

Screening for Emotional and Behavioral Delays: The Early Screening Project

EDWARD G. FEIL AND HERBERT H. SEVERSON

Oregon Research Institute

HILL M. WALKER

University of Oregon

Over 2500 preschool children from across the United States were screened for emotional and behavioral problems using the 3 stage, multiple-gating procedures of The Early Screening Project (ESP). Preschool teachers were asked to nominate, rank, and rate children in their classroom on both externalizing and internalizing behavioral dimensions. The behaviors were then independently confirmed through direct observations in natural preschool environments. Gender disparities also were addressed. Significant differences were found between children who did and did not exceed the ESP's normative referral criteria. The ESP can minimize both the time and cost of preschool assessments while providing accurate screening for emotional and behavior problems with less gender bias. Additionally, the ESP can promote early prevention interventions for all young children.

With the enactment of the P.L. 99-457 amendments to IDEA, two major changes occurred in the field of special education. First, all children from birth to 21 years who needed special education and related services became eligible for federal and state funding for special education and related services. Second, P.L. 99-457 further defined and delineated the early identification and assessment requirements with a mandate to initiate a comprehensive child-find system. The law stated, "The child-find system must include the policies and procedures that the State will follow to ensure that ... an effective method is developed and implemented to determine which children are receiving needed services and which children are not receiving those services" (Federal Register, v54(119), p. 26319, §303.321). Although each state has begun to implement child-find systems for young children needing special education and related

services, there are limited tools for the child-find tasks (Beare & Lynch, 1986; Martin, 1986). Concurrent with increasing mandates, state funds for education are often declining. In Oregon, for example, the effect of a property tax limitation (Measure 5) reduced K-12 school funding by up to 16% (Hill, 1993). Practitioners are looking for screening measures for the social and emotional domain that are not only time and cost efficient, but also are accurate and without a gender bias, especially for the identification of behavioral problems where overrepresentation by boys is common. This study reports on the evaluation of a child-find system that meets these requirements of accuracy, efficient use of resources, and gender equity.

Screening for Behavior Problems

Beare and Lynch (1986) proposed that screening in natural, early childhood environments,

such as preschool classrooms, is more effective in finding eligible children than the idiosyncratic procedures currently used in public settings, such as clinics, elementary school auditoriums, and shopping centers. Naturalistic strategies stress the need for assessments to occur within a setting important to the child's development. (Barnett, Carey, & Hall, 1993; Kazdin, 1979).

The high technical adequacy of many rating scales demonstrates that teachers have a wealth of normative information regarding children's development and skills across domains. Ideally, a screening procedure should draw on teachers' normative knowledge base in a cost-effective and systematic manner. Teachers should not, however, be the sole source of information in a screening assessment. Traditional teacher referral procedures depend on the idiosyncratic behavioral tolerance of the individual teacher. In research on ecological assessments of elementary regular and special education environments, Walker and Rankin (1983) found that teachers' tolerance levels, expectations, and standards vary widely concerning child behavior.

A common bias in screening occurs across gender. Current referral practices focus primarily on externalizing behavior patterns, which teachers find salient and obnoxious. In particular, teachers are more likely to view boys as being behaviorally disturbed and as needing mental health services (Callahan, 1994). The DSM-IV estimates the ratio of males to females in clinical samples described as conduct disordered or aggressive is 3:1 (APA, 1994), other estimates place the ratio as high as 6:1 (Callahan, 1994). This is evidenced by the overrepresentation of males in programs for students with emotional and behavioral disorders (Caseau, Luckasson, & Kroth, 1994). Screening for both internalizing and externalizing patterns might provide a more equal and balanced distribution of services across genders. Additionally, children

exhibiting internalizing behaviors are often not referred for services yet represent serious long-term risks for school failure (Achenbach & McConaughy, 1992).

Assessment should not be made on the basis of a single developmental assessment or screening device, but should include direct observations of a child's behavior in the context of interacting with peers. Multiple-gating screening procedures integrate varying sources of assessment information into a multi-method, agent and setting approach. This approach requires that the child must exceed preset criteria at each stage to continue being screened, and these criteria or gates can involve different methods and raters. Multiple-gating procedures, such as those used in the Early Screening Project (ESP), allow for class-wide screening of all children while maximizing accuracy and efficiency. Use of multiple-gating procedures within a large group setting also minimizes teacher and special services consultant time by expending more specific assessment time on a small number of students who pass earlier "gates" or criteria. Because screening and assessment for young children with behavior disorders is relatively new, we can only project that multiple-gating procedures (as used in the ESP) will reduce the time required for assessment.

Convergent Validity

The key to assessment using multiple data sources is convergent validity (Achenbach, McConaughy, & Howell, 1987). Convergent validity refers to the degree of overlap of variance between different measures of the same construct (Heppner, Kilvighan, & Wampold, 1991). The synthesis of information gathered from several sources, methods, settings, and occasions produces the most valid appraisal of developmental status (Bagnato & Neisworth, 1991). Figure 1 provides a diagram illustrating convergent validity for three sources of information on a child's behavior

(Cairns & Green, 1989). In practice, a child's assessment would include several types of measures, such as parent and teacher ratings and direct observations. If there is agreement across measures, then the results have high credibility. In eligibility decisions, for example, if a child is rated as exhibiting antisocial behavior in both home and school and the behavior's severity is independently confirmed with direct observations, one could be confident that the child should be eligible for special services. During treatment, agreement among measures showing a decrease in antisocial behavior confirms the intervention's effectiveness. Nomothetically, if there are moderate to high correlations between instruments that measure the same construct, then convergent validity exists. Nevertheless, correlations among measures of different constructs should be smaller than correlations among measures of the same constructs; if this pattern is found, then there is evidence for discriminant validity (Heppner et al., 1991).

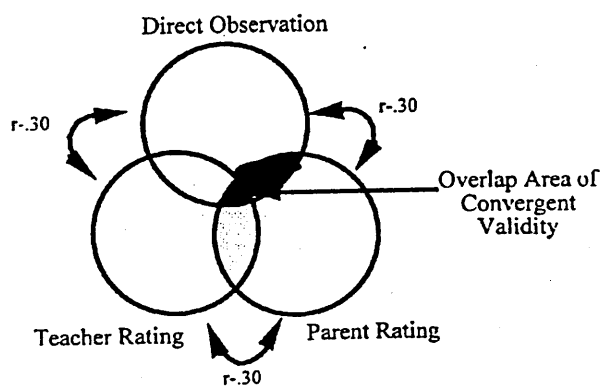
Investigators have studied the use of multi-dimensional assessment to more accurately assess a child's true behavior. Patterson, Reid, and Dishion (1992) found a better statistical fit when multiple sources of data were used: Correlations from data in which the depen-

dent variables were rated by one individual were not adequate for building or evaluating a performance theory. For example, the researchers asked a mother to describe her parenting practices and her child's antisocial behavior. Although the correlations satisfied the requirements for a performance theory, it was not possible to portion out the confound of shared method variance. Furthermore, when both the dependent and independent variables were generated by the same data source (i.e. parent interviews), the correlations were disappointingly low and a systematic effort to replicate the results failed. When data from multiple sources and methods were used to define the constructs, all path coefficients were significant and the model accounted for two to three times more variance in the criteria being predicted. (Patterson et al., 1992.)

Schaughency and Rothlind (1991) concluded that systematic multi-method assessment and application of diagnostic criteria can produce reliable diagnosis. Multiple informants contributed unique types of information about the child's behavior that were superior to a score provided by a single informant (Meisels, 1991; Meisels & Provence, 1989). In their review of the literature, Gresham and Elliot (1984) found that a multi-method approach yielded the most accurate assessment in terms of discriminative and convergent validity. In one study, four methods (teacher, peer, self-ratings, and analog situations) all shared approximately equal amounts of variance on three classes of social behavior: (a) positive interaction, (b) receiving positive interaction, and (c) assertion. Thus, it would appear that an assessment is most accurate if it incorporates information from multiple methods and sources.

Combining observational data and informant ratings may provide better prognostic information about a child's behavior than either measure alone (Cairns & Green, 1989;

FIGURE 1.
Convergent Validity Diagram



Campbell, 1994; McMahon, 1984). Observations have lower rater bias, but test-retest reliability is low and time demands for recording are high. Informant ratings have lower cost and better test-retest reliability, but are more susceptible to rater bias. Using the convergent validity of observations and ratings can increase diagnostic accuracy and improve treatment outcomes for young children (Hinshaw, Lahey, & Hart, 1993; Patterson et al., 1992). To date, few assessment systems have been created that demonstrate convergent validity across methods, settings, and raters in order to make valid generalizations.

Development of the Early Screening Project

Walker and Severson (1990) developed and validated the Systematic Screening for Behavior Disorders (SSBD), a three-stage multiple-gating system for use with elementary-age children. Based on empirical findings that behavior disorder characteristics are divided into externalizing and internalizing dimensions (Ross, 1980), Stage One of the SSBD has teachers rank children in their classroom on the externalizing and internalizing dimensions. At Stage Two, teachers complete behavior checklists and adaptive behavior rating scales using items that have been trial tested, refined, and socially validated. Teacher ratings are then independently confirmed in the SSBD Stage Three, using observations of pupil behavior in instructional and playground settings. Overall, the SSBD's percentage of false positives is very small, reported at 0% and .5% for externalizing and internalizing criteria respectively (Walker & Severson).

Perhaps, the greatest strength of the SSBD is the multiple-gating procedure. An instrument such as the SSBD utilizes teachers' expertise and controls for teacher bias via independent reliable observations conducted by others. The use of a gating system within a large-group setting minimizes teacher and special

services consultant time. The gating procedure was found to reduce assessment time up to 16% over other procedures for elementary school children (Walker et al., 1994). Furthermore, Walker et al. found that it was more cost efficient and often desirable to conduct a group administration of Stages One and Two of the SSBD procedure. Stage One and Two screening procedures can be completed for an entire school in 90 minutes.

Based on positive results with the SSBD, Sinclair, Del'Homme, and Gonzalez (1993) responded to the need for an effective screening measure for preschool children and conducted a pilot study using the SSBD. To use the SSBD with preschoolers, Sinclair et al. made some adaptations such as eliminating direct observations of academic time and increasing playground (social behavior) observations. In addition, the cut-off criteria for defining problem children were adjusted for the behavior of younger children. The results were encouraging: Means between groups of children on one measure of externalizing and internalizing behaviors (i.e. Critical Events Index) were significantly different, although means on another Stage Two measure (Adaptive Behavior Scale) were not significantly different.

In developing the ESP, further revisions of the SSBD were required to make it even more appropriate for preschool children. Because of smaller class sizes typical in preschools, teachers at Stage One rank only five children on each dimension (externalizing and internalizing). At Stage Two, more substantial adaptations were made for the ESP. The Critical Events Index of the SSBD consists of a checklist of 33 occurrence or non-occurrence items. Because most preschool children exhibit some problem behaviors at one time or another (Campbell, 1990; Paget, 1990), the more important discriminative features of the behaviors are frequency and intensity. Thus, 16 items from the Critical Events Index were

converted to a 5-point Likert scale to rate frequency and intensity. These 16 items are: steals, set fires, vomits after eating, tantrums, physically assaults an adult, physically aggressive with other children, damages property, has nightmares, exhibits inappropriate sexual behaviors, cries suddenly, has physical complaints, ignores teacher warnings, makes lewd gestures, swears, teased or neglected by peers, enuretic, and encopretic.

Another revision of the SSBD was on the Adaptive Behavior Scale which evaluates prosocial behavior (e.g., cooperation and positive social interactions). For the ESP, four items were omitted because they were considered developmentally inappropriate for preschool children: The omitted items are: (a) considerate of others' feelings, (b) produces acceptable work, (c) socially perceptive, and (d) does seat-work assignments. The SSBD Maladaptive Behavior Scale, representing anti- or non-social behavior also was revised. Two items from this scale, "Uses coercive tactics to force the submission of peers; manipulates, threatens, etc." and "Manipulates other children and/or situation to get his/her own way" were combined into one item, "Manipulates or threatens other children to follow him/her," which is a better representation of preschool behavior. Other adaptations included rewording items to increase readability.

For Stage Three, the observational measures were redesigned to reflect early childhood development and the contextual features of early childhood education settings. First, the Academic Engaged Time coding measure in SSBD Stage Three was omitted. Second, the Peer Social Behavior code, a solitary-play category, was added to more adequately reflect typical preschoolers' social behavior within free play settings. Eisert, Walker Severson, and Block (1989) found this measure discriminated reliably between preschool children with externalizing characteristics, internalizing characteristics, and controls. And third, a

duration method, utilizing a stopwatch was used to record the amount of time a child exhibited anti- or non-social behavior during a specified interval. The stopwatch was turned on when the child exhibited antisocial or nonsocial behavior and was turned off when the child displayed pro-social behaviors.

METHOD

This research evaluated the Early Screening Project (ESP), a three-stage, multiple-gating procedure designed to screen for behavior disorders among preschool children 3 to 5 years old (Walker, Severson, & Feil, 1995).

Participants

Participants consisted of 2,797 children, aged 2.4 to 6 years old (mean age = 4.28 years, SD = .68) who were enrolled in 64 preschool and kindergarten programs (see Table 1). The participants were from 160 preschool classrooms (see Table 2) in the following states (number of participating children in parentheses): California (507), Kentucky (677), Louisiana (310), Nebraska (65), New Hampshire (25), Oregon (220), Texas (555), Utah (296), and Washington (142). The educational status reported by teachers indicated that 21% held a master's degree, 59% a bachelor's degree, 14% had some college, and 5% completed high school. The sample of children consisted of 45% females; none of the children were receiving special education services. Racial and ethnic data was reported by teachers: seventy-one percent of the children were white, and 15%, 10%, and 3% reported as Hispanic, African-American, and Native American or Asian, respectively. Family income, also reported by teachers, was 50% middle income (\$15,000 – \$75,000 yearly), and 48% low income (less than \$15,000 yearly, or Head Start eligible). Of the 1,198 families with low incomes, 1,016 (85%) had children enrolled in Head Start. Community

TABLE 1.
Number and Age of Children

Age	Stage One Teacher Ranking	Stage Two Teacher Rating	Stage Three Direct Observations
3 years old	328	166	79
4 years old	1,487	728	288
5 years old	907	443	177
6 years old	75	34	18
Total	2,797	1,371	562

size was 16% urban (over 1 million), 21% semi-urban (between 250,000 and 1 million), 62% suburban, and 1% small town/rural (less than 100,000). Because this research involved a gating procedure (see Figure 2) and a comparison group, a decreasing number of children participated from Stage One to Stage Three. Stage One rankings of externalizing and internalizing characteristics resulted in six students being identified and rated in Stage Two (three of each category). The children who were the top-ranked in the externalizing and internalizing behavioral dimensions, respectively, were subsequently observed in Stage Three. Also, to provide a comparison group, a boy and girl (neither of whom was nominated or ranked on either of

the externalizing and internalizing dimensions) were identified for ratings in Stage Two and observations in Stage Three. This procedure resulted in 2,797, 1,371, and 562 subjects on whom completed data were collected for ESP Stages One, Two, and Three, respectively.

The Early Screening Project (ESP)

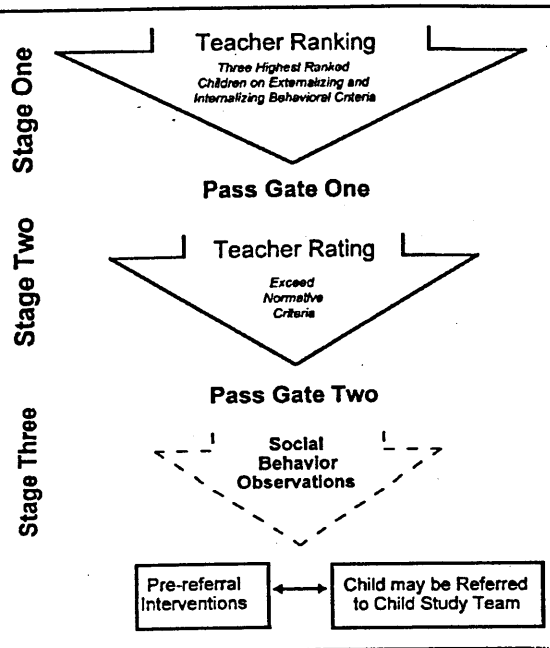
Stage One. Stage One was based on teachers' rankings of their students on externalizing and internalizing behavior dimensions. Teachers were asked to list the five children in their classroom who best exemplified a description of externalizing characteristics, and the five children who best fit the description of internalizing characteristics. The two lists were mutually exclusive, so a child could only be put on one list. Then the teachers ranked the children on each list from most characteristic (highest rank) to least characteristic, in relation to the externalizing or internalizing dimension.

The descriptions of externalizing behavior states: "Externalizing refers to all behavior difficulties that are directed outwardly by the child, toward the external social environment. Externalizing behavior usually involves behavior excess (i.e., too much behavior) and is considered inappropriate by teachers. Non-examples of externalizing behavior would include all behavior that is appropriate for their age and the school." The definition of

TABLE 2.
Number of Classrooms by State

State	Classrooms
California	28
Kentucky	42
Louisiana	16
Nebraska	5
New Hampshire	1
Oregon	9
Texas	27
Utah	18
Washington	14
Total	160

FIGURE 2.
ESP Procedure



internalizing behavior states: "Internalizing refers to all behaviors that are directed inwardly (i.e., away from the external environment) and that represent problems *within* the child. Internalizing behavior frequently involves behavioral deficits and patterns of social avoidance and withdrawal. Non-examples of internalizing behavior problems would be social behavior that shows social involvement with other children in expected social development."

Stage Two. In Stage Two, teachers rated the children who were either (a) the top three of the five ranked for externalizing behavior, (b) the top three of the five ranked for internalizing behavior, or (c) in the comparison group of one male and one female not listed on either the externalizing and internalizing behavioral dimension. Stage Two consisted of four behavioral measures: (a) Critical Events Index, (b) Aggressive Behavior Scale, (c) Adaptive Behavior Scale, and (d) Maladaptive Behavior Scale. The Critical Events Index con-

tained 16 occurrence or non-occurrence items. The Aggressive Behavior Scale contained nine items and was only used with children ranked on the externalizing dimension. The Adaptive Behavior Scale contained eight items representing overall prosocial behavior. The Maladaptive Behavior Scale consisted of nine items representing overall antisocial or nonsocial behavior. Items were carefully worded in order to facilitate the completion of the ESP by a diverse group of preschool teachers who may have limited experience in assessment.

Stage Three. Stage Three observations were conducted on children who were either the (a) top-ranked child on the externalizing behavioral dimension, (b) top-ranked child on the internalizing behavioral dimension, or (c) from the comparison group of one male and one female not listed on either the externalizing or internalizing behavioral dimension. Stage Three measures were based on direct observations of a child's social behavior in the classroom or on the playground. The Social Behavior observation procedure provides a record of the quality, level, and distribution of a child's social behavior during free play settings. Target child behavior was also coded as either antisocial or nonsocial. Antisocial or nonsocial behavior was defined as: (a) a *negative* exchange involving either verbal or physical interaction, (b) disobeying established classroom rules, (c) tantrumming, and (d) solitary play. If the target child engaged in social behavior other than antisocial or nonsocial, it was coded as social (appropriate).

Each child was observed for 20 minutes, 10 minutes on two separate occasions. If the total time was under 20 minutes, a third observation was conducted to bring the time up to 20 minutes. As noted above, in recording these observations, the stopwatch ran when the child exhibited antisocial or nonsocial behavior and was turned off when the child displayed prosocial behavior. The stopwatch was

then restarted whenever the child again exhibited antisocial or nonsocial behavior, and the process was repeated throughout the observation session. The amount of time (antisocial or nonsocial behavior) recorded on the stopwatch was divided by the total time, resulting in a percent of time engaged in antisocial or nonsocial behavior.

Observations were conducted in Oregon by a group of trained personnel (e.g., graduate students) and elsewhere by local professionals trained by project personnel on site. For Oregon observers, training sessions were conducted until each observer could code videotaped scenes at a 90% accuracy (average training time was 6 hours). The observers were blind to subject ranking, history, and accuracy; a random 20% of observations were completed in pairs to ascertain interrater reliability. For sites outside of Oregon, observers were professionals in the community (e.g., Special Education consultants) who attended a day-long inservice training on ESP administration procedures, but were not blind to subject ranking or history.

Technical adequacy of ESP. Beginning in 1991, studies on the ESP have been conducted to assess its reliability and validity. These findings have been very promising (Feil & Becker, 1993; Feil, Walker, & Severson, 1995). The interrater reliability coefficients of most ESP measures are at least .80, which meets Salvia and Ysseldyke's (1988) guidelines for a screening instrument. Good psychometric standards were attained despite the difficulties inherent in the assessment of young children (Martin, 1986). ESP validity studies show consistently high relationships to assessment validity measures (i.e., Conners' Teacher Rating Scales, Conners, 1989; Preschool Behavior Questionnaire, Behar & Stringfield, 1974). Correlations resulted in highly significant coefficients ranging from .34 to .87, with most above .70. The consistency of scores across ESP measures illustrates

the potential utility of the screening system, providing additional evidence that behavior problems may be identified accurately among preschool children.

Concurrent measures. Due to the variety of sites and programs and the serial nature of the data collection, concurrent validity data were collected on either the Child Behavior Checklist-Teacher Report Form (Achenbach & Edelbrock, 1983) or the Conners' Teacher Rating Scale (Conners, 1989). While providing data on two measures makes a more persuasive study, the lack of consistency in the administration of the concurrent measure is a limitation. The Child Behavior Checklist (CBC) and Conners' Rating Scale are widely-used teacher-completed behavior checklists for young children. In a review of the CBC, Wilson and Bullock (1989) cite a .89 test-retest reliability and note that the instrument correctly identified 76% of children referred for emotional and behavior problems from non-referred children. One-month test-retest reliability of the Conners' questionnaire ranged from .71 to .92 for a sample of 578 children 3 to 17 years of age (Goyette Conners, & Ulrich, 1978). Teachers in the ESP completed one of the concurrent measures at the same session as the completion of Stages One and Two instruments. They rated the children who were either (a) in the top three of the five ranked on the externalizing behavioral dimension, (b) in the top three of the five ranked on the internalizing behavioral dimension, or (c) in the comparison group of one male and one female not listed on the externalizer or internalizer list.

Results of Screening Procedures

The ESP provides normative criteria of "at-risk status" (one standard deviation from the mean; Walker et al., 1995). A child can only progress from Stage Two to Stage Three if his or her scores are within the risk range (e.g., at least one standard deviation from the mean) on the

TABLE 3.
ESP Criteria and Cut-off Scores per Gender and Stage

Stage	Criteria	Boys' Cut-off Scores	Girls' Cut-off Scores
One: Teacher Ranking	Ranked as one of top three on externalizing or internalizing procedures.	Ranked in top three on externalizing or internalizing ranking.	Ranked in top three on externalizing or internalizing ranking.
Two: Teacher Ratings	Scores exceeding 1 <i>SD</i> from <i>M</i> on Stage Two measures.	Critical Events: ≥ 2 Aggressive: ≥ 15 Adaptive: ≤ 25 Maladaptive: ≥ 20	Critical Events: ≥ 2 Aggressive: ≥ 14 Adaptive: ≤ 27 Maladaptive: ≥ 20
Three: Direct Observation	Observations of anti- or nonsocial behavior exceeding 1 <i>SD</i> from <i>M</i> .	$\geq 40\%$ observation time engaged in aggressive or solitary play.	$\geq 37\%$ observation time engaged in aggressive or solitary play.

relevant measures. Table 3 provides the criteria and cut-off scores for each stage of the ESP. Using these methods of screening and cutoff points, we obtained a complete picture of the number, gender, and behavioral characteristics (i.e., externalizing and internalizing) for children at each stage of ESP assessment.

Table 4 gives number and gender distribution at each stage of the ESP assessment. The first row shows the sample of 2,797 children

by gender. The next row profiles the 947 boys and girls that were ranked by their teachers as children exhibiting high levels of (a) aggressive and externalizing behaviors or (b) withdrawn and related internalizing behaviors. Even at the first stage of simply ranking by behavioral dimension, gender differences quickly became a factor: Boys were overly represented in the externalizing classification (12.6% compared to 4.3% for girls). In the

TABLE 4.
Results of Gating Procedure: Numbers and Percent of Children by Gender and Behavioral Grouping per ESP Stage

Children ranked in Stage One	Males <i>n</i> = 1,528 (55%)		Females <i>n</i> = 1,269 (45%)		Total <i>N</i> = 2797
	Externalizing <i>n</i> (%)	Internalizing <i>n</i> (%)	Externalizing <i>n</i> (%)	Internalizing <i>n</i> (%)	
Top three rankings (Stage One)	353 (12.6%)	255 (9.1%)	121 (4.3%)	218 (7.8%)	947
Exceeding normative criteria in Stage Two ratings	167 (6.0%)	101 (3.6%)	53 (1.9%)	89 (3.2%)	410
Exceeding normative criteria in Stage Three observation	23 (0.8%)	15 (0.5%)	7 (0.3%)	17 (0.6%)	62

TABLE 5.
Means and Standard Deviations of the Child Behavior Checklist Aggression Subscale by Gender per ESP Stage

Children/Stage	Male		Female		Total	
	<i>n</i> (<i>SD</i>)	<i>M</i>	<i>n</i> (<i>SD</i>)	<i>M</i>	<i>n</i> (<i>SD</i>)	<i>M</i>
Ranked in Stage One	319 (12.5)	11.8	226 (10.6)	07.6	545 (11.9)	10.1
In top three Stage One rankings	242 (12.9)	14.7	142 (11.8)	10.4	384 (12.7)	13.1
NOT in top three Stage One rankings	77 (4.0)	02.8	84 (5.7)	03.0	161 (4.9)	02.9
Exceeding normative criteria in Stage Two ratings	122 (13.1)	21.9	64 (13.6)	14.4	186 (13.7)	19.3
NOT exceeding normative criteria in Stage Two ratings	197 (6.8)	05.6	162 (7.7)	04.9	359 (7.2)	05.3
Exceeding normative criteria in Stage Three observation	26 (13.5)	22.1	12 (15.9)	14.2	38 (14.5)	19.6
NOT exceeding normative criteria in Stage Three observation	293 (12.0)	10.9	214 (10.2)	07.3	507 (11.4)	09.4

internalizing dimension, girls have a more substantial representation, with 7.8% versus 9.1% for boys. The third row shows the number of children ($n = 410$) exceeding normative criteria in Stage Two ratings and thus observed in free play settings. The last row shows the number of children who exceed normative criteria on Stage Three of the ESP and would be eligible for further assessment and, if indicated, a pre-referral intervention. At Stage Three, gender predominance is most evident by behavioral dimension. Twice as many boys were in the externalizing dimension, whereas near gender parity (15 boys and 17 girls) existed in the internalizing dimension. A series of Chi-square analyses showed a significant difference for gender and children in the highest Stage One rankings $\chi^2(1, N = 2797) = 52.94, p < .001$ and Stage Two ratings $\chi^2(1, N = 2797) =$

22.34, $p < .001$, but there was not a significant gender difference for Stage Three observations $\chi^2(1, N = 2797) = 1.13, p = .29$.

A clear majority of the externalizing ranking group was composed of boys while girls were more represented in the internalizing ranking group. This gender split is characteristic of past research indicating that females are less likely to manifest aggressive behavior and more likely to manifest internalizing problems (Achenbach, 1993). The gender dominance effects on each behavioral dimension increased with each successive assessment stage. At the final referral stage (i.e., passing gate three), males accounted for 77% of the externalizers and females accounted for 53% of the internalizers.

Confirmation of referral status using the ESP system was examined by t-tests between

TABLE 6.*Means and Standard Deviations of the ESP measures by gender and referral status*

Measure Referral Status	Male		Female		Total	
	<i>n</i> (<i>SD</i>)	<i>M</i>	<i>n</i> (<i>SD</i>)	<i>M</i>	<i>n</i> (<i>SD</i>)	<i>M</i>
Critical Events Checklist						
Exceeding normative Stage Three referral criteria	38 (1.9)	03.5	24 (1.7)	02.7	62 (1.9)	3.2
NOT exceeding Stage Three referral criteria	720 (1.5)	01.1	589 (1.3)	01.0	1309 (1.4)	1.1
Aggressive Behavior Scale						
Exceeding ESP referral criteria	38 (8.0)	22.9	24 (9.4)	17.0	62 (9.0)	20.6
NOT exceeding ESP referral criteria	720 (6.7)	14.7	589 (4.9)	12.0	1309 (6.2)	13.5
Adaptive Rating Scale						
Exceeding ESP referral criteria	38 (4.1)	19.8	24 (5.2)	20.6	62 (4.6)	20.1
NOT exceeding ESP referral criteria	720 (7.8)	28.1	589 (7.4)	31.3	1309 (7.8)	29.4
Maladaptive Rating Scale						
Exceeding ESP referral criteria	38 (8.2)	28.8	24 (9.5)	22.7	62 (9.2)	26.4
NOT exceeding ESP referral criteria	720 (8.8)	19.4	589 (8.0)	16.1	1309 (8.6)	18.0
Social Behavior Observation						
Exceeding ESP referral criteria	38 (0.2)	0.6	24 (0.2)	0.6	62 (0.2)	0.6
NOT exceeding ESP referral criteria	271 (0.2)	0.2	229 (0.2)	0.2	500 (0.2)	0.2

referred and non-referred students on concurrent measures; statistically significant differences were found. Results of t-tests showed highly significant differences on both measures between children who passed Gate Three criteria and children who did not meet criteria for Gate Three and were therefore referred by the ESP system: Conners'

Hyperactive subscale $t(179) = 3.90, p < .001$; Conners' Inattention subscale $t(179) = 3.61, p < .001$; Child Behavior Checklist Aggression subscale $t(41) = 4.23, p < .001$; and Child Behavior Checklist Withdrawal subscale $t(59) = 3.97, p < .001$. Descriptive data by gender on the Child Behavior Checklist Aggression subscale, comparing children exceeding

criteria and those not exceeding criteria per ESP stage, are found in Table 5. Further descriptive data of ESP measures by gender comparing children exceeding Stage Three referral criteria and children not exceeding referral criteria are found in Table 6.

DISCUSSION

In order to meet the needs of practitioners, an effective child-find screening system should be cost-effective, proactive, universal in nature, accurate, and able to identify children displaying both externalizing (e.g., aggressive) and internalizing (e.g., socially withdrawn) behaviors. The ESP procedure represents three innovations in screening for preschoolers with behavior problems: (a) a time- and cost-efficient procedure, (b) a proactive, universal screening system, and (c) screening of internalizing behavioral characteristics to include a more representative number of girls.

Martin's (1986) review of literature on the assessment of social and emotional functioning of preschool children shows that some technically and financially feasible screening instruments are available, but are not cost-effective and require lengthy administration time. To provide quality early screening and intervention, the educational system must confront increasing costs and dwindling budgets. A teacher with 20 students could spend as much as 5 hours to screen all children in the classroom with technically adequate rating scales. Because this is impractical, most rating scales are utilized on a reactive rather than a proactive basis, thus action is taken only after a teacher referral is made (Walker & Severson, 1990). The Early Screening Project is an easy to use assessment that has received positive reports from users (Yoshikawa & Knitzer, 1997). For example, a preschool director stated that she expects the ESP will increase the credibility of her staff when they make referrals to the local early childhood

special education program. The ESP thus may make a difference in obtaining timely referrals, diagnoses, and follow-through for preschool children showing behavior and emotional problems (Yoshikawa & Knitzer).

A proactive system offers several advantages over currently used, reactive referral procedures. As children with emotional or behavioral problems develop, their problematic behaviors become more severe (Loeber, 1991; Patterson et al., 1992). For many children, behavior problems follow a progression of: (a) disobedience in the home, (b) temper tantrums, and (c) teacher reports of fighting and stealing. Longitudinal research has found that behavior problems in preschool is the single best predictor of antisocial disorders at age 11 (White, Moffitt, Earls, Robins, & Silva, 1990). Youth with early childhood onset of antisocial behavior patterns (i.e., the most severe) account for approximately 3% to 5% of the school-age population, but they account for nearly 50% of all crimes committed by children (Hinshaw et al., 1993). Prevention is predicated on a proactive, early detection strategy: Early intervention costs less and can prevent future escalation in serious behavior problems. Compensatory programs, such as Head Start, follow such a proactive model. The early identification model funnels energy toward prevention instead of toward more costly and less successful adolescent and adult remediation.

As shown in Table 3, males dominated the ESP externalizing dimension (as rated in Stage One), while gender was more equally distributed in the internalizing rankings. Throughout the screening stages, the ratio of boys to girls for externalizing remained close to 3:1, but for children ranked in the internalizing dimension the ratio was closer to 1:1. This also occurred with a nationwide psychopathology study in which teachers rated 5 to 8 year-old males as significantly more aggressive than females, but not different from

females on social avoidance (internalizing) behaviors (McDermott, 1996).

Information collected using the ESP measures at Stages Two and Three can be used for several purposes. These include (a) determining eligibility for special services, (b) intervention planning, and (c) monitoring progress and evaluating outcomes of intervention. The ESP should, however, be used in conjunction with other additional quantitative and qualitative assessment information. Even when a child exceeds normative criteria on the ESP measures, the child's behavior could be due to a variety of factors, such as activity levels, poor supervision and instruction, speech and language delays, parental divorce or stress, immaturity, and cognitive delays. The ESP provides a broad range of information, yet a complete assessment should include other types and sources of information, including professional (clinical) judgment. Likewise, parent and teacher interviews are highly recommended for students exceeding ESP criteria.

A limitation of the ESP is its dependence on children already in a preschool program. While this does not present problems for Head Start centers, it is a disadvantage for professionals working in home settings. The ESP is an effective screening and assessment measure only for children who attend preschool in a group setting with a teacher. Other measures to assess a child's behavior might include parent ratings (Achenbach & Edelbrock, 1983) or parent-child observations (Barkley, 1990; Mash & Barkley, 1986).

Overall, the ESP can be utilized as part of best and preferred practices for early intervention programs where systematic screening for school adjustment problems occurs on a regular basis. Preschool programs, facing increasing child accountability requirements (e.g., child-find) need to maximize their resources (e.g., teachers' knowledge and experience) within a proactive and equitable system. The universal nature of the ESP pro-

vides assessment of adjustment problems for all children in a classroom, not simply those children with externalizing behaviors that are so disruptive a teacher will refer them for evaluation. The ESP also screens for internalizing characteristics (e.g., socially withdrawn behaviors) which are frequently overlooked because they do not disrupt classroom activities or pressure teachers. We believe the ESP can minimize the time invested in preschool assessments while providing accurate identification of students exhibiting problem behaviors that put their school success at risk.

REFERENCES

- Achenbach, T. M. (1993). Taxonomy and comorbidity of conduct problems: Evidence from empirically-based approaches. *Development & Psychopathology, 5*, 61-64.
- Achenbach, T. M., & Edelbrock, C. (1983). *Manual for the Child Behavior Checklist and Revised Child Behavior Profile*. Burlington, VT: University of Vermont, Department of Psychiatry.
- Achenbach, T. M. & McConaughy, S. H. (1992). Taxonomy of internalizing disorders of childhood and adolescence. In W. M. Reynolds (Ed). *Internalizing disorders in children and adolescents*. New York: John Wiley & Sons.
- Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101*(2), 213-232.
- American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.
- Bagnato, S. J. & Neisworth, J. T. (1991) *Assessment for early intervention: Best practices for professionals*. New York: Guilford.
- Barkley, R. A. (1990) *Attention deficit hyperactivity disorder: A handbook for diagnosis and treatment*. New York: Guilford.
- Barnett, D. W., Carey, K. T., & Hall, J. D. (1993). Naturalistic Intervention Design for Young Children: Foundations, Rationales, and Strategies. *Topics in Early Childhood Special Education, 13*, 430-444.

- Beare, P. L., & Lynch, E. C. (1986). Under identification of preschool children at risk for behavioral disorders. *Behavioral Disorders, 11*, 177-183.
- Behar, L., & Stringfield, S. (1974). *Manual for the preschool behavior questionnaire*. Durham, NC: Behar.
- Cairns, R. B. & Green, J. A., (1989). How to assess personality and social patterns: Observations or ratings. In R. B. Cairns (Ed). *The analysis of social interaction: Methods, issues, and illustrations*. Hillsdale, NJ: Lawrence Earlbaum and Associates.
- Callahan, K. (1994) Wherefore art thou, Juliet? Causes and implications of the male dominated sex ratio in programs for students with emotional and behavioral disorders. *Education and Treatment of Children, 17*(3), 228-243
- Campbell, S. B. (1994) Hard-to-manage preschool boys: Externalizing behavior, social competence, and family context at two-year follow-up. *Journal of Abnormal Child Psychology, 22*(2), 147-166.
- Caseau, D. L., Luckasson, R., & Kroth, R. L. (1994). Special education services for girls with serious emotional disturbance: A case of gender bias. *Behavioral Disorders, 20*(1), 51-60.
- Conners, C. K. (1989). *Manual for the Conners' rating scales*. North Tonawanda, NY: Multi-Health Systems.
- Eisert, D. C., Walker, H. M., Severson, H., & Block, A. (1989). Patterns of social-behavioral competence in behavior disordered preschoolers. *Early Childhood Development & Care, 41*, 139-152.
- Feil, E. G. & Becker, W. C. (1993) Investigation of a multiple-gated screening system for preschool behavior problems. *Behavioral Disorders, 19*(1), 44-55.
- Feil, E. G., Walker, H. M., & Severson, H. H. (1995). Young children with behavior problems: Research and development of the early screening project. *Journal of Emotional and Behavioral Disorders, 3*(4), 194-202.
- Goyette, C. H., Conners, C. K., & Ulrich, R. F. (1978) Normative data on revised Conners Parent and Teacher Rating Scales. *Journal of Abnormal Child Psychology, 6*, 221-236.
- Gresham, F. M., & Elliot, S. N. (1984) Assessment and classification of children's social skills: A review of methods and issues. *School Psychology Review, 13*, 292-301.
- Heppner, P. P., Kilvighan, D. M., & Wampold, B. E. (1991) *Research Design in Counseling*. Pacific Grove, CA: Brooks/Cole.
- Hill, G. K. (1993, January 6). A fight for funding. *Oregonian*. p. D4.
- Hinshaw, S. P., Lahey, E. B., & Hart, E. L. (1993) Issues of taxonomy and comorbidity in the development of conduct disorders. *Development & Psychopathology, 5*, 31-49.
- Kazdin, A. E. (1979) Situational specificity: The two-edged sword of behavioral assessment. *Behavioral Assessment, 1*, 57-75.
- Loeber, R. (1991). Antisocial Behavior: More Enduring than Changeable? *Journal of American Academy of Child and Adolescent Psychiatry, 30*, 393-397.
- Martin, R. P. (1986). Assessment of the social and emotional functioning of preschool children. *School Psychology Review, 15*, 216-232.
- Mash, E. J. & Barkley, R. A. (1986) Assessment of family interaction with the Response Class Matrix. In R. Prinz (Ed.), *Advances in behavioral assessment of children and families* (Vol. 2, pp. 29-67). Greenwich, CT: JAI Press.
- McDermott, P. A. (1996). A nationwide study of developmental and gender prevalence for psychopathology in childhood and adolescence. *Journal of Abnormal Child Psychology, 24*(1), 53-66.
- McMahon, R. J. (1984) Behavioral checklists and rating scales. In T. H. Ollendick & M. Hersen (Eds.) *Child behavioral assessment: Principals and procedures* (pp. 80-105). New York: Pergamon Press.
- Meisels, S. M. (1991). Dimensions of Early Identification. *Journal of Early Intervention, 15*(1), 26-35.
- Meisels S. J. & Provence, S. (1989). *Screening and assessment: Guidelines for identifying young disabled and developmentally vulnerable children and their families*. Washington, DC: National Center for Clinical Infant Programs.
- Paget, K. D. (1990). Best practices in the assessment of competence in Preschool-age-children. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology-II*, (pp. 107-119). Washington, DC: NASP.
- Patterson, G. R., Reid, J. B., & Dishion, T. J. (1992). *Antisocial boys*. Eugene, OR: Castalia .
- Ross, A. O. (1980). *Psychological disorders of child-*

- hood, New York: McGraw-Hill.
- Salvia, J., & Ysseldyke, J. E. (1988). *Assessment in special & remedial education*. Boston: Houghton Mifflin Company.
- Schaughency, L. A., & Rothlind, J. (1991). Assessment and classification of attention deficit hyperactive disorders. *School Psychology Review, 2*(20), 187-202.
- Sinclair, E., Del'Homme, M., Gonzalez, M. (1993). Systematic screening for preschool behavior disorders. *Behavioral Disorders, 18*(3), 175-185.
- Walker, H. M. & Rankin, R. (1983). Assessing the behavioral expectations and demands of less restrictive settings. *School Psychology Review, 12*(3), 274-284
- Walker, H. M. & Severson, H. H. (1990). *Systematic screening for behavior disorders user's guide and administration manual*. Longmont, Colorado: Sopris West.
- Walker, H. M., Severson, H. H., & Feil, E. G. (1995). *Early screening project: A proven child-find process*. Longmont, CO: Sopris West.
- Walker, H. M., Severson, H. H., Nicholson, F., Kehle, T., Jenson, W. R., & Clark, E. (1994). Replication of the Systematic Screening for Behavior Disorders (SSBD) procedure for the identification of at-risk children. *Journal of Emotional and Behavioral Disorders, 2*(2), 66-77.
- White, J. L., Moffitt, T. E., Robins, L., Silva, P.A., & Earls, F. (1990). How Early Can We Tell?: Predictors of Early Childhood Conduct Disorder and Adolescent Delinquency. *Criminology, 28*, 507-533.
- Wilson, M. J., & Bullock, L. M. (1989). Psychometric characteristics of behavior rating scales: Definitions, problems, and solutions. *Behavioral Disorders, 14*(3), 186-200.
- Yoshikawa, H. & Knitzer, J. (1997) *Lessons from the Field: Head Start Mental Health Strategies to Meet Changing Needs*. NY: National Center for Children in Poverty.

The authors would like to thank Laura Akers for her editorial expertise and invaluable assistance in the preparation of this manuscript. This research was supported in part through grants from the U.S. Department of Education, Office of Special Education and Rehabilitative Services, Research in Education of the Handicapped Programs: Student-Initiated and Field-Initiated Research, and the U.S. Department of Health and Human Services, Administration for Children and Families: Head Start Research Fellows Program. The views expressed are not necessarily those of the U.S.D.O.E. or U.S.D.H.H.S. Requests for reprints and/or more information on the ESP should be sent to Edward G. Feil, Oregon Research Institute, 1715 Franklin Blvd., Eugene, OR 97403, or e-mail: edf@ori.org.