Running head: EFFECT OF PROBLEM-SOLVING

# Effect of a Problem-Solving Model (STEEP) on Accurate Identification of Children

Amanda M. VanDerHeyden

Vail School District

Joseph C. Witt

Louisiana State University

Donna Gilbertson

Utah State University

The authors wish to express their tremendous admiration for the talented individuals working in the Vail Unified School District outside of Tucson, Arizona where these data were collected. Please direct correspondence to Amanda VanDerHeyden, who is now at University of California at Santa Barbara, at amanda@education.ucsb.edu.

## Abstract

The purpose of this study was to examine the effect of implementation of a problem-solving model of assessment on the identification and evaluation of children for special education. Using a multiple baseline design, a problem-solving model of assessment was introduced in consecutive years for all elementary schools (N=5) in the district. Effect of the model on number of evaluations conducted, percentage of evaluated children who qualified for services, proportion of identified children by sex and ethnicity before and after implementation of the problem-solving model was examined. Additionally, outcomes for children who did not have an adequate response to intervention versus those who were at-risk but responded successfully to short-term intervention were examined. A cost-benefit analysis of use of the model was provided. The degree to which data obtained were used by the decision-making team was also examined. The problem-solving model procedures, decision rules, and schoolwide training procedures are described in detail and practical implications are discussed.

# Effects of a Problem-Solving Model on Accurate Identification of Children

A confluence of forces emanating from practitioners, researchers, and policy makers is challenging the traditional practices surrounding the classification of a student as having a learning disability (LD). At the center of this growing debate is a polemical issue of ancient origins typified within the history of science by, for example, the distinction between Galilean and Aristotelian modes of thinking pertaining to the behavior of objects. The Aristotelian view of objects is that they fall because they are "heavy." Hence the behavior of Newton's apple falling through the air is a function of a property of the apple (i.e., its weight). In contrast Galileo's view was that the behavior of the falling apple is a function of the apple (e.g., weight, mass) *interacting* with the environment (i.e., gravity, wind resistance). By analogy, similar controversies in psychology have centered on attributions of causality for human functioning (e.g., behavior, psychopathology, disability). The issue reduces to the question of whether the locus of a psychological problem resides primarily within the person, the environment, or in person/environment interaction. With some exceptions, most contemporary psychological theories adopt some version of the latter person/environment locus to explain human functioning.

A notable exception to this general trend has been the practice of classification in LD. By law, current practice usually consists of defining LD as a severe discrepancy between ability and achievement. In practice, this consists of administering a standardized achievement test, a standardized measure of intelligence, and determining if actual achievement levels deviate significantly from what would be expected based upon intellectual functioning. Implicit in this process is a presumption that the locus of the problem resides within the student because the student is the focus of measurement efforts. Presumably the contribution of environmental factors (e.g., educational or cultural disadvantage) are “considered” and “ruled out” but practitioners are provided little guidance in how to operationalize these variables.

The challenges to an ability/achievement discrepancy definition focus on several inter-related issues pertaining to the role of the environment (i.e., lack of appropriate instruction) as an explanatory factor in understanding “why” a student may be of normal ability but have significantly below average achievement (Gresham, VanDerHeyden, & Witt, in submission). For example, Johnny is referred because of reading problems and he exhibits a severe discrepancy because he is in fifth grade, reads at the third grade level, and has normal intelligence. The well-regarded achievement test administered to Johnny cannot “know” anything outside the room in which the testing occurs. It cannot “know” that Johnny may be in a low-performing school where 70% of his peers in the same class are also reading two or more years below grade level. Such a scenario would indicate that the core curriculum in this school is not effective. The test cannot “know” if Johnny’s previous teachers were “highly qualified” to teach reading. The test cannot “know” either that Johnny is still eager, motivated and ready to learn to read but current classroom instructional objectives are at too high of a level to benefit Johnny. What the test cannot tell them and what the school-based team does not know can be problematic for Johnny because he is likely to be placed in special education (Ysseldyke, Vanderwood, & Shriner, 1997) and remain there for the duration of his school career. If Johnny does not have LD, then why does he have low achievement? Because the assessments are child focused, the school-based team cannot evaluate the most prominent of the rival hypotheses that may explain why Johnny is two years below grade level: lack of high quality instruction.

Why is instruction given short shrift in attempts by schools to understand the low achievement of an individual student, especially when given the magnitude of the decision to diagnose LD? It is difficult to imagine a more significant or life changing act that can be committed by a school district than placing a student in special education; Yet, this act has become almost as commonplace as writing an ordinary hall pass (US Department of Education, 1998). Placement has become a routine mechanism for the school-based team to respond to the problem of low achievement because viable alternatives (e.g., effective interventions) are not obvious, available, or easily implemented. It is also difficult for some educational professionals, whose responsibility it is to provide high quality instruction, to critically evaluate the possibility that their actions, or lack thereof, are responsible for the absence of reading skills with a particular student (Ysseldyke, Pianta, Christenson, Wang, & Algozzine, 1983). Many school-based professionals, however, are eager to evaluate instruction as contributory but they have lacked the tools for systematic evaluation of a Galilean interactionist premise. How does one properly evaluate the possibility that a student is a "normal learner" but has not received adequate instruction?

The ability to evaluate academic outcomes as a function of the student interacting with the environment/instruction is available in the form of a Response to Instruction (RTI) model. This model represents a set of systematic processes that have become available over the last several years for determining student need for services of increasing intensity based upon a review of student progress data as various instructional options are implemented with a student. Instructional options are organized and presented sequentially via a 3-tiered model (Tilly, Reschly, & Grimes, 1999) such that movement through the tiers is based upon inadequate response to intervention and, importantly, intensity of intervention is increased as the student fails to respond to successive tiers. The student who remains unresponsive to instruction and/or who requires intensive services is considered for special education. Eligibility determination is informed by student progress data that have been collected as the student moves through the instructional tiers and is exposed to instruction that has systematically increased in intensity. What RTI adds to the composite picture of a student is that "data from estimates of needed intervention quality and intensity can be used to help answer service delivery questions as opposed to the use of estimates of child deficits from test-based results" (p. 4, Barnett, VanDerHeyden & Witt, in submission).

Research pertaining to RTI has generally indicated that this model can result in greater decision accuracy in the classification of LD than the ability-achievement discrepancy model (Gresham, 2001). Further, it results in few students who would be classified as LD because it allows professionals to assess whether a student *can* learn in response to correctly implemented, effective instructional practices as opposed to *has* learned in response to whatever previous instruction the student may have received (Fuchs & Fuchs, 1997; 1998). In addition, referrals are reduced because progress monitoring of students at Tier 1 (i.e., the core curriculum) results in early detection of students not making adequate progress so that the "achievement gap" can be eliminated while it is narrow (Donovan & Cross, 2002). Using RTI in the identification of LD students is purported to have several additional advantages (Gresham, VanDerHeyden, & Witt, in submission; VanDerHeyden & Witt, in press; VanDerHeyden, Witt, & Barnett, in submission) including reduction of identification biases, viewing learning problems in terms of a risk versus deficit model, and focusing on student progress and outcomes versus diagnosis. The empirical basis for RTI has been reviewed extensively elsewhere (Grehsam, et. al., in submission).

In addition to considerable empirical support, policy makers have integrated RTI into the reauthorization of IDEA 2004. IDEA builds upon NCLB and there are no fewer than 20 references to NCLB in IDEA. A child in general education should not be considered for special education if there are problems with the core curriculum which derive from instruction which is not “evidence based,” is not delivered by a “highly qualified” teacher, or is not producing results for certain subgroups. Logically how could one expect a student to be achieving if these elements were missing?

One challenge with RTI is that it is not merely one thing. Instead, RTI is a process consisting of an integrated set of tools, procedures and decisions. To utilize RTI, the school-based team will need to define a problem appropriately, select an intervention that is likely to be effective, implement the intervention, evaluate the effects, and make changes if needed. Proponents of RTI point to a large and growing body of research supporting the various components within an RTI model. Clearly this research has provided evidence to guide the difficult choices that must be made within an RTI model pertaining to which students need intervention, what type of intervention is needed, delivered with what intensity, integrity, and duration so that a determination can be made as to whether the student improved “enough” or requires more intensive services. There are at least two problems with the research thus far conducted in support of RTI. Much of the research heretofore conducted suffers from one or both of these problems. First, implementing RTI means implementing an *integrated set* of procedures or components while correctly applying sequenced decision rules (Barnett, Daly, Jones, & Lentz, 2004; Barnett et al., in submission). The research conducted to date has focused primarily on the efficacy of the components *individually* but not on the efficacy of the *RTI process as an integrated whole*. In theory, if the components are effective, then the overall process would be expected to produce results; However, the question of whether the overall process is efficacious must also be addressed. The second issue is that most of the research has been conducted by well-funded research centers (Ardoin, Witt, Connell, & Koenig, in submission). Hence, for the intervention component, data suggest that evidence based interventions can markedly decrease the need for special education services *when implemented with high integrity by a research associate who is paid for to do that job* (Torgesen, Alexander, Wagner, Rashotte, Voeller, & Conway, 2001); Vellutino, Scanlon, & Tanzman, 1998). The question is whether these components can be effective when implemented by front line educational professionals. In particular, implementation is the linchpin of RTI. If there is to be an evaluation of RTI, an intervention must be implemented correctly and monitored. Whereas such a statement appears self-evident and parsimonious, the extent to which practitioners can implement these procedures with fidelity remains unknown and in actuality, is not parsimonious (Noell et al., 2005). The research on intervention integrity has shown uniformly dismal results with implementation of only the intervention component (Noell, et al., 2005). Fidelity to the RTI process will almost certainly be reduced when implemented in schools, the question is whether such inevitable degradation can still produce results.

The purpose of this study was to evaluate a research-based RTI process, Screening to Enhance Equitable Placement (STEEP). STEEP consists of a series of problem-solving assessment procedures with specific decision rules to identify children who might benefit from an eligibility evaluation. STEEP was built upon the research in CBA/CBM (Shinn, 1989) and problem-solving (Good & Kaminski, 1996; Fuchs & Fuchs, 1998; Shinn, 1989; 1998). Children are screened using CBM probes, a subset are identified to participate in a brief assessment of the effect of incentives on child performance, and a smaller subset are then identified to participate in individual intervention. Research has found STEEP to be highly accurate in identifying which children should and should not be considered for special education eligibility (VanDerHeyden, Witt, & Naquin, 2003) and has shown a positive effect on disproportionate identification by ethnicity, sex, and achievement level (VanDerHeyden & Witt, in press). Prior research with STEEP occurred with researchers implementing most of the procedures (e.g., intervention). The purpose of this study was to evaluate STEEP implementation and its effect in a district. Research questions were (1) What effect would STEEP implementation have on total number of evaluations and percentage of evaluations that qualified for services? (2) To what degree would the decision-making teams utilize STEEP data to determine whether or not an evaluation should be conducted? How often did the outcome of STEEP match with the outcome of evaluation? (3) How did the use of STEEP reduce assessment and placement costs for the district and how were these funds re-allocated? (4) What effect did STEEP implementation have on identification rates by ethnicity, sex, free or reduced lunch status, and primary language status? (5) What were the outcomes for children judged to have an adequate response to intervention relative to those children who were judged to have an inadequate response to intervention?

## Method

### *Participants and Setting*

A rapidly growing suburban district in the southwestern US served as the site for this project. Vail School District is a district outside of Tucson, Arizona that had previously been a small, rural district but had recently experienced substantial growth. From April 2002 to April 2004 (the school years during which this study occurred), number of children enrolled in the primary grades increased 30% districtwide. The STEEP model was implemented in each of the five elementary schools (grades 1 through 5) beginning with two schools in 2002-2003, adding one additional school in 2003-2004 and two schools in 2004-2005. Demographic data, obtained from the census data provided to the Office of Civil Rights, for each of the schools is presented in Table 1. The first two participating schools volunteered to participate (these sites had the highest number of referrals and evaluations). The third site was a new school that opened with STEEP in place. STEEP was introduced simultaneously to schools four and five because those schools were staffed by the same school psychologist.

Overall, the district was one in which mostly middle-class families lived and worked. Student to teacher ratio was about 23:1 for all primary grade classes in the district. The highest percentage of children who received free or reduced lunch was enrolled at School 2 where 40% received this benefit. School 3 was the second lowest SES school with 26% of children receiving free or reduced lunch.

Four female Caucasian school psychologists were trained to coordinate STEEP activities at each school given their existing role and responsibilities in the prereferral process. The same school psychologist remained at each site through baseline and STEEP implementation with one exception. At school 1, STEEP was withdrawn at the end of the first year of implementation by replacing the trained school psychologist with an untrained school psychologist. The following year (2004-2005), the untrained school psychologist remained at school 1, but was trained to use STEEP. Four school psychologists were trained. The first school psychologist had a specialist degree in school psychology and had been working in the district as a school psychologist for about 20 years. This psychologist was trained by the first author (see procedures below) to coordinate activities at schools 1 and 3. The second school psychologist had a PsyD degree in child clinical psychology and had been working in the district as a school psychologist for one year prior to STEEP implementation at her school. This psychologist was trained by the first trained school psychologist with assistance from the first author to coordinate activities at school 2. The third school psychologist had a specialist degree in school psychology and had worked in the district for one year prior to STEEP implementation at her school. This psychologist worked at school 1 when STEEP was withdrawn and was trained the following semester by the first school psychologist to coordinate activities at school 1 when STEEP was re-instated. The fourth school psychologist had a specialist degree in school psychology and had worked in the district for two years prior to STEEP implementation. This psychologist was trained by the first school psychologist to coordinate activities at schools 4 and 5. Prior to STEEP implementation, psychologists attended the meeting at which a decision was made to refer a child for evaluation, performed evaluations, and conducted Individualized Education Plan meetings. None of the psychologists had experience using curriculum-based measurement or performing functional academic assessment prior to STEEP implementation.

*Description of Instructional Setting and Teacher Preparation*. Instruction was provided to students according to a set of standards specified by the state. A specific curriculum calendar was used to ensure that all essential standards were introduced in a similar timely fashion across all schools. Multiple sources of assessment (e.g., standard tests, curriculum-based assessment probes) were used to routinely track individual student, class, and school performance on the essential standards. Any student who obtained a failing grade in any of the district's core content areas was required to participate in 12 hours of remediation either through private (i.e., parent-funded) or district-provided (i.e., no cost to parents) tutoring. Additionally, any child with a failing grade or who was below the instructional minimal standard during schoolwide screening (see STEEP Implementation below) was required to participate in 12 hours of tutoring conducted during school breaks. Additionally, each school provided supplemental instructional services to children at-risk for early reading failure through Title I funding.

All children were screened for participation in the ELL program whose parents indicated in their registration packet that any language other than English was spoken in the home. The district used the SELP (Harcourt) to screen all children and identify children as non-English proficient, limited English proficient, or fluent English proficient. Once identified, ELL services were continued for a minimum of 2 years with annual monitoring on the SELP. Services were discontinued after a minimum of 2 years and the child scored in the fluent English proficient range. Each child had an individual needs plan which specified goals and objectives for the child and this plan was reviewed and revised each quarter. Children who scored in the non English proficient range received a minimum of 1 hour per day of instructional assistance in their classroom with a ELL site coordinator. Children who scored in the limited English proficiency range received a minimum of 30 minutes each day, three days per week of instructional assistance in their classroom with the ELL site coordinator. ELL coordinators all spoke Spanish, which was the most frequent primary language of ELL students in the district.

To promote effective instruction, new teachers participated in an induction program (Wong & Wong, 1998) that included seven days of all-day training the first year of service. Teachers were assigned two coaches, a literacy coach and an instruction and classroom management coach. These coaches worked with new teachers for the first two years of service and completed a total of nine observations per year followed by reflective feedback with the new teacher.

### *Baseline Referral Process*

Each school used a school-based pre-referral team to consider whether or not children referred by their teachers or parents might be in need of a special education eligibility assessment. The pre-referral process involved submitting questionnaire and student record information to a multi-disciplinary team for consideration. A formal meeting was scheduled with this team that included special and regular education teachers, the school psychologist, and an administrative representative. At this meeting, the team reviewed available data (e.g., report card grades, standardized test scores, work folder, grade book, and a teacher-completed questionnaire of strategies attempted). This team met with the teacher and the child's parents to review existing information, discuss concerns, and provide recommendations to attempt to address the problem in the regular setting. The team agreed upon a time to reconvene and determine whether or not the problem had been adequately resolved or whether the problem was persisting and an eligibility evaluation should be recommended. Written records were maintained by the team and special education department specifying the names of children who were referred to the team for consideration for evaluation, whether or not the team decided to refer the child for evaluation, and evaluation results.

The existing team decision-making process remained in place throughout the years of this study. When the STEEP model was introduced at each site, the STEEP data (see STEEP Implementation below) were offered by the school psychologist to the team for consideration in determining whether or not to refer a child for evaluation.

### *STEEP Implementation*

Screening to Enhance Equitable Placement (STEEP; Witt, Daly, & Noell, 1999) is a problem-solving model of assessment that can be used to identify children who might benefit from eligibility assessment. STEEP has similarities to the problem-solving (Good & Kaminski, 1996), problem certification (Shinn, 1985) and dual discrepancy models (Fuchs & Fuchs, 1998) previously described in the literature. STEEP uses curriculum-based assessment and measurement to obtain data in the child's classroom concerning absolute level of performance relative to same-class peers and an instructional standard to proactively identify performance problems, to plan remediation efforts to resolve those problems, and to evaluate the effectiveness of the solutions. The process yields data that can be used by the school's assessment team to determine whether or not intervention services are needed, and if so, whether those interventions would most appropriately be provided through special education (Reschly, Tilly, & Grimes, 1999).

Trained consultants worked with teachers and students to complete a series of procedures and sequentially apply a series of decision rules to resulting data at each stage of the process. The four sequential stages described below are (1) universal screening, (2) classwide intervention, (3) brief assessment of the effect of incentives on performance, and (4) assessment of the child's response to short-term standardized intervention delivered with integrity in the regular classroom setting. Decision rules are summarized in Table 2.

*Universal Screening*. Curriculum-based assessment and measurement (CBM) probes were administered classwide in reading and math twice each year following standardized procedures (Shinn, 1989) and using a commercially-available set of content-controlled materials. During screening, two types of data were collected. First, oral reading fluency and computation fluency scores were obtained to assess children's performance relative to their classmates and relative to instructional standards on a task that represented current grade level difficulty (a skill that students would be expected to do well that time of year to benefit from the instruction being provided in their classes). Second, a task that reflected skills that would be learned throughout the year was also used so that periodic probes could reflect growth toward year-end goal of competence in a broad array of computational tasks. For reading, a single grade-level probe was selected for screening at all grade levels. At first and second grades, the same probe was re-administered monthly to track growth in oral reading fluency until the class median reached the mastery range. When the class median reached mastery range, a more difficult probe was selected to track growth from that point forward until the end of the year. For math, one probe was administered at each grade level at each screening that reflected current instructional placement for screening. A second probe was administered monthly to all children at all grade levels to track progress in math. The rationale for monitoring math more frequently is that the district had targeted math achievement as a problem area, whereas fewer than five classwide reading problems were detected during all the years of this project districtwide and no classwide reading problems ever occurred above grade 2. Reading probes were scored as words read correctly per minute (wc/min) and math probes were scored as digits correct per two minutes (dc/2 min). The instructional standard applied for reading was 40-60 wc/min for grades 1 and 2, 70-100 wc/min for grades 3-5. The instructional standard applied for math was 20-40 dc/2 min for grades 1-3, and 40-80 dc/2 min for grades 4-5.

Teachers were trained to reliably administer CBM probes and administration required no more than 1 hour per class. Following the screening, the teacher received a graph showing the performance of all children in the class relative to an instructional standard (Deno & Mirkin, 1977). Following schoolwide screening, problems were categorized as classwide (class median score falls below the instructional range described by Deno & Mirkin, 1977) or individual child problem (classwide median falls at or above the instructional range and individual child scores below the 16th percentile for his or her class and in the frustrational range described by Deno & Mirkin, 1977). Thus, two anchors were applied initially to define the problem (local anchor was classwide performance, broader anchor is instructional level performance that has been linked to functional competence, Deno & Mirkin, 1977).

*Classwide Intervention*. When a classwide problem was identified (class median score fell below the instructional range described by Deno & Mirkin, 1977), a classwide intervention was implemented. The first step in performing classwide intervention involved finding the instructional level of the *class* by administering a series of easier CBM probes until the class median reached the instructional range. Classwide intervention can take many forms but the STEEP model has used the following protocol most frequently: modeling the target skill, guided practice with frequent opportunities to respond and immediate feedback, timed independent practice to yield a score for progress monitoring, and use of delayed error correction with a verbal rehearsal strategy. Classwide intervention was delivered at a difficulty level that matched the instructional level of the majority of students in the class using paired peer practice (e.g., classwide peer tutoring, peer-assisted learning strategies; Fuchs, Fuchs, Mathes, & Simmons, 1997; Greenwood, 1991). The intervention required about 10 minutes daily. The classwide intervention was performed for 10 consecutive school days. Following the data decision rules in Table 2, the children who continued to perform below the instructional standard and demonstrate poor growth relative to peers in the same class (i.e., children who were not learning when other children were learning at a rapid pace) were identified and referred for the next phase, the performance/skill deficit assessment.

If a classwide problem was ruled out following the classwide assessment, then children who performed below the 16th percentile for their classes (i.e., approximately 1 SD below the mean) and fell below the instructional range participated in the next stage of assessment, a brief assessment of the effect of powerful incentives upon performance (i.e., performance/skill deficit assessment). In prior studies when STEEP was used, approximately 15% of children were identified through the schoolwide screening to participate in further assessment (VanDerHeyden et al., 2003). The school psychologist conducted the next stage of assessment outside of the classroom using scripted administration procedures.

*Performance/Skill Deficit Assessment*. During the performance/skill deficit assessment, the school psychologist provided the student with a copy of the classwide academic assessment probe that had been previously administered. Students were told that they could earn a reward of their choice from the treasure chest by "beating their last score." This score was written in the top left-hand corner of the student's paper. Students were allowed to sample briefly the items in the treasure chest. The treasure chest was a small transparent box containing several small tangible items (e.g., pencils, balls, stickers, bracelets, coupons for free time). The probe that was used during the classwide screening was then re-administered. The performance/skill deficit assessment for math was administered to groups of three to five students simultaneously, whereas the performance/skill deficit assessment of reading was administered individually in a quiet space on the school campus. This component required no more than five minutes per assessment. Children whose performance improved to the instructional range (Deno & Mirkin, 1977) to earn an incentive did not participate in further assessment. Children whose performance did not improve to the instructional range participated in an individual intervention in their classrooms. Prior research found that approximately 11% of the total cases screened were found to exhibit a skill deficit that merited individual intervention or the third tier of assessment (VanDerHeyden et al., 2003).

*Individual Intervention*. At this point, those children exhibiting skill deficits, in classes where the majority of the class was performing at or above the instructional range, participated in daily individual intervention performed by the classroom teacher (or teacher designee) in the regular classroom setting during the regular school day. In this stage, a standard protocol-based intervention that required approximately ten minutes was applied. The school psychologist or consultant worked individually with the student to determine intervention task difficulty (i.e., the student’s instructional level) and to identify an appropriate intervention. The student’s instructional level was determined by sampling backward through successively lower level materials until the student scored in the instructional range. Protocol-based interventions shared four common basic components: modeling, guided practice with immediate error correction (to improve accuracy), independent timed practice with slightly delayed error correction (to build fluency), and the opportunity to earn a reward for “beating the last highest score” (to maximize motivation to respond and build fluency). The interventions were protocol-based and designed to produce evidence (i.e., permanent products) that they occurred to allow for estimation of treatment integrity.

The school psychologist collected the intervention data weekly, quantifying two critical variables: the degree to which the intervention occurred correctly and the child’s performance on a novel, instructional-level probe of the target skill and a novel, criterion-level probe of the target skill. Similar to prior studies (e.g., Witt, et al., 1997), intervention integrity was evaluated based on the production of permanent products generated as the intervention was implemented. The school-based consultant entered the data into the database and graphing tools automatically generated graphs for the teacher, principal, and consultant. If problems occurred in implementing the intervention, the consultant provided performance feedback to the teacher and re-trained the teacher to implement the intervention correctly for the following week.

The purpose of the brief intervention was to measure the child’s RTI. To measure RTI, 10 to 15 consecutive intervention sessions, conducted with integrity, were required. Additionally, a similar but unpracticed probe (the probe used at screening) was administered each week to track progress. The intervention was determined to have been successful if the child performed in the instructional range on the grade-level screening probe following intervention. Intervention trend data were also examined to ensure that growth was occurring each week and to determine when to increase the difficulty level of the materials used during intervention sessions. Data showing a lack of response to short-term intervention were made available to the school-based team to assist in determining whether or not a child should receive an eligibility evaluation. These data were graphically presented to the team with a recommendation to obtain more information through a full psychoeducational evaluation. Estimates from a highly controlled study indicated that about 3 to 5% of children failed to respond sufficiently to brief intervention performed with integrity for five to nine days (VanDerHeyden, et al., 2003).

*Training the School Psychologist to Implement STEEP*. The first author trained a school psychologist at the first school to implement STEEP late in the second semester of the 2002-2003 school year. Once per week, the first author spent the school day with the school psychologist teaching the psychologist to implement STEEP procedures. Scripted instructions were provided to the school psychologist and performance coaching was used to train all required components of STEEP in the actual setting where STEEP was being implemented. When a new component was introduced, the first author would describe the steps, provide scripted instructions, and model correct performance. The school psychologist would then implement the new component with immediate assistance from the first author. Finally, the school psychologist implemented procedures independently with delayed feedback from the first author. The first school psychologist and the first author together trained the school psychologist at the second school in the fall semester of 2003 using similar procedures. The first school psychologist trained the third school psychologist at schools four and five during 2004-2005 using the training procedures just described.

*Preparing the School for Implementation*. A detailed presentation including rationale, basic procedures, how data link to decision-making, and anticipated benefits of the program was given to the principal of the school about 3-4 weeks prior to implementation at each school. At this meeting, principal agreement to implement the model was obtained, needed materials were specified and arranged to be ordered, additional presentations were scheduled with the general education site leadership team, the special education site leadership team, and the parent advisory group for the school. The schoolwide screening activities were placed on the master schedule for the school. Additionally, at this meeting, the principal identified an on-site person to assist with copying probe materials and entering data.

Within one week, detailed presentations were given to the general and special education leadership teams at the site and the principal or principal designee was asked to attend each meeting. These presentations were the same as the one given to the principal individually. Within one week, a faculty overview presentation was given to the entire faculty that lasted from 15 to 60 minutes depending on the number of questions. The presentation described the rationale, general overview of procedures, and what to expect as the next step at the school (date, time, and activity).Within one week, the school psychologist attended grade-level planning meetings to select appropriate grade-level probes for screening, overview the procedures for screening, and provide the scripted instructions for schoolwide screening to each teacher. Within two days, screening occurred for the entire school. Each grade was scheduled to conduct all screening activities within the same 1-hour time period and a trained coach was present in each classroom to monitor for integrity of implementation. For reading, the trained coach administered probes while the teacher simultaneously scored until the teacher reached 100% scoring agreement on two consecutive trials. Once the teacher scored two consecutive probes in 100% agreement with the trained coach, the coach observed the teacher administering a reading probe to ensure that the teacher administered the probe using the scripted instructions. The coach then either assisted the teacher by reading with half of the remaining students individually or by managing classroom activity while the teacher read with all remaining students. The school psychologist assisted the teachers to score their math probes during the next grade-level planning meeting. By the end of that week, the school psychologist delivered graphs to teachers showing the performance of all children in the class relative to standards for frustration, instructional, and mastery level performance.

The school psychologist conducted the skill/performance deficit procedures in a small office on the school's campus and delivered graphs to teachers showing the in-class performance of all children with a second bar showing performance during the skill/performance deficit assessment for the lowest performing students. For children who received individual intervention, the school psychologist met briefly with the teacher to summarize assessment procedures up to that point, and to show an example of an intervention script that would be recommended for that problem. The teacher was given an opportunity to modify variables of the intervention not thought to be related to intervention strength (i.e., time of day the intervention would be conducted, whether a peer tutor or the teacher would conduct the intervention). All interventions shared the following key components: were implemented daily, occurred in the regular classroom by the teacher or a peer tutor, utilized instructional level materials, included modeling correct responding, guided practice, timed independent practice for a score, and incentives for improvement. All interventions produced a daily score on a CBM probe to track growth. All interventions were protocol-based and could be monitored for integrity. The school psychologist prepared all needed materials to run the intervention for one week, delivered the materials to the classroom, and trained the person who would be conducting the intervention using a "tell-show-do" format. The school psychologist described the intervention and provided a written script of observable steps of the intervention. The school psychologist modeled intervention implementation, and then observed the teacher or peer and student complete the intervention with coaching. Once the teacher or peer could complete the intervention 100% correctly without prompting from the school psychologist, training was complete. The school psychologist then picked up intervention materials once each week, performed a generalization probe with the student outside of the classroom, placed student data on a graph, and provided feedback to the teacher about student performance and accuracy of intervention implementation. If the intervention was to be continued another week, new materials were provided at the appropriate instructional level for the following week. Decisions about intervention responsiveness could generally be made once 10 consecutive intervention sessions had occurred with integrity.

*Procedural Integrity of STEEP Procedures*

*Implementation of Screening Procedures.* An integrity checklist that specified each observable step of the classwide screening was provided to a trained observer. The trained observer noted the occurrence of each step with a checkmark. Teachers were reminded to follow the scripted instructions when conducting the screening and were told that the trained observer would follow along on a separate copy of the script to note correct implementation of steps in the script and interrupt the teacher with a prompt to complete any incorrectly implemented steps in the script. The total number of correctly (i.e., unprompted) implemented steps was divided by the total number of steps possible and multiplied by 100% to estimate integrity of procedures. For all schools, 54 observations were conducted and average integrity for screening procedures was

98.76%. Of 54 observations, three teachers required 1-2 prompts to correctly complete missed steps.

*Individual RTI judgment agreement*. On average, 6.68 number of intervention sessions (range, 3 to 15) occurred before a decision was reached about whether the response was adequate and 12.41 number of sessions (range, 4 to 19) occurred before a decision was reached that a response was inadequate. The criterion applied to determine intervention success was provided to an untrained observer along with the children's individual intervention data for 56 cases (44% of total intervention cases) and agreement exceeded 87%.

## Results

### *Collection and Calculation of Dependent Measures*

The primary dependent measures included evaluations, demographic information for students, and outcome of evaluations. These data were maintained by the referral and evaluation decision-making team at each school and by the district special education office. Names of students who were considered for referral for evaluation were obtained from the team chairperson at each site. These names were then cross-referenced with the data maintained by the district special education office. Each student was coded by gender, ethnicity, free lunch status, and ELL status using district data. The individual files were then checked for each student at district office to verify that (a) an evaluation was conducted and (b) whether or not the child qualified and if so, the qualifying category. This process was conducted for each year of the study. Once STEEP was underway, all assessment data were maintained in a centralized database. All children for whom STEEP data indicated that an evaluation should be considered were discussed by the school's decision-making team. At this meeting, the school psychologist used a summary sheet to report the child's performance at each stage of the assessment and attached a graph showing the child's progress during individual intervention. The summary sheet indicated that the recommendation of the STEEP assessment data was to (a) consider a full psychoeducational evaluation, or (b) not refer for evaluation. Additionally, any teacher or parent could place a child on the team's agenda for discussion at the meeting at any time. If a child was placed on the discussion list, the school psychologist either provided the completed STEEP summary sheet with graphs as applicable to the team or indicated that the STEEP process was not complete and summarized existing data with a recommendation to finish the STEEP process prior to making a decision to refer for evaluation.

*Design*

Effects were examined within a multiple baseline across schools. The baseline and STEEP procedures experimental conditions were sequentially introduced and evaluated for their effects on initial evaluation, percentage of children evaluated who qualified, and the degree to which teams used STEEP data for decision-making for all students referred to the school-based team. STEEP effects were also evaluated for differences by gender, ethnicity, and SES level. In addition to the multiple baseline, a reversal was implemented at school 1. At school one, STEEP was withdrawn from the school near the end of 2003-2004. Specifically, the school psychologist who had been trained to use STEEP worked at the school for the first eight months of the school year. After 8 months, the principal and decision-making team expressed concerns about the accuracy of STEEP data. The director of special education agreed to remove the trained school psychologist as a "test" of the accuracy and utility of the STEEP process and send an un-trained school psychologist from within the district to work for the remainder of the school year. Hence, all screening data that had been obtained prior to the trained school psychologist's departure were available to the untrained school psychologist but no specific instructions for how to use the data (or not use the data) were provided to the untrained school psychologist and the decision-making team. All other members of the decision-making team remained the same during this time period. The untrained school psychologist worked for the last two months at school one in 2003-2004. To permit a comparison across psychologists with baseline and subsequent years, dependent measures for school 1 during the year 2003-2004 were converted to a rate estimate by dividing each dependent measure estimate (e.g., number of evaluations) by the total number of months the school psychologist worked in the school and then multiplied the rate by the total number of months in the year (i.e., 10 months) to estimate what the value would have been if that school psychologist had worked there the entire year (see Figure 1).

*Overall Findings*

*Initial evaluations*. Total number of initial evaluations for each site during consecutive years is presented in Figure 1. Average total number of evaluations for school 1 during baseline years was 19.5 evaluations. The trained school psychologist conducted 7 evaluations in 8 months which computed to an estimate of 9 evaluations for the entire school year (7/8 * 10) whereas the untrained school psychologist performed 10 evaluations in two months which computed to an estimate of 50 evaluations (10/2 * 10) for the year (similar to baseline level) in 2003-2004. In 2004-2005, 7 evaluations were conducted for the entire school year. At school 2, there were 30 evaluations during the baseline year and 9 during the first year of STEEP implementation. In the second year of implementation (2004-2005), 7 evaluations were conducted. School 3 had no baseline data because it was a new school that opened with STEEP in place with the school psychologist from school 1 staffing school 3 as well. Average total number of evaluations at baseline for school 4 was 12.33. School 4 had 7 evaluations during the first year of STEEP implementation (2004-2005). Average total number of initial evaluations at baseline for school 5 was 10.5. Six evaluations were conducted during the first year of STEEP implementation (2004-2005).

*Percentage of children evaluated who qualified*. Figure 1 also shows the number of children who qualified for services at each site. This number was computed by dividing the total number of children who qualified for services by the total number of children evaluated at each site across years. During baseline years at school 1, on average, 41% of children evaluated qualified for services. With STEEP, this percentage increased to 71% and then reversed to 40% when STEEP was removed in 2003-2004. During the second year of implementation (2004-2005), 57% of children evaluated qualified. At school 2, the percentage of evaluated children qualifying for services increased from 70% at baseline to 100% with STEEP in 2003-2004. In 2004-2005, 83% of children evaluated qualified at school 2. At school 3, during the first year of STEEP implementation, 80% of children evaluated qualified for services. During the second year of implementation at school 3, 77% of children evaluated qualified for services. On average at school 4 during baseline, 43% of children evaluated qualified for services. During the first year of implementation at school 4, 67% of children evaluated qualified for services. On average at school 5 during baseline, 53% of children evaluated qualified for services. During the first year of implementation at school 5 40% of children evaluated qualified for services.

For schools 1 through 4, the percentage of children evaluated who qualified for services increased with STEEP implementation. Interestingly, for schools 1 through 3, for which more than one year of implementation data were available, the percentage of children evaluated who qualified in the second year of implementation decreased. Because fewer children were being evaluated, those who were evaluated and did not qualify in subsequent years of implementation had a stronger effect on the computed percentage. In other words, the number of children who were evaluated and did not qualify decreased substantially for all schools with STEEP implementation and remained low in subsequent years for schools 1 through 3. At each school, the MDT could elect to recommend evaluation irregardless of STEEP data. At school 1, the one case that was evaluated and did not qualify in 2004-2005, had participated in STEEP and had an adequate response to intervention during two consecutive years (i.e., STEEP recommended twice that evaluation not be conducted). At school 2, 3 children were evaluated and did not qualify. Of these 3, two had participated in STEEP procedures and had an adequate response to intervention (i.e., STEEP recommended that evaluation not be conducted). One child did not have an adequate response to intervention and thus was recommended by STEEP for evaluation. At school 3, all of the cases (N=3) who were evaluated and did not qualify had participated in STEEP and had an adequate response to intervention (i.e., STEEP recommended that evaluation not be conducted). At school 4, two children were evaluated and did not qualify. One child had participated in STEEP and had an adequate response to intervention (i.e., STEEP recommended that evaluation not be conducted) and one child had not participated in STEEP because the child was referred for evaluation prior to STEEP being underway for the 2004-2005 school year (in effect, the decision to refer for evaluation had occurred during the preceding school year and the referral occurred on the first day of school in 2004-2005). At school 5, three children were evaluated who did not qualify for services. Two of these children had an adequate response to intervention with STEEP (i.e., STEEP recommended evaluation not be conducted). One child had not participated in STEEP because the child was referred on the first day of school similar to the case at school 4.

*Degree to which teams used STEEP data for decision-making*. Seventy-two percent of children evaluated in 2003-2004 (when STEEP was being implemented) actually had completed STEEP data (all four stages of assessment had been completed) across the three schools using STEEP that year. This value varied across sites, however. At school 1, 72% of children who were evaluated had completed STEEP prior to the reversal. At school 1, 60% of children evaluated after STEEP was withdrawn had completed STEEP data (i.e., these data had been collected prior to withdrawal of STEEP conditions). At school 2, 100% of evaluated children had completed STEEP. At school 3, 60% of children evaluated had completed STEEP. Hence, overall, nearly 30% of evaluations did not have completed STEEP data. Of those who did not have completed STEEP data, 91% of these children qualified for services, a qualify rate that was much higher than the baseline average. Most of these children qualified under speech and language impairment (SLI; 46%), but 23% qualified under SLD and 15% qualified under SLD/SLI. In 2004-2005, 63% of children who were evaluated across all schools had completed STEEP data (57% at school 1, 71% at school 2, 81% at school 3, 29% at school 4, and 50% at school 5). Again, of those who did not have completed STEEP data, a high percentage (81%) qualified for services. Schools 4 and 5 contributed the greatest number of children who were evaluated without completed STEEP data. Most of these children qualified under SLD (71%), 14% qualified under ED, and 7% qualified under autism and SLI respectively.

Because STEEP was conducted as a pre-referral process entirely whereby the data were provided to the school's decision-making team for consideration in determining whether or not an evaluation should occur, an analysis of the team's decision-making behavior was permitted. This analysis permitted some idea of the degree to which the decision-making teams gave credence to the data provided them through the STEEP process and an interesting trend emerged.

Across the three schools in which STEEP was being used in 2003-2004, the team's decision to evaluate a child matched with STEEP findings about 62% of the time (i.e., STEEP recommended do not evaluate and team decided not to evaluate or STEEP recommended evaluation and the team decided to evaluate. Specifically, counting only those children for whom STEEP data had been completed, 9 children were recommended for evaluation by STEEP. All nine of these children were subsequently recommended for evaluation by the decision-making team and eight of these children qualified for services. However, 17 children were not recommended for evaluation at the team decision-making meeting based on STEEP findings, but the teams decided to evaluate ten of these children anyway. Of the ten children who were evaluated, 50% qualified for services. Specifically, three children qualified under SLD, one qualified under Speech and Language Impairment, and one qualified under Other Health Impairment. Recall that the percentage of children qualifying for services prior to STEEP implementation was about 40% at school 1 and 70% at school 2. Hence, the qualify rate for children recommended for evaluation by the decision-making team when the team decided to evaluate when STEEP recommended against evaluation was comparable to baseline rates. The qualify rate for children who were recommended for evaluation by both STEEP and the decision-making team was 89%. In 2004-2005, 106 cases were not recommended for evaluation based on their having had an adequate response to intervention. The team decided to evaluate 14 of these children anyway. Of these 14 children evaluated, 29% of children qualified for services, 64% did not, and 1 case was pending at the completion of this study. Twelve children were recommended for evaluation by STEEP. Of these 12 children, 7 qualified for services, 1 did not qualify, and 4 cases were pending at study completion.

### *Effects on Disproportionate Identification*

*Ethnic disproportionality*. Ethnic minority evaluation was examined in two ways. First, the percentage of children of minority ethnicity who were evaluated at each site was examined relative to the number of minority children who were expected to be evaluated at each site based on base rates alone. These results are shown in Figure 2. Second, the percentage of evaluations that were conducted with children who were of minority ethnicity was examined relative to the number of evaluations that could be expected to be of minority children at each site during each year. These results are shown in Figure 3.

To determine if proportion of identification was approximately correct, expected numbers of evaluations were compared to observed numbers of evaluations by race. Chi-square analyses were performed to determine whether or not there was a significant difference between expected evaluations of students by race and observed (actual) evaluation rate by race with STEEP in 2004-2005 across all schools. A finding of no statistical difference between expected evaluation of minority students and observed evaluation of minority students would be interpreted to support proportionate identification for evaluation by race. Conversely, disproportionate evaluation by race may indicate bias. Given a normal distribution of performance, 26% of students scoring below the 16th percentile would be expected to be minority students (population base rate of minority students was .26). The observed proportion of evaluated minority students (.31) did not differ significantly from the expected proportion of .26, two-tailed $p$ >.01 during baseline years for each site. The observed proportion of evaluated minority students in 2005 (.37) did not differ significantly from the expected proportion of .26, two-tailed $p$> .01. Thus, minority children were not disproportionately identified for evaluation relative to their classmates at baseline or with STEEP implementation.

*Gender disproportionality.* Figure 4 shows the number of males evaluated and placed across all sites and all years. The number of males evaluated and placed was reduced with STEEP relative to baseline. To determine if proportion of identification was approximately correct, expected numbers of evaluations were compared to observed numbers of evaluations by sex for all schools in during baseline years and when STEEP was being implemented. Figure 5 shows these results. Given normal distributions of performance across sex, 50% of students referred by their teachers would be expected to be male. During baseline years, expected proportion of evaluated males (.50) significantly differed from observed proportion of evaluated males (.62), two-tailed $p < .01$ (observed-expected = 15 males). That is, more males were evaluated than would be expected by base rate occurrence of males in the population. Expected proportion of evaluated male cases (.50) did not differ significantly from the observed proportion of evaluated male cases (.59), two-tailed $p > .01$ (observed-expected = X) with STEEP in place.

*Performance differences by ethnicity, gender, SES, and primary language*. Presented in Table 3 are average scores on screening and growth rates in reading and math for children overall and by ethnicity, gender, free or reduced lunch status, and English Language Learner status in 2004-2005 when all children participated in STEEP. The percentage of students identified at each stage of STEEP were also calculated (e.g., percent of students who scored in the bottom 16% during schoolwide screening, percent of students who failed the skill/performance deficit assessment, and percent of students had an inadequate response to intervention). These data are presented in Table 3.

### *Cost-Benefit Analysis for STEEP*

For this analysis, schools 1 and 2 were included for the 2003-2004 school year because baseline data were available for comparison for both schools and STEEP was underway at each site during 2003-2004. At schools 1 and 2, a total of 51 evaluations were conducted during the 2002-2003 school year. In 2003-2004, a total of 16 evaluations were conducted. These numbers permitted a comparison of assessment costs prior to STEEP and assessment costs following STEEP implementation. If full psychoeducational evaluations were valued at $3000 each then total assessment costs in the year preceding STEEP implementation were $153,000. Assessment costs during the first full year of STEEP implementation (2003-2004) were $48,000. At schools 1 and 2, during the 2002-2003 year (prior to STEEP implementation), 29 children were placed in special education. During 2003-2004, 14 children were placed in special education. In the school district in which this study was conducted, the average expenditure per student placed into special education was $5246 per student (computed by total budget divided by number of students served in special education). Hence, placement costs for new students placed in special education were reduced from $152,138.08 in 2002-2003 to $73,556.00 in 2003-2004. Following the 2003-2004 school year, the district dropped four full time equivalent resource teacher positions because of the reduction in newly identified students for special education and observed a district-wide reduction from 6% of children in the district being identified with SLD to 3.5% of children in the district being identified with SLD. The district re-allocated the monies saved and matched them 100% to create a full-time intervention support teacher at each elementary school, 2 middle schools, and 1 high school in the district for the 2004-2005 school year. In 2004-2005, at schools 1 and 2, 9 children were placed in special education indicating that the cost savings were maintained.

Because RTI affects the SLD category most strongly and the SLD category is the disability category for which districts receive very little federal funding to offset the costs of serving these children, reduction of SLD numbers produces compelling cost savings to a district.

The district in which this study was conducted received only $8.57 per SLD-identified student to provide special education services to children in this category; yet, the cost of providing adequate specialized services to these children far exceeded that amount.

*Outcomes for those exposed to STEEP*

Descriptive data are provided in Table 4 for children who ultimately had a failed response to intervention (STEEP +) and for children who had a successful response to intervention (STEEP -). During intervention, the same probe (criterion-level) that had been administered during the classwide screening and performance/skill deficit assessment was re-administered each week by the school psychologist to track growth. The child was not exposed to that particular probe during intervention and the psychologist did not provide feedback when the probe was administered each week. Given the challenge of equating task difficulty across probes, this approach was used to judge intervention response (permitting the use of a single probe at each decision point). If a child scored in the instructional range, the child was determined to have shown an adequate response to intervention. If the child did not score in the instructional range on this particular probe, the child was determined to have had an inadequate response to intervention. Psychologists were asked to continue intervention until (a) the child scored in the instructional range on the grade-level unpracticed probe, or (b) 15 sessions had occurred. In general, the children who were determined to have had an inadequate response to intervention scored lower and grew at a slower pace. They were more likely to be instructed on materials below their current grade placement during intervention (67% of children who did not successfully respond were instructed on below grade level materials whereas 7% who had an adequate response were instructed on below grade level materials). Additionally 41% of children who were determined to have had an adequate response to intervention were detected during classwide screening the following year (2004-2005). The degree to which subsequent detection (following an adequate response to intervention) may indicate additional risk, but this question exceeded the scope of this paper.

## Discussion

This study aimed to evaluate the effects of a RTI approach to screening and eligibility determination (i.e., STEEP) on various outcomes leading up to and including evaluation and placement in special education. The purpose of STEEP was to identify early those students at-risk for academic problems and to attempt to rule out educational or cultural disadvantage, lack of motivation, and lack of instruction as contributors to a student's academic difficulties. STEEP data were presented by the school psychologist as a member of the school-based team to enable teams to more accurately determine who should be referred for evaluation and eligibility determination. This study extends the small but growing literature on RTI in applied school settings. Based upon previous research with STEEP (VanDerHeyden, et al, 2003), we hypothesized that use of STEEP would reduce the number of special education evaluations and improve indicators of disproportionality by increasing decision accuracy.

Fewer evaluations were conducted and evaluated students were more likely to qualify for services when STEEP data were included in the team decision-making process. Whereas baseline data were slightly variable within schools across years, the total initial evaluations and total qualified when STEEP was implemented fell below any data point collected during baseline. Percent of children evaluated who qualified was consistently higher when examining differences for male students and to a lesser degree for female and minority students. Practically, these effects reduce time spent on unnecessary eligibility testing and reduce costs to a district.

Based on results of prior STEEP studies demonstrating more accurate identification of minority and male students with RTI relative to other methods of identification such as teacher referral (VanDerHeyden & Witt, in press), we also expected to find positive effects on proportionate representation of children of minority ethnicity and sex. These expectations were partially realized. The percentage of minority students at each school ranged between 20 and 34 percent. Expected proportions of minority students evaluated could be computed two ways. First the percentage of evaluations that occurred for children of minority ethnicity ranged from 20% to 65% across schools and years. In any given year, one would expect that the percentage of evaluations that were conducted with children of minority ethnicity would roughly match the percentage of children in the school who were of minority ethnicity (e.g., if 34% of children enrolled in the school were of minority ethnicity, then 34% of the evaluations would be expected to be performed with children who were of minority ethnicity). Hence, in some years, before and after STEEP implementation, the proportion of children evaluated who were of minority ethnicity deviated substantially from the expected proportion but no particular pattern emerged. Another way to look at proportionality is to consider the percentage of minority students who were evaluated. If 5% of children in a school are evaluated, then it would be expected that 5% of children irrespective of minority ethnicity would be evaluated. The percentage of minority students evaluated ranged between 2-5% for all schools during baseline years. Thus, there did not appear to be a racial disproportionality problem prior to STEEP and proportions remained at approximately 3% at all schools once STEEP was implemented. Because decision scores become more strongly affected by base rates when the base rates diverge from .50 (Meehl & Rosen, 1955), such analyses may be more relevant with datasets where very small proportions or very large proportions of minority students are enrolled. Indeed others have written about this possibility (Finn, 1982; Gottlieb & Alter, 1994; Harry, 1992). An evaluation of STEEP (VanDerHeyden & Witt, in press) revealed a reduction in minority overrepresentation in schools with higher percentages of minority students and less stringent assessment and curriculum requirements. More research is needed in districts with different percentages of ethnic minorities and curricular programs to further examine the effect of RTI on minority representation in low-performing, high-risk groups and accurate identification of students of minority ethnicity in special education programs (Hosp & Reschly, 2004; VanDerHeyden & Witt, in press).

With respect to gender, the data indicated that the number of males evaluated and placed in special education exceeded the number expected to be evaluated and placed by base rates. Further, STEEP positively affected disproportionate identification of males by reducing the number of children who were evaluated overall and achieving a stronger reduction for males than females. This finding is consistent with previous findings related to the positive affect of RTI data-based decision models on disproportionate identification by sex (VanDerHeyden & Witt, in press) and calls for reform (Holtzman & Messick, 1982).

The effect of STEEP on children of minority ethnicity whose primary language is not English is another important consideration that merits further scrutiny. Sixty-nine percent of students from minority backgrounds were Latino and 17% of these students were provided with ELL services at the time of their evaluations. Prior to STEEP, about half of the evaluated students qualified for services. Testing accuracy increased with STEEP, and 83% of the evaluated Latino students qualified for services when STEEP was introduced. Interestingly, there were no ELL students evaluated when STEEP was used. Due to the small population of Latino students at each school in this study and the extreme variation in language experiences among these students, a more in depth analysis of the performance of ELL children during screening activities relative to their peers and progress over time was needed and exceeded the scope of this paper.

It is important to emphasize that effects on evaluations reflect conservative findings for several reasons. First, *all* evaluations for classifications were included in these analyses due to the lack of reliability and validity in classification categories. That is, students who qualified due to speech, cognitive, or behavior problems in addition to academic concerns were included in the numbers of initial evaluations. The inclusion of all categories helps to mitigate the possible confound of teams classifying children who were responsive to intervention under categories other than SLD (e.g., SLI, ED). This approach could cause the number of students qualifying for SLD to decrease but produce a simultaneous increase in the number of students qualifying under other categories. Moreover, because an adverse impact on educational performance is an important indicator of the need for special education services, over-identification of students under any category can result from insufficient academic assessment of prior instruction and motivation. Because overall evaluations and qualifications decreased, these results may suggest that successful responsiveness assessments could potentially reduce the number of students who receive special education services. Following only one year of STEEP implementation, SLD diagnosis decreased from 6% of elementary school children to 3.5% of elementary school children district-wide. The cost-benefit analyses presented indicate that resources devoted to traditional assessment are consequently reduced and replaced by direct assessment, intervention, and consultation services in classrooms. Qualification is certainly not a "perfect" standard indicating the presence or absence of a disability. It could be considered a conservative measure because final qualification decisions did not always correspond with data collected for special education eligibility assessment. For example, a review of evaluation reports of students whose STEEP data indicated that no evaluation was warranted revealed that at least two of the five students who qualified for services in 2003-2004 did not meet the technical requirements for SLD, a finding that has been frequently reported elsewhere (Macmillan, 1998; Macmillan & Speece, 1999). Whereas the trend of overidentification (as indicated by overreferral for evaluation) continued at the team decision-making level, fewer children were evaluated because fewer children were discussed by the decision-making team. Hence, the effect was truly a pre-referral effect on overidentification in general and disproportionate identification by sex. Whether or not children were evaluated and qualified for services are not “pure” dependent measures. Functionally, they are “messy” with many factors affecting whether or not evaluation or qualification occurs. However, they were selected as the primary dependent measures for this study because they reflect the “diagnostic realities” that exist in schools. That is, these dependent measures were selected because they were strongly linked to outcomes for children (i.e., placement into special education) were functionally meaningful, consistent with the values prompting the research in the first place, and considered to be reflective of “real” change in the system.

Research has yet to determine which set of procedures paired with what set of decision rules and measurement technologies will best identify children for specialized assistance. Part of the challenge in answering these questions requires articulating what the characteristics of the resulting group of non-responders should be (e.g., likely to not acquire functional skills without special assistance, “true LD,” requiring resources that are too cumbersome for general education to provide but are effective at promoting learning when used). Articulating this goal also requires identifying what purpose RTI models are intended to serve in schools for which there are many possibilities (VanDerHeyden, Witt, & Barnett, in submission). With STEEP, teams were simply presented with students whose performances fell below the criterion at each stage of assessment (screening, classwide intervention, performance/skill deficit assessment, and individual intervention) and resulted in identification of about 3% of the population as ultimately being detected as at-risk by the STEEP screening. Fuchs, Fuchs, & Compton (2004), however, identified benchmark criteria as a criterion that identifies too many nonresponders when applied. Slope of high frequency words or reading passage words read correctly per minute more consistently discriminated non-responders from responders to intervention on various types of reading assessments than other criteria including a benchmark criterion, post-intervention standardized test scores, and slope for nonsense words (Fuchs, et al., 2004). To date, no study has reported the accuracy and utility of various responsiveness criteria with children in older grades or in any grade level for math. However, the present data are encouraging in that a relatively simple criterion used at four separate stages was effective in reducing total number of evaluations.

### *Practical Implications*

In addition to the criteria used to judge intervention responsiveness, other variables that control the intervention response include variables related to the intervention itself. Future research is needed to examine whether the responses obtained with STEEP given relatively simple, short term intervention as part of a larger package of scripted assessment procedures, is meaningfully related to child outcomes and replicable in sites with other characteristics (e.g., weaker core instructional procedures), particularly in light of reports of higher nonresponsiveness rates given much more intensive and individualized intervention (Torgeson, et al., 2001). Two differences in intervention are worth highlighting. Interventions conducted as part of STEEP were conducted in the student's regular classroom with the regular classroom teacher. Second, intervention integrity was monitored. One could hypothesize that each of these variables could account for enhanced intervention effectiveness over time. Given the diversity of teacher characteristics, achieving adequate intervention integrity required more than a didactic "train-and-hope" (Stokes & Baer, 1977) approach to teacher support. When treatment implementation was not sustained after training, feedback with retraining was effective to re-establish correct intervention implementation.

The return to baseline levels for dependent measures when a STEEP trained school psychologist was removed from the school highlights the importance of support and training for the system to maintain RTI effects. The school psychologists coordinated STEEP activities at the schools and presented the data to the decision-making team when a child was referred for that team's consideration. With STEEP, the psychologists spent less time engaged in traditional assessment related to eligibility and placement considerations and more time engaged in direct functional academic assessment in the classroom, consultation with teachers and principals, and translating data to inform instructional practices in the classroom and identification practices at the schools. The data obtained with the reversal of conditions at school 1 illustrates the pivotal role of the psychologist in assisting the team to consider data when reaching decisions about individual student progress and whether or not to refer for evaluation. Even when STEEP data were available for individual children, the team did not consider those data in making a decision when the untrained psychologist was present. Without a trained school psychologist in place, the implementation of program procedures and use of STEEP data did not generalize to other decision-makers.

One disappointing finding that may have important practical implications of RTI effectiveness in applied settings was the degree to which the team followed the available STEEP data. The effectiveness of any RTI program will rely on decisions based on interpretations of data. Improvements in reducing the number of children who are exposed to the school-based team and decision-making process produced improved accurate evaluation testing results. However, when STEEP results were reviewed by the school-based preferral teams, 67% of students who had a successful response to intervention were recommended for full psychoeducational evaluation despite the data. Alternatively, teams evaluated 100% of the students when STEEP data suggested additional testing for students during the first year of STEEP implementation. The lack of correspondence between the team's decision and assessment data is consistent with previous findings (Macmillan, 1998; Macmillan & Speece, 1999). Because RTI relies on data-based decisions to improve outcomes, investigations of extraneous factors influencing team decisions are important lines of future research. During the second year of STEEP implementation, only 13% of children who had an adequate response to intervention were referred for evaluation and 92% of children who did not have an adequate response to intervention were referred for evaluation.

### *Limitations*

Several potential limitations of these findings are worth noting. Order of STEEP implementation across sites was not randomly determined. Schools were given an opportunity to volunteer and the first two schools to volunteer were the schools where STEEP was first implemented. These two schools also had the highest number of evaluations, which perhaps accounted for their interest in participating. STEEP was implemented at school 3 without conducting a baseline year first. School 3 opened with the psychologist from school 1 and an assistant principal from school 1 taking the principal position at school 3, and these individuals wished to open school 3 using the problem-solving model with which they had become familiar. Whereas data were collected in different schools, the discussions of STEEP procedures and effects within the district may have interfered with baseline data collected in schools in prolonged baselines and may have underestimated the effect of STEEP procedures on traditional team evaluation decisions. Replication of the effect across schools is a strength of the study as each school had different demographic characteristics. The reversal at school 1 also served to further verify the effect of STEEP when dependent measures returned to baseline levels when the school psychologist trained to facilitate STEEP procedures was removed.

However, the rate calculation for the reversal to baseline in school 1 may not have been a fair estimate of what the total numbers would have been for an entire year for each psychologist (e.g., the rate may not have been constant for all 10 months). Without monitoring of initial evaluations per month, it was not possible to determine any distinct trends in high assessment times throughout the year which may have accounted for the high initial rates calculated during the reversal to baseline conditions.

Because schools were evaluated in one district, these results may not generalize to other districts with different demographic characteristics or that does not provide the strong district administration support that was given to implement STEEP in this study. Moreover, only a few years were included in this project. Interestingly, these results differed from Minnesota's 10-year study (Marston, Muyskens, Lau, & Canter, 2003) which found no differences over time in the count of students qualified for special education. Thus, these findings require validation with larger samples with additional longitudinal data to further investigate long-term outcomes.

Research has yet to sort out how response to intervention will best be implemented in schools. Emerging from largely grass roots efforts in behavior analysis, curriculum-based assessment and measurement, and functional academic assessment (e.g., brief experimental analysis), RTI may have many futures. The advantage of this evolution or iterative process is that many "models" might emerge and over time evolve for greater effectiveness and efficiency for children. Fuchs, Mock, Morgan, and Young (2003) recognized two potential types of models, standard protocol approaches and problem-solving models, but more variations are likely. Will

RTI function as a screening approach that informs the team's decision to refer for a psychoeducational evaluation? If so, what will the additional components of a psychoeducational evaluation be? Will RTI evolve into a full eligibility approach typified by the Heartland model of problem-solving (Tilly, et al., 1999) or the dual discrepancy or treatment validity model described by Fuchs and Fuchs (1998)? Will RTI operate primarily in general education (Speece, Case, & Molloy, 2003) or somewhere in between general and special education, for example, in tracking supplemental services provided to at-risk students (Vaughn, Linan-Thompson, & Hickman, 2003)? Each approach necessarily requires different decision criteria, cut-scores, and results in different numbers and types of children served (Fuchs, 2003). How will research in brief experimental analyses of academic responding (Daly, Martens, Hamler, Dool, & Eckert, 1999) combined with basic research in how to promote robust and functional skill sets informed by the effective teaching literature inform existing (or evolve into new) approaches to measuring and judging intervention responsiveness? Many futures of RTI are possible, including a vulnerable future if empiricism slows. In addition to the operational variables that merit investigation, examining the technical adequacy of RTI which involves sequenced procedures and correct application of sequenced decision rules to reach defensible conclusions, will present new challenges (Barnett, VanDerHeyden, & Witt, in submission; Barnett, Daly, Jones, & Lentz, 2004; Hintze, Owen, Shapiro, & Daly, E., 2000).

To whatever new horizons research in RTI leads, the potential for assessment and intervention science to grow in ways that positively affect student outcomes is exciting. We believe critical components of evolved RTI for decision-making must include a keen focus on efficiency and parsimony. There are certainly more complicated ways than less complicated ways to solve problems, but complicated methods are not likely to be implemented or implemented with integrity in schools with many competing responsibilities, demands, and contingencies that often do not support correct implementation of intervention in classrooms. In politically charged environments such as has often been the case in education, empiricism has much to offer as a vehicle for evaluating the utility of what will surely be different applications in evolving models of identification, service provision, and outcome analysis.

# References

Ardoin, S., Witt, J., Connell, J., & Koenig, J. (submitted for publication). Application of a three-tiered response to intervention model for instructional planning, decision making, and the identification of children in need of services.

Barnett, D. W., Daly, E. J. III, Jones, K. M., & Lentz, F. E. Jr. (2004). Response to intervention: Empirically-based special service decisions from increasing and decreasing intensity single case designs. *The Journal of Special Education, 38*, 66-79.

Barnett, D., VanDerHeyden, A. M., & Witt, J.C. (in submission). Achieving science-based practice through response to intervention: What it might look like in preschools. Manuscript submitted for publication.

Daly, E.J.,III, Martens, B.K., Hamler, K.R., Dool, E.J., & Eckert, T.L. (1999). A brief experimental analysis for identifying instructional components needed to improve oral reading fluency. *Journal of Applied Behavior Analysis,* 32, 83-94.

Donovan, S. & Cross, C., eds. (2002). *Minority Students in Special and Gifted Education*, Washington, DC: National Academy Press.

Fuchs, L., & Fuchs, D. (1997). Use of curriculum-based measurement in identifying students with disabilities. *Focus on Exceptional Children, 30*, 1-16.

Fuchs, L., & Fuchs, D. (1998). Treatment validity: A unifying concept for reconceptualizing the identification of learning disabilities. *Learning Disabilities Research & Practice, 13,* 204-219.

Fuchs, D., Fuchs, L. S., Mathes, P. G., & Simmons, D. C. (1997). Peer-assisted learning strategies: Making classrooms more responsive to diversity. *American Educational Research Journal, 34*, 174-206.

Greenwood, C. R. (1991). Longitudinal analysis of time, engagement, and achievement in at-risk versus non-risk students. *Exceptional Children*, 57, 521-535.

Gresham, F. M. (2001). *Responsiveness to intervention: An alternative approach to the identification of learning disabilities (2nd draft)*. Paper presented for the OSEP Learning Disabilities Initiative, Office of Special Education Programs, U. S. Department of Education, Washington, DC.

Gresham, F., VanDerHeyden, A., & Witt, J. (in press). Response to Intervention in the Identification of Learning Disabilities: Empirical Support and Future Challenges. Empirical Support and Future Challenges. *School Psychology Review.*

Hintze, J. M., Owen, S. V., Shapiro, E. S., & Daly, E. J. (2000). Generalizability of oral reading fluency measures: Application of G theory to curriculum-based measurement. *School Psychology Quarterly, 15*, 52-68.

Marston, D., Muyskens, P., Lau, M., & Canter, A. (2003). Problem-solving model for decision-making with high-incidence disabilities: The Minneapolis experience. *Learning Disabilities Research and Practice, 18,* 187-200.

Noell, G., Witt, J., Slider, N., Connell, J., Gatti, S., Williams, K., Koenig, J., & Resetar J. (2005). Treatment implementation following behavioral consultation in schools: A comparison of three follow-up strategies. *School Psychology Review, 34,* 87-106.

Tilly, W.D., Reschly, D.J., & Grimes, J. (1999). Disability determination in problem solving systems: Conceptual foundations and critical components. In D. Reschly, W.D. Tilly, & Grimes (Eds.), *Special education in transition: Functional and noncategorical programming* (pp. 285-301). Longmont, CO: Sopris West.

Torgesen, J., Alexander, A., Wagner, R., Rashotte, C., Voeller, K., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities, 34*, 33-58.

U.S. Department of Education. (1998). *To assure the free appropriate public education of all children with disabilities: Twentieth annual report to Congress on the implementation of the Individuals with Disabilities Education Act*. Washington, DC: U.S. Government Printing Office.

VanDerHeyden, A. M. & Witt, J. C. (in press). Quantifying the context of assessment: Capturing the effect of base rates on screening accuracy. *School Psychology Review.*

VanDerHeyden, A. M., Witt, J. C., & Barnett, D. W. (in submission). The emergence and possible futures of response to intervention. Manuscript accepted with revisions. *Journal of Psychoeducational Assessment*.

VanDerHeyden, A. M., Witt, J. C., & Naquin, G. (2003). The development and validation of a process for screening and referrals to special education. *School Psychology Review, 32,* 204-227.

Vellutino, F. R., Scanlon, D. M., & Tanzman, M. S. (1998). The case for early intervention in diagnosing specific reading disability. *Journal of School Psychology, 36,* 367-397.

Wong, H. K., & Wong, R. T. (1998). How to be an effective teacher: The first days of school. Mountain View, CA: Harry K. Wong Publications, Inc.

Ysseldyke, J. E., Pianta, R., Christenson, S., Wang, J., & Algozzine, B. (1983). An analysis of pre-referral interventions. *Psychology in the Schools, 20,* 184-190.

Ysseldyke, J. E., Vanderwood, M. L., & Shriner, J. (1997). Changes over the past decade in special education referral to placement probability: An incredibly reliable practice. *Diagnostique, 23*, 193-201.

Table 1

*Demographics of Elementary Schools in District*

| | | School 1 | | School 2 | | School 3 | | School 4 | | School 5 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 2001-2002 | 2003-2004 | 2001-2002 | 2003-2004 | 2001-2002 | 2003-2004 | 2001-2002 | 2003-2004 | 2001-2002 | 2003-2004 |
| Total Number of Children | | 706 | 781 | 638 | 583 | --- | 595 | 647 | 562 | 586 | 580 |
| Race | Caucasian | 71% | 74% | 81% | 76% | --- | 75% | 67% | 70% | 67% | 71% |
| | Hispanic | 18% | 19% | 15% | 17% | --- | 21% | 24% | 21% | 22% | 21% |
| | African American | 6% | 4% | 3% | 4% | --- | 2% | 5% | 4% | 6% | 5% |
| | Other | 4% | 3% | 2% | 3% | --- | 2% | 5% | 5% | 4% | 3% |
| Sex | Male | 52% | 52% | 52% | 52% | --- | 51% | 53% | 55% | 53% | 51% |
| Free Lunch | | 16% | 14% | 37% | 28% | --- | 26% | 18% | 21% | 22% | 19% |
| Mean SAT-9 Percentile Grades 2-5 | Reading | 73 | 71 | 62 | 60 | --- | 62 | 62 | 73 | 58 | 62 |
| | Math | 76 | 80 | 60 | 70 | --- | 66 | 67 | 82 | 62 | 73 |

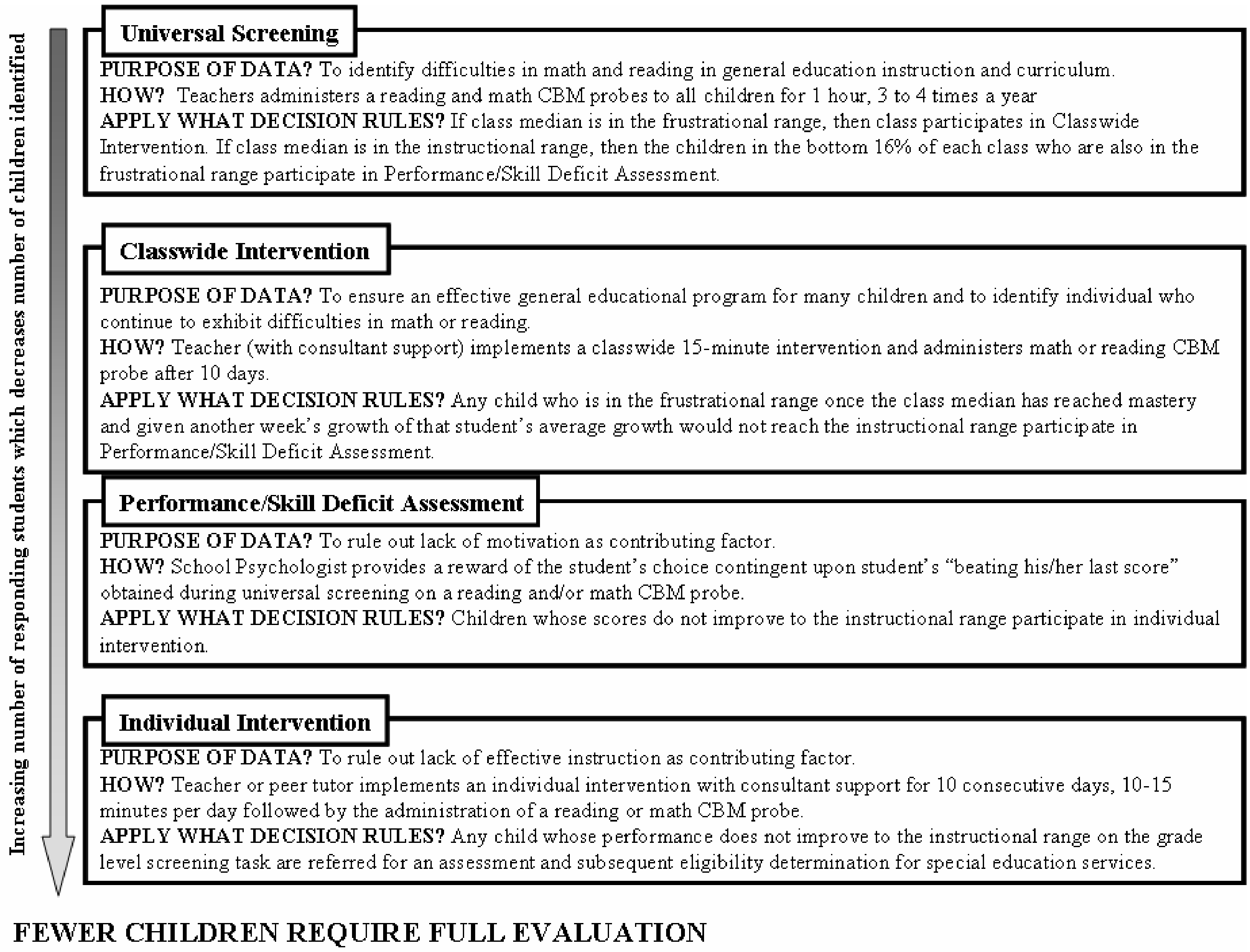
| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Language Arts | 69 | 66 | 57 | 54 | --- | 54 | 57 | 67 | 56 | 58 |
| ELL | | 3% | 3% | | 0% | --- | 4% | 0% | <1% | 4% | 4% |
| Special Education | Total | 11.6% | 10.8% | 15.4% | 18.0% | -- | 11.4% | 12.1% | 11.2% | 13.5% | 14.5% |
| | SLD | 5.8% | 3.7% | 4.7% | 3.6% | -- | 2.5% | 6.5% | 3.3% | 6.7% | 3.4% |

*The opening of an additional school in 2003-2004 resulted in reductions in percentages of children served across all sites.

All estimates were obtained from the census data provided to the Office of Civil Rights.

Table 2

*Summary of Decision Rules Applied at Each Tier*



Increasing number of responding students which decreases number of children identified

**Universal Screening**

**PURPOSE OF DATA?** To identify difficulties in math and reading in general education instruction and curriculum.

**HOW?** Teachers administers a reading and math CBM probes to all children for 1 hour, 3 to 4 times a year

**APPLY WHAT DECISION RULES?** If class median is in the frustrational range, then class participates in Classwide Intervention. If class median is in the instructional range, then the children in the bottom 16% of each class who are also in the frustrational range participate in Performance/Skill Deficit Assessment.

**Classwide Intervention**

**PURPOSE OF DATA?** To ensure an effective general educational program for many children and to identify individual who continue to exhibit difficulties in math or reading.

**HOW?** Teacher (with consultant support) implements a classwide 15-minute intervention and administers math or reading CBM probe after 10 days.

**APPLY WHAT DECISION RULES?** Any child who is in the frustrational range once the class median has reached mastery and given another week's growth of that student's average growth would not reach the instructional range participate in Performance/Skill Deficit Assessment.

**Performance/Skill Deficit Assessment**

**PURPOSE OF DATA?** To rule out lack of motivation as contributing factor.

**HOW?** School Psychologist provides a reward of the student's choice contingent upon student's "beating his/her last score" obtained during universal screening on a reading and/or math CBM probe.

**APPLY WHAT DECISION RULES?** Children whose scores do not improve to the instructional range participate in individual intervention.

**Individual Intervention**

**PURPOSE OF DATA?** To rule out lack of effective instruction as contributing factor.

**HOW?** Teacher or peer tutor implements an individual intervention with consultant support for 10 consecutive days, 10-15 minutes per day followed by the administration of a reading or math CBM probe.

**APPLY WHAT DECISION RULES?** Any child whose performance does not improve to the instructional range on the grade level screening task are referred for an assessment and subsequent eligibility determination for special education services.

**FEWER CHILDREN REQUIRE FULL EVALUATION**

Table 3

*Performance Differences and Identification Overall and by Ethnicity, Gender, SES, and Primary Language in 2004-2005*

| | | | Total | Male | Ethnic Minority | Free or Reduced Lunch | ELL |
|---|---|---|---|---|---|---|---|
| CBM Data | Mean Spring level WC/Min | 1st-2nd | 80<br>(SD=33) | 72<br>(SD=31) | 82<br>(SD=23) | 66<br>(SD=54) | 60<br>(SD=21) |
| | | 3rd-5th | 127<br>(SD= 34.9) | 125<br>(SD= 35) | 124<br>(SD=22) | 119<br>(SD=31) | 107<br>(SD=18) |
| | Mean Spring level DC/Min | 1st-3rd | 38<br>(SD= 14) | 37<br>(SD= 21) | 39<br>(SD= 9) | 36<br>(SD= 13) | 37<br>(SD= 10) |
| | | 4th-5th | 92<br>(SD= 24) | 88<br>(SD= 24) | 87<br>(SD= 15) | 79<br>(SD= 24) | 48<br>(SD= 16) |
| | Mean Spring Growth WC/Min | 1st-2nd | 0.8<br>(SD= 0.9) | 0.7<br>(SD= 0.9) | 2.8<br>(SD= 1.4) | 0.7<br>(SD= 0.5) | 0.6<br>(SD= 0.3) |
| | | 3rd-5th | 0.05<br>(SD= 0.7) | 0.5<br>(SD= 0.6) | 0.5<br>(SD= 0.8) | 0.5<br>(SD= 2.0) | 0.5<br>(SD= 0.1) |
| | Mean Spring Growth DC/Min | 1st-2nd | 0.7<br>(SD= 0.8) | 0.7<br>(SD= 0.8) | 2.6<br>(SD= 1.1) | 0.6<br>(SD= 0.5) | 1<br>(SD= 0.3) |
| | | 3rd-5th | 1.7<br>(SD= 1.5) | 1.4<br>(SD= 1.4) | 2.8<br>(SD= 1.9) | 1.5<br>(SD= 0.8) | 1.5<br>(SD= 0.8) |

| | | | | |
|---|---|---|---|---|
| STEEP | At-risk during universal screening (Tier 1) | 10% | | |
| | At-Risk Skill/ Performance Deficit Assessment (Tier 2) | 104 | 50 | 38 |
| | At-Risk Individual Intervention (Tier 3) | 14 | 5 | 5 |
| | Number evaluated | 37 | | |

Note. Interpretation is provided in each cell.

Table 4

*Outcome Data for Children who have a Successful RTI versus Unsuccessful RTI*

| | | 2003-2004 | 2003-2004 | 2004-2005 | 2003-2004 |
|---|---|---|---|---|---|
| | Subject and Grades | Inadequate Response to Intervention | Adequate Response to Intervention | Inadequate Response to Intervention | Adequate Response to Intervention |
| Percentage of Cases | | 13 inadequate responses | 68 adequate responses | 4 inadequate responses | 39 adequate responses |
| | | 81 Interventions<br>1591 children screened | | 43 Interventions<br>3101 children screened | |
| Mean In-Class Screening Score | Reading (wc/min) | | | | |
| | 1 to 2 | 0 (n=1) | 27.05 (SD=6.69; n=20) | No cases | 20.12 (SD=5.97; n=17) |
| | 3 to 5 | 21.69 (n=9; SD=9.14) | 51.09 (n=36; SD=8.75) | 27.25 (n=4 ; SD=15.91) | 52.38 (SD=10.66; n=19) |
| | Math (dc/2 min) | | | | |
| | 1 to 3 | 13 (n=1) | 12.25 (n=7; SD= 3.65) | No cases | No cases |
| | 4 to 5 | 13 (n=1) | No cases | No cases | No cases |
| Mean Score with Incentives | Reading (wc/min) | 2 (n=1) | 36.1 (n=20; SD= 8.34) | No cases | No cases |
| | 1 to 2 | 35.35 (n=10; SD=14.78) | 60.55 (n= 37; SD= 8.58) | 30.25 (n=4; SD= 13.08) | 25.19 (n = 17; SD = 5.93) |
| | 3 to 5 | No cases | No cases | No cases | 57.88 (n= 19; SD = 11.41) |
| | Math (dc/2 min) | | | | |
| | 1 to 3 | 9 (n=1) | 13.59 (n=7; SD=4.74) | No cases | No cases |
| | 4 to 5 | 13 (n=1) | No cases | No cases | No cases |
| Mean Final Criterion | Reading (wc/min) | | | | |

| | | | | | |
|---|---|---|---|---|---|
| Score | 1 to 2 | 4 (n=1) | 57.55 (SD=12.05; n=20) | No cases | 46.18 (n=17; SD=11.10) |
| | 3 to 5 | 38.22 ( n=10; SD=9.96) | 79.47 (SD= 8.58; n=37) | 39.75 (n=4; SD=24.40) | 85.25 (SD=10.28; n=19) |
| | Math (dc/2 min) | | | | |
| | 1 to 3 | 12 (n=1) | 26.42 (n=7; SD= 3.91) | No cases | No cases |
| | 4 to 5 | 0 (n=1) | No cases | No cases | No cases |
| Mean Score First Intervention Session (Instructional Level Task) | Reading (wc/min) | | | | |
| | 1 to 2 | 17 (n=1) | 45.95 (n=20; SD 17.19) | No cases | 39.80 (SD=12.33; n=17) |
| | 3 to 5 | 41.5 (n=10; SD=9.14) | 80.52 (n=37; SD=14.75) | 1.75 (n=4; SD=14.49) | 80.74 (SD=24.78; n=19) |
| | Math (dc/2 min) | | | | |
| | 1 to 3 | 32 (n=1) | 21.34 (n=7 SD=2.84) | No cases | No cases |
| | 4 to 5 | 32 ( n=1) | No cases | No cases | No cases |
| Mean Intervention Slope | Reading (wc/min/session) | | | | |
| | 1 to 2 | -.05 (n = 1) | .59 (SD= 4.79; n=20) | | .49 (n=17; SD=2.53) |
| | 3 to 5 | .69 (n=10; SD=1.00) | .77 (SD=2.97; n=37) | -.37 (n=4; SD= 3.42) | 4.62 (n=19; SD= 14.62) |
| | Math (dc/2 min/session) | | | | |
| | 1 to 3 | -.18 | No cases | No cases | No cases |
| | 4 to 5 | 11.3 | 1.2 (n=7 SD=1.18) | No cases | No cases |
| Mean Generalization Slope | Reading (wc/min/session) | | | | |
| | 1 to 2 | .25 (n=1) | 3.97 (n=20; SD= 2.31) | | 5.00 (n=17; SD= 3.46) |
| | 3-5 | .22 (n=10; SD=.81) | 3.45 (n=37; SD= 2.35) | .44 (n =4 ; SD = .93) | 5.60 (n=19; SD=2.85) |
| | Math (dc/2min/session) | | | | |
| | 1 to 3 | .43 | 2.49 (SD=1.05 n=7) | No cases | No cases |
| | 4 to 5 | -4.33 | No cases | No cases | No cases |

| | | | | | |
|---|---|---|---|---|---|
| Detected in Subsequent Screenings | | | 28 | | |
| SAT-9 Score | Reading | 567.64 | 572.53 | | |
| | Math | 569.73 | 589.19 | | |
| | Language Arts | 553 | 559.79 | | |

Figure Captions

*Figure 1.* Total number of initial evaluations and total number of students who qualified for services during baseline and STEEP implementation conditions at each participating school during each school year.

*Figure 2.* Percentage of children of minority ethnicity who were evaluated at each site relative to the number of minority children who were expected to be evaluated at each site during each baseline and STEEP implementation school year.

Figure 3. Percentage of evaluations that were conducted with children who were of minority ethnicity was examined relative to the number of evaluations that could be expected to be of minority children at each site during each baseline and STEEP implementation school year.

*Figure 4.* Number of male student initial evaluations and total number of male students who qualified for services during baseline and STEEP implementation conditions at each participating school during a school year.

*Figure 5.* Percent of expected numbers of male evaluations compared to observed percent of male students evaluated in each school during each baseline and STEEP implementation school year.

Figure 1

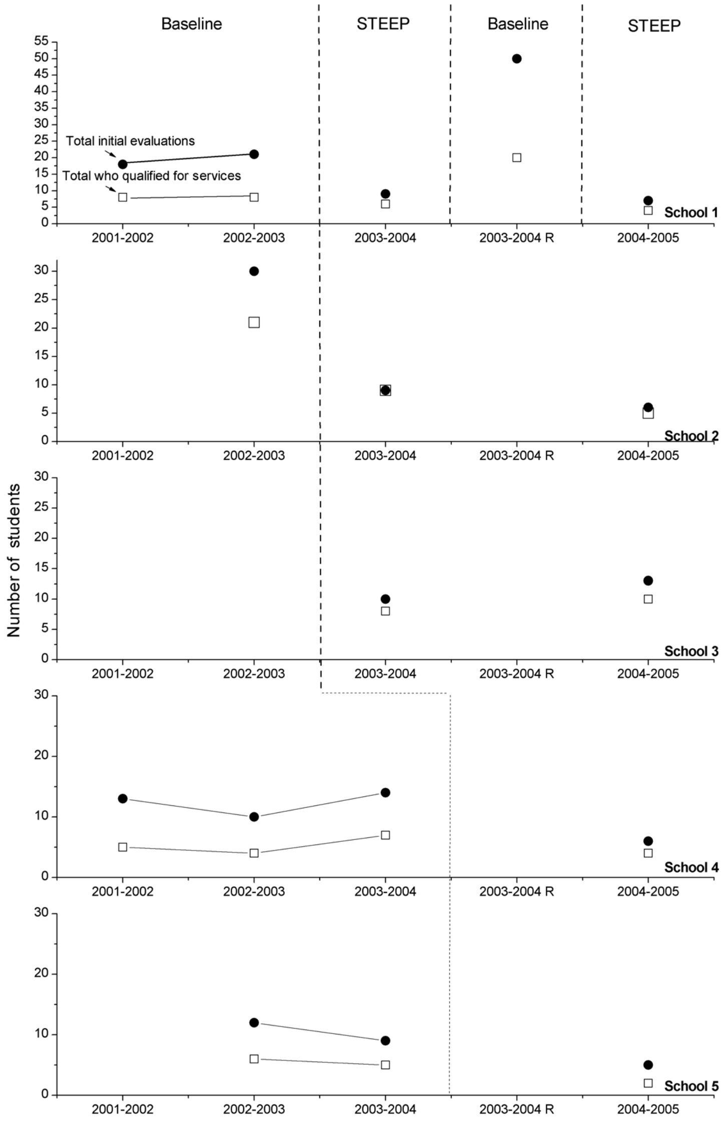

Baseline
STEEP
Baseline
STEEP
Total initial evaluations
Total who qualified for services
Number of students
School 1
School 2
School 3
School 4
School 5
0
5
10
15
20
25
30
35
40
45
50
55
2001-2002
2002-2003
2003-2004
2003-2004 R
2004-2005

Figure 2

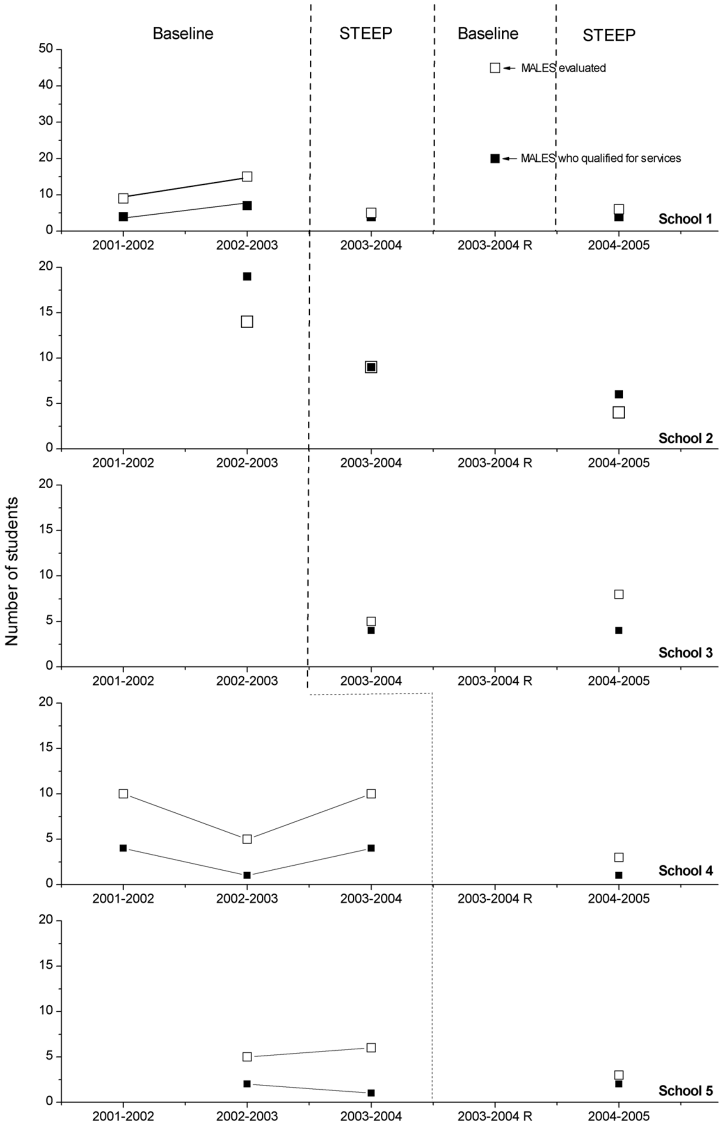

Baseline
STEEP
Baseline
STEEP
□ MALES evaluated
■ MALES who qualified for services
Number of students
School 1
School 2
School 3
School 4
School 5
50
40
30
20
15
10
5
0
2001-2002
2002-2003
2003-2004
2003-2004 R
2004-2005

Figure 3

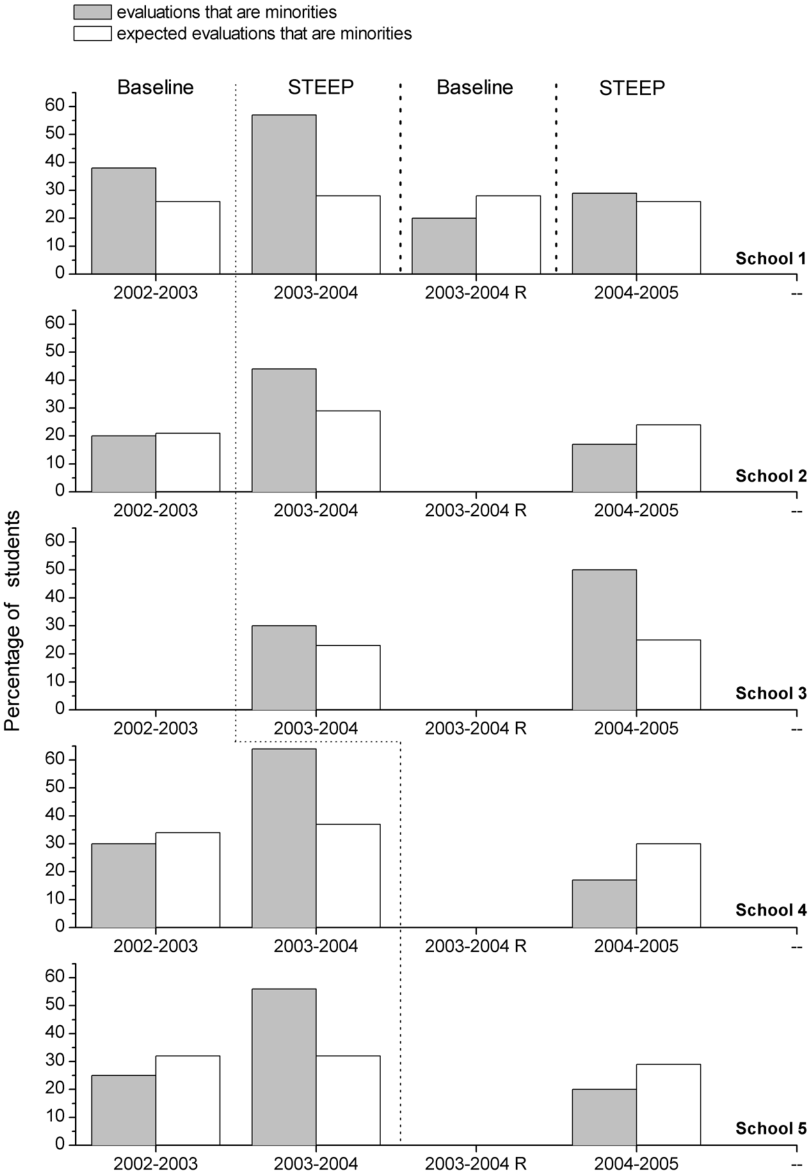

evaluations that are minorities
expected evaluations that are minorities
Baseline
STEEP
Baseline
STEEP
Percentage of students
0
10
20
30
40
50
60
School 1
School 2
School 3
School 4
School 5
2002-2003
2003-2004
2003-2004 R
2004-2005
--

Figure 4

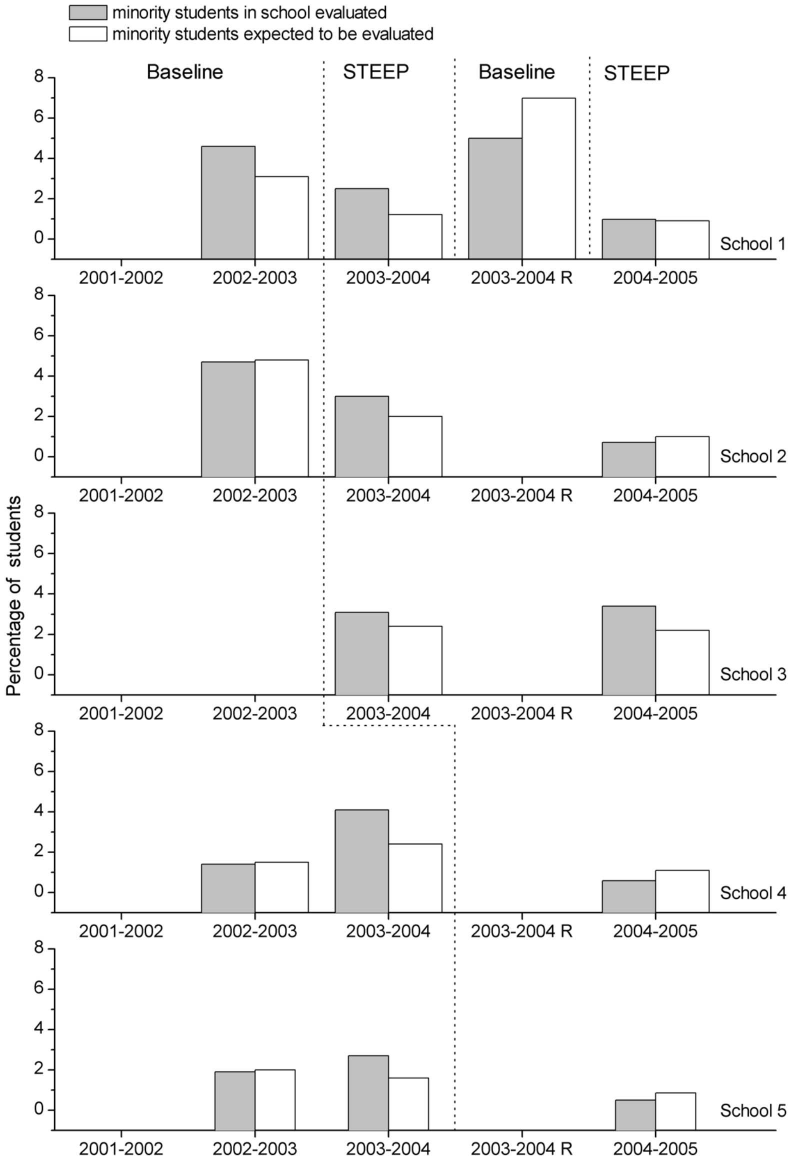
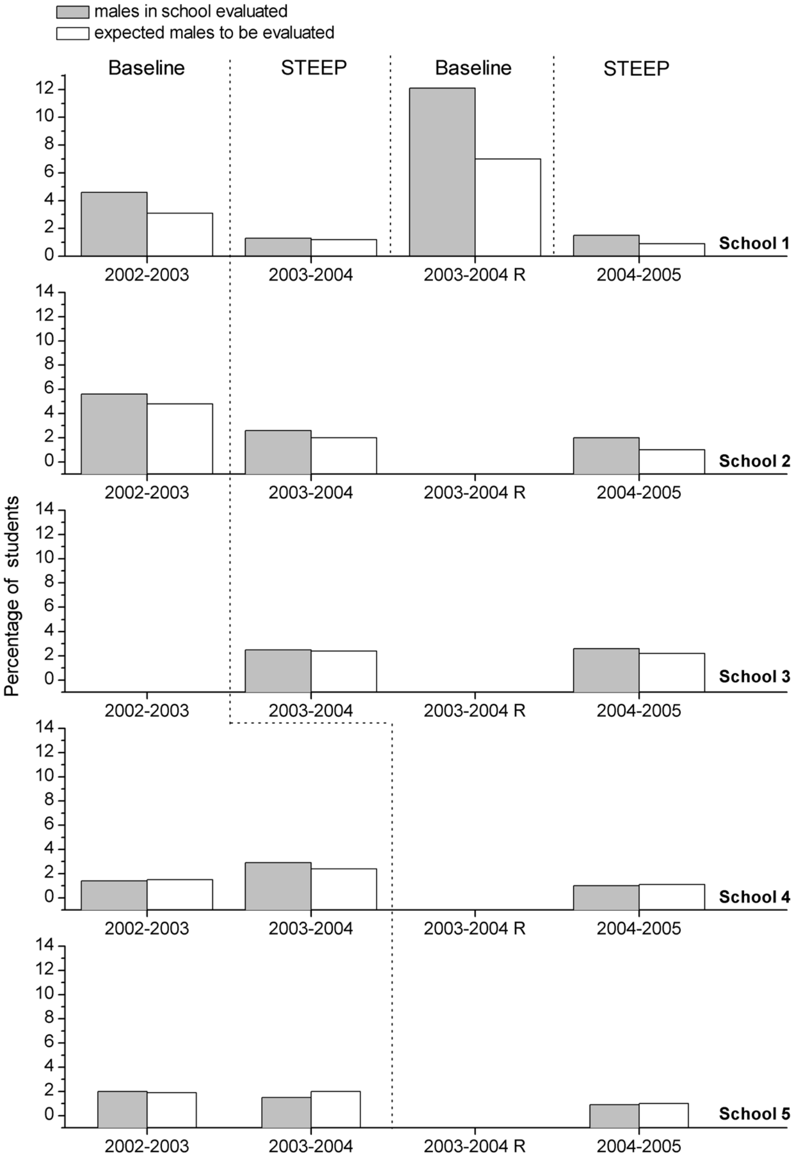

minority students in school evaluated
minority students expected to be evaluated
Baseline
STEEP
Baseline
STEEP
Percentage of students
0
2
4
6
8
2001-2002
2002-2003
2003-2004
2003-2004 R
2004-2005
School 1
School 2
School 3
School 4
School 5

Figure 5


males in school evaluated
expected males to be evaluated
Baseline
STEEP
Baseline
STEEP
Percentage of students
0
2
4
6
8
10
12
14
School 1
School 2
School 3
School 4
School 5
2002-2003
2003-2004
2003-2004 R
2004-2005