ELSEVIER

Journal of School Psychology
45 (2007) 225–256

Journal of
School
Psychology

A multi-year evaluation of the effects of a Response to Intervention (RTI) model on identification of children for special education ☆

Amanda M. VanDerHeyden [a,*], Joseph C. Witt [b], Donna Gilbertson [c]

[a] *University of California at Santa Barbara, United States*
[b] *Louisiana State University, United States*
[c] *Utah State University, United States*

Received 5 July 2005; received in revised form 1 June 2006; accepted 2 November 2006

Abstract

The purpose of this study was to examine the effects of implementation of a systematic response to intervention (RTI) model on the identification and evaluation of children for special education. Using a multiple baseline design, a systematic model of assessment and intervention was introduced in consecutive years for all elementary schools ($N=5$) in the district. Effect of the RTI model on number of evaluations conducted, percentage of evaluated children who qualified for services, and proportion of identified children by sex and ethnicity before and after implementation of the 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 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 assessment and intervention procedures, decision rules, and schoolwide training methods are described in detail and practical implications are discussed.

Keywords: Response to intervention; Early intervention; Problem solving model; Standard protocol model

☆ 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.

* Corresponding author.
E-mail address: amanda@education.ucsb.edu (A.M. VanDerHeyden).

0022-4405/$ - see front matter © 2006 Society for the Study of School Psychology. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.jsp.2006.11.004

Response to Intervention (RTI) refers to a particular criterion for decision-making and does not denote a particular set of procedures (Christ, Burns, & Ysseldyke, 2005). Several types of procedures have been developed and studied that generate datasets upon which RTI judgments can be made. The basic concept of RTI is that when provided with effective intervention, a student can be determined to have responded or not responded adequately to that intervention and such information can be used to guide service delivery decisions. RTI requires that teams make a series of data-based decisions. Frequently this decision making is facilitated by the problem-solving model of assessment. Problem-solving models evolved from the work of curriculum-based measurement (CBM) researchers who sought to develop systems of decision-making that would promote effective use of the data collected through CBM and enhance outcomes for children. Problem-solving models of assessment have been implemented widely in many states with promising results including Iowa (Tilly, 2003) and the Minneapolis public schools (Marston, Muyskens, Lau, & Canter, 2003).

One challenge with many of the procedural models of RTI is that they are not merely one activity. Instead, RTI decisions are made based upon a process consisting of an integrated set of tools, procedures, and decisions (VanDerHeyden, Witt, & Barnett, 2005). To utilize the problem-solving model, the school-based team must 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 problem-solving and RTI decision-making point to a large and growing body of research supporting the various components of RTI models. Clearly this research has provided evidence to guide the series of decisions about 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 models. First, implementing RTI means implementing an *integrated set* of procedures or components while correctly applying sequenced decision rules (Barnett, Daly, Jones, & Lentz, 2004; VanDerHeyden et al., 2005). The research conducted to date with few exceptions (Gravois & Rosenfield, 2002) 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 effective must also be addressed. The second issue is that most of the research has been conducted by well-funded research centers. 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 to do that job* (Torgesen et al., 2001; Vaughn, Linan-Thompson, & Hickman, 2003; Vellutino, Scanlon, & Tanzman, 1998). The question is whether these components can be effective when implemented by front line educational professionals. Implementation is the linchpin of RTI. If there is to be an evaluation of RTI, a series of interventions 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 (Kovaleski, Gickling, Morrow, & Swank, 1998; 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 (Baer, Wolf, & Risley, 1987).

The purpose of this study was to evaluate the referral, identification process, and student outcomes. Specifically, this study evaluated the use of a systematic research-based RTI model, System to Enhance Educational Performance (STEEP). STEEP consists of a series of assessment and intervention procedures with specific decision rules to identify children who might benefit from an eligibility evaluation. STEEP was built upon the research in curriculum-based assessment (CBA), CBM, (Shinn, 1989) and problem-solving (Fuchs & Fuchs, 1998; Good & Kaminski, 1996; Shinn, 1989). 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. Standard, protocol-based interventions are delivered for a specified number of consecutive sessions and monitored for integrity. Progress monitoring data is used to determine whether or not the intervention response was adequate or inadequate. Children who show an inadequate RTI are recommended for full psychoeducational evaluation by the multi-disciplinary team. Hence, STEEP is a set of procedures that function as a screening device to identify children who might benefit from special education services.

Well-controlled studies have demonstrated preliminary evidence for the technical merits of STEEP. Strong correlation values were reported between scores obtained across two consecutive trials of classwide-administered CBM probes in reading (words read correctly per minute) and math (digits correctly computed in 2 min on single-skill computation probes) with first and second grade students (VanDerHeyden, Witt, & Naquin, 2003). Strong concurrent correlation values have been reported between scores obtained using CBM as part of the STEEP process and other commonly used achievement scores including the Iowa Test of Basic Skills (ITBS) and the Woodcock–Johnson Test of Psychoeducational Achievement (VanDerHeyden et al., 2003). Predictive power estimates were obtained by concurrently exposing first and second grade participants to a series of criterion measures including ITBS in reading and math and CBA with extended individual intervention. Standardized decision rules used at each tier of STEEP resulted in positive predictive power of .53 and negative predictive power of .95 relative to teacher referral positive predictive power of .19 and negative predictive power of .89 (VanDerHeyden et al., 2003). These predictive power estimates were found to be superior to teacher identification (as a potential screening source) and use of STEEP was found to maximize the number of correctly identified children in the sample (i.e., “hit rate”) whereas teacher identification was not. Additionally, predictive power estimates were found to remain relatively stable across classrooms that varied substantially in their demographic make-up and general level of achievement (VanDerHeyden & Witt, 2005), whereas teacher identification was found to be unstable. Use of STEEP was found to identify children at rates proportionate with their sample base rates and also to correctly identify children at comparable rates across racial and gender categories. Yet, each of these studies of STEEP occurred with researchers rather than school personnel implementing most of the procedures (e.g., intervention). The purpose of this study was to evaluate the effect of STEEP in a district-wide application using only district personnel to implement the model. Standardized procedures were used to implement the model across each of the elementary schools within a multiple baseline design. This study extended the previous findings on STEEP and problem-solving by

examining the actual effect on identification rates at each site, proportionality of identification rates by sex and race, reliability of decision-making by the individuals charged with determining whether or not an intervention had been successful, and the degree to which the multi-disciplinary team's decision coincided with STEEP outcome (i.e., the reliability of the team's decision-making relative to the collected data). Broadly, this study examined the usability of an RTI model in a school district where prior to its introduction, curriculum-based measurement linked to intervention had not been used.

STEEP model

System to Enhance Educational Performance (STEEP; Witt, Daly, & Noell, 2000) is a systematic model of assessment that can be used to identify children who might benefit from eligibility assessment. STEEP has similarities to the problem-solving, problem certification, and treatment validity models previously described in the literature. STEEP uses a commercially available set of CBA and CBM probes in reading and math 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 (Tilly, 2003).

Trained consultants work 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 1.

Universal screening

CBA and 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 (Basic Skill Builders, Sopris West). During screening, two types of data are collected. First, oral reading fluency and computation fluency scores are obtained to assess children's performance relative to their classmates and relative to instructional standards on a task that represents 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 reflects skills that will be learned throughout the year is also used so that periodic probes can reflect growth toward year-end goal of competence in a broad array of computational tasks. For reading, a single grade-level probe is selected for screening at all grade levels. At first and second grades, the same probe is re-administered monthly to track growth in oral reading fluency until the class median reaches the mastery range. When the class median reaches mastery range, a more difficult probe is selected to track growth from that point forward until the end of the year. For math, one

Table 1
Summary of decision rules applied at each tier

I. Schoolwide universal screening
Curriculum-based assessment and measurement (CBM) probes are administered classwide (i.e., reading, math, and writing), performance of individual students is examined. If the class median is above the instructional standard, then the bottom 16% of children who perform below the instructional standard (Deno & Mirkin, 1977) proceed to Tier II. If the class median is below the instructional standard, then classwide intervention is performed prior to identifying individual children for further assessment.
About 15% of children screened proceed to Tier II, the performance/skill deficit assessment.
II. Performance/skill deficit assessment
Students are offered a reward for exceeding their previous performance and are then re-examined using the classwide academic assessment probe that had been previously administered. Children who perform below the instructional standard (Deno & Mirkin, 1977) proceed to Tier III.
About 11% of children proceed to Tier III, individual intervention.
III. Individual intervention
Students exhibiting skill deficits (in classes where the majority of the class is performing above the instructional standard) participate in daily individual intervention performed by the classroom teacher in the regular classroom setting during the regular school day. Progress is monitored to determine whether other intervention strategies or assessments are warranted.
About 3–5% of children do not respond successfully to short-term protocol-based intervention delivered with integrity in the regular classroom setting.

Note. For reading, the instructional standard is 40 wc/min at grades 1–2 and 70 wc/min in grades 3–5. For math, the instructional standard is 20 digits correct/2 min at grades 1–3 and 40 digits correct/2 min at grades 4–5.

probe is administered at each grade level at each screening that reflects current instructional placement for screening. A second probe is administered monthly to all children at all grade levels to track progress in math[1]. Reading probes are scored as words read correctly per minute (wc/min) and math probes are scored as digits correct per 2 min (dc/2 min). The instructional standard applied for reading is 40–60 wc/min for grades 1 and 2, 70–100 wc/min for grades 3–5 (Deno & Mirkin, 1977). The instructional standard applied for math is 20–40 dc/2 min for grades 1–3, and 40–80 dc/2 min for grades 4–5 (Deno & Mirkin, 1977).

Teachers are trained to reliably administer CBM probes and administration requires no more than 1 h per class. Following the screening, the teacher receives 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 standard described by Deno & Mirkin, 1977) or individual child problem (classwide median falls at or above the instructional standard and individual child scores below the 16th percentile for his or her class and below the instructional standard described by Deno & Mirkin, 1977). Thus, two anchors are applied initially to define the problem (local anchor is classwide performance, broader anchor is instructional level performance that has been linked to functional competence, Deno & Mirkin, 1977). Student performance is monitored separately for reading and math. Hence, a

[1] 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.

single child may participate in the entire process twice, once for reading and once for math following each screening.

Classwide intervention

When a classwide problem is identified (class median score falls below the instructional standard described by Deno & Mirkin, 1977), a classwide direct-instruction intervention[2] is implemented. The first step in performing classwide intervention involves finding the instructional level of the *class* by administering a series of easier CBM probes until the class median reaches 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 is delivered at a difficulty level that matches 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 requires about 10 min daily. The classwide intervention is performed for 10 consecutive school days. Following the data decision rules in Table 1, the children who continue to perform below the instructional standard and demonstrate poor growth relative to peers in the same class (i.e., children who are not learning when other children are learning at a rapid pace) are identified and referred for the next phase, the performance/skill deficit assessment.

If a classwide problem is 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 standard participate in the next stage of assessment, a brief assessment of the effect of 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). Typically the school psychologist conducts the performance/skill deficit assessment outside of the classroom using scripted administration procedures.

Performance/skill deficit assessment

During the performance/skill deficit assessment, the school psychologist provides the student with a copy of the classwide academic assessment probe that had been previously administered. Students are told that they can earn a reward of their choice from the treasure chest by "beating their last score." This score is written in the top left-hand corner of the student's paper. Students are allowed to sample briefly the items in the treasure chest. The treasure chest is 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 is then re-administered. The performance/skill deficit assessment for math is administered to groups of three to five students simultaneously, whereas the performance/skill deficit assessment of reading is administered individually in a quiet space

[2] Sample intervention protocols can be obtained from http://www.gosbr.net/.

on the school campus. This component requires no more than 5 min per assessment. Children whose performance improves to the instructional range (Deno & Mirkin, 1977) to earn an incentive do not participate in further assessment. Children whose performance does not improve to the instructional range participate 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 is performing at or above the instructional range, participate 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 requires approximately 10 min is applied. The school psychologist works 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 is determined by sampling backward through successively lower level materials until the student scores in the instructional range. The difficulty level at which the student scores in the instructional range (Deno & Mirkin, 1977) is the difficulty level at which the intervention is conducted. Protocol-based interventions share 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 are protocol-based and designed to produce evidence (i.e., permanent products) that they occurred to allow for estimation of treatment integrity. Permanent products typically include correctly scored worksheets.

The school psychologist collects 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. Performance on both the instructional-level and criterion-level probes is needed because the student may be instructed using task materials that are easier than task materials being used in his/her classroom at that time in the year. Performance on the instructional-level probe reflects performance changes due to intervention. Performance on the criterion-level probe reflects generalization or the degree to which performance changes might generalize to classroom performance. Intervention integrity is evaluated based on the production of permanent products generated as the intervention is implemented (Noell et al., 2005). Permanent products are by-products that are generated when an intervention is used (e.g., a correctly scored worksheet). The school psychologist enters the data into the database and graphing tools automatically generate graphs for the teacher, principal, and psychologist. If problems occur in implementing the intervention, the school psychologist provides performance feedback to the teacher and re-trains the teacher to implement the intervention correctly for the following week.

The purpose of the brief intervention is to measure the child's RTI. To measure RTI, 10 to 15 consecutive intervention sessions, conducted with integrity, are required. Additionally,

Table 2
Demographics of elementary schools in district

		School 1		School 2	
		2001–2002	2003–2004	2001–2002	2003–2004
Total enrollment		706	781	638	583
Classrooms (range of number of students per classroom)		3027 (18–25)	3226 (16–25)	27 (16–26)	27 (18–25)
Race	Caucasian	71%	74%	81%	76%
		504	577	515	445
	Hispanic	18%	19%	15%	17%
		128	149	94	97
	African	6%	4%	3%	4%
	American	45	30	18	21
	Other	4%	3%	2%	3%
		29	25	11	20
Sex	Male	52%	52%	52%	52%
		369	405	333	305
Free lunch		16%	14%	37%	28%
		113	109	236	163
Mean SAT-9 Percentile Grades 2–5	Reading	73	71	62	60
	Math	76	80	60	70
	Language arts	69	66	57	54
ELL		3%	3%	1%	0%
		23	20	6	0
Special education	Total	11.6%	10.8%	15.4%	18.0%
		82	84	98	105
	SLD	5.8%	3.7%	4.7%	3.6%
		41	29	30	21

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.

a similar but unpracticed probe (the probe used at screening) is administered each week to track progress. The intervention is determined to have been successful if the child performs above the instructional standard on the grade-level screening probe following intervention. Intervention trend data are also examined to ensure that growth is 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 are made available to the school-based team to assist in determining whether or not a child should receive an eligibility evaluation. These data are graphically presented to the team with a recommendation to obtain more information through a full psychoeducational evaluation. Estimates from a well-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).

Research questions

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

School 3		School 4		School 5 (CW)	
2001–2002	2003–2004	2001–2002	2003–2004	2001–2002	2003–2004
–	595	647	562	586	580
–	2016 (17–30)	26 (17–26)	25 (18–25)	2420 (18–25)	24 (17–22)
–	75%	67%	70%	67%	71%
	447	431	391	393	409
–	21%	24%	21%	22%	21%
	124	155	117	131	123
–	2%	5%	4%	6%	5%
	11	31	25	38	31
–	2%	5%	5%	4%	3%
	13	30	29	24	17
–	51%	53%	55%	53%	51%
	304	343	307	308	296
–	26%	18%	21%	22%	19%
	155	116	118	129	110
–	62	62	73	58	62
–	66	67	82	62	73
–	54	57	67	56	58
–	4%	0%	<1%	4%	4%
	25	0	1	26	23
–	11.4%	12.1%	11.2%	13.5%	14.5%
–	68	78	63	79	84
	2.5%	6.5%	3.3%	6.7%	3.4%
	15	42	19	39	20

degree would the decision-making teams utilize STEEP data to determine whether or not an evaluation should be conducted? (3) What effect did STEEP implementation have on identification rates by ethnicity, sex, free or reduced lunch status, and primary language status? (4) How did the use of STEEP reduce assessment and placement costs for the district and how were these funds re-allocated? (5) What were the outcomes for children judged to have an adequate RTI relative to those children who were judged to have an inadequate RTI?

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 2. 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.

Because school psychologists played a pivotal role in the existing prereferral process, the four female Caucasian school psychologists assigned to each school were trained to coordinate STEEP activities at their schools. 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 twenty years. 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. 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 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. 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

The procedures described in this section were in place at each school during baseline and remained in place throughout the course of the study. 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. Children who performed poorly on these measures were provided with supplemental services by the district.

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 a commercially-available screening measure to screen all children,

identify children as non-English proficient, limited English proficient, or fluent English proficient, determine type of supports needed, and monitor progress.

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.

Design

Effects were examined within a preliminary multiple baseline across schools. The baseline and STEEP procedures experimental conditions were sequentially introduced and evaluated for their effects on initial evaluation and percentage of children evaluated who qualified for services (an estimate of diagnostic efficiency). 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 1, 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. 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 1 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 Fig. 1). School 3 was a new school that opened in 2003–2004 and opened with STEEP in place because the school psychologist and the principal had previously worked at school 1 and wished to use STEEP in the school during its first year. Data obtained at School 3 were comparable to the other schools on all dependent measures when STEEP was implemented (data available from the first author upon request). School 3 was excluded from the multiple-baseline data because no baseline data were available for the school.

Procedures

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. When a teacher or parent had a concern about a student, the teacher completed

a brief pre-referral questionnaire that included previous grades, attendance history, and a written summary of the teacher and/or parent's concerns. 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 were offered by the school psychologist to the team for consideration in determining whether or not to refer a child for evaluation.

Training the school psychologists 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. Training occurred on-site, one full day each week for one semester. When a new component was introduced, the first author described the steps, provided scripted instructions, and modeled correct performance. The school psychologist then implemented the new component with 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 school 2 the fall semester of 2003 using similar procedures. The first school psychologist trained the third school psychologist at schools 4 and 5 during 2004–2005 using the same training procedures.

STEEP implementation

The schoolwide screening occurred a minimum of three times per year at each site beginning 4–6 weeks following the start of the school year, and some measures (e.g., reading for first graders and math for all students) were repeated more frequently to monitor progress. Each time the screening occurred the following procedures were followed to identify children who might need intervention. Schools were encouraged to consider STEEP data in making evaluation decisions, but school-based multi-disciplinary teams were free to (a) reach a decision that did not correspond with STEEP data, (b) collect additional data, or (c) refer a child for evaluation prior to completion of the STEEP process. Each grade was scheduled to conduct all screening activities within the same 1-hour time period and a trained coach (4–5 coaches were identified and trained at each school and generally included the special education professionals who routinely worked on that

campus) 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 (Deno & Mirkin, 1977).

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. Training was complete when the teacher or peer could complete the intervention 100% correctly without prompting from the school psychologist. 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 RTI could generally be made once 10–15 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 for correct implementation.

Individual RTI judgment agreement

On average, 6.68 number of intervention sessions (range, 3 to 15) occurred before a decision was reached about whether RTI was adequate and 12.41 number of sessions (range, 4 to 19) occurred before a decision was reached that RTI 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%.

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.

Number of evaluations

Number of evaluations was computed as the number of children who were evaluated for special education eligibility under any category at each school. 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. The individual files were 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 RTI. 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.

Demographics of children evaluated and placed in special education

Each student was coded by sex, ethnicity, free lunch, and ELL status using district data. Expected proportions for race and sex were computed for each school as an anchor against

which to compare the proportionality of minority and male students evaluated prior to and during implementation of STEEP. Expected proportions for race were computed by dividing the total number of children identified as being of minority ethnicity by the total number of students in the school. Expected proportions for sex were computed by dividing the total number of males at each school by the total number of students at the school.

Outcome of evaluation

For each child who was evaluated, district records were obtained and each child was coded in the database as having been found eligible to receive special education services and if so, the eligible category was specified.

Results

What effect did STEEP implementation have on the total number of evaluations and percentage of evaluations that qualified for services?

Initial evaluations

Total number of initial evaluations for each site during consecutive years is presented in Fig. 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. Because the total number of evaluations were interpolated based upon rate of evaluations over a shorter period in 2003–2004 for School 2, these data may have over-or under-estimated actual number of evaluations had each psychologist worked at that site the entire year (i.e., rate of referral may not have been stable across all months of the school year). However, the number of estimated evaluations (i.e., 50) obtained during the reversal was not inconsistent with the number of evaluations performed during baseline years. Further, the reduction to an actual number of evaluations of 7 in the 2004–2005 school year was consistent with number of evaluations at other sites were STEEP was implemented and was obtained with the same psychologist who had evaluated at a rate equivalent to 50 evaluations per year when STEEP was withdrawn from the school. In other words the obtained difference between baseline and STEEP implementation was replicated across two psychologists at the same school. School 3 was excluded from the multiple baseline. 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

Fig. 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

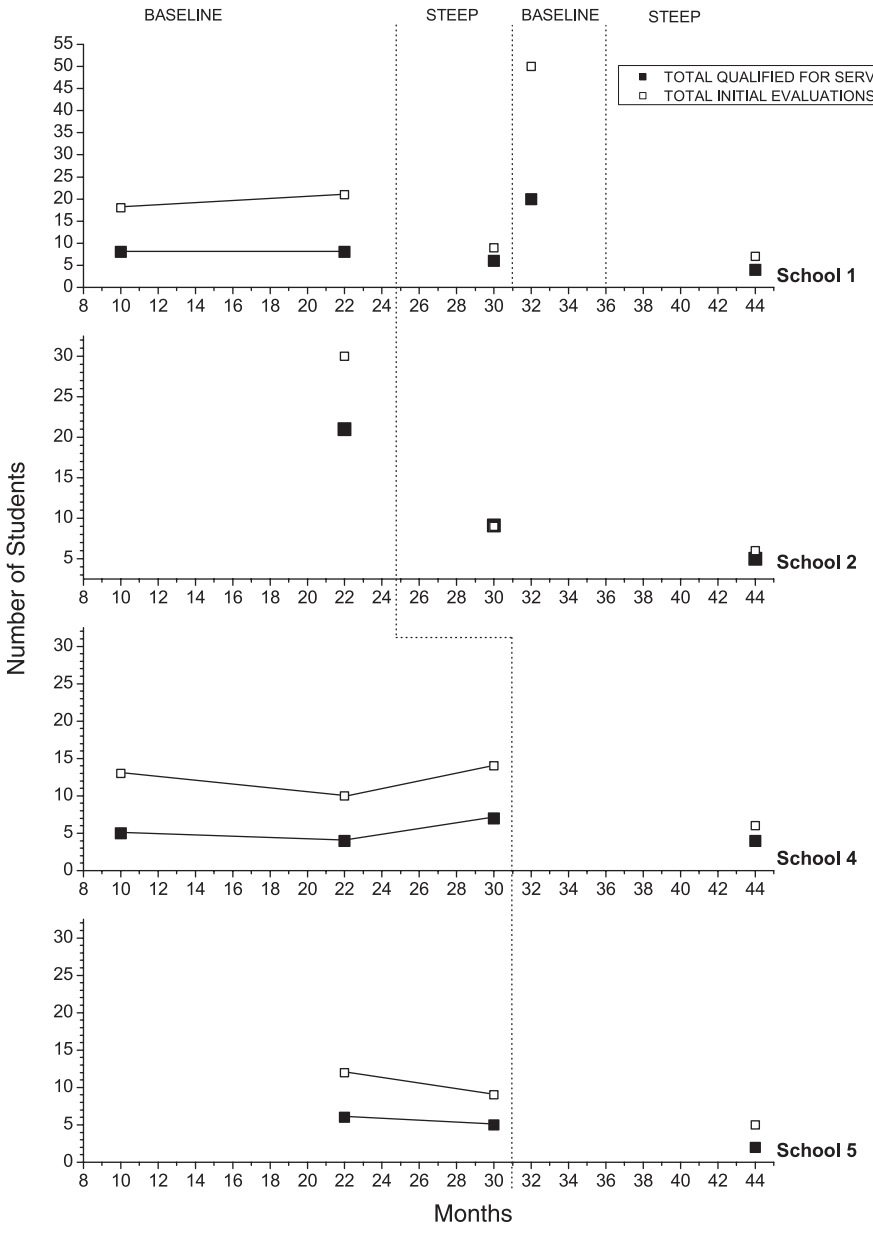

Fig. 1. Total number of initial evaluations and total number of students who qualified for services during baseline and STEEP implementation conditions at each participating school during each school year.

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. School 3 was excluded from the multiple baseline analysis. 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 decision-making team could elect to recommend evaluation irrespective 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 RTI 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 RTI (i.e., STEEP recommended that evaluation not be conducted). One child did not have an adequate RTI 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 RTI (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 RTI (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 RTI 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.

To what degree did the decision-making teams utilize STEEP data to determine whether or not an evaluation should be conducted?

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 rate of qualification that was much higher than the baseline average. Most of these children qualified under speech and language impairment (SLI; 46%), but 23% qualified under Specific Learning Disability (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. These data are presented in Table 3. Hence, the rate of qualification 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 rate of qualification for children who were recommended for evaluation by both STEEP and the decision-making team was 89%.

Table 3
Percent of evaluated cases that qualified based on team decision to refer

	Baseline (%)	STEEP+ and team decided to evaluate (%)	STEEP− and team decided to evaluate (%)
2003–2004 Cases, Schools 1–3	55	89 [a]	50 [b]
2004–2005 Cases, Schools 1–5	52	88 [c]	29 [d]

[a] Counting only those children for whom STEEP data had been completed, 9 children were recommended for evaluation by STEEP. All 9 were subsequently recommended for evaluation by the decision-making team and 8 of these children qualified for services.

[b] However, 17 children were not recommended for evaluation at the team decision-making meeting based on STEEP findings, but the teams decided to evaluate 10 of these children anyway. Specifically, 3 children qualified under SLD, 1 qualified under Speech and Language Impairment, and 1 qualified under Other Health Impairment.

[c] In 2004–2005, 14 children were recommended for evaluation by STEEP. Of these 14 children, 12 were evaluated and 7 qualified for services, 1 did not qualify, and 4 cases were pending at study completion.

[d] 106 cases were not recommended for evaluation based on their having had an adequate RTI. 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.

What effect did STEEP implementation have on identification rates by ethnicity and sex?

Ethnic proportionality

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 Fig. 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 Fig. 3.

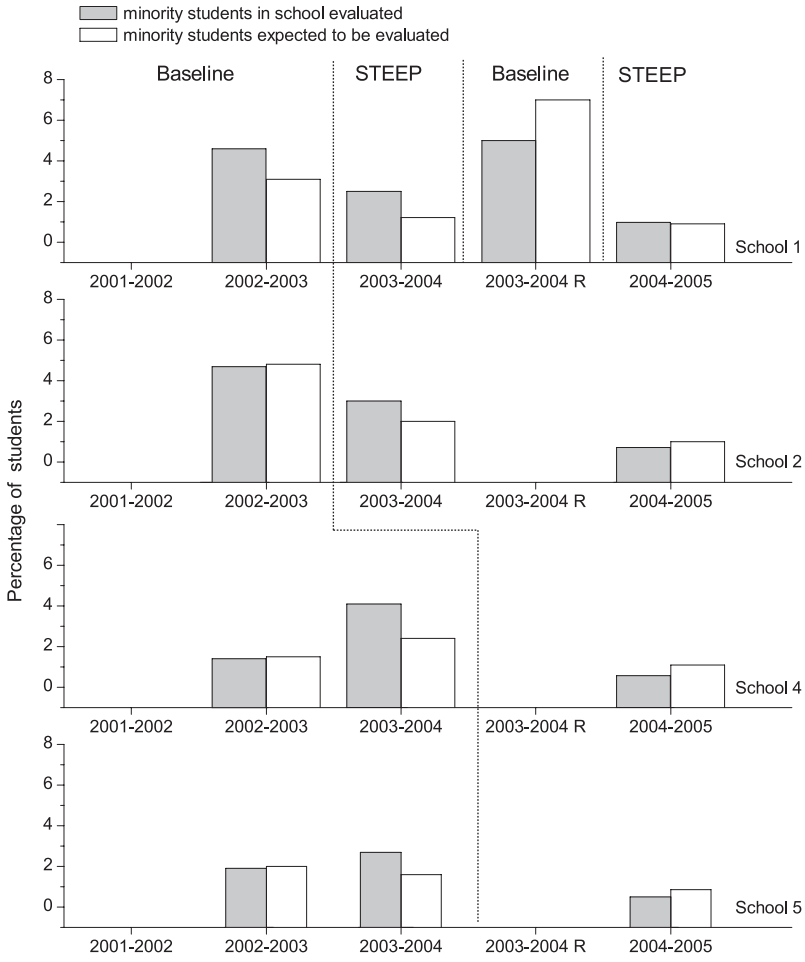

Fig. 2. Number of children of minority ethnicity who were evaluated at each site and number who qualified during baseline and STEEP implementation.

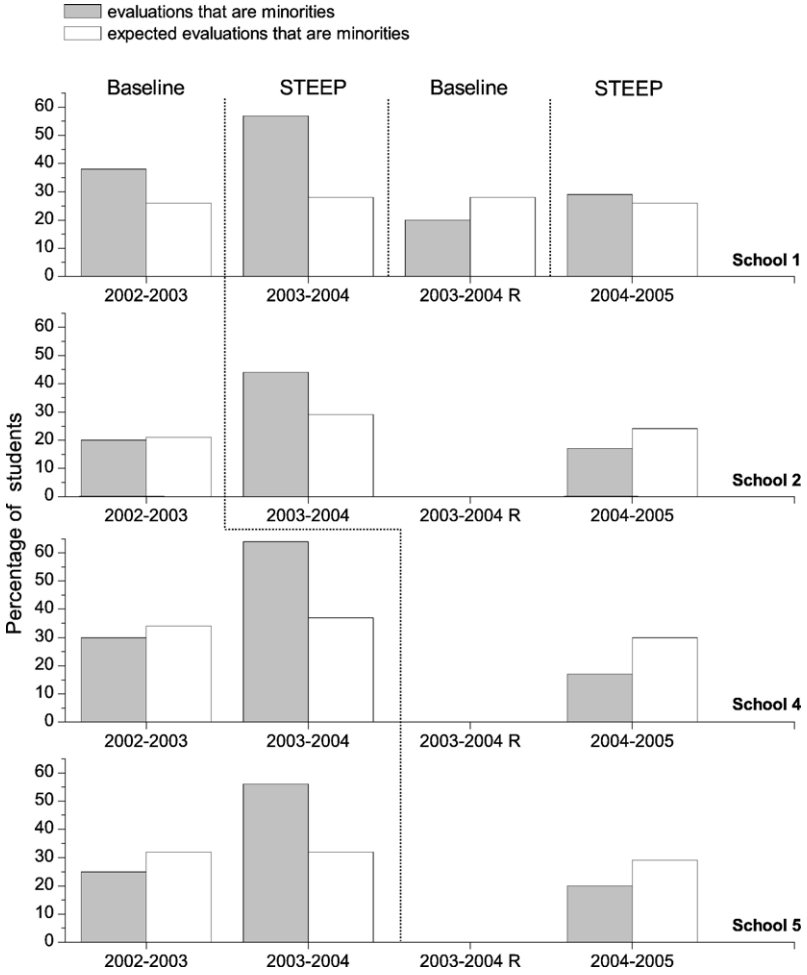

Fig. 3. Percentage of evaluations that were conducted with children who were of minority ethnicity was examined relative to the number of evaluations that could be expected to be of minority children at each site during baseline and STEEP implementation.

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

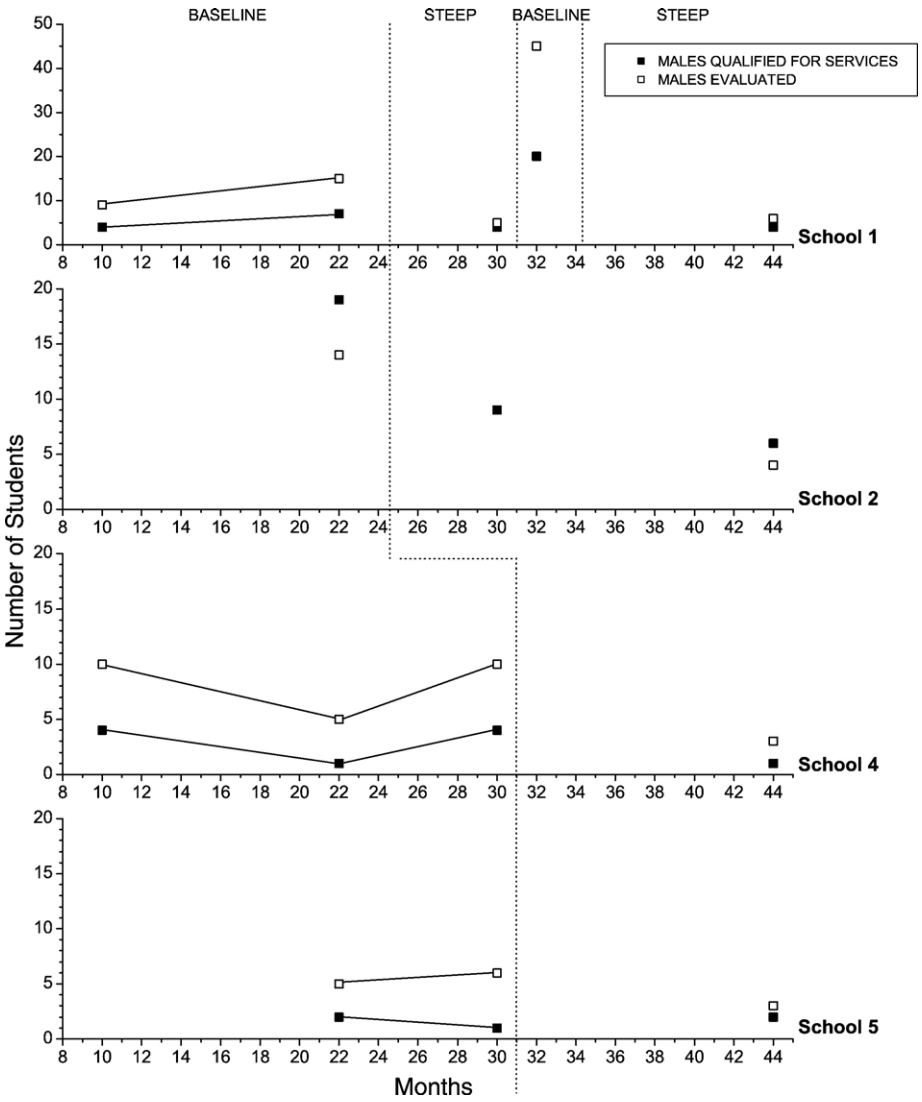

Fig. 4. Number of male student initial evaluations and total number of male students who qualified for services during baseline and STEEP implementation.

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 2004–2005 (.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.

Sex disproportionality

Fig. 4 shows the number of males evaluated and placed across all sites and all years. At baseline, on average 1.52 males were evaluated for each female. Following STEEP implementation, 1.35 males were evaluated for each female. 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 during baseline years and when STEEP was being implemented. Fig. 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)

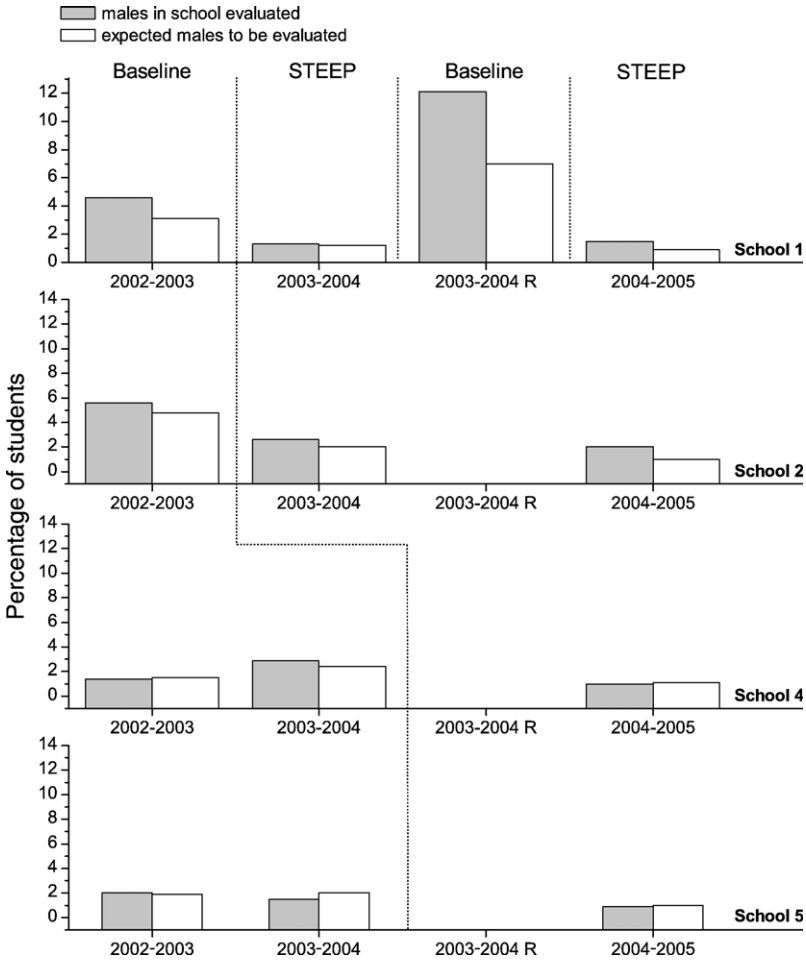

Fig. 5. Percentage of evaluations that were conducted with male children was examined relative to the number of evaluations that could be expected to occur with male children at each site during baseline and STEEP implementation.

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 $p > .01$ when STEEP was implemented.

Performance differences by ethnicity, gender, SES, and primary language

Presented in Table 4 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

Table 4
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 (SD=33)	72 (SD=31)	82 (SD=23)	66 (SD=54)	60 (SD=21)
		3rd–5th	127 (SD=34.9)	125 (SD=35)	124 (SD=22)	119 (SD=31)	107 (SD=18)
	Mean spring level DC/2 min	1st–3rd	38 (SD=14)	37 (SD=21)	39 (SD=9)	36 (SD=13)	37 (SD=10)
		4th–5th	92 (SD=24)	88 (SD=24)	87 (SD=15)	79 (SD=24)	48 (SD=16)
	Mean spring growth wc/min change per week of instruction	1st–2nd	0.8 (SD=0.9)	0.7 (SD=0.9)	2.8 (SD=1.4)	0.7 (SD=0.5)	0.6 (SD=0.3)
		3rd–5th	0.05 (SD=0.7)	0.5 (SD=0.6)	0.5 (SD=0.8)	0.5 (SD=2.0)	0.5 (SD=0.1)
	Mean spring growth DC/2 min change per week of instruction	1st–2nd	0.7 (SD=0.8)	0.7 (SD=0.8)	2.6 (SD=1.1)	0.6 (SD=0.5)	1 (SD=0.3)
		3rd–5th	1.7 (SD=1.5)	1.4 (SD=1.4)	2.8 (SD=1.9)	1.5 (SD=0.8)	1.5 (SD=0.8)
STEEP identification	At-risk during universal screening (Tier 1)		265 9% of total population	165 62% of risk group	73 28% of risk group	51 19% of risk group	16 6% of risk group
	At-risk skill/performance deficit assessment (Tier 2)		129 4% of total population	73 57% of risk group	34 26% of risk group	24 19% of risk group	7 5% of risk group
	At-risk individual intervention (Tier 3)		14 0.5% of total population	6 43% of risk group	7 50% of risk group	7 50% of risk group	5 36% of risk group
Evaluated	Number evaluated		43 1.4% of total population	24 56% of those evaluated	14 33% of those evaluated	17 40% of those evaluated	10 23% of those evaluated

who failed the skill/performance deficit assessment, and percent of students who had an inadequate RTI). These data are presented in Table 4.

Did the use of STEEP reduce assessment and placement costs for the district?

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 each full psychoeducational evaluation were valued at $3000 (Mirkin & Potter, 1983) 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. This cost reduction represents a smaller number of psychoeducational evaluations; however, additional assessment costs occurred in conducting the schoolwide screening three times per year at each school, performance/skill deficit assessments for children identified during the schoolwide screening, and individual assessment and intervention for a subset of the most at-risk children. During schoolwide screening, 1176 children were screened on three occasions at these two schools requiring 59 h of total reading assessment time and 15 min of math assessment time for each of 52 classrooms for a total of 13 h of math assessment. Total time invested in schoolwide screening was approximately 72 h of time with two adults in each classroom (the classroom teacher and one coach). Scoring required one additional hour of teacher time following each screening for a total of 156 h of scoring time. Data entry required about 4–6 h per site per screening occasion for a total of 12–18 h of data entry time for each of the two schools included in this analysis or 24–36 h total for both schools. One hundred and seventy eight performance/skill deficit assessments were conducted requiring 3–5 min per assessment for a total of about 15 h. Individual intervention was performed for 66 children requiring about 100 h of total individual intervention time. Total "new" assessment and intervention time was about 379 h of time across these two schools in year one of implementation. Hence, if a full psychoeducational evaluation were estimated to require about 15 h per child, then this equates to about 26 evaluations indicating that assessment costs were reduced with STEEP implementation by about 50% (i.e., 51 evaluations were conducted in these two schools the preceding year). Schools experienced more of a shift in assessment costs instead of a "real" reduction since school psychologists' roles expanded to include more instructional consultation and "extra" data collection requested by teachers who valued the data. No school psychologist positions were cut during the years of this project and on the contrary, the district hired an additional school psychologist each year to facilitate the school psychologists' evolving role as instructional consultants.

Actual placement costs were reduced with STEEP implementation. 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.

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 & Witt, 2005; VanDerHeyden et al., 2005; 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, 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.

The percentage of minority students at each school ranged between 20 and 34%. 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 evaluated minority students 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 ethnicity would be evaluated. The percentage of minority students evaluated ranged between 2 and 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.

With respect to gender, a disproportionate number of males were evaluated and placed during the baseline years (i.e., 44 males: 29 females or 1.52 males: 1 female). 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 (i.e., ratio was reduced to 23 males: 17 females or 1.35 males:1 female). This finding is consistent with previous findings related to the positive effect of RTI data-based decision models on disproportionate identification by sex (VanDerHeyden & Witt, 2005).

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 across all years of the study. The inclusion of all categories helps to mitigate the possible confound of teams classifying children who had an adequate RTI 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 RTI 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 analyses presented indicate that resources devoted to traditional assessment were reduced and replaced by direct assessment, intervention, and consultation services in classrooms. 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 overidentification of males. 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 (VanDerHeyden et al., 2005). 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 (Messick, 1995).

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 (Fuchs, 2003; VanDerHeyden et al., 2005). With STEEP, teams were simply presented with students whose performances fell below the criterion at each stage of assessment and resulted in identification of about 3% of the population as ultimately being detected as at-risk by the STEEP screening which is consistent with identification rates reported by other models (Case, Speece, & Molloy, 2003; Torgesen, 2000).

Practical implications

In addition to the criteria used to judge RTI, 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). These studies are needed to provide additional evidence of construct and external validity (Glover and Albers, this issue).

The reversal data obtained at School 1 are limited by several potential confounds including the time of school year when the reversal occurred and the introduction of a new psychologist. These potential confounds are mitigated somewhat since the dependent variable estimates were consistent with baseline estimates and return to STEEP implementation conditions the following year produced a change in the dependent variables replicating previous STEEP effects when the new school psychologist was trained. These data illustrate the potential 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 at school 1. At the time of the reversal, the school had no backlog of cases and was handling a number of intervention cases that was similar to the number handled that time of year at school 2. Whereas these reversal data are limited, they provide preliminary evidence that (a) use of STEEP data may require specific on-site training, (b) that correct use of STEEP data was responsible for the decline in number of evaluations and increase in the percentage of children qualifying for services who were evaluated, and (c) that the school psychologist may play a pivotal role in correct use of the data.

One 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 model will rely on decisions based on interpretations of data. Hence, the degree to which decisions correspond with data will be a critical component of validity of RTI models of decision-making (VanDerHeyden et al., 2005) and represents a serious challenge to successful use of RTI models in practice (Macmillan, 1998). 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 RTI were recommended for full psychoeducational evaluation despite the data during 2003–2004. 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). 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 RTI were referred for evaluation and 92% of children who did not have an adequate RTI were referred for evaluation.

Limitations

Several additional 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. School 3 was excluded from the multiple baseline analyses because 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 RTI model with which they had become familiar. Importantly, dependent variable estimates were tracked at school 3 and included in the analyses conducted separately from the multiple baseline. Data at school 3 were consistent with all other implementation sites with STEEP in place. The multiple baseline used in this study is limited by the small number of datapoints within each phase. Given that

an entire year's worth of data collection was required to yield a single datapoint, collecting more data was not feasible for this study. Nonetheless, the small number of datapoints increases the possibility that chance variation may have accounted for observed effects. This possibility is mitigated somewhat by the reversal manipulation at school 1 and the replication of effects across schools. Ideally future studies will utilize other design options (e.g., randomized controlled trial) to investigate the effectiveness of RTI procedures and decision-making models on a large scale. 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. 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. The same assessment probe was used for screening and weekly progress monitoring during intervention to avoid the problem of equating task difficulty across conditions and to ensure a constant criterion for decision-making. A student may have been exposed to the same task a maximum of four times (at screening, during the performance deficit/skill deficit assessment, and twice in a period of two weeks during individual intervention). Truly sophisticated ways of handling stimulus material equation exist (Daly, Bonfiglio, Hauger, Persampieri, & Yates, 2005) but were impractical for use. Use of the same materials across conditions creates the potential for inflated performance due to repeated exposure to the same materials and false negative identification errors. This possibility should be scrutinized in future research. Identification rates reported in this study were similar to those reported in previous studies but estimates of predictive power were not obtained in this study.

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. Thus, these findings require validation with larger samples with additional longitudinal data to further investigate long-term outcomes. Finally, the cost analyses presented in this paper do not account for the costs associated with STEEP implementation. Future analyses should attempt to quantify the cost of assessment and intervention services provided through RTI models and obtain concurrent measures of student performance to evaluate the degree to which child outcomes are improved (VanDerHeyden & Burns, 2005). Such analyses would permit a more balanced estimate of costs and benefits associated with the use of RTI decision-making models. In the district where this project was conducted, a new psychologist was hired each year of the project (to provide services at middle school, high school, and preschool to permit stability of staffing during the years of this study at the elementary schools). The hiring of a new psychologist each year provides preliminary evidence that psychologists were busier than ever, but were providing services that were qualitatively different from those previously provided, and were services that the district seemed to value. The reduction in costs associated with serving fewer children in special education allowed the district to

reallocate funds to intervention efforts directed at enhancing the core instructional experience of most students.

Research has yet to sort out how RTI 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, 2003) or the treatment validity model described by Fuchs and Fuchs (1998)? Will RTI operate primarily in general education (Case et al., 2003) or somewhere in between general and special education, for example, in tracking supplemental services provided to at-risk students (Vaughn et al., 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 et al., 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 RTI? 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 et al., 2004).

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

Baer, D. M., Wolf, M. M., & Risley, T. R. (1987). Some still-current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, *20*, 313−327.

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. *Journal of Special Education*, *38*, 66−79.

Case, L. P., Speece, D. L., & Molloy, D. E. (2003). The validity of a response-to-instruction paradigm to identify reading disabilities: A longitudinal analysis of individual differences and contextual factors. *School Psychology Review*, *32*, 557−582.

Christ, T. J., Burns, M. K., & Ysseldyke, J. E. (2005). Conceptual confusion within response-to-intervention vernacular: Clarifying meaningful differences. *Communique*, *34*, 6−8.

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.

Daly, E. J., III, Bonfiglio, C. M., Hauger, T., Persampieri, M., & Yates, K. (2005). Refining the experimental analysis of academic skill deficits, Part I: An Investigation of variables affecting generalized oral reading performance. *Journal of Applied Behavior Analysis*, *38*, 485−498.

Deno, S. L., & Mirkin, P. K. (1977). *Data-based program modification: A manual.* Reston, VA: Council for Exceptional Children.

Fuchs, L. (2003). Assessing intervention responsiveness: Conceptual and technical issues. *Learning Disabilities Research and Practice*, *18*, 172−186.

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

Fuchs, D., Mock, D., Morgan, P. L., & Young, C. L. (2003). Responsiveness-to-intervention: Definitions, evidence, and implications for the learning disabilities construct. *Learning Disabilities Research and Practice*, *18*, 157−171.

Good, R. H., & Kaminski, R. A. (1996). Assessment for instructional decisions: Toward a proactive/prevention model of decision-making for early literacy skills. *School Psychology Quarterly*, *11*, 326−336.

Gravois, T. A., & Rosenfield, S. A. (2002). A multi-dimensional framework for evaluation of instructional consultation teams. *Journal of Applied School Psychology*, *19*, 5−29.

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

Kovaleski, J. F., Gickling, E. E., Morrow, H., & Swank, P. R. (1998). High versus low implementation of instructional support teams: A case for maintaining program fidelity. *Remedial and Special Education*, *20*, 170−183.

MacMillan, D. L. (1998). Unpackaging special education categorical variables in the study and teaching of children with conduct problems. *Education and Treatment of Children*, *21*, 234−245.

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.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741−749.

Mirkin, P. K., & Potter, P. L. (1983). *A survey of program planning and implementation practices of LD teachers. Research Report*, *Vol. 80*. Minneapolis, MN: University of Minnesota Institute for Research on Learning Disabilities.

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

Shinn, M. R. (1989). Identifying and defining academic problems: CBM screening and eligibility procedures. In S. N. Elliot & J. C. Witt (Series Eds.) & M. R. Shinn (Volume Ed.), *Curriculum-Based Measurement: Assessing Special Children* (pp. 90–129). New York: Guilford Press.

Tilly, D. (2003, December). Heartland Area Education Agency's evolution from four to three tiers: Our journey — Our results. Paper presented at the National Research Center on Learning Disabilities Responsiveness-to-Intervention Symposium, Kansas City, MO.

Torgesen, J. (2000). Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. *Learning Disabilities Research and Practice*, *15*, 55−64.

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.

VanDerHeyden, A. M., & Burns, M. K. (2005). Using curriculum-based assessment and curriculum-based measurement to guide elementary mathematics instruction: Effect on individual and group accountability scores. *Assessment for Effective Intervention*, *30*, 15−31.

VanDerHeyden, A. M., & Witt, J. C. (2005). Quantifying the context of assessment: Capturing the effect of base rates on teacher referral and a problem-solving model of identification. *School Psychology Review*, *34*, 161−183.

VanDerHeyden, A. M., Witt, J. C., & Barnett, D. A. (2005). The emergence and possible futures of response to intervention. *Journal of Psychoeducational Assessment*, *23*, 339−361.

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.

Vaughn, S., Linan-Thompson, S., & Hickman, P. (2003). Response to instruction as a means of identifying students with reading/learning disabilities. *Exceptional Children*, *69*, 391−409.

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.

Witt, J. C., Daly, E., & Noell, G. (2000). *Functional Assessments.* Sopris West: Longmont, CO.

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.