Comparison of Four Screening Procedures for Identifying Students for Special Education: Validation of Problem Validation Screening.

Amanda Mathany Vanderheyden

Louisiana State University and Agricultural & Mechanical College

Follow this and additional works at: https://digitalcommons.lsu.edu/gradschool_disstheses

Recommended Citation

https://digitalcommons.lsu.edu/gradschool_disstheses/325
INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

ProQuest Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
COMPARISON OF FOUR SCREENING PROCEDURES FOR IDENTIFYING STUDENTS FOR SPECIAL EDUCATION: VALIDATION OF PROBLEM VALIDATION SCREENING

A Dissertation
Submitted to the Graduate Faculty of the Louisiana State University and Agricultural and Mechanical College in partial fulfillment of the requirements for the degree of Doctor of Philosophy
in
The Department of Psychology

by
Amanda Mathany VanDerHeyden
B.A., Tulane University, 1995
M.A., Louisiana State University, 1998
May 2001
# Table of Contents

Abstract ................................................................................................................................ iii

Introduction ........................................................................................................................... 1
  Review of the Literature ................................................................................................. 2
  Problems with Current Classification System ............................................................... 2
  Reliability of Teacher Referral ..................................................................................... 9
  Problem Validation ......................................................................................................... 13
  The Problem of Base Rates ............................................................................................ 18
  Purpose of Study ............................................................................................................. 20

Method .................................................................................................................................. 22
  Participants......................................................................................................................... 22
  Measures ............................................................................................................................ 22
  Procedure ........................................................................................................................... 32

Results ................................................................................................................................... 49
  Reliability .......................................................................................................................... 51
  Overall Descriptive Findings .......................................................................................... 53
  Validity ............................................................................................................................... 54
  Endnote .............................................................................................................................. 80

Discussion ............................................................................................................................. 82

References ............................................................................................................................. 94

Appendix A: Sequence of Experimental Procedures and Screening Criteria .............. 101
Appendix B: Instructions for Administration of CBM Probes ...................................... 104
Appendix C: Average Slopes for Reading and Math ..................................................... 106
Appendix D: Tables of Results ........................................................................................ 110
Vita...................................................................................................................................... 119

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Abstract

Four screening methods were compared to determine the extent to which each method accurately identified students who exhibited math and reading problems severe enough to warrant a full psycho-educational assessment for special education eligibility determination. Methods examined were teacher referral, two subtests from the Comprehensive Inventory of Basic Skills, Revised, the Developmental Reading Assessment, and a new screening measure called “Problem Validation Screening” (PVS). PVS consisted of three components: classwide curriculum-based measurement probes in math and reading, performance/skill deficit assessment, and a brief instructional session.

All students enrolled in first and second grade classes at a participating school were exposed to each of the four screening measures. Students who met criteria as potentially exhibiting a serious problem in reading or math were then exposed to a second measure, termed the “Criterion Assessment” along with two traditional measures of student achievement (i.e., Woodcock-Johnson Psychoeducational Battery, Revised and Iowa Test of Basic Skills). The Criterion Assessment consisted of individually administered curriculum-based assessment and individual intervention. Sensitivity, specificity, as well as positive and negative predictive power were calculated for each screening measure using the Criterion Assessment as the outcome standard. The predictive accuracy estimates of each screening method were compared in classrooms where base rates of academic problems were high and classrooms where base rates of academic problems were low. In all cases, PVS achieved the highest predictive accuracy estimates of all the screening measures regardless of the sample base rates. Implications of the study for
practice are discussed along with the importance of measuring both performance levels and trend when evaluating student performance.

Key Words: resistance to intervention, screening, curriculum-based measurement, overidentification, functional assessment, learning slopes
Introduction

The percentage of children labeled with a disability and placed in special education has increased dramatically over the last 20 years (Algozzine, Ysseldyke, & Christenson, 1983; Ysseldyke, Vanderwood, & Shriner, 1997). Whereas some may interpret this increase in numbers as a positive trend indicating more precise methods of identifying and serving students, others (Algozzine et al., 1983; Ysseldyke & Thurlow, 1984) are alarmed by this large increase, citing reports of negative outcomes associated with placement in special education (e.g., decreased probability of high school graduation, decreased levels of academic achievement) (Reynolds, 1991). Initially, those who viewed increased identification and labeling as a problem attributed it to a failure of diagnostic procedures (e.g., Intelligence tests) and diagnosticians, such as school psychologists, to properly discriminate children who have disabilities from children who do not have disabilities. This issue of inappropriate, invalid, or biased identification has been shown to be most acute for minority children whose numbers are often over-represented in special education. More recently, the responsibility for too many children being inappropriately placed in special education has shifted away from the diagnostic process and, instead, has moved “upstream” to question many of the variables that impact whether or not a child is ever tested in the first place (Maheady, Algozzine, & Ysseldyke, 1984). Specifically, many are attributing inappropriate identification of students for special education to a lack of appropriate instruction in the regular education setting (Ysseldyke & Thurlow, 1984), to inappropriate, biased, and inaccurate teacher referrals (Shinn, Tindal, & Spira, 1987), and to a failure of school systems to afford children mandated protections against being labeled, such as documented use of effective pre-
referral interventions (Gresham, 1991a). Much of the variance in whether a child is classified or not is accounted for by initial teacher referral (Algozzine et al., 1983; Ysseldyke et al., 1997). The purpose of this study was to investigate the validity of a new process for screening and referring children to special education. This process was compared to other methods of identification, such as teacher referral and curriculum-based assessment. In addition to providing basic psychometric data on the new screening procedure, the goal of this study was to better understand how decision-making about referral was influenced by classroom context and the use or non-use of appropriate pre-referral intervention procedures.

**Review of the Literature**

Student referral, diagnosis, and subsequent placement in special education are topics that have been discussed extensively in the special education and school psychology literature. This review is organized around an analysis of some major issues affecting referral, diagnosis, and placement. Specifically, this review describes problems with the current classification system and introduces a data-based screening process to identify students needing special services called Problem Validation Screening.

**Problems with Current Classification System**

Several variables contribute to the difficulty of making reliable decisions regarding student educational progress. One problem with the current classification system is that diagnostic categories are ill-defined and fail to demonstrate that symptoms (e.g., poor reading) are unique to a given diagnosis (e.g., learning disability) (Gresham & Gansle, 1992; Reynolds, 1991). If variables other than learning disability may account for or produce reading problems, then diagnosis based on poor reading becomes unreliable.
Further, the reading problem (i.e., symptom) exhibited by the student suspected of having a learning disability must exceed the base rate occurrence of reading problems in a non-learning disabled sample for the reading problem to be considered clinically significant and worthy of diagnosis (Gouvier, Uddo-Crane, & Brown, 1988). Ysseldyke and colleagues (1997) found that 51% of students receiving services in districts responding to their survey, were classified as learning disabled. This finding could indicate that learning disability has a high prevalence in the population. Alternatively, this finding could indicate that learning disability is commonly misapplied or misdiagnosed. Multiple studies have demonstrated that standardized assessment instruments fail to reliably distinguish learning disabled (LD) students from low achieving students (Algozzine & Ysseldyke, 1983; Ysseldyke, Algozzine, Shinn, & McGue, 1982), leading some authors to suggest that other factors account for the decision to diagnose a learning disability (Marston, Mirkin, & Deno, 1984; Shinn, Ysseldyke, Deno, & Tindal, 1986). For example, Ysseldyke et al. (1982) applied the federal LD criteria to existing groups of classified LD students and non-identified students and found that 40 of the 99 students were misclassified. In another study, Algozzine and Ysseldyke (1983) demonstrated that achievement scores for students identified as LD and students referred for possible LD and diagnosed as not exhibiting LD according to the school's assessment team did not differ. That is, achievement scores for both groups fell in the average range. Ability estimates (i.e., IQ scores) similarly fell in the average range for all students. Algozzine and Ysseldyke point out that the critical criterion in many cases is a discrepancy between ability and achievement, yet, little is known about the distribution of discrepancies in a
normal population, and even this questionable criterion was inconsistently applied in their sample.

Because criteria for classification are frequently unclear and confusing (Algozzine & Ysseldyke, 1983), many of those critical to the decision making process may employ different criteria at each stage of the process (e.g., referral, classification) (Thurlow, Ysseldyke, & Casey, 1984). Thus, unclear categories also result in sub-standard interclinician reliability.

Standardized tests have traditionally been considered the standard upon which diagnoses are made and the validity of screening and referral measures (e.g., teacher referral) are evaluated. However, studies yielding a match between teacher referral and ultimately placement in special education (Algozzine et al., 1983; Marston, Mirkin, et al., 1984; Ysseldyke et al.; 1997) do not necessarily validate the accuracy of teacher referral. Standardized testing may be a fallible basis for making determinations required to classify a student as needing special services (Gresham & Witt, 1997). MacMillan (1998) administered a commonly used standardized battery of instruments to all 150 students who were referred to the school pre-referral committee. Students were followed and their eventual classification was recorded, allowing for a direct comparison of the concordance between empirically-based diagnosis according to state criteria and school-level classification according to the same criteria. MacMillan found very poor concordance between empirically-based identification and school-based classification. For example, he found that only 6 of the 43 students diagnosed with a mental disability using the standardized assessment battery were identified as having a mental disability by the school teams. He also found that many students were identified as learning disabled when
they did not meet district diagnostic criteria. Finally, Ysseldyke, Algozzine, Regan, and McGue (1981) found that “expert” diagnosticians identified students as eligible for special education despite those students having normal psychometric profiles, and cited teacher referral as the primary basis for their decision. These findings and others highlight a pattern of consistency in teacher referral, standardized assessment, and expert interpretation that may lead to an invalid or faulty decision to place a child in special education. That is, once the teacher has identified a student as needing help, the goal of the decision-making team is to determine not if, but which problem a student has, and classification and placement are almost certain outcomes (Ysseldyke & Thurlow, 1984). Thus, the accuracy of teacher referral should be held to the same measurement standards as other identification sources.

Another problem with the current classification system is the requirement that the student be provided with a prereferral intervention. Designed to control for lack of educational opportunity, prereferral interventions are not always implemented with integrity (Gresham, 1991a). Gresham, Gansle, Noell, Cohen, and Rosenblum (1993) conducted an extensive review of the behavioral literature over a ten-year period and found that only 16% of studies systematically assessed and reported the reliability with which experimental manipulations were conducted as planned. Because adequate treatment integrity is critical to demonstrating that behavior change occurred as a result of changes in the IV (i.e., treatment) and ruling out confounds (i.e., establishing internal validity), Gresham et al. have recommended requiring systematic assessment and reporting of treatment integrity estimates when making student eligibility decisions. Yet,
systematic monitoring of pre-referral interventions does not routinely occur in most places.

Similarly, the Behavioral Consultation literature has been criticized for relying solely on teacher report to determine identification of targets for behavior change, integrity of interventions, and outcome of behavior change efforts (Watson & Robinson, 1996). Studies have shown in each case (targets, integrity, outcomes) that teacher report may be unreliable and lead to erroneous conclusions (Witt, Gresham, & Noell, 1996). Further, several studies have demonstrated that interventions are frequently not implemented with integrity (Happe, 1982; Wickstrom, Jones, LaFleur, & Witt, 1998). Happe (1982) found that teachers implemented plans only 50% of the time after verbally having agreed to do so. Wickstrom et al. (1998) found that all 29 teachers in their sample reported that they had conducted an intervention with integrity, whereas direct observation revealed that teachers had implemented interventions on only 4% of the prescribed occasions. Fuchs and Fuchs (1987) conducted a large-scale analysis of a pre-referral intervention program termed “Mainstream Assistance Teams.” These authors found that all participating teachers reported improved student behavior, whereas direct observation indicated that no significant change in student behavior was obtained. Thus, direct observation failed to corroborate teacher report. Fuchs and Fuchs attributed their lack of treatment effect to weak interventions, poor implementation/integrity, and poor monitoring. Pre-referral interventions can only serve their intended purpose (i.e., control for poor instruction, lack of educational opportunity, ensure least restrictive placement) if they are implemented and implemented correctly. Studies have shown that treatment integrity may be insufficient. Thus, assessment teams must find alternative ways to
ensure that pre-referral interventions occur as planned. Gresham (1991a) has recommended assessing and considering the integrity with which mandated pre-referral interventions were conducted when making eligibility and placement decisions.

Finally, the current classification system has been criticized as lacking treatment utility and social validity (Barnett, Macmann, & Carey, 1992; Gresham & Witt, 1997). Traditionally, school psychologists have focused on the assessment of child variables to the exclusion of antecedent or consequent variables that may affect student responding. Frequently, treatment has been guided by pragmatism, resource availability, or skill level of the treatment agent as opposed to adjusting child-environment fit. The goal, traditionally, has been classification. Multiple studies have demonstrated that placement does not systematically relate to quality, quantity, or type of instructional activity (Ysseldyke & Christenson, 1987; Ysseldyke, Thurlow, Christenson, & Weiss, 1987; Ysseldyke, O’Sullivan, Thurlow, & Christenson, 1989), and one study found that those classified as EMR spent less time in instructional activity and more time in free time than did their peers (Ysseldyke, Thurlow, Christenson, & Muyskens, 1991). Multiple negative outcomes may be associated with this type of child-centered assessment (Maheady et al., 1984). One possible outcome is overidentification of minority students. Several authors have framed the problem of overidentification as a consequence of a medical model driven “search for pathology” (Sarason & Doris, 1979) that begins when a student is identified (typically by his/her teacher) as exhibiting behaviors discrepant from those expected for the class. Many authors argue that assessment teams should attempt to identify deficiencies in the student’s environment prior to individual assessment of the
learner (Maheady, et al., 1984), and that the first step of screening ought to be alteration of the student's regular environment (Adelman, 1982; Shinn, 1989).

Standard assessment batteries have been criticized as lacking treatment utility and social validity. Following assessment and classification, the familiar scenario follows that the student has been labeled, but has received no systematically linked instructional programming changes designed to remediate deficits and train skills. This outcome raises multiple ethical and possibly legal concerns regarding the practice of identifying and placing students in programs that do not improve student outcomes (e.g., Marshall et al. v. Georgia, 1984).

Messick (1995) has defined validity not as a property of an assessment instrument, but rather as the validity of the meaning or interpretation of the test scores and the associated intended or unintended consequences for the individual who was subjected to the assessment. In its most basic sense, if testing fails to result in the outcomes for which it was intended (i.e., a placement that provides improved habilitation) then the test is invalid. The decision to evaluate a child carries with it certain intended consequences. Several studies have demonstrated that placement does not reliably result in these intended outcomes (Ysseldyke & Christenson, 1987; Ysseldyke, et al., 1987; Ysseldyke, Christenson, Thurlow, & Bakewell, 1989; Ysseldyke, O'Sullivan, et al., 1989). Additionally, unintended consequences are possible and have not been routinely operationalized and measured. For example, to what degree does previous referral influence teacher expectations regarding student ability, and to what extent do teacher expectations influence student performance? From Brophy and Good’s (1970) work in this area, we may infer that teacher and student behavior may be influenced by
teacher expectations, and we may infer that teacher expectations may be influenced by
other teachers having identified the student as exhibiting learning problems (i.e., previous
referral). When the very construct of learning disability, for example, has failed to meet
the basic measurement standards of reliability and validity, the application of this
diagnosis or topography must be questioned. Messick (1995) defined validity broadly to
encompass all stages and outcomes of the testing process. In sum, the special education
referral, identification, and placement process has not been held to the rigor of validity
criteria as Messick (1995) has defined them, especially with respect to demonstrating
treatment validity.

Reliability of Teacher Referral

Perhaps because diagnostic categories are somewhat fluid, and interclinician
reliability weak, some diagnostic criteria may be assessed by teacher report. For example,
the assessment team may ask the teacher “Do you think this student’s behavior
performance is influenced by lack of motivation?” in an effort to determine a student’s
“true” ability. Direct measurement of the accuracy with which teachers evaluate student
performance with respect to specific classification criteria has not been examined.
Algozzine et al. (1983) conducted a large-scale survey study examining referral and
placement rates and reported that 92% of referred students were tested and 73% of those
students qualified and were placed in special programs. During the next decade, most
school districts implemented programs designed to reduce the number of students
receiving a formal psycho-educational evaluation to determine eligibility (e.g., pre-
referral problem solving committees, mandated pre-referral interventions). Despite the
implementation of these programs, Ysseldyke et al. replicated Algozzine et al.’s study in

9

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
1997 and obtained remarkably similar results. In one study, 72% of referred students were placed in some form of special education, and most were placed in the category for which they were referred (Foster, Ysseldyke, Casey, & Thurlow, 1984). These findings could indicate that teachers fairly reliably identify students in need of special services. Alternatively, these findings could indicate that teacher report is an integral component of the identification and classification decision-making process. That is, teacher report may not only influence the decision to evaluate a child, but also whether or not to qualify a student as needing special services and what categorical placement is most appropriate. Thus, it is not possible to assess the accuracy of teacher referral based on evaluation and classification numbers alone.

Nonetheless, several studies have attempted to evaluate the accuracy of teacher referral within these described measurement constraints. Gresham, Reschly, and Carey (1987) found that teachers could reliably and accurately discriminate children in need of special services from typically developing students. However, these findings were limited because teachers who provided the ratings were aware of the student’s previous classification of LD. Ten years later, Gresham and colleagues examined this question again (Gresham, MacMillan, & Bocian, 1997). The authors improved upon the first study by having separate teachers rate control and classified students and ensuring that teachers were blind to the purpose of the study. Yet, these findings were limited by the fact that teachers nominated the control group. Results indicated that teachers could discriminate students in need of special services from “normals,” but could not distinguish between low achieving (LA), learning disabled (LD), and mildly mentally retarded (MMR)
students. The research regarding the accuracy of teacher referral is largely equivocal due to methodological limitations in the research. In short, the issue remains undecided.

The reliability with which teachers make the series of judgments that result in referral, and more times than not classification, remains unclear. What variables may influence teacher self-report about a student? Tolerance, perceptions of normality, parent and colleague influence, access to reinforcing properties of referral (i.e., attention, escape from difficult child), and system resource constraints are factors that may introduce bias to the decision-making process (Kauffman, Lloyd, & McGee, 1989). Gresham (1991b) defined tolerance as the extent to which a behavior disturbs others in the child’s environment and the probability that adults in the environment will implement strategies to reduce the occurrence of the troubling behavior. Among the least-tolerated behaviors are disruptive, externalizing behaviors (Nelson, Rutherford, Center, & Walker, 1991). Further, these low tolerance behaviors may be more likely to be inadvertently reinforced by the teacher, resulting in increased occurrence (e.g., child who gets out of his or her seat is reminded to sit back down, child who tantrums may escape math). The intervention most frequently attempted by many teachers is referral to special education (Witt & Martens, 1988), and the resulting placements are among the most restrictive for these students (Ramsey & Walker, 1988). Brophy and Good (1970) demonstrated the influence of teacher perception when they used direct observation to quantify the interaction between teacher expectation, student behavior, and teacher response to student behavior. Teachers were asked to rank their students in descending order according to achievement. The authors then observed teacher interaction with the five highest ranked students and the five lowest ranked students in four first grade classrooms. Results
indicated that teachers were more persistent in obtaining responses from the high-achieving students than the low. That is, teachers were more likely to provide a second opportunity to respond (e.g., rephrasing the question, offering a clue) for the high students compared to the low students. In contrast, teachers were more likely to supply the answer or call on another student when responding to the lows compared to the highs. Teachers more frequently praised and less frequently criticized the highs, and teachers were much more likely to provide no feedback at all to the lows (14.75% versus 3.33% for the highs). Additionally, high students were much more likely to engage in behaviors designed to set the occasion for teacher feedback (e.g., showing their work, asking questions about their work) and were more likely to initiate response opportunities (e.g., hand raising). Finally, low boys were much more prone to teacher criticism than high boys, while boys in general received much more attention than did girls overall.

These findings are critical because they highlight an interactive pattern that may contribute to the shaping of poor academic performance in students perceived as low-achieving by their teachers. That is, teachers may unwittingly differentially enhance the reinforcing value of a correct answer provided by a student whom he or she perceives as high-achieving as compared to a student whom he or she perceives as low-achieving. That student, in turn, receives a greater number of opportunities to respond and a higher rate of feedback, and is likely to engage in behaviors (e.g., correct responding, active participation) that are simultaneously reinforcing to the teacher and set the occasion for teacher provided reinforcement to the student.

In a direct comparison of teacher-based referral and CBM-based referral, Marston et al. (1984) noted an interesting trend. These authors found that females in the teacher-
referred group tended to exhibit problem behaviors more frequently than did females in a
CBM-referred group. Varying teacher tolerance levels for attributes or behaviors may
result in varying degrees of error in the teacher’s decision to refer a student for special
education (Shinn et al., 1987). For example, Shinn et al. found that teachers tended to
refer the lowest achieving students for special services generally speaking. However,
results indicated disproportionate referral of black students and male students. That is,
teachers referred more black students and more male students than would be expected
according to base rate occurrence of low-achieving black students and low-achieving
males in the sample. Marston et al. (1984) found that the performance of teacher-referred
students did not significantly differ from CBM referred students. However, males and
females were more equally represented when referral was based on CBM scores. Finally,
64% of teacher referred students were labeled learning disabled despite not having met
district criteria (compared to 20% of CBM referred students). Thirty-six students were
referred based on CBM performance. Yet, only 5 of these students qualified as L.D. an
extremely low hit rate. Given that some estimate the cost of an educational evaluation to
be as much as $3000 per child (Ysseldyke et al., 1983), CBM referrals did not increase
the efficiency of the identification process in this case. Certainly teacher perception is
important. However, opinion is fallible. Wherever classification and placement decisions
are based upon opinion, those decisions are equally fallible. Yet, basing referral upon
CBM scores alone may not increase the efficiency of the referral process.

**Problem Validation**

The limitations associated with traditional methods of identifying students in need
of special services have led some to propose a functional assessment or problem solving
approach to identification (Good & Kaminski, 1996; Shinn, 1989). A functional assessment (FA) approach emphasizes assessment of the instructional environment. The FA approach is predicated upon frequently repeated direct measurement of academic skills for goal setting and progress monitoring purposes. Additionally, manipulating variables known to affect student learning are primary goals (e.g., increasing academic engaged time, working on instructional level). In short, the FA approach is a problem solving approach that focuses on altering the instructional environment rather than assigning a student to a category and then basing treatment on placement or topography.

Function based treatment has several advantages over treatment by topography or classification. Imagine, for example, that a child exhibits some of the characteristics associated with a given diagnostic category and some of the characteristics associated with a second diagnostic category, making the determination of one category over the other a potentially unreliable distinction. If category membership is an unreliable process, then treatment by category is likely to be ineffective. Further, treatment by category assumes that all students assigned to a given category exhibit similar deficit and skill patterns. This assumption may lead to greater environmental restriction (and diminished habilitation) for the student who falls at the high end of the curve (i.e., exhibits more skills and fewer deficits compared to peers). Alternatively, the student who falls at the low end of the curve (i.e., exhibits fewer skills and more deficits) is likely to be underserved.

Iwata, Vollmer, Zarcone, and Rodgers (1993) emphasize that no specific form of treatment has been shown to be sufficiently effective across behavioral topographies, contingencies, and situations. Therefore, function-based treatments may be most effective
in reducing behavioral excesses (e.g., errors) and behavioral deficits (e.g., words read correctly). The purpose of functional assessment is to identify the antecedent or subsequent variables that are consistently associated with the occurrence of the behavior targeted for assessment. These assessment data link directly to treatment by identifying potential predictor or maintaining variables that can be manipulated to alter the response-reinforcer relationship. Because FA by its very nature is idiographic, function-based treatments are tailored to individual needs, resulting in more precise treatments that avoid many of the limitations associated with classification or topography-based treatment.

One type of screening approach based on functional assessment is called problem validation. Problem validation, as described by Witt, Daly, and Noell (2000) consists of brief classwide CBM screenings, direct observation, comparison to same-class peers, assessment of the effect of providing an incentive for improved performance, and intervention conducted in the natural setting and monitored for treatment integrity. Referral to formal evaluation is based on resistance to intervention (i.e., failure to make sufficient growth given a documented assessment-based intervention conducted in the regular education setting).

The purpose of the problem validation model is to provide systematic, fairly immediate modifications in students' environments to increase academic performance and learning. This model is currently being implemented in several pilot schools in a participating district. During the first full year of implementation, 63% of the referrals in the three participating schools were determined to be validated problems. Of these validated problems, 53% were successfully resolved with intervention. Thus, approximately 70% of referrals to the school level committee were resolved to the
satisfaction of the committee so that they did not proceed for a full eligibility evaluation. The second year of implementation, 64% of the referrals to the school level committee were found to be validated problems. Of these validated problems, 44% were successfully resolved with intervention. Thus, 64% of the referrals to the school level committee were resolved to the satisfaction of the committee and did not proceed for a full eligibility evaluation.

Following the first full year of implementation, 33% of the students who were referred to the school committee, were validated minority cases at a school where 16% of the student population were minority students. At another school, 40% of the students who were referred to the school committee were validated minority cases, whereas the school population consisted of 75% minority students. At the third participating school, 31% of the students who were referred to the school level committee were validated minority cases, whereas the school population consisted of 30% minority students. The following year, considering the same three schools, 25% of the referrals at the first school were found to be validated minority problems. At the second school, 71% of the referrals were found to be validated minority problems. At the third school, 33% of the referrals were found to be validated minority problems.

Interestingly, from year one to year two, a decreasing trend in overall number of referrals was observed. At school one, referrals decreased from 15 to 6. At school two, referrals decreased from 31 to 7. At school three, referrals decreased from 30 to 3. Finally, the proportion of referrals that represented minority students for the participating schools was as follows. For the first school, minority referrals decreased from 33% to 16%. For the second school, minority referrals increased from 81% to 100%. For the
third school, minority referrals increased very slightly from 60% to 67% (Pupil Appraisal Model, Louisiana Department of Education, Grant # 324 50 4067).

The Problem Validation Screening procedure is derived from the behavior analysis literature and emphasizes direct observation and measurement of the component behaviors that eventually lead to a decision to qualify or not to qualify a student as disabled. Problem Validation Screening provides direct measurement of academic performance compared to same-class peers and a national normative sample, disruptive behavior compared to same-class peers, and influence of motivation. This screening procedure is designed to yield data upon which referral decisions may be based. Thus, Problem Validation Screening is a data-based procedure for identifying students in need of further assessment and possible referral for formal evaluation. The goal of Problem Validation Screening is to improve the efficiency of support service resources. That is, Problem Validation Screening proposes to provide a data-based means of identifying students who may need help, provide specific (i.e., function-based) help as a first step, and then refer for special education evaluation those students who do not respond sufficiently to intervention conducted in the regular education setting. This model is predicated upon the idea that the first step of any screening procedure should be instructional modification of the current environment (Adelman, 1982; Shinn, 1989). Although this model has been evaluated to satisfy the requirements of funding agencies, such evaluation efforts have not been subjected to peer review. Hence, data pertaining to the psychometric adequacy of the process were needed. Collection of data for this purpose was a goal of this investigation. Toward this goal, data were collected to examine the technical adequacy of Problem Validation Screening.
The Problem of Base Rates

Any screening procedure must meet certain technical adequacy standards. Typically, this goal is accomplished in part by comparing the outcomes of the experimental screening measure to the outcome of some criterion measure. A more stringent assessment of the utility of a screening tool, however, involves the demonstration that the instrument improves identification accuracy over accurate identification obtained by chance alone (i.e., base rate occurrence in the population) (Meehl & Rosen, 1955). Further, because referral and identification occur within a classroom context, we wondered if base rate occurrence of academic problems may affect accuracy of various identification methods (e.g., problem validation, teacher referral). That is, it may be possible that students in low-achieving classrooms may be more or less likely to be identified than students in high-achieving classrooms. Further, students in low-achieving classrooms/schools may be more likely to meet state and federal criteria as needing special services using traditional psycho-educational batteries, when in reality, the finding is a confound of poor instruction. Thus, this study attempted to assess the validity of Problem Validation Screening by considering base rate occurrence of academic problems on a class-by-class basis. A brief description of the base rate literature as it pertains to this goal is provided below.

Paget and Barnett (1990) described several important qualities for screening assessments, including that procedures should have demonstrated reliability and validity, data specifying the sensitivity and specificity of the screening instrument, demonstrated acceptability, and include careful observation of child behavior. Further, to prove that a screening mechanism has utility, it must be demonstrated that the test improves the
predictive accuracy rate more so than the chance accuracy rate derived from antecedent probabilities (or base rates) (Meehl & Rosen, 1955). In quantifying the utility of a predictive tool, statistically significant differences between groups is less important than number of correct decisions made for individuals within the groups (Meehl & Rosen, 1955). It is important to quantify the sensitivity (i.e., power of the test to identify true positives) and specificity (i.e., power of the test to identify true negatives). In this case, the sensitivity of the Problem Validation Screening procedure was defined as the probability that a student exhibiting a serious learning problem resulted in a validated problem during Problem Validation Screening. Specificity was defined as the probability that a student who did not exhibit a serious learning problem resulted in a non-validated problem. To enhance the accuracy of prediction, one would want to maximize both sensitivity (to avoid false negatives) and specificity (to avoid/minimize false positives). Dawes (1962) proposed that consideration of conditional probability is a more clinically applicable and meaningful criterion. That is, determining the probability of category membership (e.g., true disability) given a test response (e.g., validated problem) is more useful to those who are charged with making category assignments (i.e., diagnoses). Given the current research question, positive predictive power was defined as the probability that a child has a disability given a validated problem, and negative predictive power was defined as the probability that a child does not have a disorder given the absence of a validated problem (i.e., a non-validated problem) (Milich, Widiger, & Landau, 1987). Several authors argue that understanding and quantifying positive predictive power and negative predictive power are critical to developing diagnostic criteria (Elliot, Busse, & Gresham, 1993; Milich et al., 1987). For example, Milich et al.
conducted structured parent interviews in a clinic setting to measure the utility of the
reported occurrence of various problem behaviors in differentiating conduct disorder and
ADHD. Their results identified the reported behaviors that had the strongest positive and
negative predictive power for each disorder, thus identifying the symptom behaviors that
could serve as useful inclusionary or exclusionary criteria in differentiating between
ADHD and conduct disorder. The accuracy and utility of the Problem Validation
Screening measure will be evaluated in terms of its positive and negative predictive
power. That is, this study will examine the extent to which Problem Validation Screening
accurately identifies students in need of special services.

**Purpose of Study**

One of the problems associated with the current classification system is an
overreliance on teacher referral. The reliability with which teachers identify students
needing a formal psycho-educational assessment has not been adequately examined.
Another problem is the tendency to rely on a single piece of information that may or may
not have been validated as an accurate screening mechanism to support referral or non-
referral of a child. For example, the Developmental Reading Assessment is a state-
mandated assessment administered annually to first, second, and third grade students in
Louisiana. Low scores are frequently cited as the reason for referral or as support for the
teacher’s concerns once a referral has been made. This study examined the accuracy of
four screening measures by comparing them to a criterion measure (i.e., a combination of
traditional curriculum-based assessment and resistance to intervention). Specifically, this
study examined the accuracy of two screening measures commonly employed by schools
(i.e., teacher referral and the Developmental Reading Assessment) and two additional
screening measures (i.e., subtests from the Comprehensive Inventory of Basic Skills, Revised, 1999 and the Problem Validation Screening procedure developed by Witt, et al., 1999). This process allowed a direct comparison of the accuracy of Problem Validation Screening and teacher referral. Additionally, base rate occurrence of low-achieving students by race was calculated to examine the question of bias for both Problem Validation Screening and teacher referral. Finally, accuracy of Problem Validation Screening and teacher referral were examined in low-achieving classrooms (high base rate of academic problems) and high-achieving classrooms (low base rate of academic problems) to examine the influence of achievement base rates on accuracy of teacher referral and Problem Validation Screening.
Method

Participants

All assessment and intervention activities were conducted in a public school in a rural community in southern Louisiana. All first and second grade classes, including approximately 200 students, participated. This school consisted of 15% Minority and 85% Caucasian students. Approximately 46% of students who attended this school received a free or reduced price lunch.

Measures

This study compared four screening procedures to determine the extent to which each accurately identified children who were strongly suspected of having a disability and who should be referred for a complete evaluation for special education eligibility. Another goal of this study was to provide some validity data on a new measure called Problem Validation Screening. In addition to this measure, the Comprehensive Inventory of Basic Skills, Revised, (Brigance, 1999), the Developmental Reading Assessment, and teacher referral were used as alternative screening methods to identify students who may have potentially exhibited a reading or math deficit that warranted possible assessment and placement in special education. Each measure is described below. The accuracy of these screening measures were compared to a more intensive package of assessment and intervention activities, labeled the “Criterion Assessment” as well as outcomes from the Woodcock-Johnson Psychoeducational Battery-Revised (Woodcock & Johnson, 1989).

Problem Validation Screening (PVS)

The first measure, Problem Validation Screening (PVS), consisted of three separate procedures: classwide academic assessment, performance/skill deficit
assessment, and a brief instructional session. The outcome of this process was that the child either was or was not considered to be a valid referral. Ordinarily a child who was considered to be a valid referral would have been referred for a formal comprehensive special education evaluation.

Classwide Academic Assessment. As the first step of PVS, curriculum-based measurement probes (CBM) were administered to all students enrolled in regular education first and second grade classes at the participating school. CBM measures of reading and math are highly reliable and have demonstrated content, criterion, and construct validity (Marston, 1989). Probes were selected from the Basic Skill Builders series (Beck, Conrad, & Anderson, 1997-1998) in grade-level meetings. The Basic Skill Builders math probe series consists of multiple controlled content worksheets assessing specific skills (e.g., sums to 5, sums to 12, subtraction with answers to 9). Task difficulty is controlled within each worksheet so that a given worksheet samples a single target skill. Worksheets are arranged in increments, progressing from less difficult to more difficult. Similarly, the Basic Skill Builders reading probe series consists of multiple controlled content reading passages of approximately 150 words designed to assess certain reading levels (e.g., 1st grade-2nd semester). Teachers were asked to select the probe for each grade level that best reflected current instruction in their classroom or current placement in their curriculum for reading and math. For math, first grade teachers selected addition with answers to 12. Second grade teachers selected subtraction with answers to 9. For reading, first grade teachers selected a first grade second semester probe. Second grade teachers selected a second grade second semester probe.
**Performance/Skill Deficit Assessment.** The second component of PVS involved a performance/skill deficit assessment. In this phase, participating students were offered a reward from a “treasure chest” to improve their scores on the classwide reading or math probe by one point. The treasure chest was a small transparent box containing several small tangible items. Typical items included in the box were pencils, pens, small toys (e.g., cars, balls, yo-yos), stickers, bracelets, and hair jewelry.

**Brief Instructional Session.** In the final component of PVS, students participated in a three-minute session during which the consultant re-administered the classwide probe, reviewed task directions with the student, briefly modeled correct responding, and allowed the student an opportunity to correct previously made errors. Following this brief session during which task instructions were clarified, the student repeated the task for a score. The purpose of this phase was to identify students who may have earned low scores because they did not understand the task requirements or directions, and who were likely to respond quickly to intervention.

**Psychometric Characteristics of PVS.** Although the purpose of this study was to examine the technical adequacy of PVS, some preliminary data do exist. First, PVS is based upon traditional CBM methods that have been shown in multiple studies to be reliable and valid indicators of learning (Shinn, 1989). Second, previously conducted studies have demonstrated that some combination of the classwide assessment, performance/skill deficit assessment, and brief intervention indicated that approximately two-thirds of referrals to the school committee did not exhibit serious learning problems that warranted comprehensive eligibility evaluation. That is, the problem was sufficiently resolved to the satisfaction of the committee (that included the referring teacher) so that a
referral for evaluation was not made. Because these data are incomplete (i.e., PVS outcomes were not compared to full evaluations for all students referred to the committee) and have not been subjected to peer review, they must be regarded as potentially promising but not conclusive evidence of the psychometric adequacy of PVS.

**Comprehensive Inventory of Basic Skills, Revised (CIBS-R)**

The Addition Facts subtest of the Comprehensive Inventory of Basic Skills, Revised (CIBS-R; Brigance, 1999) was administered to all classes to identify students scoring in the bottom 16% of their classes in math performance. This subtest consists of seven columns of ten horizontally presented addition problems. Answers range from 1 to 19. This test is group-administered and students are allowed three minutes to complete as many problems as possible. This subtest yields a score of total number of problems correct. The Word Recognition Grade Placement Test from the CIBS-R was used to identify students exhibiting potential reading difficulties. This subtest is an individually administered word recognition task of demonstrated reliability and validity (Glascoe, 1999). This subtest consists of approximately 100 words arranged in ten columns of ten words each ranging from pre-primer to eighth grade level of difficulty. On this test, the student is required to read each word aloud to the examiner within three seconds of presentation. Five consecutive correct responses constitutes the basal and ten consecutive incorrect responses formulates the ceiling. This subtest yields a raw score of number of words read correctly. Reliability and validity data for the CIBS-R are reported for most of the subtests (including the Word Recognition Grade Placement Test) and all the composite scores. Reported test-retest and alternate form correlations are generally above .80. The median reported test-retest correlation was .90, and the median reported alternate
form correlation was .94. Additionally, the CIBS-R has been reported to correlate well with multiple achievement tests (Glascoe, 1999).

**Developmental Reading Assessment Scores**

The Developmental Reading Assessment is a statewide measure administered by teachers in the first and second semester to first, second, and third grade students in Louisiana. The Developmental Reading Assessment is a qualitative measure that was designed to indicate whether students were reading approximately on, below, or above grade level. There are no published studies investigating the technical properties of the Developmental Reading Assessment. However, one unpublished study indicated that trained teachers obtained the same result as those obtained by state-identified Developmental Reading Assessment “experts” in 51% of cases (T. Buchanon, personal communication, October, 2000). That is, independent observers agreed on the DRA-indicated reading level of a child in only 51% of cases. Thus, the psychometric properties of the DRA are unsubstantiated by research. The DRA was included in this project because the state mandates its use to monitor reading progress statewide and teachers frequently cite low DRA scores as their reason for referral. On the DRA, students are required to read aloud to the teacher a story that is controlled to reflect grade-level content. The teacher marks words that are read incorrectly on a separate copy of the story. This measure yields a score of percent of words correctly read on an untimed passage. Scores below 89% accuracy are considered unacceptable or below level by the state. For this study, the most recent Developmental Reading Assessment that all teachers had completed was used.
Teacher Referral

Teacher referral is not technically an assessment device. However, it is included here because it is one of the primary “screening devices” used by school-based professionals to determine whether or not a child should be referred for evaluation. Teacher referral was operationalized as a teacher contacting the chairperson of the school committee to request a meeting to discuss his or her academic concerns for a student. The chairperson of the school committee is responsible for documenting all teacher referrals to the school committee. A referral to the committee is the first, required step toward identifying a student as needing special services. During the school committee meeting, the school committee chairperson documents the name of the referring teacher, the name of the referred student, and the reason for referral. The consultant obtained a copy of the committee logs each week to document students referred to the school committee and the reasons for referral throughout the course of this study (i.e., an entire academic year).

The adequacy of teacher referral as a screening source remains unclear. Whereas some studies have indicated that teachers could reliably discriminate learning disabled students from typically-performing students (Gresham et al., 1987; Gresham et al., 1997), other studies have called into question the reliability of teacher referral (Marston et al., 1984; Shinn et al., 1987; Thurlow et al., 1984; Ysseldyke et al., 1981). Gresham et al. (1997) found that although teachers could discriminate typically-performing students from those students exhibiting some sort of learning problem, the teachers could not reliably discriminate between low achieving, learning disabled, and mildly mentally retarded students. Thus, the validity of teacher referral has not been demonstrated.

27

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Criterion Assessment

Overview. The Criterion Assessment consisted of curriculum-based assessment in the problem area (i.e., reading or math), and brief intervention conducted daily for five to nine sessions. These components were selected because the combination of these two practices represent "best practice" (Shinn, 1998) and because together they provide strong evidence about whether the child has an academic deficit, whether that deficit is mild or transient, or whether the deficit is more serious requiring a full special education evaluation for the purposes of determining special education eligibility. One purpose of this study was to determine the validity of PVS in comparison to other methods such as teacher referral and traditional curriculum-based assessment. The "Criterion Assessment" was developed for this study to serve as a standard against which to assess the validity of all measures. For the purpose of this study, the Criterion Assessment was used to determine "true positives" and "true negatives." True positives represent those children who exhibited deficits severe enough to warrant referral for a full and complete evaluation. True negatives were children whose results suggest that they did not exhibit problems severe enough to warrant a full assessment for possible placement in special education. To be considered a true positive, a child must have exhibited a marked deficit in an academic area that was not improved during brief intervention (i.e., 5 sessions for reading and 9 sessions for math). A brief intervention was included in the Criterion Assessment because so called pre-referral interventions are required by law to be used in order to eliminate the possibility that the child's problems are attributable to lack of appropriate instruction.
To determine length of intervention in the Criterion Assessment, all issues of the *Journal of Applied Behavior Analysis* published from 1998 to Fall 2000 were reviewed. All articles pertaining to instructional modifications designed to improve student reading or math performance in the elementary grades, that presented data on a session-by-session basis (as opposed to aggregate data representing blocks of sessions), and used an outcome measure of units of academic work correct during the intervention session or per unit of time (e.g., words read per minute) were selected for review. Each study was required to have implemented a skill-training strategy as opposed to a reinforcement contingency or punishment strategy that would have been likely to produce a more immediate effect. Seven studies met inclusion criteria. Three studies targeted reading skills, whereas four studies targeted math skills. Although the number of studies reviewed was limited, the results were very consistent both within and across studies. An upward trend was defined as three consecutive data points where the first of the three treatment points was greater than the last baseline data point and the second and third treatment data points were equal to or greater than the first treatment point. When a trend was identified, the treatment session number at which the trend began was recorded for all units of analysis in the study (i.e., across children, skill levels). The average of these treatment session numbers was calculated for each study. For math, the average number of treatment sessions required to produce the onset of an upward trend was 1.5 (1-2). For reading, the average number of treatment sessions required to produce the onset of an upward trend was 2.3 (2-3). Thus, it was determined that a minimum of five intervention sessions would be conducted for both reading and math. That is, based on our review of previous studies, we felt confident that if an upward trend were going to occur it would
occur within the first five sessions. Generalization sessions were conducted following each reading intervention session. Generalization sessions were conducted following every third math session in an attempt to limit practice effects occurring as an effect of repeated practice on the generalization probes (i.e., math problems were more frequently repeated due to the limited sample of problems for the age group such as sums to 12). Reading probes contained a greater number of novel measurable units (i.e., novel words appropriate for grade level) so the concern of enhanced reading performance occurring as an effect of daily generalization reading probes was minimized. Thus, nine math intervention sessions were conducted to allow measurement of generalization on three occasions during intervention. Five reading intervention sessions were conducted with daily generalization probes.

**Individual Curriculum-Based Assessment (CBA).** For the individual math assessment, several parallel forms of the math probes were created using the Basic Skill Builders probe series. For the individual reading assessment, the Standard Reading Passages (Children’s Educational Services, Inc., 1987) were used. Two levels of the Standard Reading Passages were used (i.e., Level A and Level B). Level A corresponds to 1st/2nd grade level for reading ability, whereas Level B corresponds to 2nd/3rd grade level for reading ability. Each level contains 18 stories of approximately 200 words each.

**Individual Academic Intervention.** CBM yields data that link directly to instructional planning. Practitioners can use CBM to identify potential causes for insufficient academic performance and alter variables demonstrated to affect student learning (e.g., maximizing academic engaged time, provides multiple opportunities to respond, immediate corrective feedback, sequencing of instruction, pacing, and progress.
monitoring; Daly, Witt, Martens, & Dool, 1997). Additionally, CBM measures are sensitive to instruction (Marston, 1989) and when used formatively can lead to significantly improved academic responding (i.e., improved academic achievement) (Fuchs, 1989). During intervention, students were instructed on their instructional level (Gickling & Armstrong, 1978). Interventions were designed to increase accuracy and build fluency (Daly, Lentz, & Boyer, 1996; White & Haring, 1981). In general, interventions included review of task directions and strategies, modeling of correct performance of the task, guided practice with prompting and immediate and delayed feedback, independent task performance for a score, and performance contingent access to the treasure chest.

Generalization probes were used to measure the student’s progress. Generalization probes were probes of the target skill (the skill that students were expected to perform in the regular classroom) using materials on which the student had not previously been instructed during intervention. For example, a first grade student was instructed using a passage from Level A. The student received instruction on the first 60 words in the passage, whereas the generalization probe was administered using a novel portion of the same story (i.e., the first paragraph following the first 60 words of the story). That is, the generalization probe was conducted using a part of the story to which the student had not been exposed during instruction.

Woodcock-Johnson Psychoeducational Battery-Revised (WJ-R)

The Woodcock-Johnson Psychoeducational Battery, Revised (WJ-R; Woodcock & Johnson, 1989) is an individually administered battery of tests with demonstrated sound psychometric properties. Portions of the WJ-R were administered to students
whose parents gave consent for their children to participate. Specifically, the test was used to obtain the math and reading cluster scores. These included the Letter Word Identification, Passage Comprehension, Calculation, and Applied Problems subtests. Median split-half reliabilities for the WJ-R have been reported as above .86 for subtests and .94 for cluster scores. Additionally, the WJ-R correlates well with multiple intellectual measures and achievement tests (Woodcock & Johnson, 1989).

**Iowa Test of Basic Skills (ITBS)**

The Iowa Test of Basic Skills (ITBS; Hoover, Hieronymus, Frisbie, & Dunbar, 1993) is a group-administered battery of achievement tests with demonstrated psychometric properties. The ITBS is administered annually statewide to second grade students in Louisiana. These data were collected to determine the degree to which classwide probe scores (collected during PVS) correlated with ITBS Math and Reading Total scores.

**Procedure**

Figure 1 (see Appendix A) represents the sequence of experimental procedures. All students in the first and second grade potentially could have participated in the Criterion Assessment. A child was identified to participate in the Criterion Assessment based upon meeting the criteria of any one of the four screening measures (i.e., PVS, Teacher Referral, CIBS-R math and reading screenings, and Developmental Reading Assessment). Table 1 (see Appendix A) depicts the decision criteria for each screening device. Hence, all students enrolled in first and second grade classes were administered a math and reading probe, as well as the CIBS-R Addition Facts subtest and Word Recognition Grade Placement subtest. Additionally, the state mandated Developmental
Reading Assessment data were collected for each student. Students who scored in the bottom 16% of their class and in the frustrational range (Deno & Mirkin, 1977) on the classwide reading and/or math probes were exposed to both the remaining PVS assessment procedures (i.e., performance/skill deficit assessment and the brief instructional session) and the Criterion Assessment. Students who scored in the bottom 16% of their class on the Word Recognition Grade Placement subtest or the Addition Facts subtest of the CIBS-R were administered the Criterion Assessment for their problem area (i.e., math and/or reading). Students who failed to read 89% of the words correctly on the standardized reading passage of the Developmental Reading Assessment were administered the Criterion Assessment for reading. Students referred by their teacher to the school committee for academic concerns were exposed to the Criterion Assessment for math and reading irrespective of the teacher’s reason for referral to allow for a comparison of accuracy of referral reason as well as referral alone.

**Problem Validation Screening**

**Classwide Academic Screening.** Following a meeting with the school administrator, the consultant and administrator attended a grade-level planning meeting for each grade to explain the classwide screening procedures. Teachers were asked to select probe materials during this meeting that best reflected current curriculum content in math and reading. The administration procedures were explained and teachers were provided with scripted instructions for administering the probes. Teachers were told that a consultant would be present in their room during probe administration to ensure that probes were properly administered. The classwide screenings for reading and math problems were conducted using the teacher selected CBM probes, the two subtests of the
CIBS-R, and the state mandated Developmental Reading Assessment. Order of administration of experimenter-administered screening measures (i.e., CBM probes and CIBS-R subtests) was counterbalanced across classes. Classwide probes for PVS in math were administered according to the procedures described by Witt et al. (2000). The CIBS-R addition facts subtest was administered to the class following the procedures described by Brigance (1999). Following the math assessment, classwide reading probes for PVS were administered individually according to the procedures described by Witt et al. (2000). The CIBS-R Word Recognition Grade Placement Test was administered individually according to the procedures described by Brigance (1999). Teachers were asked to submit their most recent Developmental Reading Assessment scores for their classes.

Following administration of all the screening measures in each class, several decision rules were applied to the data to determine which students would proceed for further assessment. Students were ranked according to their scores in descending order on each measure. The lowest 16% of students on the CBM probes and the CIBS-R subtests were identified. Students scoring in the bottom 16% of their class and in the frustrational range (Deno & Mirkin, 1977) on classwide CBM probes were exposed to remaining PVS assessment procedures (i.e., performance/skill deficit assessment and brief instruction session) followed by the Criterion Assessment. Students scoring in the bottom 16% of their class on the CIBS-R subtests, students scoring below 89% accuracy on the Developmental Reading Assessment, and students referred by their teachers, were administered the Criterion Assessment.
Performance/Skill Deficit Assessment. Students whose scores fell within the bottom 16% of the class on classwide probes and scored in the frustrational range according to national norming criteria proceeded to the second phase of PVS, the performance/skill deficit assessment. This assessment was conducted in an administrative office on the school campus three weeks following administration of classwide probes. The performance/skill deficit assessment for math was administered to groups of three to five students simultaneously, whereas performance/skill deficit assessment of reading was administered individually. During the performance/skill deficit assessment, the consultant provided the student with a new copy of the classwide probe on which that student had previously scored in the frustrational range. Students were told that they could earn a reward of their choice from the treasure chest by “beating their previous score.” This score was written in the top left-hand corner of the student’s paper. Students were allowed to briefly sample the items in the treasure chest. The probe was then re-administered using the same directions as were used to administer the classwide probe. A score that increased 20% over baseline (i.e., score on classwide probe) to fall in the instructional range constituted a pure performance deficit. Students whose scores did not improve 20% and did not reach the instructional range for their grade level were considered to exhibit a skill deficit. Students whose scores improved 20% but still fell below the instructional range were considered to exhibit a combined performance/skill deficit. Students whose scores improved to the class median or the instructional range were coded as PVS-negative problems. Students coded as exhibiting a PVS-negative problem participated in the Criterion Assessment. Students exhibiting a skill deficit or a combined performance/skill deficit were coded as PVS-positive problems and proceeded
to the next phase of PVS. Those students found to exhibit a skill or a combined performance/skill deficit (i.e., PVS-positive) participated in the brief instructional session prior to the Criterion Assessment.

**Brief Instructional Session.** A single brief session (i.e., 5 minutes) was conducted individually in an administrative office on campus. The performance/skill deficit assessment and the brief instructional session were not conducted on the same day in an attempt to minimize the effect of immediate repeated exposure to the task. For reading problems, the consultant re-administered the same CBM probe, modeled correct reading to 20% beyond where the student read in one minute, allowed the student to verbally correct previously made errors, and then re-administered the probe for a score. For math problems, the consultant re-administered the same CBM probe, scored the probe, allowed the student to correct errors, and then re-administered the probe for a score. Following this session, students whose scores did not improve to the instructional range were coded as exhibiting a PVS-positive problem. Students whose scores improved to the instructional range were coded as PVS-negative. Figure 2 (see Appendix A) depicts each component of PVS and the criteria for being coded as a PVS-positive problem.

**Teacher Referral**

Currently, the first required step toward being formally evaluated for special education is referral to the school level committee by the teacher or parent. Typically, the teacher contacts the committee chairperson and describes his or her concerns regarding student performance in the classroom. The chairperson then gives the teacher a set of forms to complete and schedules a meeting to discuss the teacher’s concerns with the parents and the school committee. In this study, teacher referral was a screening
mechanism that was compared to the Criterion Assessment to determine accuracy. Thus, all students who were referred by their teachers for an academic problem participated in the Criterion Assessment in both reading and math regardless of referral reason. To monitor teacher referral, the consultant obtained from the school committee chairperson a copy of the school committee logs each week. The consultant reviewed the logs to identify students who were referred by their teachers. Teachers had access to the classwide probe data when making the decision to refer students to the school committee.

**Criterion Assessment**

All students who scored in the bottom 16% of their class and in the frustrational range (Deno & Mirkin, 1977) on classwide probes, students who scored in the bottom 16% of their class on the CIBS-R subtests, students who scored below the acceptable range on the Developmental Reading Assessment (i.e., below 89%), and students who were referred by their teachers participated in the Criterion Assessment. Given a normal distribution of scores, 16% of students' scores were expected to fall at least one standard deviation below the mean. These students were targeted as students most likely to be at risk for or in need of special services. The Criterion Assessment consisted of individual curriculum-based assessment and intervention if the student did not score in the instructional range during the assessment. The curriculum-based assessment conducted as part of the Criterion Assessment differed from the classwide probes in several ways. Specifically, the Criterion Assessment CBA was individually-administered, involved repeated trials using parallel forms of probe materials, involved taking a median score to estimate student performance, comparing student progress during intervention to the average slopes of same grade average-achieving peers, and measuring generalization to
monitor progress and determine intervention effectiveness. The purpose of the Criterion Assessment was to serve as an index of whether or not a student exhibited a problem severe enough to warrant a full and complete psycho-educational evaluation. That is, a child who was classified as having a problem would have been performing poorly and would not have responded to a short-term intervention. The Criterion Assessment was conducted as follows.

**Reading Assessment.** For reading problems, the consultant individually administered the first three probes in the *Standard Reading Passages* on Level A (for first grade students) or Level B (for second grade students) based on the procedures described by Shinn (1989). Students were told that they could earn a reward from the treasure chest by earning a median score of 30 words correct per minute with fewer than 6 errors on the last passage read. First grade students were required to read a minimum median score of 30 words correct per minute on Level A (i.e., instructional range) to be coded as a non-validated problem (Children’s Educational Services, Inc., 1987). Second grade students were required to read a minimum median score of 30 words correct per minute on Level B (i.e., instructional range) to be coded as a non-validated problem (Children’s Educational Services, Inc., 1987). In addition to the fluency requirement, students were required to make fewer than six errors (Deno & Mirkin, 1977) on the final passage read to have been considered in the instructional range and coded as a non-validated problem. For students scoring in the instructional range (i.e., median score greater than 30 words correct per minute), the consultant asked three comprehension questions about the content of the scored passage. First, the consultant asked the student to recall the main idea of the story. Second, the consultant asked the student to identify the main character.
of the story. Third, the consultant asked either when or where the story took place depending on the content. Comprehension scores were calculated as the percentage of questions answered correctly. Students were required to achieve a comprehension score of 100% accuracy to pass the Criterion Assessment of reading. Individual interventions were conducted for those students found to exhibit academic problems during the individual curriculum based assessment (i.e., score in the frustrational range).

**Reading Intervention.** When individual intervention was indicated, instructional level was determined by sampling prerequisite skills (e.g., lower level passages) until the student scored in the instructional range. When determining instructional level, students were offered a reward contingent upon achieving an instructional score on each probe. The consultant conducted all individual assessment and intervention sessions in an administrative office on the school campus. The consultant administered a generalization probe each session to monitor progress of the reading intervention.

Reading interventions were conducted as follows. The first intervention session began with passage 4 on Level A for first grade students and second grade students scoring below 12 words correct per minute on Level B readings (Children’s Educational Services, Inc., 1987). Second grade students scoring greater than 12 words correct per minute on Level B were instructed and monitored on Level B passages (Children’s Educational Services, Inc., 1987). Students proceeded linearly through the passages, being exposed to a new passage each day. Students were exposed to individual reading passages only once during intervention. In intervention, the consultant modeled approximately the first 60 words of the passage. The student then read the passage while the consultant provided immediate corrective feedback (Barbetta, Heward, Bradley, &
Miller) and prompting as needed (i.e., guided practice). The consultant prompted certain strategies (e.g., decoding, blending sounds). Mispronounced words were reviewed with the student following guided practice. The consultant pointed to the missed word and asked the student to read the word aloud. Then the consultant asked the student to repeat the missed word three times in quick succession (Singh, 1987). The student then read the same passage for one minute for a score. The number of words read correctly was recorded on the student’s monitoring chart and graph. If the student increased his/her score above the previous day’s score, he/she was allowed to choose a reward from the treasure chest. Whereas intervention sessions were conducted on a student’s instructional level, generalization probes were administered using grade-level materials. Thus, for first grade students, Level A passages were used for all intervention and generalization probes. For second grade students it was possible that intervention could be conducted using a Level A passage and generalization probes could be conducted using the same number passage from Level B. Therefore, to assess generalization, the student read either a novel section of the same story or the same numbered passage from the criterion level for one minute at the end of each session, with access to the treasure chest contingent on achieving a score higher than the score achieved during the previous generalization probe (e.g., baseline, first generalization probe). Students were allowed to select items from the treasure chest at the end of the session just prior to returning to class.

Students were required to achieve growth on generalization probes commensurate with that of their average-achieving same-grade peers. To determine average growth, individual slopes were calculated for each student in each grade whose spring classwide reading probe score fell in the instructional range. Students who scored in the
frustrational range on the spring probe may have been students who were at risk for or in need of special services and thus, may not have represented “average achievement.” These students were excluded from this analysis. Individual slopes were calculated by subtracting a student’s fall probe score from the spring probe score and dividing that number by the number of intervening weeks (i.e., 12 weeks). The average of these students’ slopes was calculated and was considered to represent an average student’s rate of learning in the target grade and school. Interventions were considered successful if the student reached the instructional level for the target skill and produced a slope from the first to the final generalization data point that was at least that of the average same-grade student (See Tables 1-4 in Appendix C). One exception was made to this rule. One student was only able to participate in three intervention sessions due to his being absent the last week of school. Following the third intervention session, he read 26 words correctly in one minute on the generalization probe and achieved a slope of generalization data of 3.3 words per session. If this student had continued with the same slope over the next two intervention sessions, then the student easily would have met the criterion necessary to be coded as having had a successful reading intervention. Thus, in this one case, this student was coded as having a successful intervention based on slope alone. Interventions continued until the student reached the instructional range for the target skill (i.e., 30 words correct per minute on the generalization probe) and answered the comprehension questions with 100% accuracy or until five generalization data points were obtained.

**Math Assessment.** Math probes were created from the Basic Skill Builders Series. For each skill, five parallel forms of the math probe were created so that the student was
exposed to a given probe only once during the CBA portion of the Criterion Assessment and only once per week while in intervention. The consultant individually administered three parallel forms of the math probe for the criterion skill. Specifically, the consultant administered three parallel forms of an addition probe with answers to 12 to first grade students. The consultant administered three parallel forms of a subtraction probe with answers to 9 to second grade students. Probes were administered according to the directions described by Shinn (1989). Students were told that they would earn a reward from the treasure chest if their middle score occurred in the instructional range. The median score was calculated and recorded. If the median score fell in the instructional range for a student’s grade level, the student was coded as exhibiting a non-validated problem. Students whose median score fell below the instructional range (i.e., fewer than 20 digits correct in two minutes) were exposed to a math intervention.

Math Interventions. Interventions were conducted to increase correct responding on an instructional level skill. To determine instructional level, the consultant administered probes of prerequisite skills until the student met instructional criteria (i.e., 20 digits correct in two minutes; Deno & Mirkin, 1977). On each trial, the student was offered a reward from the treasure chest for meeting instructional criteria. All math interventions involved guided practice followed by two minutes of independent work on an instructional level skill. Specifically, the student was guided to complete the first two rows of the worksheet (i.e., approximately 40 problems, mastery for the skill) as the consultant provided immediate feedback and reviewed decision rules (e.g., “any number plus 0 is that number”) as needed. Additionally, the consultant prompted correct responding as needed. The consultant then covered the practice problems with a piece of
paper and allowed the student to complete as many problems as possible in two minutes. The number of digits correct in two minutes was written on the student’s progress monitoring chart and graph. If the student scored at least one point above the previous day’s score, he or she was allowed to choose an item from the treasure chest. Every third session, the student completed a probe of the criterion skill to assess generalization. For example, intervention sessions may have been conducted on the skill of addition facts to 5 (instructional skill), whereas the generalization probe would have sampled addition facts to 12 (criterion skill). In cases in which the student was being instructed on the criterion-level skill (i.e., at least one of the student’s scores fell in the instructional range on the Criterion Assessment, but the median score did not), a generalization probe was conducted each time the student met instructional criteria (i.e., 20 digits correct) in an intervention session. Intervention sessions continued until the student reached mastery criteria for that skill (i.e., 40 digits correct) or reached instructional criteria for the criterion skill as measured by a generalization probe. Once the student met mastery criteria, the level was increased toward the criterion skill. For first grade students, the progression of skills from least difficult to most difficult was sums to 5, sums to 7, sums to 9, and the criterion skill of sums to 12. For second grade students, the progression of skills from least difficult to most difficult was subtraction with answers to 5, subtraction with answers to 7, and the criterion skill of subtraction with answers to 9. Students were required to achieve growth on generalization probes commensurate with that of their average-achieving same-grade peer. To determine average growth, individual slopes were calculated for each student in each grade who scored in the instructional range on the spring math probe. Students who scored in the frustrational range on the spring probe
may have been students who were at risk for or in need of special services and thus may not have represented “average achievement.” These students were excluded from this analysis. Individual slopes were calculated by subtracting a student’s fall probe score from the spring probe score and dividing that number by the number of intervening weeks (i.e., 12 weeks). The average of these students’ slopes was calculated and was considered to represent an average student’s rate of learning. Interventions were considered successful if the student reached the instructional level for the criterion skill (i.e., 20 digits correct) as measured by generalization probes and produced a slope from the first to the final generalization data point that was at least that of the average same-grade student. Interventions continued until the student reached the instructional range for the criterion skill or three generalization data points were obtained (i.e., 9 intervention sessions).

Following intervention, several criteria were applied to determine whether or not the intervention was successful. Students who failed to achieve a slope (across generalization data points) equal to or greater than that of his or her average achieving classroom peers and continued to score in the frustrational range on the generalization probes were coded as validated problems. Students whose performance on the criterion skill (as measured by generalization probes) fell in the instructional range were coded as non-validated problems.

**Individual Traditional Assessment with the Woodcock-Johnson Psychoeducational Battery-Revised**

All testing sessions were conducted by two doctoral level students in clinical psychology and one doctoral level student in school psychology. Testing sessions took
place in an administrative office on campus. Subtests were administered according to the procedures described by Woodcock and Johnson (1989) and each session required approximately 30 minutes. Each student was administered all the subtests that comprised both the reading and the math cluster scores. These data were evaluated to determine whether or not any of the students met assessment criteria as Learning Disabled (LD) as specified in the Pupil Appraisal Handbook (Louisiana Department of Education, 2000). These criteria specify that students who demonstrate a strength (defined as no more than 1 standard deviation below the mean) and a deficit (defined as at least 1.5 standard deviations below the mean) with at least 1 standard deviation between the strength and deficit may meet criteria as Learning Disabled in the state of Louisiana. The identified areas to be assessed for first and second grade students include basic reading skills, reading comprehension, mathematics calculations, mathematics reasoning, oral expression, listening comprehension, and written expression. The data collected as part of this study were evaluated to determine if participating students met criteria as demonstrating an academic deficit or strength in any of the areas assessed (i.e., reading and math). These results were compared to the results of PVS and the Criterion Assessment.

Iowa Test of Basic Skills

Classroom teachers administered the ITBS according to the procedures described by Hoover, Hieronymus, Frisbie, and Dunbar (1993) during March. These data were collected to determine the degree to which spring classwide CBM scores in reading and math (collected during PVS) correlated with ITBS Math and Reading Total scores, and
the degree to which each of the screening methods identified the same students who performed poorly on the ITBS.

Calculation of Base Rates

Base rates of classwide academic problems were estimated as .07. Given a normal distribution of scores, 7% of the scores were expected to occur at least 1.5 standard deviations below the mean. To qualify as an academic weakness toward the diagnosis of learning disability, a score was required to fall at least 1.5 standard deviations below the mean (Pupil Appraisal Handbook, Louisiana Department of Education). Base rates were calculated to determine the proportion of students scoring in the frustrational range by race and gender for each grade (Shinn et al., 1987). Finally, the proportion of students scoring in the frustrational range on classwide probes in reading and math were calculated separately for reading and math in each classroom to allow a comparison of base rate occurrence of academic problems across classrooms.

Teacher Acceptability Ratings

An off-site consultant distributed the Intervention Rating Profile-20 (Witt & Martens, 1983) to all nine participating teachers during a faculty meeting. The Intervention Rating Profile sampled perceived effectiveness, practicality, ease of implementation, potential risks, and a teacher’s willingness to use recommended interventions. Each item was modified to read “PAM intervention...” as opposed to “this intervention.” PAM stands for Prereferral Assessment Model and is the title of the pilot project operating in this school district. Two items were removed because school administrators felt they were redundant items and desired to minimize the amount of time required for each teacher to complete the scale. Teachers were asked to rate each item on
a likert scale ranging from 1 to 6. Lower scores indicated poor acceptability. Higher scores indicated greater acceptability. Item ratings were summed and then divided by the total number of items (i.e., 18) to yield an average acceptability rating for each teacher. Average ratings lower than 3 suggest that the procedures were unacceptable to the teacher, whereas average ratings greater than 3 suggest that the procedures were generally acceptable (Witt & Martens, 1983). The purpose of the scale was described to teachers as an opportunity for teachers to give the consultants feedback about their opinion of the utility and validity of the PVS process. The school administrator attended the meeting and encouraged teachers to provide honest feedback without fear of repercussion. The consultant asked that teachers not provide any potentially identifying information on the forms to ensure anonymity. The consultant provided an opportunity for teachers to ask questions or express concerns. After designating a person to collect the scales and deliver them to the consultant, the consultant left the room so that teachers could complete the rating scales.

Reliability of Measures and Procedural Integrity

Classwide math probes were administered by the classroom teacher using scripted instructions (adopted from Shinn, 1989). An independent observer (i.e., doctoral student in school psychology, an educational diagnostician, or trained practicum students enrolled in a graduate level course on educational assessment) observed math probe administration in all classrooms and noted the occurrence of each scripted step of administration procedures. The consultant monitored each component and prompted any missed components, noting on the data sheet any particular steps that required prompting. These data yielded a percent integrity score for probe administration (i.e., percent of steps
independently and correctly completed) (see Appendix B). Reading probes were administered by consultants (i.e., a doctoral student in school psychology or a trained practicum student enrolled in a graduate level course in educational assessment). All consultants had taken or were currently enrolled in coursework emphasizing CBM. Probes were administered using scripted instructions (see Appendix B). Consultants were required to demonstrate 100% procedural integrity on three consecutive trials prior to administering probes independently. Additionally, consultants were required to demonstrate 100% inter-scorer agreement on three consecutive trials with the primary consultant prior to administering or scoring probes independently. Approximately 25% of all probes, evenly distributed across grades, were scored by two independent scorers to allow for estimation of inter-scorer reliability. The two independent scorers were blind to the purpose and phases of the study. Specifically, interobserver agreement (IOA) was calculated by dividing the number of agreements over the number of agreements plus disagreements on a case-by-case basis. Average IOA was estimated by calculating the average of the percent agreement scores across all cases. Approximately 33% of all probes conducted during PVS and the Criterion Assessment, evenly distributed across session type (performance/skill deficit assessment, brief instructional session, Criterion Assessment, and intervention) and probe type (intervention and generalization) were scored by two independent scorers to allow for estimation of inter-scorer reliability. Some reading sessions were audio-recorded to allow for subsequent reliability scoring during the PVS and Criterion Assessment.
Results

The Criterion Assessment was used in this study to determine and code for study purposes the “true” state of each child. That is, the Criterion Assessment was used to determine whether or not a child truly exhibited a learning problem. Each screening measure was then compared to the Criterion Assessment to judge the validity of the various screening measures. For all screening measures, participating students were coded twice. In the case of the PVS, for example, students were coded as exhibiting a PVS-positive or PVS-negative problem. Second, students were coded as exhibiting a validated or non-validated problem following the Criterion Assessment. This coding system allowed for a direct comparison of percentage agreement between the four screening methods and the Criterion Assessment.

As described, PVS consisted of three components to which a student may have been linearly exposed depending upon the student’s score on each component of PVS. That is, to receive the second component of PVS, a student must first have scored in the bottom 16% of his or her class on the classwide probes and in the frustrational range for his or her grade level. To be exposed to the third component of PVS, the student must have performed insufficiently during the performance/skill deficit assessment and so on. Specifically, students who scored in the bottom 16% of their class and in the frustrational range on classwide probes, and failed to obtain a score in the instructional range during the performance/skill deficit assessment and the brief instructional session were coded as exhibiting PVS-positive problems. Students who were referred based on alternate referral sources (i.e., teacher referral, DRA, or CIBS-R subtests) were coded as exhibiting PVS-negative problems if they did not meet the criteria just described. Students may, however,
have been referred by multiple sources simultaneously. For example, a student may have been referred by the teacher, met PVS criteria, and scored below the criterion on the DRA. In these cases, students were coded as exhibiting both a PVS-positive problem and a positive referral for the other screening methods.

All students who were identified as exhibiting a potential problem according to one of the screening measures (i.e., teacher referral, classwide probes, CIBS-R subtests, and the Developmental Reading Assessment) were exposed to the Criterion Assessment and were again independently classified as exhibiting a validated or non-validated problem according to the Criterion Assessment. The accuracy of each screening measure was compared to the Criterion Assessment.

Specifically, students who failed to meet instructional criteria on the Criterion Assessment were coded as exhibiting a validated problem. Students who met instructional criteria were coded as exhibiting a non-validated problem. Each of the screening methods used was then compared to the Criterion Assessment outcome to determine the extent to which the two outcomes matched. That is, the accuracy of each screening measure (i.e., PVS, teacher referral, subtests of the CIBS-R, and the Developmental Reading Assessment) was determined by the percentage agreement between the screening measure and the Criterion Assessment across students. The greater the percentage agreement between a screening measure and the Criterion Assessment, the more accurate the measure was judged to be.

Finally, the base rate of math and reading problems were calculated for each classroom as follows. In each classroom the number of students scoring in the frustrational range on classwide CBM probes was divided by the total number of students.
in that class to yield a proportion of students scoring in the frustrational range for both reading and math. High achieving classrooms were defined as classrooms with a base rate lower than .2 (i.e., fewer than 20% of students scored in the frustrational range). Low achieving classrooms were defined as classrooms with a base rate equal to or surpassing .5 (i.e., 50% or more of students scored in the frustrational range). The data collected were subjected to a series of analyses to determine reliability and validity.

Reliability

Prior to beginning the validity phase of this project, classwide probes were administered on two successive trials, one immediately following the other, to estimate temporal stability in 16 first and second grade classrooms, including all first and second grade classrooms at this participating school. Students were ranked in descending order according to their probe scores on the first trial and again, according to their scores on the second trial for reading and math. A Kendall’s tau b correlation was calculated to determine the degree to which student rankings were associated across trials. For math, probes were administered to 273 students in 16 first and second grade classrooms at two schools. Rank-order correlations were significant at the p<.01 level for all 16 classrooms. Correlations ranged from .577 (for one classroom) to .930 (for another classroom) with a mean across classes of .726. A Pearson r correlation was calculated between the raw scores on trial one and the raw scores on trial two for math yielding an r value of .946 (significant at the p<.01 level). A within subjects analysis of variance was conducted to determine the degree to which scores changed (improved) on the second trial. Results indicated that scores did significantly improve on the second trial for both grades. Specifically, the mean math score on the first trial was 15.97 for first grade and 40.13 for
second grade and the mean math score on the second trial was 20.16 for first grade and 42.46 for second grade. The within subjects analysis of variance yielded an F-value of 80.61 for first grade (p<.000) and 9.78 (p<.002) for second grade. Thus, whereas scores across trials improved significantly, classwide rankings remained roughly the same.

For reading, probes were administered to 162 students in all 9 first and second grade classrooms at the participating school. Kendall’s tau b correlation was calculated to determine the degree to which student rankings remained the same across trials. All rank order correlations were significant at the p<.000 level. Rank order correlations ranged from .727 to .904 with a mean correlation of .845. A Pearson r correlation was calculated to determine the degree of association between the two sets of raw scores yielding an r-value of .974 (significant at the p<.01 level). A within subjects analysis of variance was performed to determine the degree to which students’ scores changed between trials. Results indicated that scores improved significantly on the second trial for both grades. The mean score on the first reading trial was 34.4 for first grade and 63.78 for second grade. The mean reading score on the second trial was 45.66 for first grade and 80.88 for second grade. The within subjects analysis of variance yielded an F-value of 260.79 (p<.000) for first grade and 235.02 (p<.000) for second grade. Thus, whereas scores across trials improved significantly, classwide rankings remained roughly the same.

These results support using a single opportunity math and reading probe to identify students for further assessment.

Procedural reliability of classwide math probe administration and CIBS-R math subtest administration was estimated by calculating the percentage of correctly completed steps observed by independent, trained observers. Procedural reliability for the
administration of classwide math CBM probes and CIBS-R math subtest were both 100% across all classrooms. Interscorer reliability was calculated as the percentage agreement for individual items on math and reading probes, and was calculated by counting the number of agreements and number of disagreements for all attempted items, dividing the number of agreements by the number of agreements plus disagreements, and multiplying the resulting number by 100%. Approximately 25% of the schoolwide probes were scored for reliability. For reading, average interscorer agreement for schoolwide probes was 97.5% (84-100%). For math, mean interscorer agreement for schoolwide probes was 90.4% (80-100%). Interscorer reliability was calculated for at least 33% of all sessions in each phase, counterbalanced across grade, session type (e.g., intervention and generalization sessions during the intervention phase), skill area, and student in the intervention phase (i.e., 33% of each student’s intervention sessions were scored for reliability). Mean interscorer agreement for Performance/Skill Deficit Assessment sessions was 99.4% (95-100%). Mean interscorer agreement for Brief Instructional Sessions was 99.8% (98-100%). Mean interscorer agreement for Criterion Assessment Sessions was 99.5% (94-100%). On average, independent scorers agreed on the accuracy of 99.5% (91-100%) of student responses during intervention sessions, conducted as part of the Criterion Assessment.

**Overall Descriptive Findings**

**Demographics**

Table 1 in Appendix D depicts the number of students identified as exhibiting a valid problem at each phase of the analysis and by each screening method. A total of 182 students participated in the screening activities. Because each student participated in
screening activities in two academic areas (i.e., math and reading), the total number of screened cases was 364. Of these students, 101 cases were identified for further assessment based upon meeting at least one of the screening criteria. The classwide probes identified 55 cases for further PVS. Teachers referred 31 students. The CIBS-R subtests identified 64 cases for further assessment. The Developmental Reading Assessment scores identified 17 cases for further assessment in reading. Each of the 55 cases identified for further PVS participated in a skill/performance deficit assessment. Forty of these cases participated in a Brief Instructional session in reading or math. Twenty-two of these 40 cases met criteria as exhibiting a PVS-positive problem following the Brief Instructional session, and thus were coded as PVS-positive. All of the 101 identified cases participated in a Brief Instructional session and Criterion Assessment. The Brief Instructional session did not sufficiently resolve the problem in 34 of the 101 cases. Seventeen cases met criteria as exhibiting a validated problem following the Criterion Assessment.

Validity

Accuracy of Each Screening Method

Table 2 in Appendix D depicts the predictive power estimates for each screening method using the Criterion Assessment as the standard for comparison. Tables 3 and 4 in Appendix D depict the predictive power estimates for each screening method using the ITBS and WJ-R as the standards for comparison. PVS and the CIBS-R obtained the strongest estimates of predictive power. Overall, PVS obtained a much higher percentage agreement with the Criterion Assessment (87% agreement compared to 66% for Teacher Referral, 51% for CIBS-R, and 68% for DRA). The CIBS-R subtests were more sensitive
(.94) than PVS (.76) but were less specific (i.e., produced a high number of false positive errors) (.43) than PVS (.89). The CIBS-R was the most sensitive. PVS was the most specific. PVS also obtained the highest degree of positive predictive power. The CIBS-R obtained the highest negative predictive power (.97), compared to .95 for PVS, and .89 for both Teacher Referral and DRA. Each screening method is described in detail in the following sections.¹

**Problem Validation Screening.** The percentage of cases in which a child was coded PVS-positive for a problem and was also found to be positive for a problem according to the Criterion Assessment was calculated. PVS corresponded with the Criterion Assessment in 88 of 101 cases or 87% of cases. Sensitivity was calculated as the proportion of true positives (i.e., percent of PVS-positive problems that were found to be validated problems on the Criterion Assessment). Specificity was calculated as the proportion of true negatives (i.e., the percent of PVS-negative problems that were found to be non-validated problems on the Criterion Assessment). Positive predictive power was calculated to determine the utility of a PVS-positive finding toward identifying students who may be at risk for or in need of special services. Positive predictive power was calculated as the probability the Criterion Assessment indicated a validated problem given a PVS-positive finding. Negative predictive power was calculated to determine the utility of the finding of a PVS-negative finding as an exclusionary criterion (indicating which students are not at risk for or in need of special services). Negative predictive power was calculated as the probability that the Criterion Assessment indicated a non-validated problem given PVS-negative problem using those participants for whom these data were available (i.e., those referred by teacher, the CIBS-R sub-tests, or the DRA that
were found to be PVS-negative upon initial problem validation assessment). Using the Criterion Assessment as the standard for comparison, the sensitivity of PVS was .76. The specificity of PVS was .89. The positive predictive power of PVS was .59. The negative predictive power of PVS was .95. Phi was .596 (p<.000). For math only, the sensitivity of PVS was 1. Specificity was .85. Positive Predictive Power of PVS was .46. Negative Predictive Power was 1. Phi was .626 (p<.000). For reading only, sensitivity of PVS was .64. Specificity was .95. Positive predictive power was .78. Negative predictive power was .9. Phi was .629 (p<.000).

Sensitivity, specificity, positive predictive power, and negative predictive power were also calculated using ITBS scores as the criterion measure. The degree to which PVS accurately identified those students who would score more than one standard deviation below the mean on the ITBS reading or math tests was calculated. Sensitivity was 1. Specificity of PVS was .99. Positive predictive power was .67. Negative predictive power was 1. Phi was .811, p< .000. For math only, sensitivity was .5. Specificity was .91. Positive predictive power was .33. Negative predictive power was .95. Phi was .342 (p<.094). For reading cases only, sensitivity was 1. Specificity was .9. Positive predictive power was .33. Negative predictive power was 1. Phi was .548 (p<.102). A stepwise discriminant function analysis was conducted to determine the degree to which ITBS scores could predict a PVS-positive problem in reading or math. Total Math ITBS scores entered the analysis and accounted for 13% of the variance. ITBS math scores correctly classified students as having a valid problem in 3 of 3 or 100% of cross-validated cases. ITBS math scores correctly classified students as PVS-negative in 68 of 78 or 87% of cases. Thus, ITBS scores correctly classified cases as PVS-positive or negative in
approximately 88% of cross-validated cases. Kappa was computed at .335 (p<.000) for this analysis.

A Discriminant Function Analysis was performed to determine the degree to which WJ-R math and reading cluster scores could predict a PVS-positive or negative problem. Reading and math WJ-R cluster scores accounted for 17% of the variance and correctly classified students as having a valid problem in 1 of 5 cross-validated cases (20%) and as not having a valid problem in 24 of 25 or 96% of cases. Hence, WJ-R scores correctly predicted PVS-positive or negative problems in approximately 83% of cross-validated cases. Kappa was computed as .211 (p<.190) for this analysis. Predictive power estimates were calculated for PVS using WJ-R scores as the standard for comparison to determine the degree of match between students coded as PVS-positive and students scoring at least 1½ standard deviations below the mean on the WJ-R (criteria for demonstrated weakness). Sensitivity of PVS for math was .56. Specificity was .74. Positive predictive power was .38. Negative predictive power was .62. Phi was .265 (p<.093). Sensitivity of PVS for reading was .38. Specificity was .88. Positive predictive power was .5. Negative predictive power was .82. Phi was .289 (p<.092). Predictive power estimates were calculated to determine the degree of agreement between PVS-positive and negative problems and students scoring 1.5 standard deviations below the mean on either the reading or math cluster scores of the WJ-R. Sensitivity of PVS was .58. Specificity was .77. Positive predictive power was .44. Negative predictive power was .86. Phi was .322 (p<.021).

Several outcome variables may help quantify the utility of PVS. Specifically, the degree to which PVS identified students who were referred by the school-level committee
for formal special education evaluation, students who were classified as exceptional, and students who were retained in their grade at the end of the year may reflect the utility of PVS. Using referral for special education evaluation as the outcome standard, sensitivity, specificity, and positive and negative predictive power were calculated. PVS sensitivity in identifying students who would be referred for special education evaluation was .4. Specificity was .83. Positive predictive power was .21. Negative predictive power was .92. Phi was calculated as .173, p<.090, but this value should be interpreted with caution since samples may have been dependent (the degree to which the committee’s decision to refer for evaluation was influenced by PVS data). For math only, sensitivity of PVS was .5. Specificity was .77. Positive predictive power was .15. Negative predictive power was .95. Phi was .167 (p=.229). For reading only, sensitivity was .33. Specificity was .84. Positive predictive power was .22. Negative predictive power was .9. Phi was .144 (p=.312). Using qualification for special education services as the outcome or criterion standard, sensitivity, specificity, and positive and negative predictive power were calculated for problem validation. Problem validation sensitivity in identifying students who would qualify for special education was .8. Specificity was 1.0. Positive predictive power was 1.0. Negative predictive power was .83. Phi was calculated to estimate effect size. Phi was .82, p=.010, but this value must be interpreted with caution as the sample size was small and the samples may have been dependent (to the degree that problem validation data may have influenced the assessment team’s decision to qualify a child). For math only, sensitivity, specificity, positive predictive power, and negative predictive power were 1. Phi was 1 (p=.046). For reading only, sensitivity was .67. Specificity was 1. Positive predictive power was 1. Negative predictive power was .75. Phi was .707
Sensitivity, specificity, positive predictive power, and negative predictive power of PVS was calculated using retention as the criterion. Sensitivity was .39. Specificity was .95. Positive predictive power was .63. Negative predictive power was .88. Phi was calculated to estimate effect size. Phi was .419, p<.000. For math only, sensitivity of PVS was .5. Specificity of PVS was .82. Positive predictive power was .6. Negative predictive power was .75. Phi was .334 (p<.017). For reading cases only, sensitivity of PVS was .42. Specificity was .83. Positive predictive power was .71. Negative predictive power was .58. Phi was .265 (p<.069).

Which component of PVS best predicted a validated problem? A stepwise discriminant analysis was performed to determine which component of PVS accounted for the most variance toward identifying students who were found to have a validated problem on the Criterion Assessment. Specifically, the classwide probe score, the performance/skill deficit assessment score, and the post-instruction score in the brief instructional session were entered into the analysis to determine which was the most powerful predictor of a valid problem on the Criterion Assessment. This analysis is limited by the fact that the cases included in the analysis were students who met initial criteria as potentially exhibiting a PVS-positive problem and thus, were exposed to the remaining components of PVS. Fifty-five cases contained scores on all three variables and were included in the analysis. That is, 55 students participated in the PVS. The reinforcement probe score was the only variable to enter the analysis. This variable accounted for 18% of the variance and on average correctly classified 76.4% of cases. Specifically, of the 16 Criterion Assessment validated problems, 6 were correctly classified based only on the reinforcement probe score. Of the 39 Criterion Assessment...
non-validated problems, 36 were correctly classified based on the reinforcement probe score alone. Kappa was computed for this analysis at .342 (p<.007). Twenty-nine cases contained scores on all three variables for math. For math, the reinforcement probe score and post brief instruction score were retained in the analysis. Together, these variables accounted for 72% of the variance and correctly classified 13 of the 13 valid math problems and 14 of the 16 non-valid math problems or 93% of cases using a jackknife procedure. Kappa was computed at .931 (p<.000). For reading, twenty-six cases contained scores on all three variables. Brief instruction scores entered the analysis and accounted for 53% of the variance. Brief instruction correctly classified approximately 85% of cases using a jackknife procedure. Kappa was computed at .675. (p<.001).

Specifically, Brief Instruction scores correctly classified students as exhibiting a valid problem in 8 of 10 cases, and correctly classified students as not exhibiting a valid problem in 14 of 16 cases.

**Teacher Referral.** The percentage agreement between teacher referred problems and Criterion Assessment validated problems was calculated. Teachers correctly identified students as having an academic problem in 66% of cases. This value was compared to the percent agreement between PVS-positive problems and Criterion Assessment validated problems. The sensitivity of teacher referral was .46. The specificity of teacher referral was .69. The positive predictive power of teacher referral was .19. The negative predictive power of teacher referral was .89. Phi was .108 (p<.292). For math problems only, sensitivity of teacher referral was .4. Specificity was .7. Positive predictive power was .13. Negative predictive power was .91. Phi was .061 (p<.662). Teachers were given credit as having referred the student irrespective of
whether or not they identified the correct problem area. For example, if a teacher referred a student for reading problems, but the student was found to exhibit a valid math problem, teachers were given credit as having correctly referred a student with a valid problem. For reading problems only, sensitivity was .5. Specificity was .68. Positive predictive power was .25. Negative predictive power was .82. Phi was .140 (p<.347).

Using referral for special education evaluation as the outcome or criterion variable, sensitivity of teacher referral was .9. Specificity was .74. Positive predictive power was .41. Negative predictive power was .98. Phi was calculated at .421, p<.000, but this value should be interpreted with caution because samples were not independent and sample size was smaller than is recommended. For math only, sensitivity of teacher referral was 1. Specificity was .74. Positive predictive power was .25. Negative predictive power was 1. Phi was .431 (p<.002). For reading problems only, sensitivity was .83. Specificity was .72. Positive predictive power was .31. Negative predictive power was .97. Phi was .391 (p<.009). Using qualification for special education as the outcome variable, sensitivity of teacher referral was 1.0. Specificity was .2. Positive predictive power was .56. Negative predictive power was .5. Phi was calculated at .333, p<.292. This analysis could not be conducted separately for math as no students exhibiting a math problem were evaluated for special education placement who were not also referred by their teachers. For reading problems only, sensitivity was 1. Specificity was .33. Positive predictive power was .6. Negative predictive power was 1. Phi was .447 (p<.273).

Using poor performance on the ITBS as the criterion (i.e., one standard deviation below the mean on ITBS reading or math), sensitivity of teacher referral was .33. Specificity was .94. Positive predictive power was .17. Negative predictive power was...
.97. Phi was .194 (p<.081). Using retention as the criterion, sensitivity of teacher referral was .32. Specificity was .97. Positive predictive power was .67. Negative predictive power was .87. Phi was .396 (p<.000). Predictive power estimates were calculated for teacher referral using WJ-R scores as the standard for comparison to determine the degree of match between students referred by their teachers and students scoring at least 1 ½ standard deviations below the mean on the WJ-R (criteria for demonstrated weakness). Sensitivity of teacher referral using WJ-R scores as the criterion was .42. Specificity was .85. Positive predictive power was .45. Negative predictive power was .83. Phi was .271 (p<.053).

A discriminant function analysis was performed to determine the degree to which ITBS scores could predict whether or not a teacher would refer a student. ITBS scores accounted for 6% of the variance and correctly classified students as being referred by their teachers in 4 of 6 or 67% of cross-validated cases and not being referred by their teachers in 53 of 75 or 71% of cases. Kappa was computed at .207 (p<.007) for this analysis. A discriminant function analysis was performed to determine the degree to which WJ-R scores could predict whether or not a teacher would refer a student. WJ-R scores accounted for 8% of the variance and correctly classified students as being referred by their teachers in 6 of 11 or 55% of cross-validated cases and not being referred by their teachers in 27 of 40 or 68% of cross-validated cases. Kappa was computed at .303 (p<.016) for this analysis.

CIBS-R Referral. Overall, the CIBS-R subtests correctly identified students as exhibiting an academic problem in 51% of cases. Using the Criterion Assessment as the standard for comparison, the sensitivity of the CIBS-R subtests was .94. The specificity
of the CIBS-R subtests was .43. The positive predictive power of the CIBS-R subtests was .25. The negative predictive power of the CIBS-R subtests was .97. Phi was .287 (p<.004). For math only, sensitivity of the CIBS-R math subtest was .83. Specificity was .46. Positive predictive power was .17. Negative predictive power was .95. Phi was .187 (p<.176). For reading problems only, sensitivity was 1. Specificity was .39. Positive predictive power was .32. Negative predictive power was 1. Phi was .357 (p<.012). Using referral for special education evaluation as the outcome variable, CIBS referral sensitivity was .6. Specificity was .38. Positive predictive power was .1. Negative predictive power was .89. Phi was calculated at -.010, p<.920. For math cases only, sensitivity was .5. Specificity was .42. Positive predictive power was .07. Negative predictive power was .9. Phi was -.045 (p<.746). For reading cases only, sensitivity was .67. Specificity was .3. Positive predictive power was .12. Negative predictive power was .87. Phi was -.022 (p<.877). Using qualification for special education as the outcome variable, CIBS referral sensitivity was 1.0. Specificity was .8. Positive predictive power was .83. Negative predictive power was 1.0. Phi was calculated at .816, p<.010, but this value should be interpreted with caution due to extremely limited sample size (i.e., n=10). For math cases only, sensitivity was 1. Specificity was 1. Positive predictive power was 1. Negative predictive power was 1. Phi was calculated at 1.0 (p<.046), but this value should be interpreted with caution due to extremely limited sample size (i.e., n=4). For reading cases only, sensitivity was 1. Specificity was .67. Positive predictive power was .75. Negative predictive power was 1. Phi was .707 (p<.083), but this value should be interpreted with caution due to extremely limited sample size (n=6).
Using poor performance on the ITBS as the criterion for comparison (i.e., one standard deviation below the mean on ITBS reading or math), sensitivity of the CIBS-R was 1. Specificity was .72. Positive predictive power was .12. Negative predictive power was 1. Phi was .294 (p<.008) for this analysis. Using retention as the criterion for comparison, sensitivity of the CIBS-R was .84. Specificity was .82. Positive predictive power was .49. Negative predictive power was .96. Phi was .546 (p<.000). Predictive power estimates were calculated for the CIBS-R using WJ-R scores as the standard for comparison to determine the degree of match between students referred as a result of low CIBS-R scores and students scoring at least 1 ½ standard deviations below the mean on the WJ-R (criteria for demonstrated weakness). Sensitivity of the CIBS-R was .83. Specificity was .36. Positive predictive power was .29. Negative predictive power was .88. Phi was .176 (p<.209).

A discriminant function analysis was performed to determine the degree to which ITBS scores would predict CIBS-R referral. ITBS scores accounted for 25% of the variance and correctly classified students as being referred by the CIBS-R in 22 of 25 or 88% of cross-validated cases and as not being referred by the CIBS-R in 40 of 56 or 71% of cross-validated cases. Kappa was .520 (p<.000). A discriminant function analysis was performed to determine the degree to which WJ-R scores would predict CIBS-R referral. WJ-R scores accounted for 4% of the variance and correctly classified students as being referred by the CIBS-R in 17 of 35 or 49% of cross-validated cases and as not being referred by the CIBS-R in 10 of 16 or 63% of cross-validated cases. Kappa was computed at .092 (p<.461) for this analysis.
Developmental Reading Assessment. Overall, the DRA correctly identified students as exhibiting a reading problem in 68% of cases. Using the Criterion Assessment as the standard for comparison, sensitivity of the DRA was .67. Specificity of the DRA was .69. Positive predictive power of the DRA was .35. Negative predictive power of the DRA was .89. Phi was .292 (p<.053).

Using referral for special education as the outcome variable, DRA sensitivity was .2. Specificity was .59. Positive predictive power was .06. Negative predictive power was .85. Phi was calculated as -.137, p<.363, but this value should be interpreted with caution due to small sample size and possibly dependent samples. Using qualification for special education as the outcome variable, sensitivity of DRA was .5. Specificity was 1.0. Positive predictive power was 1.0. Negative predictive power was .75. Phi was .612 (p<.171), but this value should be interpreted with caution due to the extremely limited sample size (n=5) and the possibility dependent samples.

Using poor performance on the ITBS (i.e., one standard deviation below the mean on ITBS reading or math) as the criterion for comparison, sensitivity of DRA was .67. Specificity was .92. Positive predictive power was .25. Negative predictive power was .99. Phi was .373 (p<.001) for this analysis. Using retention as the standard for comparison, sensitivity of DRA was .35. Specificity was .96. Positive predictive power was .65. Negative predictive power was .88. Phi was .407 (p<.000). Predictive power estimates were calculated for the DRA using WJ-R scores as the standard for comparison to determine the degree of agreement between students referred as a result of substandard performance on the DRA and students scoring at least 1½ standard deviations below the mean on the WJ-R (criteria for demonstrated weakness). Sensitivity of the
DRA was .5. Specificity was .85. Positive predictive power was .5. Negative predictive power was .85. Phi was .346 (p<.013).

Several outcome variables may further quantify the utility of DRA as a screening measure. A discriminant function analysis was performed to determine the degree to which WJ-R scores would predict DRA referral. WJ-R scores accounted for approximately 9% of the variance in whether or not a student would be referred based upon low DRA scores. Reading and math WJ-R cluster scores correctly classified students as being referred by DRA in 7 of 12 cross-validated cases (58%) and as not having a valid problem in 27 of 39 or 69% of cases. Kappa was .265 (p<.040). A discriminant function analysis was performed to determine the degree to which ITBS scores would predict DRA referral. ITBS scores accounted for 15% of the variance and correctly classified students as being referred by DRA in 6 of 8 or 75% of cross-validated cases and as not being referred by DRA in 57 of 73 or 78% of cross-validated cases. Kappa was .225 (p<.008).

Accuracy of Screening Methods Compared to Base Rate Accuracy

Problem Validation Screening. Base rates of academic problems were calculated two ways. First, base rates were calculated using general population estimates (i.e., 1.5 standard deviations below the mean of a normally distributed population) and second, based upon actual performance of students in the sample (i.e., percentage of students scoring in the frustrational range on classwide CBM probes). Base rate occurrence of academic problems was estimated as 7%. Given a normal distribution of estimates of academic performance, approximately 7% of students would be expected to score at least 1.5 standard deviations below the mean. Using the procedures described by Meehl and
Rosen (1955), the accuracy of PVS was compared to base rate accuracy by comparing the percentage agreement between PVS and Criterion Assessment outcomes to the accuracy rate obtained if no students were assumed to exhibit a validated academic problem. If no students were assumed to exhibit a validated problem, the error rate would be 7%. Thus, to conclude that PVS procedure has utility, it was required to achieve an accuracy rate of greater than 93% (or an error rate lower than 7%). A finding of statistical association between validated problems obtained in PVS and expected referrals (calculated as base rate occurrence of 7%) would be interpreted as supporting the accuracy of PVS. A Chi-Square analysis was performed to determine the frequency of PVS referrals versus referrals expected by base rate alone. Chi-square analysis yielded 22 observed referrals compared to 7.1 expected referrals and 79 observed non-referrals compared to 93.9 expected non-referrals yielding a chi-square value of 33.901 (significant at p<.000). That is, observed and expected values were significantly different (i.e., more students were referred than would be expected by base rate alone). Additionally, proportion of correct decisions made for individual cases was calculated. Proportion of correct decisions based on PVS was compared to proportion of correct decisions based on base rate alone (i.e., assuming no students needed special services) to determine which method was most accurate. A chi-square analysis was performed. Observed correct referrals for problem validation were 88 compared to 93.9 correct referrals expected by base rate. Observed incorrect referrals for problem validation were 13 compared to expected incorrect referrals of 7.1. Chi-square was calculated at 5.35 (p<.02).

Teacher Referral. Teacher referral was required to exceed 93% accuracy to be considered useful in identifying students at risk for or in need of special services.
Statistically significant differences between expected referral accuracy (based on base rate) and observed teacher referral accuracy may indicate inaccuracy in teacher referral. A chi-square analysis was performed. Observed correct teacher referrals were 64 compared to expected correct referrals of 89.3. Observed incorrect referrals were 32 compared to expected incorrect referrals of 6.7. Chi-square was calculated at 102.26 (p<.000) indicating a significant difference between expected and observed correct teacher referrals (i.e., teachers were significantly less accurate than base rate accuracy).

**Stability of Referral Sources Under Varying Classroom Conditions**

The effect of base rates on the accuracy of the different screening methods was determined by comparing the accuracy (i.e., percentage agreement with the Criterion Assessment) of each referral method in classrooms where base rates of students scoring in the frustrational range on classwide probes were high (i.e., above 50% occurrence) and classrooms where base rates were low (i.e., below 20% occurrence). These findings are depicted in Tables 5 and 6 in Appendix D. Overall, PVS obtained the highest predictive power estimates across both settings, and particularly when base rates were low (i.e., high-achieving classrooms). The predictive power estimates are described in detail in the following section.

**Problem Validation Screening.** In classrooms where base rate occurrence of academic problems was high (i.e., greater than or equal to .5), the following accuracy estimates were obtained. The sensitivity of PVS was .75. The specificity of PVS was .88. The positive predictive power of PVS was .69. The negative predictive power of PVS was .91. Phi was .197 (p<.187). In classrooms where base rates were greater than or equal to .5 and for math cases only, sensitivity of PVS was 1. Specificity was .89. Positive
predictive power was .67. Negative predictive power was 1. Phi was .772 (p<.000). In classrooms where base rates were greater than or equal to .5 and for reading cases only, sensitivity was .63. Specificity was .87. Positive predictive power was .71. Negative predictive power was .81. Phi was .509 (p<.015). In classrooms where the base rate occurrence of academic problems was low (i.e., lower than .2), the following accuracy estimates were obtained. The sensitivity of PVS was .67. The specificity of PVS was 1.0. The positive predictive power of PVS was 1.0. The negative predictive power of PVS was .97. These analyses are not reported separately for math problems, because the Criterion Assessment did not validate any math problems in high-achieving classrooms.

For reading cases in classrooms where the base rate was lower than .2, sensitivity was .5. Specificity was 1. Positive predictive power was 1. Negative predictive power was .95. Phi was .689 (p<.002).

**Teacher Referral.** In classrooms where base rate occurrence of academic problems was high (i.e., greater than or equal to .5), the following accuracy estimates were obtained. The sensitivity of teacher referral was .55. The specificity of teacher referral was .68. The positive predictive power of teacher referral was .35. The negative predictive power of teacher referral was .82. Phi was .197 (p<.187). For math cases only and in classrooms where the base rate was equal to or greater than .5, sensitivity was .5. Specificity was .68. Positive predictive power was .25. Negative predictive power was .87. Phi was .147 (p<.482). In classrooms where base rates were equal to or greater than .5 and for reading problems only, sensitivity was .57. Specificity was .67. Positive predictive power was .44. Negative predictive power was .77. Phi was .226 (p<.290). In classrooms where the base rate occurrence of academic problems was low (i.e., lower
than .2), the following accuracy estimates were obtained. The sensitivity of teacher referral was 0. The specificity of teacher referral was .67. The positive predictive power of teacher referral was 0. The negative predictive power was .95. These analyses are not reported separately for math problems, because the Criterion Assessment did not validate any math problems in high-achieving classrooms. For reading cases in classrooms where the base rate was lower than .2, sensitivity was 0. Specificity was .67. Positive predictive power was 0. Negative predictive power was .92. Phi was -.160 (p<.485).

**CIBS-R Referral.** In classrooms where base rate occurrence of academic problems was high (i.e., greater than .5), the sensitivity of the CIBS-R subtests was 1.0. The specificity of the CIBS-R subtests was .53. The positive predictive power of the CIBS-R subtests was .43. The negative predictive power of the CIBS-R subtests was 1.0. Phi was .476 (.001). For math cases only and in classrooms where the base rate was greater than or equal to .5, sensitivity was 1. Specificity was .53. Positive predictive power was .31. Negative predictive power was 1. Phi was .402 (p<.054). For reading cases in classrooms where base rates equaled or exceeded .5, sensitivity was 1. Specificity was .53. Positive predictive power was .53. Negative predictive power was 1. Phi was .533 (p<.011). In classrooms where base rate occurrence of academic problems was low (i.e., lower than .2), the sensitivity of the CIBS-R subtests was 1.0. The specificity of the CIBS-R subtests was .32. The positive predictive power of the CIBS-R subtests was .13. The negative predictive power of the CIBS-R subtests was 1.0. These analyses are not reported separately for math problems, because the Criterion Assessment did not validate any math problems in high-achieving classrooms. For reading cases in classrooms where the
base rate was lower than .2, sensitivity was 1. Specificity was .32. Positive predictive power was .13. Negative predictive power was 1. Phi was .205 (p<.347).

**Developmental Reading Assessment.** In classrooms where base rate occurrence of academic problems was high (i.e., greater than or equal to .5), the sensitivity of the Developmental Reading Assessment was .67. Specificity of the Developmental Reading Assessment was .69. The positive predictive power of the Developmental Reading Assessment was .5. The negative predictive power of the Developmental Reading Assessment was .82. Phi was .338 (p<.141). In classrooms where base rate occurrence of academic problems was low (i.e., lower than .2), sensitivity of the Developmental Reading Assessment was 1.0. Specificity of the Developmental Reading Assessment was .67. Positive predictive power was .25. Negative predictive power was 1. Phi was .408 (p<.068).

**The Relationship of Race to Teacher Referral and Problem Validation Screening Accuracy.** The probability of referral, the accuracy of referral, and the effect of race were calculated for classrooms exhibiting high base rates of academic problems for Teacher Referral and PVS and these results are described in the following sections. Table 7 in Appendix D depicts PVS and Teacher Referral Accuracy by race overall, in low-achieving (i.e., high base rate of academic problems), and high-achieving (i.e., low base rate of academic problems) classrooms. Table 8 in Appendix D depicts the types of errors (e.g., false positive) made by race for both PVS and Teacher Referral.

In classrooms where base rate occurrence of academic problems was high (i.e., greater than or equal to .5) white students were correctly identified as having or not having valid problems using the PVS in 29 of 33 or 88% of cases. Minority students were
correctly identified in 10 of 13 or 77% of cases. For math cases only, white students were correctly identified as having or not having valid problems using the PVS in 14 of 16 or 88% of cases. Minority students were correctly identified in 7 of 7 or 100% of cases. For reading cases only, white students were correctly identified using PVS in 15 of 17 or 88% of cases. Minority students were correctly identified in 3 of 6 or 50% of cases. White students were correctly identified as having or not having valid problems using teacher referral in 18 of 28 or 64% of cases, whereas minority students were correctly identified in 7 of 11 or 64% of cases. For math cases only, white students were correctly identified as having or not having valid problems using teacher referral in 10 of 16 or 63% of cases. Minority students were correctly identified in 5 of 7 or 71% of cases. For reading cases only, white students were correctly identified using teacher referral in 11 of 16 or 69% of cases. Minority students were correctly identified using teacher referral in 3 of 6 or 50% of cases. In classrooms where base rate occurrence of academic problems was low (i.e., lower than .2), white students were correctly identified as having or not having valid problems using PVS in 23 of 24 or 96% of cases. Minority students were correctly identified as having or not having valid problems in 9 of 9 or 100% of cases. For math cases only, white students were correctly identified in 9 of 9 or 100% of cases. Minority students were correctly identified in 2 of 2 or 100% of cases. For reading cases only, white students were correctly identified using PVS in 13 of 14 or 93% of cases. Minority students were correctly identified in 7 of 7 or 100% of cases using PVS. White students were correctly identified using teacher referral in 12 of 22 or 55% of cases, whereas minority students were correctly identified using teacher referral in 7 of 8 or 88% of cases. For math cases only, white students were correctly identified in 5 of 9 or 56% of
cases. Minority students were identified in 2 of 2 or 100% of cases. For reading cases only, white students were correctly identified using teacher referral in 7 of 13 or 54% of cases. Minority students were correctly identified in 5 of 6 or 83% of cases.

Further Analysis of the Utility of PVS and Teacher Referral as Screening Devices

According to Meehl and Rosen (1955), for a positive result to be more likely true than false, the ratio of the positive to negative base rates in the examined population must exceed the ratio of the false positive rate to the valid positive rate. We examined these ratios for both high-achieving (i.e., low base rates) and low-achieving (i.e., high base rates) classrooms. High-achieving classrooms were defined as classes in which fewer than 20% of the students scored in the frustrational range on classwide probes in a given subject area (i.e., math or reading). Thus, the ratio of positive to negative base rates in high-achieving classrooms was .20/.80 or .25. The ratio of false positives to valid positives in high-achieving classrooms was 0/.67 or 0. Because .25 was greater than 0, a positive finding on PVS (i.e., validated problem) is more likely true than false. This formula could not be computed for teacher referral, because teachers did not correctly identify a single student as exhibiting a learning problem in a high-achieving classroom.

Low-achieving classrooms were defined as classes in which more than 50% of students scored in the frustrational range on classwide probes. The lowest base rate in this group was .53. Thus, the smallest proportion of positive to negative base rates in low-achieving classrooms was .53/.47 or 1.13. This proportion was calculated for each base rate in the range considered low-achieving (i.e., .55, .67, .76, .78, .79, .82) in an attempt to identify the point at which the predictor variables were no longer an efficient estimate. In all cases, this proportion was greater than the ratio of false positives to valid positives for
PVS, .12/.75 or .16. Similarly, the ratio of false positives to valid positives for Teacher Referral was .32/.55 or .58, a value that was exceeded by the proportion of positive to negative base rates in all low-achieving classrooms.

This formula was adapted to further consider the value of a negative finding on the predictor variables. Thus, for a negative finding to be more likely true than false, the ratio of negative to positive base rates in the examined population must exceed the ratio of false negative rates to valid negative rates. The proportion of negative to positive base rates in high-achieving classrooms was .80/.20 or 4. The ratio of false negative rates to valid negative rates for PVS was .33/1 or .33. Thus, a negative finding on PVS was more likely true than false in classrooms with low base rates of valid problems, because 4 exceeded .33. For Teacher Referral, the ratio of false negative rates to valid negative rates was 1/.67 or 1.49. Thus, a non-referral was more likely correct than incorrect in classrooms with few valid academic problems. In low-achieving classrooms where many students exhibited valid academic problems, the ratios of negative to positive base rates were compared to the ratio of false negative rates to valid negative rates for both PVS and Teacher Referral. For PVS, at each base rate (i.e., .53 to .82), a negative finding on PVS was more likely true than false because the ratio of false negative rates to valid negative rates (.25/.88 or .28) was always exceeded by the ratio of negative to positive base rates in low-achieving classrooms. In contrast, when teachers did not refer a student in a low-achieving classroom, they were more likely to be incorrect than correct in doing so. That is, the ratio of false negatives to valid negatives (.45/.68 or .66) was exceeded by the base rates ratio in classrooms where 53 to 67% of students exhibited valid academic problems. In classrooms where more than 67% of students exhibited valid academic problems, the
ratio of false negative rates to valid negative rates was no longer exceeded by the population base rates ratio.

Further, Meehl and Rosen (1955) stated that when the base rate of valid negatives (students not exhibiting an academic problem) is greater than the number of valid positives (students do exhibit an academic problem), the ratio of the valid positive rates (i.e., sensitivity) to the valid positive rates plus the false positive rate should exceed the base rate of valid negatives in order for the test to have utility. These numbers were examined for both PVS and Teacher Referral. For PVS, the base rate of valid negatives (i.e., .83) was less than the ratio of sensitivity (.76) to sensitivity plus false positive rate (.11) or .87. Thus, using PVS maximizes the number of “hits” in the population being examined in this study. For Teacher Referral, the ratio of sensitivity (.46) to sensitivity plus false positive rate (1.92) or .19 was not greater than the base rate of valid negatives or .83. Thus, the use of Teacher Referral does not maximize the number of correct findings in the population examined in this study.

Race Effects

In general, teachers accurately referred 61% of Caucasian students compared to 78% accurate referral of minority students. PVS accurately referred 86% of Caucasian students compared to 90% accurate referral of minority students as measured by the Criterion Assessment. The accuracy of teacher referral, PVS, and the other screening measures was compared for Caucasian students and minority students. Table 9 in Appendix D depicts the probability of teacher referral, PVS validated problems, and Criterion Assessment validated problems overall and by race. The probability of teacher referral was .16. The probability of teacher referral for a Caucasian student was .16. The
probability of teacher referral for a minority student was .21. The probability of referral based on PVS was .11. The probability of Problem Validation referral for a Caucasian student was .10. The probability of Problem Validation referral for a minority student was .21. The probability of obtaining a validated problem on the Criterion Assessment was .17. The probability of a Criterion Assessment validated problem for a Caucasian student was .17. The probability of a Criterion Assessment validated problem for a minority student was .17.

The base rate occurrence of academic problems by race was used to make the following comparisons. Base rate of minority students exhibiting academic problems was calculated in two ways. Expected referral of minority students was calculated in two ways (i.e., based upon a normal distribution of scores in a population, and based upon the actual sample). Chi-square analyses were performed to determine whether or not there was a significant difference between expected referral of students by race (using both base rate calculations) and observed (actual) referral rate by race for problem validation and teacher referral. A finding of no statistical difference between expected teacher referral of minority students (using both base rate calculations) and observed teacher referral of minority students would be interpreted to support teacher referral as an unbiased identification method. Similarly, a finding of no statistical difference between expected minority student validated problems (using both base rate calculations) and observed minority validated problems on the PVS would be interpreted to support PVS as an unbiased identification method. Conversely, disproportionate referral of race may indicate bias. First, the 16% estimate of students scoring at least 1 standard deviation below the mean was multiplied by the base rate of minority students in the population.
(15) to yield a percent of minority students expected to fall at least 1 standard deviation below the mean. Thus, 2% of validated problems and teacher referrals should be minority students given an approximately normal distribution of minority student performance. Teachers would be expected to refer approximately 1 minority student out of their 32 referrals. Teachers actually referred 26 Caucasian students and 6 minority students. Chi-square was 645.16, p<.000. Thus, teachers referred significantly more minority students than would be expected by base rate alone. The same analysis was performed for PVS. Given the base rate calculation described above, it was expected that 1 validated problem would involve a minority student, whereas the remaining 21 validated problems would involve Caucasian students. PVS actually yielded 17 validated cases for Caucasian students and 6 validated cases for minority students. Chi-Square was 255.08, p<.000. Thus, PVS resulted in significantly greater number of validated problems for minority students than would be expected by base rate alone.

Second, the percent of minority students scoring in the bottom 16% of their class was calculated. Specifically, 18 minority students scored in the bottom 16% of their classes on classwide probes. This number (18) was divided by the total number of students who scored in the bottom 16% of classes (32) indicating that 56% of the bottom 16% of scorers on classwide probes were minority students. A chi-square analysis was performed. Teachers were expected to refer 18 minority cases and 14 Caucasian students based on these base rates. Teachers actually referred 6 minority cases and 26 Caucasian cases. Chi-square was 18.29, p<000. Teachers referred significantly fewer minority cases and significantly more Caucasian students than would be expected by base rate alone. Given the base rate occurrence of minority students scoring in the bottom 16% on
of the 22 validated problems (56%) would be expected to be minority students. Of the 22 validated problems, 17 cases involved Caucasian and 6 involved minority students. Chi-square was 7.51, p<.006. PVS validated significantly more Caucasian students than would be expected and fewer minority students than would be expected by base rate alone.

Referral Source Measures Compared to Other Measures of Student Achievement

Table 10 in Appendix D depicts the correlations of reading and math measures administered during this project. A Pearson correlation coefficient was calculated to determine the degree of association between the classwide reading probes administered to first and second-grade students and the CIBS-R Word Recognition Grade Placement subtest. Classwide reading probes correlated with the CIBS-R subtest at .902 (N=174; p<.000). A Pearson correlation was calculated to determine the degree of association between the classwide addition probe administered to first grade students in two schools and the CIBS-R addition subtest yielding a correlation coefficient of .768 (N=168; p<.000). A Pearson correlation was performed to determine the degree of association between the classwide subtraction probe administered to second grade students in two schools and the CIBS-R addition subtest yielding a correlation coefficient of .612 (N=147; p<.000). The CIBS-R reading scores correlated with ITBS total reading standard scores at .733 (N=77; p<.000). The CIBS-R math scores correlated with ITBS total math standard scores at .600 (N=76; p<.000). A Spearman’s rho correlation was performed to determine the degree of association between whether or not a student scored in the unacceptable range on the DRA and the classwide reading probes administered to students who participated in experimental procedures, yielding a non-significant
correlation coefficient of .066 (N=44; p<.669). A Spearman’s rho correlation was
performed to determine the degree of association between whether or not a student scored
in the unacceptable range on the DRA and the classwide reading probes administered
schoolwide, yielding a significant correlation coefficient of .293 (N=175; p<.000). ITBS
scores correlated with whether or not a student scored in the unacceptable range on the
DRA at .370 (N=81; p<.001). CIBS-R reading subtest scores correlated with whether or
not a student scored in the unacceptable range on the DRA at .271 (N=172; p<.000).

The ITBS total reading standard scores correlated with classwide reading probe
scores at .701 (p<.000; N=77). The ITBS total math standard scores correlated with
classwide math probes consisting of subtraction and representing current placement in the
curriculum at .421 (p<.000; N=76). ITBS total math standard scores correlated with
classwide math probes consisting of addition problems and representing mastery level
material for that grade at .545 (p<.000; N=78). ITBS total language standard scores
correlated with classwide reading probe scores at .420 (p<.000 N=77).

Additionally, ITBS data were collected for all second-grade students and a
regression analysis was performed to determine the degree to which probe scores
predicted performance on the ITBS. A regression analysis was performed to determine
the degree to which probe scores could accurately predict ITBS scores in each academic
area (i.e., reading and math). A linear relationship was found between probe scores in
reading and ITBS reading scores (see Figure 1 in Appendix D). This linear relationship
accounted for 49% of the variance in ITBS reading scores. The correlation between
reading probe scores and ITBS reading scores was F=72.61, p<.000. Two types of
classwide math probes were examined. Specifically, students were administered probes
sampling a mastery level math skill (i.e., addition) and an instructional level skill (i.e., subtraction). Linear relationships were found for both probe types with ITBS total math standard scores (See Figures 2-3 in Appendix D). Approximately 30% of the variance in the ITBS math scores was accounted for by its linear relationship with classwide math probes of the mastery level skill (i.e., addition). The correlation between classwide addition probes and ITBS math scores was $F=31.28$, $p<.000$. Approximately 18% of the variance in the ITBS math scores was accounted for by its linear relationship with classwide math probes of the instructional level skill (i.e., subtraction). The correlation between classwide subtraction probes and ITBS math scores was $F=15.94$, $p<.000$.

**Teacher Acceptability of Problem Validation Screening**

Scales with more than two missing or non-completed items were excluded. Five teachers returned the modified Intervention Rating Profile (Witt & Martens, 1983) as requested at the participating school. Ratings greater than 3 indicate acceptability. Ratings lower than 3 indicate unacceptability of the process. The average of the teachers' ratings was 4.12 (2.7-5.11; SD=.95). Because so few scales were returned at the participating school, the same scale was administered at the two remaining pilot schools. At these two schools, 22 teachers were asked to complete the scale if they had referred students to the school level committee and thus, had been exposed to the project. Thirteen teachers returned the scales as requested. One scale was excluded because the second page was not completed. Including these scales in the analysis, the average rating across teachers was 3.9 (2.7-5.11; SD=.64), with only one rating lower than 3.

**Endnote**

1 Three identified special education students participated in screening activities. These students’ data are included in all the analyses reported below, except for teacher
referral. Teachers did not have an opportunity to refer these students for special education assessment because students were already identified. Thus, these students were not included in teacher referral calculations.
Discussion

Each year, significant numbers of students are identified as exceptional learners and placed in special programs. Some researchers have suggested that students who exhibit mild learning problems are not receiving accurate and complete assessments prior to being labeled and placed in special programs (Maheady et al., 1984). Given the outcome data associated with placement in special education, it has been suggested that students exhibiting mild problems that respond to intervention in the regular setting, can achieve greater habilitation in regular education (Fuchs & Fuchs, 1998). The current classification system relies heavily on teacher referral. That is, the teacher is typically the person who identifies who will and won’t be discussed at the school level committee meeting where it is determined who will and will not be evaluated. Once children are referred to the school level meeting, their chances of being evaluated are great (approximately 90%) and when evaluated, their chances of qualifying are also great (approximately 70%). These estimates have been highly stable over the preceding decade (Algozzine et al., 1983; Ysseldyke et al., 1997), despite multiple programs being implemented across the country to provide pre-referral services. Multiple studies have indicated that pre-referral interventions are not routinely implemented as planned and agreed to by the committee (Fuchs & Fuchs, 1987; Gresham, 1991a; Happe, 1982). Failure to properly implement powerful intervention strategies prior to referral for evaluation creates a bias against students without adequate preparation or support from other sources (e.g., parents who read). Complicating data interpretation and possibly compromising validity of this actuarial process is an overreliance on teacher self-report. Typically, the type of intervention recommended depends entirely on the teacher’s
description or definition of the problem. The teacher does not typically receive hands-on help or training in the classroom to properly implement recommended interventions. Finally, the teacher's perception also formulates the basis for deciding whether or not an intervention was successful. The reliability with which teachers make the series of judgments necessary to result in an accurate referral for evaluation have not been systematically examined. We attempted to do so in this study, and our findings indicate that teacher referral may not be a reliable identification method.

The unsubstantiated reliability and validity of the actuarial process of classification combined with the poor outcomes associated with placement in special education, have led some to question the adequacy and appropriateness of current screening methods, such as teacher referral, and eligibility criteria, such as discrepancy formulas (Ysseldyke & Thurlow, 1984). CBM has been proposed as an alternative screening mechanism (Elliot & Fuchs, 1997; Marston, Tindal, & Deno, 1984). This study described a model based largely upon the principles of curriculum-based measurement called Problem Validation Screening. This study then compared the predictive accuracy of Problem Validation Screening to other commonly used screening mechanisms with the most important of those being teacher referral.

First, this study compared PVS, teacher referral, and two additional screening methods to a “Criterion Assessment” to determine which method most accurately identified students exhibiting serious learning problems. PVS produced slightly (i.e., CIBS-R) or much better (i.e., teacher referral and DRA) predictive power estimates when compared to the other screening methods included in this study. Next, for second grade students, the screening methods were compared to ITBS scores to determine which
method most accurately identified those students who would perform poorly on the ITBS. Again, PVS accurately identified students who would and would not score greater than one standard deviation below the mean on the ITBS math or reading subtests. As a screening method, negative predictive power was considered to be of great importance. PVS achieved excellent negative predictive power while achieving reasonable levels of positive predictive power. Thus, most of the errors were false positive errors (i.e., student did not truly exhibit a serious learning problem according to the Criterion Assessment). Moreover, PVS was also time-efficient, requiring approximately 45 minutes to screen an entire class, and approximately 5 minutes for each follow up assessment that was necessary (e.g., performance/skill deficit assessment and brief instructional session). In comparison, DRA and the CIBS-R reading screenings required approximately twice the amount of time as PVS. Interestingly, PVS predictive accuracy estimates were much lower when referral for evaluation, qualification for special education, and retention were each employed as the criterion for comparison, a finding that coincides with the findings of Macmillan (1998). That is, the school committee has been shown to arrive at conclusions that are not supported by the data collected.

Second, this study compared the stability of predictive power estimates of all the screening methods across varying environmental conditions. Specifically, we wondered if accuracy of the screening methods might differ as a function of the number of high or low-achieving students in the same classroom. PVS was accurate in 85% of cases in low-achieving classrooms where many students exhibited academic problems. PVS was accurate in 97% of cases in high-achieving classrooms where few students exhibited academic problems. Teacher Referral was 64% accurate and 65% accurate respectively.
Accuracy estimates were slightly better in high-achieving classrooms compared to low-achieving classrooms for PVS, possibly an effect of restricted variation in scores in the low-achieving classrooms. Additionally, PVS yielded higher predictive power estimates (i.e., combined positive and negative predictive power) than all other screening methods in both high-achieving and low-achieving classrooms. Most notably, dramatic differences occurred in the accuracy of teacher referral in high-achieving versus low-achieving classrooms. Further, Meehl and Rosen’s (1955) formula was applied to examine the utility of PVS and teacher referral. Teacher referral was not found to maximize the number of “hits” in the sample, whereas PVS was. Thus, PVS was supported as a useful screening device, whereas teacher referral was not.

Because the initial criterion for PVS was in part influenced by the performance of classmates (i.e., bottom 16% and in the frustrational range), agreement was examined using the class as the level of analysis (i.e., bottom 16% of the class and in the frustrational range) and then using the grade as the unit of analysis (i.e., bottom 16% of the grade and in the frustrational range). That is, to what degree would the sample vary if students were selected based on scoring in the bottom 16% of their grade and scoring in the frustrational range? Overall, in 86% of cases, the same students were identified. The following differences were observed. When students were selected using the entire grade as the unit of analysis, one additional false negative error in math and one additional false negative error in reading were obtained. One false negative error in reading was prevented. An additional 11 false positive errors were observed for math and 3 additional false positive errors were observed for reading. Hence, the most accurate level of analysis appears to be the student’s classroom.
Third, the experimental measures were compared to other measures of student achievement. Specifically, classwide probe scores were compared to scores obtained by the same students on the ITBS. Classwide CBM probes in reading correlated very well with ITBS total reading scores. Classwide CBM probes in math correlated moderately well with ITBS total math scores. Classwide CBM probes in math and reading correlated with the CIBS-R subtests in math and reading respectively. Interestingly, CBM reading probes and DRA scores (state-mandated reading assessment scores) were negligibly correlated.

Fourth, teacher and administrator acceptability was estimated. Treatment acceptability was used to indicate the social validity of the process. Unfortunately, these findings must be interpreted with caution, as not all teachers returned the rating scales as requested. It is possible that the sample of scales that were returned represented a biased sample. Nonetheless, the scales that were returned indicated that teachers and administrators found the PVS process to be generally acceptable. Further, several teachers indicated that they particularly liked the classwide CBM probes that produced student rankings in math and reading for their classrooms.

Practitioners frequently rely on teacher report to make diagnostic decisions (Marston et al., 1984). Teacher perception is important. For example, it has been demonstrated that teachers may provide inferior instruction to students whom they perceive as low-achieving (Brophy & Good, 1970). Yet, the accuracy with which teachers make the series of judgments that result in referral, and more times than not classification, does not appear promising. Previous studies have demonstrated that teacher referral may be subject to bias (Marston et al., 1984; Shinn et al., 1987) and that
teachers do not have a consistent, clear definition, either across or within states, of the behaviors that constitute a serious learning problem or "learning disability" (Thurlow et al., 1984). The results of this study are similar to other studies that have called into question the reliability and validity of teacher referral as an identification method (Marston et al., 1984; Shinn et al., 1987). The results of this study indicate that teachers were not able to accurately identify students who did (i.e., poor positive predictive power) and did not (i.e., poor negative predictive power) exhibit serious learning deficits.

This study is also similar to previous studies that have demonstrated the utility of CBM methods as screening devices (Deno, Marston, & Tindal, 1986; Elliot & Fuchs, 1997; Fuchs & Fuchs, 1998; Marston, Mirkin, et al., 1984; Marston, Tindal et al., 1984). Finally, this study extends the literature by employing resistance to intervention as the outcome standard (Fuchs & Fuchs, 1998), using classwide measures of math and reading performance to allow for comparison to class mean, class trend, and national standards of performance, and judging the validity of PVS and other screening devices employed in this study by the proportion of correct decisions made and in light of base rates. PVS provides a standardized, objective assessment of student performance in the context of that student's environment. PVS proactively screens all the students in a school to identify students who may be at risk for serious learning problems, and results in more accurate and time-efficient identification of students exhibiting potentially serious learning problems than other commonly used identification methods.

Although these findings must be considered tentative pending replication, the potential of PVS appears promising. However, several limitations of this study are worth noting. First, predictive power estimates are only generalizable to the conditions under
which data were collected. The school system where these data were collected was being monitored by the Office of Civil Rights due to overrepresentation of minority students in special education classrooms. Thus, race data may have been influenced to the extent that teachers were more hesitant to refer minority students to the school committee. The year in which this study was conducted was also the third year of participation for the participating school in a pilot project implementing PVS procedures. The degree to which teachers were familiar with PVS and the resistance to intervention outcome standard may have influenced the probability and accuracy with which teachers at this participating school identified students for possible formal psycho-educational evaluations. Similarly, the availability of classwide CBM probe scores to teachers may have influenced the probability and accuracy of teacher referral. A significant limitation of this study is that not all students in the population were administered the Criterion Assessment. Because of this limitation, these accuracy estimates (particularly specificity and negative predictive power) are incomplete and must be interpreted with caution. To attempt to ameliorate this potentially biasing effect, the same predictive power estimates were calculated and discriminant function analyses were performed using ITBS scores as the criterion for comparison (i.e., all students in the second grade were administered both PVS and the ITBS). Although power was sufficient for most of the analyses reported in this paper (Stevens, 1996), future studies should attempt to replicate these procedures with more participants. Finally, because there can be no incontrovertible index of whether or not a student has a true learning problem, we drew from the literature to create and then employed a combination of curriculum-based assessment and resistance to intervention (i.e., the Criterion Assessment). Nonetheless, the Criterion Assessment may not represent
a true measure of a student's performance capacity. The alternative measures (e.g., referral for evaluation, qualification for special education services, and retention) presented additional problems, most importantly their vulnerability to potentially biasing factors (Macmillan, 1998; Shinn, Tindal, & Spira, 1987).

Currently, the emphasis of school-based psychological services appears to be on assessing the child to determine whether or not the child meets criteria to receive special education services. If the student does not qualify under a particular category (approximately 20% of those formally evaluated according to some estimates), the child and teacher are not likely to receive any help at all. On the other hand, if the child does qualify, the research indicates that the child is not likely to receive special help as a function of the diagnostic category to which the child has been assigned anyway. Perhaps for these reasons, teachers have rated services provided by school psychologists as generally unhelpful and ineffective. In fact, teacher ratings of school psychologists' effectiveness decrease as teachers experience greater contact with school psychologists (Severson, Pickett, & Hettrick, 1985). The value of school psychological services should meet the standard of improved student outcomes as a result of instructional manipulations recommended by the school psychologist and informed by assessment data (Fuchs & Fuchs, 1998; Messick, 1995). On average, in school districts across the United States, 90% of students referred for special education services participate in a formal evaluation at an estimated cost of $3000 per child (Ysseldyke, Algozzine, & Epps, 1983). Because formal psycho-educational assessments typically do not provide data useful for instructional programming or intervention planning (Gresham & Witt, 1997), the result of all the testing is determining only whether or not a child "qualifies." Some estimates
indicate that it is not uncommon for more than 100 professional hours to be devoted to making an eligibility or placement decision (not including assessment time) (Poland, Thurlow, Ysseldyke, & Mirkin, 1982). In the end, teachers receive very little, if any, data to inform instructional planning necessary to remediate the problem that prompted the initial referral. Thus, it is not surprising that teachers rate psychological services as ineffective when after multiple hours of testing and many weeks of waiting, they receive no help concerning specific strategies to attempt to resolve the problem, regardless of where the student is placed.

Further, Macmann, Barnett, Lombard, Kocher, and Sharpe (1989) state that to truly evaluate the validity of a measure, a more in-depth analysis of each decision-making step is required. That is, practitioners must not just be concerned with the psychometric properties of an instrument used for decision-making purposes, but also with the accuracy and validity of the decision-making process itself, ultimately judged by the outcomes. This position is philosophically congruent with Messick's (1995) view that construct validity is at least in part determined by the validity of and effects of both the intended and unintended consequences of the assessment process. Just as the child's performance occurs within the context of a classroom, school, and community, so does the decision-making process. The degree to which unmeasured variables unique to a decision-making team affect decision outcomes has not been routinely operationalized or measured. For example, Ysseldyke, Pianta, Christenson, Wang, and Algozzine (1983) asked teachers to delineate the causes of the problems that prompted their initial referral of students for evaluation. Teachers overwhelmingly attributed the causes of students' problems to factors within the student (i.e., ability) or the student's home situation. Additionally,
teachers indicated that they hoped that the outcome of the referral would be assessment and ultimately placement, as opposed to obtaining strategies to help the child remain in their classroom. In fairness, it is possible that teachers have formed these expectations based upon their experiences with school-based professionals who have failed to provide data that are useful to teachers and helpful to students and indeed, some data support this interpretation (Severson et al., 1985). Regardless, the expectations and philosophies of the decision-making team may introduce error and bias to the decision-making process.

The current actuarial process for determining eligibility and placement appears to fall far short of the standard described by Messick (1995) and Macmann et al. (1989), beginning with inadequate identification methods, inaccurate and uninformative diagnostic measures and relatedly weak diagnostic categories, and concluding with nonspecific, weak, and frequently absent treatment effects.

Adequate, objective screening methods are needed to identify students exhibiting serious learning problems (Adelman, 1982). Proactive screening may decrease potential biases (e.g., teacher tolerance) that have been found to occur in the referral process (Marston et al., 1984; Shinn et al., 1987) and provide an opportunity to target intervention services to prevent future academic deficits (Good & Kaminski, 1996). PVS may contribute to a resistance to intervention model of classification (Gresham, 1991b; Fuchs & Fuchs, 1998) or a more comprehensive model of problem-solving or functional academic assessment (Lentz & Shapiro, 1986) with elementary school students.

Common problems associated with screening measures are reliance on a single response opportunity and measurement of student performance without the benefit of knowing the “context” of performance. That is, basing a screening decision on an
individual student’s score without respect to the scores of that student’s classmates may result in error. Adelman (1982) has argued that accurate screening methods require a stronger match between the predictor (screening measure) and criterion (expected performance conditions). Doesn’t this match allude to the unique differences in classroom contexts, teacher expectations, etc. that seem to render comparison to national normative data inefficient and inaccurate? The use of CBM to identify students is an important first step, but a subsequent “validation” component is needed. Specifically, time-series data, allowing for a comparison of both performance level and slope to the student’s classmates’ average level of performance and slope, are needed to determine whether or not a problem exists that warrants assessment for eligibility. Future studies are needed to determine degree of slope required to conclude that an intervention has been successful and that the student is not “resistant to intervention.”

The need to measure both performance level and trend is illustrated in the following example. One of the false negative errors obtained in this study involved a student who scored one point above the bottom 16% for his class and in the frustrational range. In this class, approximately 82% of the class scored in the frustrational range. Thus, there was a restricted range of scores in the class, potentially affecting the degree of accuracy of prediction based on the mean. One could make the argument that given the poor performance of the entire class, the absence of proper instruction could not be ruled out as a confounding variable, and further, that all students in the class may have been in need of special help. In this study, a 16% decision rule was employed irrespective of the number of students in the class who may have performed in the frustrational range. The use of this decision rule resulted in the false negative error described in the above
example. Thus, in cases like the one just described, practitioners may prefer to conduct a classwide intervention in the problem area and graph the time-series data of all students. Those students whose slopes are discrepant from the slopes of their peers, may be students in need of further assessment. A decision rule for the degree of inter-class slope discrepancy required should be empirically derived in future studies.

Finally, these assessment decisions should be judged by their treatment validity (Macmann et al., 1989). Treatment validity is defined as the extent to which the assessment process results in improved student outcomes (i.e., improved habilitation; Hayes, Nelson, & Jarrett, 1987). As Macmann et al. (1989) have pointed out, the utility of CBM methods for screening purposes merits further assessment, and their value will depend upon both the adequacy of their technical properties as well as the social and educational consequences of their use for children, just as Messick (1995) recommended providing direct evidence of applied relevance and utility when establishing construct validity.
References


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.


Appendix A: Sequence of Experimental Procedures and Screening Criteria

Classwide probe screening for reading and math

Teacher referral

CIBS-R screening for reading and math

DRA for reading

Skill/performance deficit assessment

Brief instructional session

Criterion Assessment

Figure A.1. Flowchart of Experimental Procedures.
<table>
<thead>
<tr>
<th>Screening Measure</th>
<th>Criteria</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classwide CBM Probes</td>
<td>Scored in bottom 16% of class and in the frustrational range</td>
<td>Participate in Problem Validation Screening and Criterion Assessment for problem area</td>
</tr>
<tr>
<td>Teacher Referral</td>
<td>Teacher refers to committee</td>
<td>Participate in Criterion Assessment for Reading and Math</td>
</tr>
<tr>
<td>CIBS-R</td>
<td>Scored in bottom 16% of class</td>
<td>Participate in Criterion Assessment in problem area</td>
</tr>
<tr>
<td>Developmental Reading</td>
<td>Scored below 89% accuracy on grade-level story</td>
<td>Participate in Criterion Assessment for reading</td>
</tr>
<tr>
<td>Assessment</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Classwide CBM probes administered to all first and second grade students:
Scored in bottom 16% of class and in the frustrational range?

Yes

No

Code as PVS-negative and close case

Performance/Skill Deficit Assessment:
Scored in the frustrational range?

Yes

No

Code as PVS-negative

Criterion Assessment

Brief Instructional Session:
Scored in the frustrational range?

Yes

No

Code as PVS-positive

Code as PVS-negative

Criterion Assessment

Figure A.2. Components of Problem Validation Screening and Decision Rules.
Appendix B: Instructions for Administration of CBM Probes

Instructions for Administration of Math Probes

1. Pass out papers face-down instructing students not to turn them over until you tell them to do so.
2. “Please write your first and last name on the back of your paper. Please write your teacher’s name next to your name.” Pause briefly to allow students to write their names.
3. “This is a math worksheet. All of the problems are ________ (math or subtraction). When I say ‘start,’ turn them over and begin answering the problems. Start on the first problem on the left on the top row (point). Work across and then go to the next row. Are there any questions?”
5. “Stop. Put your pencils down and hold your paper up in the air so we can pick it up.”

Be sure to monitor student performance to ensure that students work the problems in rows and do not skip around or answer only the easy problems.
(adopted from Shinn, 1989)

Integrity Checklist for Probe Administration

_____ Read scripted instructions to the class.
_____ Checked for student understanding.
_____ Allowed two minutes to answer problems.
_____ Monitored during testing, walking around the room, scanning, etc. Encouraged students only to “do their best work” in response to questions during testing.
_____ Collected papers.

Teacher Name: _______________________________________________

School: ________________________ Grade: ___________________

Administrator/Observer: _______________________________________

Total Percent Integrity: ________________
Instructions for Administration of Reading Probes

1. “We’re reading with everyone in your class/school today.” Give the student the story. Write the student’s name, teacher, and grade on the probe sheet.

2. “When I say ‘start,’ begin reading aloud at the top of the page. Read across the page (demonstrate by pointing). Try to read each word. If you come to a word that you do not know, I will tell it to you. The goal is for you to read as many words as you can correctly in one minute. Be sure to do your best reading. Do you have any questions?”

3. “Start.” Allow the student to read for one minute. Follow along on your copy, marking the words that are read incorrectly.

4. At the end of one minute, “Stop reading.” Draw a vertical line after the last word read. Thank the student for reading.

5. Count number of words read correctly and number of errors.
Appendix C: Average Slopes for Reading and Math

Table C.1

Average Slopes for Reading from Fall to Spring Probe Administration

<table>
<thead>
<tr>
<th></th>
<th>All Students</th>
<th>Students Not Referred for Reading*</th>
<th>Students Not Referred for Reading or Math**</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>First Grade</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M (range) =</td>
<td>0.51 (-2.67-8.08)</td>
<td>0.68 (-2.67-8.08)</td>
<td>0.70 (-1.67-2.67)</td>
</tr>
<tr>
<td>SD =</td>
<td>1.37</td>
<td>1.43</td>
<td>1.1</td>
</tr>
<tr>
<td>n =</td>
<td>84</td>
<td>63</td>
<td>49</td>
</tr>
<tr>
<td><strong>Second Grade</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M (range) =</td>
<td>1.12 (-3.75-5.5)</td>
<td>1.22 (-2.58-5.5)</td>
<td>1.3 (-2.58-5.5)</td>
</tr>
<tr>
<td>SD =</td>
<td>1.5</td>
<td>1.45</td>
<td>1.5</td>
</tr>
<tr>
<td>n =</td>
<td>71</td>
<td>53</td>
<td>45</td>
</tr>
</tbody>
</table>

* Students who were not identified for further assessment of reading by any of the possible referral methods including teacher referral, probe scores, CIBS-R sub-test, and Developmental Reading Assessment scores.

** Students who were not identified for further assessment of reading or math by any of the possible referral methods including teacher referral, probe scores, CIBS-R sub-test, and Developmental Reading Assessment scores.
Table C.2

Average Slopes for Reading from Fall to Spring Probe Administration

<table>
<thead>
<tr>
<th></th>
<th>Classrooms included by baserate (BR) criteria*</th>
<th>Excluding students whose spring reading probe scores fell in the frustrational range</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>First Grade</strong></td>
<td>BR = 0.5</td>
<td></td>
</tr>
<tr>
<td>M (range)</td>
<td>1.14 (-1.67-8.08)</td>
<td>1.41 (-1.67-8.08)</td>
</tr>
<tr>
<td>SD</td>
<td>2.1</td>
<td>1.61</td>
</tr>
<tr>
<td>n</td>
<td>16</td>
<td>32</td>
</tr>
<tr>
<td><strong>Second Grade</strong></td>
<td>BR = 0.07</td>
<td></td>
</tr>
<tr>
<td>M (range)</td>
<td>1.2 (-.92-3.92)</td>
<td>1.22 (-3.75-5.5)</td>
</tr>
<tr>
<td>SD</td>
<td>1.24</td>
<td>1.52</td>
</tr>
<tr>
<td>n</td>
<td>35</td>
<td>65</td>
</tr>
</tbody>
</table>

*The lowest base rate was selected for each grade that included at least one classroom. Base rates for first grade classrooms ranged from .42 to .82. Base rates for second grade classrooms ranged from 0 to .14.
Table C.3
Average Slopes for Math from Fall to Spring Probe Administration

<table>
<thead>
<tr>
<th></th>
<th>All Students</th>
<th>Students Not Referred for Math*</th>
<th>Students Not Referred for Math or Reading**</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>First Grade</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M (range) =</td>
<td>1.15 (0.16-2.53)</td>
<td>1.23 (0.16-2.53)</td>
<td>1.21 (0.16-2.42)</td>
</tr>
<tr>
<td>SD =</td>
<td>0.58</td>
<td>0.59</td>
<td>0.57</td>
</tr>
<tr>
<td>n =</td>
<td>78</td>
<td>53</td>
<td>45</td>
</tr>
<tr>
<td><strong>Second Grade</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M (range) =</td>
<td>0.39</td>
<td>0.38</td>
<td>0.46</td>
</tr>
<tr>
<td>SD =</td>
<td>0.59</td>
<td>0.64</td>
<td>0.60</td>
</tr>
<tr>
<td>n =</td>
<td>75</td>
<td>54</td>
<td>47</td>
</tr>
</tbody>
</table>

* Students who were not identified for further assessment of reading by any of the possible referral methods including teacher referral, probe scores, and CIBS-R sub-test.

**Students who were not identified for further assessment of reading or math by any of the possible referral methods including teacher referral, probe scores, and CIBS-R sub-test.
Table C.4

Average Slopes for Math from Fall to Spring Probe Administration

<table>
<thead>
<tr>
<th></th>
<th>Classrooms included by baserate (BR) criteria *</th>
<th>Including students whose spring math probe scores fell in the mastery range for second grade and instructional range for first grade</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>First Grade</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>M (range)= 1.54 (1.05-2.26)</td>
<td>1.2 (0.16-2.53)</td>
</tr>
<tr>
<td></td>
<td>SD= 0.45</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>n= 12</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>32</td>
</tr>
<tr>
<td><strong>Second Grade</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>M (range)= 0.54</td>
<td>0.60</td>
</tr>
<tr>
<td></td>
<td>SD= 0.55</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>n= 18</td>
<td>50</td>
</tr>
</tbody>
</table>

*The lowest base rate was selected for each grade that included at least one classroom. Second grade base rates ranged from 0 to .67. First grade base rates ranged from .35 to .79.
Appendix D: Tables of Results

Table D.1

Description of Students Referred for Further Assessment Following Schoolwide Screening

<table>
<thead>
<tr>
<th></th>
<th>Overall Results</th>
<th>Results for Math</th>
<th>Results for Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Referred by probe scores</td>
<td>55</td>
<td>46</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>(15%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Referred by teacher</td>
<td>31</td>
<td>70</td>
<td>16</td>
</tr>
<tr>
<td>Referred by CIBS-R subtest scores</td>
<td>64</td>
<td>37</td>
<td>30</td>
</tr>
<tr>
<td>Referred by Developmental Reading Assessment (DRA) Scores</td>
<td>17</td>
<td>27</td>
<td>NA</td>
</tr>
<tr>
<td>Referred as skill or combined skill/performance deficit</td>
<td>40</td>
<td>14</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>(11%)</td>
<td></td>
<td>(performance only)</td>
</tr>
<tr>
<td>Brief Instruction unsuccessful</td>
<td>24</td>
<td>31</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>(7%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PVS-positive</td>
<td>22</td>
<td>79</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>(6%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Criterion Assessment Validated</td>
<td>17</td>
<td>84</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>(5%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Including all cases in the analysis. In predictive power estimates, those cases in which students were already receiving special education services were excluded from analyses because teachers did not have an opportunity to refer those students (i.e., they were already identified).
**Table D.2**

**Accuracy Estimates for Each Screening Method Using the Criterion Assessment as the Standard for Comparison**

<table>
<thead>
<tr>
<th></th>
<th>Teacher Referral</th>
<th>CIBS-R math and reading subtests</th>
<th>DRA</th>
<th>Problem Validation Screening</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>.46</td>
<td>.94</td>
<td>.67</td>
<td>.76</td>
</tr>
<tr>
<td>Specificity</td>
<td>.69</td>
<td>.43</td>
<td>.69</td>
<td>.89</td>
</tr>
<tr>
<td>Positive Predictive Power</td>
<td>.19</td>
<td>.25</td>
<td>.35</td>
<td>.59</td>
</tr>
<tr>
<td>Negative Predictive Power</td>
<td>.89</td>
<td>.97</td>
<td>.89</td>
<td>.95</td>
</tr>
</tbody>
</table>

**Table D.3**

**Accuracy Estimates for Each Screening Method Using ITBS scores as the Standard for Comparison**

<table>
<thead>
<tr>
<th></th>
<th>Teacher Referral</th>
<th>CIBS-R math and reading subtests</th>
<th>DRA</th>
<th>Problem Validation Screening</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>.33</td>
<td>1</td>
<td>.67</td>
<td>1</td>
</tr>
<tr>
<td>Specificity</td>
<td>.94</td>
<td>.72</td>
<td>.92</td>
<td>.99</td>
</tr>
<tr>
<td>Positive Predictive Power</td>
<td>.17</td>
<td>.12</td>
<td>.25</td>
<td>.67</td>
</tr>
<tr>
<td>Negative Predictive Power</td>
<td>.97</td>
<td>1</td>
<td>.99</td>
<td>1</td>
</tr>
</tbody>
</table>
Table D.4

Accuracy Estimates for Each Screening Method Using WJ-R as the Standard for Comparison

<table>
<thead>
<tr>
<th></th>
<th>Teacher Referral</th>
<th>CIBS-R math and reading subtests</th>
<th>DRA</th>
<th>Problem Validation Screening</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>.42</td>
<td>.83</td>
<td>.5</td>
<td>.58</td>
</tr>
<tr>
<td>Specificity</td>
<td>.85</td>
<td>.36</td>
<td>.85</td>
<td>.77</td>
</tr>
<tr>
<td>Positive Predictive Power</td>
<td>.45</td>
<td>.29</td>
<td>.5</td>
<td>.44</td>
</tr>
<tr>
<td>Negative Predictive Power</td>
<td>.83</td>
<td>.88</td>
<td>.85</td>
<td>.86</td>
</tr>
</tbody>
</table>

Table D.5

Accuracy Estimates for Low-Achieving Classrooms

<table>
<thead>
<tr>
<th></th>
<th>Teacher Referral</th>
<th>Brigance math and reading subtests</th>
<th>Running Record or DRA</th>
<th>Problem Validation Screening</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>.55</td>
<td>1.0</td>
<td>.67</td>
<td>.75</td>
</tr>
<tr>
<td>Specificity</td>
<td>.68</td>
<td>.53</td>
<td>.69</td>
<td>.88</td>
</tr>
<tr>
<td>Positive Predictive Power</td>
<td>.35</td>
<td>.43</td>
<td>.5</td>
<td>.69</td>
</tr>
<tr>
<td>Negative Predictive Power</td>
<td>.82</td>
<td>1.0</td>
<td>.82</td>
<td>.91</td>
</tr>
</tbody>
</table>
Table D.6

Accuracy Estimates for High-Achieving Classrooms

<table>
<thead>
<tr>
<th></th>
<th>Teacher Referral</th>
<th>Brigance math and reading subtests</th>
<th>Running Record or DRA</th>
<th>Problem Validation Screening</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0</td>
<td>1.0</td>
<td>1.0</td>
<td>.67</td>
</tr>
<tr>
<td>Specificity</td>
<td>.67</td>
<td>.32</td>
<td>.67</td>
<td>1.0</td>
</tr>
<tr>
<td>Positive Predictive Power</td>
<td>0</td>
<td>.13</td>
<td>.25</td>
<td>1.0</td>
</tr>
<tr>
<td>Negative Predictive Power</td>
<td>.95</td>
<td>1.0</td>
<td>1.0</td>
<td>.97</td>
</tr>
</tbody>
</table>
Table D.7

Problem Validation Screening and Teacher Referral Accuracy by Race

<table>
<thead>
<tr>
<th></th>
<th>White (% of students correctly identified as having or not having a problem)</th>
<th>Minority (% of students correctly identified as having or not having a problem)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Overall Accuracy</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher Referral</td>
<td>61</td>
<td>78</td>
</tr>
<tr>
<td>Problem Validation</td>
<td>86</td>
<td>90</td>
</tr>
<tr>
<td>Screening</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Low-Achieving</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classrooms/High</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base Rate of</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Academic Problems</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher Referral</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>Problem Validation</td>
<td>88</td>
<td>77</td>
</tr>
<tr>
<td>Screening</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>High-Achieving</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Classrooms/Low</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Base Rate of</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Academic Problems</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher Referral</td>
<td>55</td>
<td>88</td>
</tr>
<tr>
<td>Problem Validation</td>
<td>96</td>
<td>100</td>
</tr>
<tr>
<td>Screening</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table D.8

Types of Errors by Race

<table>
<thead>
<tr>
<th></th>
<th>White</th>
<th>Minority</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher Referral</td>
<td>21 false positives</td>
<td>4 false positives</td>
</tr>
<tr>
<td></td>
<td>5 false negatives</td>
<td>2 false negatives</td>
</tr>
<tr>
<td>Problem Validation</td>
<td>7 false positives</td>
<td>2 false positives</td>
</tr>
<tr>
<td>Screening</td>
<td>2 false negatives</td>
<td>1 false negative</td>
</tr>
</tbody>
</table>

Table D.9

Probability of Identification by Screening Method and Race

<table>
<thead>
<tr>
<th></th>
<th>Overall Probability</th>
<th>Probability for Caucasian Students</th>
<th>Probability for Minority Students</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher Referral</td>
<td>.16</td>
<td>.16</td>
<td>.21</td>
</tr>
<tr>
<td>PVS Validated Problem</td>
<td>.11</td>
<td>.10</td>
<td>.21</td>
</tr>
<tr>
<td>Criterion Assessment</td>
<td>.17</td>
<td>.17</td>
<td>.17</td>
</tr>
</tbody>
</table>

Table D.10

Statistical Association between Measures of Math Performance

<table>
<thead>
<tr>
<th></th>
<th>Classwide Math CBM</th>
<th>CIBS-R Math</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Instructional Skill</td>
<td>Mastery Skill</td>
</tr>
<tr>
<td>ITBS Total Math</td>
<td>.421</td>
<td>.600</td>
</tr>
<tr>
<td></td>
<td>p&lt;.000</td>
<td>p&lt;.000</td>
</tr>
<tr>
<td></td>
<td>N=76</td>
<td>N=78</td>
</tr>
<tr>
<td>CIBS-R Math</td>
<td>.768</td>
<td>.612</td>
</tr>
<tr>
<td></td>
<td>p&lt;.000</td>
<td>p&lt;.000</td>
</tr>
<tr>
<td></td>
<td>N=168</td>
<td>N=147</td>
</tr>
</tbody>
</table>
Table D.11

Statistical Association between Measures of Reading Performance

<table>
<thead>
<tr>
<th>Measures of Reading Performance</th>
<th>Classwide Reading CBM</th>
<th>DRA</th>
<th>CIBS-R Reading</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITBS Total Reading</td>
<td>.701</td>
<td>.370</td>
<td>.733</td>
</tr>
<tr>
<td>p&lt;.000</td>
<td>p&lt;.001</td>
<td>p&lt;.000</td>
<td></td>
</tr>
<tr>
<td>N=77</td>
<td>N=81</td>
<td>N=77</td>
<td></td>
</tr>
<tr>
<td>CIBS-R Reading</td>
<td>.902</td>
<td>.271</td>
<td></td>
</tr>
<tr>
<td>p&lt;.000</td>
<td>p&lt;.000</td>
<td>N=172</td>
<td></td>
</tr>
<tr>
<td>N=174</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DRA</td>
<td>.293</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p&lt;.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N=175</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure D.1. Linear relationship between classwide CBM scores in reading and ITBS total reading scores.

Figure D.2. Linear relationship between CBM scores for addition (mastery level skill) and ITBS total math scores for all second grade students.
Figure D.3. Linear relationship between classwide CBM scores for subtraction (instructional level skill) and ITBS total math scores for all second grade students.
Vita

Amanda VanDerHeyden is currently a graduate student in the School Psychology Program at Louisiana State University where she works under the direction of Dr. Joseph Witt. She completed her bachelor of arts degree at Newcomb College of Tulane University in New Orleans, Louisiana, in 1995 with a major in psychology and an emphasis on early childhood. She completed her master of arts degree at Louisiana State University in Baton Rouge, Louisiana, in 1998 in school psychology. She is currently a candidate for the degree of Doctor of Philosophy to be awarded in May 2001.