 3:** Item was modified but student still seemed to only get the answer through a lucky guess. |
| **E** | Score reflects an error made during administration | **Score 4:** Teacher gave an additional prompt even though the student seemed to be about to respond (not enough wait time).<br>**Score 3:** Teacher *unintentionally* modified the item or the initial prompt was cut out of the tape. *<br>**Score 2:** Teacher did not give the student extra attempts to answer again or inadvertently indicated the correct answer.<br>**Score 1:** Teacher did not give the student a chance to respond, and/or the item cut out before answer and student participation was seen.<br>**Score 0:** Item was not presented. |
| **S** | Score assigned to the item does not seem correct | **If you make this determination, indicate the score you think should have been assigned to the item in the box below the codes and state the reason why.** |

* If you can't tell if the student could do the item without the modification because no opportunity was given to do it without it, use E, but make a note.

# APPENDIX E

## JUDGMENTAL TEST ITEM REVIEW INITIAL INSTRUCTIONS

This appendix contains the explanation of PASA and instructions on conducting the test item review. The two expert reviewers used these guidelines during their independent review. The framework was then revised as discussion took place during conferencing.

**Pennsylvania Alternate System of Assessment**
**Judgmental Test Item Review**
**A Level: Grades 3/4 and 7/8**

**General Description of 2005 A Level Assessment Activities in Math and Reading**

| 3/4 Description of Level A Reading Activities | 7/8 Description of Level A Reading Activities | 3/4 Description of Level A Math Activities | 7/8 Description of Level A Math Activities |
|---|---|---|---|
| This third/fourth grade student performed reading activities like matching objects; selecting an item when given the name; selecting an object based on how it is used; determining in which category an object belongs; and matching 2 objects that are used in similar ways. Other activities included showing how a common object is used; listening to a sentence and then immediately answering who or what questions. Virtually all questions were answered by making a selection from a group of 3 objects. All answer choices were very different from one another, making the correct choice obvious. | This seventh/eighth grade student performed reading activities like matching identical pictures; selecting a picture when given the name; finding pictures in an integrated display; selecting a picture based on how it is used; or selecting a picture based on a category label. Other activities included showing how an item in a picture is used; listening to a sentence and then immediately answering who, what, or where questions; or selecting the first or last event. Virtually all questions were answered by making a selection from a group of 3 pictures. All answer choices were very different from one another, making the correct choice obvious. | This third/fourth grade student performed math activities like matching objects based on length, weight, or size; matching identical sets of items; and recognizing money. All questions were answered by selecting from a group of 3 objects, in which other choices were very different from the correct choice and were clearly not correct. | This seventh/eighth grade student performed math activities like recognizing that sets increased or decreased if objects are added or subtracted; selecting a set with 1 item; matching numbers from 1-2; and recognizing money. Other activities included matching items based on length, volume, or area; identifying day- and nighttime activities; and matching digital times. All questions were answered by selecting from a group of 3 items. Answer choices involving numbers and number symbols were very different, making the correct choice obvious. Answers involving more concrete items like objects and pictures were relatively similar to the correct choice, making the decision more difficult. |

**How the Test Items are Scored:**

**2005 PASA Scoring Rubric**

| 5 | 4 | 3 | 2 | 1 | 0 |
|---|---|---|---|---|---|
| Performed **correctly** and **independently** with initial instruction only **and** demonstrated target skill | Performed **correctly** with 1 or more additional **prompts**, redirections or corrections **and** demonstrated targeted skill | Performed **correctly**, but on an easier (**modified**) version of the targeted skill | Performed **incorrectly**, or Demonstrated **skill different** from the targeted skill **or** Performed skill when the **correct response** was **ensured** | **Passively participated**; did not demonstrate targeted skill **and** Assessor ensured correct response **or** **Component not completed** by student or assessor | **Not observed**: item **omitted** or item **not recorded** |

**Task:** Using the forms provided, your job as an expert in the area of education of children with visual impairments is to serve as a judgmental test item reviewer for the 2005 A Level PASA Assessment with the focus on reviewing items for potential bias for students who have a visual impairment. You will review each item for three levels of functional vision. Are there potential problems with the test item for children with visual impairments who:

**V. Primarily use vision to perform most tasks**
**CV. Use a combination of vision and other senses (e.g. tactile, auditory) to perform most tasks**
**NV. Use other senses in place of vision (e.g. tactile, auditory) to perform most tasks**

**Considerations when reviewing:**
The main goal is to flag those items that may be more difficult (or easier) for students with visual impairments when you consider:

1. The type of question and the construct being tested
2. The materials and contexts being used
3. The spatial set-up of the item
4. The type and amount of accommodation that would give access to the test item (would it ultimately change the intent of the skill?)

**In other words, flag those items that (even when accommodated) are likely to result in a student with visual impairments scoring *differently* on the item than a comparable sighted peer of the same ability level.**

As you saw in the descriptions of the assessment activities, many of the pre-cursory math and reading literacy skills at the A Level on the PASA involve manipulation and interpretation of pictures. The underlying progression of difficulty on the PASA is based on moving from objects to pictures to words (in reading) and increasing the closeness of discrimination (that is, distracter choices become more similar). Please remember the following as you review:

1. On the PASA, **teachers are allowed to make whatever accommodations are necessary for students without penalty as long as the accommodations do not change the intent of the skill being assessed or make it easier.** For students with visual impairments, teachers can replace pictures with objects without penalty (although in terms of the PASA progression this could technically make it easier-so, if you feel as you review that this is the case, mark it as such) or substitute items for different items (as long as it doesn't make it easier). Currently, no standard adapted version of the assessment is provided to students without usable vision.

2. All students at the A Level have significant cognitive disabilities and often multiple disabilities. **So, when reviewing, try to consider each item from the perspective of visual impairment specifically** (given that the other affects almost every child taking this level of assessment) and the additional effects that visual impairment can have on the learning process and/or on accessibility to the test items. For example, you may feel a test item is outside the experience or background of any A-Level student, but to make your judgment for students who also have visual impairments, ask: "Does the visual impairment at the given functional level add additional difficulty or make the test item easier?"

3. You will see that the skill assessed area is often written in visual terms, so **think more broadly about the underlying intent of the question (construct).** The main category of the test item (e.g. reading-discrimination) is given to you on your forms. So, when accommodating the item to give the student access to demonstrate that construct, does the test item still measure the same construct at the *same* difficulty level compared to the version sighted peers would be using?

4. Since you will be considering an accommodated form to the item, you will be asked to write down the accommodations you were thinking about when making your decision about the test item since there could be a variety of accommodations made. This will help facilitate the discussion process to come to a final set of "flagged" items for each functional vision level.

**Short Form Instructions!**

Please use the excel spreadsheet provided to record your information. Make sure you are in the correct sheet that corresponds to the grade and subject (math or reading) you are reviewing. The tabs at the bottom of the spreadsheet are labeled.

**For each item:**

**1. First look at the item and its corresponding materials and consider the item *as is* for each level of functional vision** (labeled by letter in the spreadsheet- V, CV, NV):

**V Primarily use vision to perform most tasks**
**CV Use a combination of vision and other senses (e.g. tactile, auditory) to perform most tasks**
**NV Use other senses in place of vision (e.g. tactile, auditory) to perform most tasks**

If you think the item, as is, may be biased or measuring a different construct for that level of functional vision, then mark an X for that level and put your reason in the column next to it. Think about materials- type and quality, layout, visual experience, etc. when making your decisions. For some levels of functional vision *as is* might mean inaccessible without accommodations. If so, then mark it as such. (Remember, although it says biased in the column- this includes both positively and negatively biased. So, if the item might be measuring something different in an easier sense for a student at a given level of functional vision, then indicate that as well.)

**2. Now think about the item in an accommodated form.** Pretend you are the teacher of a student at each level of functional vision. Consider what you would to accommodate and *still* measure the construct of interest (intent of the test item). What

would you do to make it accessible (if needed)? When thinking about that accommodated form, decide if it solves any inaccessibility problems, or if it makes the test item measure something different and/or makes it more difficult or easier than what the sighted students are doing. If so, Mark and X in the biased column for that level of functional vision and provide the reason. *Regardless if you think the item is biased in accommodated form or not for each level, please write the accommodation you were considering for each functional vision level while reviewing, and indicate your opinion about whether or not this item is easy to accommodate (that is, easy for the teacher to make it accessible and keep the intent of the skill intact- Yes/No) in the designated columns for each functional vision level.*

**3. The last column is a column for other comments-** thoughts, challenges, suggestions for the item, etc. that you want to share as part of your going through this process.

## REFERENCES


Allman, C.B. (2002). *Results of survey on state assessment and accountability initiatives: Inclusion of students with visual impairments* [research results]. Retrieved April 24, 2005 from www.aph.org/tests/results.html.

Allman, C.B. (2004). *Test access: Making tests accessible for students with visual impairments: A guide for test publishers, test developers, and state assessment personnel* (2nd ed.). Louisville, KY: American Printing House for the Blind.

American Printing House for the Blind (2004). Distribution of eligible students based on the federal quota census of January 5, 2004 (Fiscal Year 2005). Retrieved on June 10, 2006 from http://sun1.aph.org/fedquotpgm/dist05.html.

Assessing Special Education Students (ASES) Work Group, Council of Chief State School Officers (n.d.). *Determining when accommodated test administrations are comparable to standard test administrations.* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Based on Tindal, G. (1998). *Models for understanding task comparability in accommodated testing.* Available at http://www.ccsso.org/.

Barraga, N.C., & Erin, J.N. (2001) *Visual impairments and learning (4th Ed.)* Austin, TX: Pro-Ed.

Barton, K.E., & Huynh, H. (2003). Patterns of errors made by students with disabilities on a reading test with oral reading administration. *Educational and Psychological Measurement, 63*(4).

Bennett, R.E., Rock, D.A., & Jirele, T. (1987). GRE score level, test completion, and reliability for visually impaired, physically handicapped, and nonhandicapped groups. *The Journal of Special Education, 21* (3), 9-21.

Bennett, R.E., Rock, D.A., & Kaplan, B.A. (1987). SAT differential item performance for nine handicapped groups. *Journal of Educational Measurement, 24*(1), 44-55.

Bennett, R.E., Rock, D.A., & Novatkoski, I. (1989). Differential item functioning on the SAT-M Braille Edition. *Journal of Educational Measurement, 26* (1), 67-79.

Bowen, S.K., & Ferrell, K.A. (2003). Assessment in low-incidence disabilities: The day-to-day realities. *Rural Special Education Quarterly, 22* (4), 10-19.

Bradley-Johnson, S. (1994). *Psychoeducational assessment of students who are visually impaired or blind: Infancy through high school* (2nd ed.). Austin, TX: Pro-Ed.

Brambring, M., & Troster, H. (1994). The assessment of cognitive development in blind infants and preschoolers. *Journal of Visual Impairment & Blindness, 88*(1), 9-18.

Camilli, G., & Shepard, L.A. (1994). *MMSS: Methods for identifying biased test items: Volume 4.* Thousand Oaks, CA: Sage Publications.

Caton, H. (1977). The development and evaluation of a tactile analog to the Boehm Test of Basic Concepts, form A. *Journal of Visual Impairment & Blindness,71*(9) , 382-386.

Clapper, A.T., Morse, A.B., Lazarus, S.S., Thompson, S.J., & Thurlow, M.L. (2005). *2003 state policies on assessment participation and accommodations for students with disabilities. (Synthesis Report 56).* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Coleman, P.J. (1990). Exploring visually handicapped children's understanding of length (math concepts). (Doctoral dissertation, The Florida State University, 1990). *Dissertation Abstracts International, 51,* 0071

Corn, A.L., Wall, R.S., Jose, R.T., Bell, J.K., Wilcox, K., & Perez, A. (2002). An initial study of reading and comprehension rates for students who received optical devices. *Journal of Visual Impairment & Blindness, 96*(5).

Dekker, R., Drenth, P.J.D., Zaal, J.N., & Koole, F.D. (1990). An intelligence test series for blind and low vision children. *Journal of Visual Impairment & Blindness, 84* (2), 71-76.

Destefano, L., Shriner, J.G., & Lloyd, C. A. (2001). Teacher decision making in participation of students with disabilities in large-scale assessments. *Council for Exceptional Children, 68* (1), 7-22.

Dimcovic, N., & Tobin, M.J. (1995). The use of language in simple classification tasks by children who are blind. *Journal of Visual Impairment & Blindness, 89* (5), 448-459.

Elliott, S.N., McKevitt, B.C., Kettler, R.J. (2002). Testing accommodations research and decision making: The case of "good" scores being highly valued but difficult to achieve for all students. *Measurement and Evaluation in Counseling and Development, 35,* 153-166.

Fuchs, L.S., Fuchs, D., Eaton, S.B., Hamlett, C., Binkley, E., & Crouch, R. (2000). Using objective data sources to enhance teacher judgments about test accommodations. *Exceptional Children, 67* (1), 67-81.

Geisinger, K.F. (1994). Psychometric issues in testing students with disabilities. *Applied Measurement in Education, 7* (2), 121-140.

Gersten, R., & Baker, S. (2002). The relevance of Messick's four faces for understanding the validity of high-stakes assessments. In G. Tindal & T.M. Haladyna (Eds.), *Large-scale assessment programs for all students.* (pp. 49-66). Mahwah, NJ: Lawrence Erlbaum Associates.

Gong, B., & Marion, S. (April 13, 2006). *Dealing with flexibility in assessments for students with significant cognitive disabilities.* National Center for the Improvement of Educational Assessment.

Groenveld, M., & Jan, J.E. (1992). Intelligence profiles of low vision and blind children. *Journal of Visual Impairment & Blindness, 86,* 68-71.

Hatlen, P. (December 4, 2003). *Impact of literacy on the expanded core curriculum.* Presentation at the Getting In Touch With Literacy Conference, Philadelphia, PA. Retrieved on January 8, 2005 from http://www.tsbvi.edu/agenda/literacy.htm.

Helwig, R., & Tindal, G. (2003). An experimental analysis of accommodation decisions on large-scale mathematics tests. *Council for Exceptional Children, 69*(2), 211-225.

Huebner, K.M. (2000). Visual Impairment. In M.C. Holbrook & A.J. Koenig (Eds.), *Foundations of education (2nd ed.) volume 1: History and theory of teaching children and youths with visual impairments* (pp. 55-76). New York, NY: AFB Press.

Hull, T., & Mason, H. (1995). Performance of blind children on digit-span tests. *Journal of Visual Impairment &Blindness, 89*(2), 166-168.

Individuals with Disabilities Education Improvement Act of 2004, Pub. L. No. 108-446, 118 Stat. 2648 (December 3, 2004).

Jackson, M.L. (2003). The effects of testing adaptations on students' standardized test scores for students with visual impairments in Arizona. *Dissertation abstracts International,* DAI-A 64/10. (UMI No. 3108915).

Johnstone, C.J. (2003). *Improving validity of large-scale tests: Universal design and student performance* (Technical Report 37). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved January 2, 2005 from http://education.umn.edu/NCEO/OnlinePubs/Technical37.htm.

Johnstone, C.J., Thompson, S.J., Moen, R.E., Bolt, S., & Kato, K. (2005). *Analyzing results of large-scale assessments to ensure universal design* (Technical Report 41). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved October 2, 2005 from http://education.umn.edu/NCEO/OnlinePubs/Technical41.htm

Kirchner, C., & Diament, S. (1999). Estimates of the number of visually impaired students, their teachers, and orientation and mobility specialists: Part I. *Journal of Visual Impairment & Blindness, 93*, 600-606.

Knowlton, M., Seeling, S., Martin, J., & Archer, M. (2003). Assessment review process for addressing visual impairment bias in the state of Minnesota's standardized tests. *RE:view, 35* (1), 7-13.

Koenig, A.J., & Holbrook, M.C. (1995). *Learning media assessment of students with visual impairments* (2nd ed.). Austin, TX: Texas School for the Blind and Visually Impaired.

Koretz, D., & Hamilton, L. (1999). *Assessing students with disabilities in Kentucky: The effects of accommodations, format, and subject.* (Technical Report No. 498). Los Angeles, CA: National Center for Research on Evaluation, Standards and Student Testing. Retrieved October 2005 from http://www.cse.ucla.edu/products/reports_set.htm.

Koretz, D., & Hamilton, L. (2001). *The performance of students with disabilities on New York's Revised Regents Comprehensive Examination in English* (Technical Report 540). Los Angeles, CA: National Center for Research on Evaluation, Standards and Student Testing. Retrieved October 2005 from http://www.cse.ucla.edu/products/reports_set.htm.

Linn, R.L. (2002). Validation of the uses and interpretations of results of state assessment accountability systems. In G. Tindal & T.M. Haladyna (Eds.), *Large-scale assessment programs for all students.* (pp. 27-47). Mahwah, NJ: Lawrence Erlbaum Associates.

Lussenhop, K., & Corn, A.L. (2002). Comparative studies of the reading performance of students with low vision. *RE:view, 34* (2), 57-69.

Mcloughlin, J.A., & Lewis, R.B. (1994). *Assessing special students.* New York: Macmillan College Publishing Company.

Milian, M. (1996). Knowledge of basic concepts of young students with visual impairments who are monolingual or bilingual. *Journal of Visual Impairment & Blindness, 90*(5), 386-399.

Mulford, R. (1988). First words of the blind child. In Smith, M. & Locke, J.L. (Eds.). *The emergent lexicon: The child's development of linguistic vocabulary* (pp. 293-335). San Diego: Academic Press.

National Center on Low-Incidence Disabilities (2004). *Statewide assessment results for students with low-incidence disabilities* [research results]. Retrieved on April 10, 2005 from http://nclid.unco.edu/outcomes.

Pennsylvania Alternate System of Assessment (PASA) (2005). *State report.* Retrieved on June 1, 2006 from http://www.pasaassessment.org/misc/pasareport.do

Pennsylvania Alternate System of Assessment (PASA) (2005).*Technical supplement.*

Pennfield, R.D., & Lam, T.C.M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19* (3), 5-15.

Phillips, S.E. (1994). High-stakes testing accommodations: Validity versus disabled rights. *Applied Measurement in Education, 7* (2), 93-120.

Quenemoen, R., Rigney, S., & Thurlow, M. (2002). *Use of alternate assessment results in reporting and accountability systems: Conditions for use based on research and practice* (Synthesis Report 43). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved on January 2, 2005 from http://education.umn.edu/NCEO/OnlinePubs/Synthesis43.html.

Quenemoen, R., Thompson, S., & Thurlow, M. (2003). *Measuring academic achievement of students with significant cognitive disabilities: Building understanding of alternate assessment scoring criteria* (Synthesis Report 50). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved on January 2, 2005 from http://education.umn.edu/NCEO/OnlinePubs/Synthesis50.html.

Rock, D.A., Bennett, R.E., & Jirele, T. (1988). Factor structure of the Graduate Record Examinations General Test in handicapped and nonhandicapped groups. *Journal of Applied Psychology, 73* (3), 383-392.

Roeber, E. (2002). *Setting standards on alternate assessments* (Synthesis Report 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved on January 2, 2005 from http://education.unm.edu/NCEO/OnlinePubs/Synthesis42.html.

Ryan, J.M., & DeMark, S. (2002). Variation in achievement scores related to gender, item format, and content area tested. In G. Tindal & T.M. Haladyna (Eds.), *Large-scale assessment programs for all students.* (pp. 67-88). Mahwah, NJ: Lawrence Erlbaum Associates.

*Standards for educational and psychological testing* (Second Impression 2004) (1999). Washington, DC: American Educational Research Association.

Tate, R. (2002). Test dimensionality. In G. Tindal & T.M. Haladyna (Eds.), *Large-scale assessment programs for all students.* (pp. 181-212). Mahwah, NJ: Lawrence Erlbaum Associates.

Taylor, C.S. (2002). Incorporating classroom-based assessment into large-scale assessment programs. In G. Tindal & T.M. Haladyna (Eds.), *Large-scale assessment programs for all students.* (pp. 233-259). Mahwah, NJ: Lawrence Erlbaum Associates.

Thompson, S., Blount, A., & Thurlow, M. (2002). *A summary of research on the effects of test accommodations: 1999 through 2001 (Technical Report 34).* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thompson, S.J., Johnstone, C.J., Anderson, M.E., & Miller, N.A. (2005). *Considerations for the development and review of universally designed assessments* (Technical Report 42). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved January 4, 2006 from http://education.umn.edu/NCEO/OnlinePubs/Technical42.htm.

Thompson, S. J., Johnstone, C.J., & Thurlow, M.L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved January 2, 2005 from http://education.umn.edu/NCEO/OnlinePubs/Synthesis4.html.

Thompson, S., Lazarus, S., Clapper, A., & Thurlow M. (August 2004). *Essential knowledge and skills needed by teachers to support the achievement of students with disabilities (Issue Brief Five).* College Park, MD: University of Maryland, Educational Policy Reform Research Institute.

Thompson, S., & Thurlow, M. (2002). *Universally designed assessments: Better tests for everyone!* (Policy Directions No. 14). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved January 2, 2005 from http://education.umn.edu/NCEO/OnlinePubs/Policy14.htm.

Thompson, S., & Thurlow, M.T. (2003). *2003 state special education outcomes: Marching on.* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [January 2, 2005] from http://education.umn.edu/NCEO/OnlinePubs/2003StateReport.htm./

Thurlow, M., & Bolt, S. (2001). *Empirical support for accommodations most often allowed in state policy (Synthesis Report 41).* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M.L., Elliott, J.L., & Ysseldyke, J.E. (2003). *Testing students with disabilities: Practical strategies for complying with district and state requirements.* Thousand Oaks, CA: Corwin Press.

Thurlow, M., House, A., Boys, C., Scott, D., & Ysseldyke (2000). *State participation and accommodation policies for students with disabilities: 1999 Update (Synthesis Report 33).* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M.L., McGrew, K.S., Tindal, G., Thompson, S.L., Ysseldyke, J.E., & Elliott, J.L. (2000). *Assessment accommodations research: Considerations for design and analysis (Technical report 26).* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Tindal, G. & Fuchs, L. (2000). *A summary of research on test accommodations: An empirical basis for defining test accommodations.* Lexington, KY: Mid-South Regional Resource Center. (retrieved from http://education.umn.edu/nceo/ , June 2005).

Title I: Improving the Academic Achievement of the Disadvantaged, Final Rule, 34 CFR Part 200 (Dec. 9, 2003).

U.S. Department of Education (April 7, 2005). *Raising achievement: Alternate assessments for students with disabilities.* Retrieved on April 10, 2005 from http://www.ed.gov/print/policy/elsec/guid/raising/alt-assess-long.html.

U.S. Department of Education (n.d.). Toolkit on teaching and assessing students with disabilities. Website: http://osepideasthatwork.org/toolkit/index.asp.

U.S. Department of Education, Office of Elementary and Secondary Education, *No Child Left Behind: A Desktop Reference,* Washington, D.C., 2002.

U.S. Department of Education, Office of Special Education Programs (2003). Twenty-fifth annual report to Congress on the implementation of the Individuals with Disabilities Education Act. Washington, DC: U.S. Department of Education. Retrieved on June 10, 2006 from http://www.ed.gov/about/reports/annual/osep/2003/index.html.

U.S. Department of Education, Office of Special Education Programs (2004). Twenty-sixth annual report to Congress on the implementation of the Individuals with Disabilities Education Act. Washington, DC: U.S. Department of Education. Retrieved on November 12, 2006 from http://www.ed.gov/about/reports/annual/osep/2003/index.html.

U.S. Office of Special Education Programs. *Special Education Elementary Longitudinal Study (SEELS)* [Wave 1 and 2 Data Tables]. Available from SEELS Web site, http://www.seels.net.

Wagner, M., & Blackorby, J. (2004). *Overview of findings from wave 1 of the Special Education Elementary Longitudinal Study (SEELS).* Menlo Park, CA: SRI International. Available at http://www.seels.net/designdocs/seels_wave1_9-23-04.pdf.

Warren, D.H. (1994). *Blindness and children: An individual differences approach.* New York, NY: Cambridge University Press.

Wyver, S.R., & Markham, R. (1999). Visual items in tests of intelligence for children. *Journal of visual impairment and blindness, 93*(10), 663-665.

Ysseldyke, J.E., Olsen, K.R., & Thurlow, M.L. (1997). *Issues and considerations in alternate assessments* (Synthesis Report 27). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved on January 2, 2005 from http://education.umn.edu/NCEO/OnlinePubs/Synthesis27.htm.

Zumbo, B.D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and likert-type (ordinal) item scores.* Ottawa, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zwick, R., Thayer, D.T., & Mazzeo, J. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education, 10* (4), 321-344.