



**Validation of the Nipissing District Developmental Screen
For Use With Infants and Toddlers - Working Paper**

By

V. Susan Dahinten and Laurie Ford

University of British Columbia

November 15, 2004

The authors wish to acknowledge the families and early childhood professionals in the Chilliwack community for their support in gathering the information provided in this report. We also recognize the support of our research team members, with particular thanks to Susan Anstett, Carla Merkel, Sarah Van Leeuwen, and Vanessa Lapointe.

Funded by the Social Sciences and Humanities Research Council of Canada

For more information, contact:

V. Susan Dahinten, PhD, RN
Assistant Professor, School of Nursing
University of British Columbia
T201-2211 Wesbrook Mall
Vancouver, BC V6T 2B5
604-822-7437 (Tel)
604-822-7466 (Fax)
dahinten@nursing.ubc.ca



Research Team

University of British Columbia Faculty

V. Susan Dahinten, PhD, RN
Assistant Professor
School of Nursing

Laurie Ford, PhD
Associate Professor,
Department of Educational and Counseling Psychology, and Special Education

Connie Canam, PhD, RN
Assistant Professor
School of Nursing

Community Partners

Margaret Gander
Manager Health Services
Maternal-Child/Public Health Prevention
Chilliwack Health Unit
Fraser Health Authority

Gillian Youngberg
Supervisor, Public Health Nursing
Chilliwack Health Unit
Fraser Health Authority

Susan Anstett
Consultant
Screening Project Task Group
Eastern Fraser Valley Make Children First Initiative

Research Assistants

Carla Merkel
Sarah Van Leeuwen
Vanessa Lapointe

ABSTRACT

The early identification of developmental delays is a prerequisite for early intervention. While severe developmental delays may be recognized at birth, other delays often remain undetected until the child begins school. The need for widespread screening, combined with evidence that parents can be effectively assisted to identify developmental delays using parent-completed screening tools, has led to increased interest in universal, community-based developmental surveillance throughout the early years and an increased demand for screening instruments that are relatively inexpensive and easily administered.

The purpose of this study was to evaluate the concurrent validity of the Nipissing District Developmental Screen (NDDS) for 118 children who were 4, 18, or 24 months old, by comparing NDDS results to those obtained through direct child assessments (using the *Bayley Scales of Infant Development-II*). The NDDS is a screening measure completed by the parent or service provider that has recently come into use in Canada for large-scale screening. In order to illustrate the importance of the cut-off point, selected psychometric calculations were conducted using a variety of cut-off points for both the standardized assessment and the screening tool.

The results suggest that the NDDS is effective at capturing children with severe delays. Children with mild to moderate delays were less well identified. Findings also illustrate the importance of the cut-off point for both the screening test and the criterion measure when assessing the validity of a screening instrument. The report concludes with a discussion of issues that should be taken into account when selecting a developmental screening tool and recommends that there be further psychometric validation of the NDDS, with larger samples, and among other age groups, using a psychometrically sound direct measure of child development.

VALIDATION OF THE NIPISSING DISTRICT DEVELOPMENTAL SCREEN FOR USE WITH INFANTS AND TODDLERS

The early identification of developmental delays is a prerequisite for early intervention. While severe developmental delays may be recognized at birth, other delays often remain undetected until the child begins school. Developmental screening is an important role of public health nurses and other early childhood professionals, but in many areas, only those children who are considered to be at high risk for developing delays are screened on a regular basis.

Epidemiological evidence, however, shows that the determinants of developmental delays are not singular nor easily identified at the individual level (King, Logsdon, & Schroeder, 1992; Willms, 2002), and thus there is a need for more widespread screening. This need, combined with evidence that parents can be effectively assisted to identify developmental delays using parent-completed screening tools (Glascoe & Dworkin, 1995; Glascoe, 1999; Montgomery, 1999; Squires, 1996), has led to increased interest in universal, family and community-based developmental surveillance throughout the early years and an increased demand for screening instruments that are relatively inexpensive and easily administered.

A community-based program of universal screening was implemented in the Eastern Fraser Valley in April, 2002 as part of the Make Children First Initiative. The screening program targets children who are 4, 18, and 24 months old, but accommodates children from birth to 6 years of age. Although the primary aim of the screening program is to facilitate early identification of developmental delays and more timely intervention, other important goals are to raise parents' awareness about the importance of early child development and to increase the capacity of families and the community to support healthy childhood development. The relatively new Nipissing District Developmental Screen (NDDS) was selected by the Screening Project Task Group of the Make Children First Initiative on the basis of its short length, ease of administration, potential educational benefits of the accompanying parent handout, and promising preliminary results from the Ontario Healthy Babies/Healthy Children evaluation.¹ Soon after implementation of the screening program, a team of UBC researchers was invited to partner with the community to conduct an evaluation study.

The research team developed a 5-year research plan to evaluate the concurrent and predictive validity of the NDDS, and to assess the impact of implementing an integrated and universal developmental screening system on the community's capacity to support healthy childhood development. Seven research questions will be addressed over the 5-year period, using a variety of research methods (See Appendix A). The study is funded by the Social Sciences and Humanities Research Council of Canada as part of the Consortium for Health, Intervention, Learning and Development (CHILD) project.

The purpose of this part of the study was to evaluate the concurrent validity of the NDDS for children 4, 18, and 24 months of age, by comparing NDDS results to the results of standardized child assessments. One component of this was an evaluation of the benefits of using a *one-flag* or *two-flag* criterion for identifying children at risk for developmental delay.

¹ At that time, the Ages and Stages Questionnaire (ASQ) was being used by public health nurses in the area to screen children with suspected delays and to monitor the development of at-risk children.

Nipissing District Developmental Screen™

The NDDS is a parent-completed measure that has recently come into use and is rapidly gaining popularity for large-scale screening in British Columbia (<http://www.ndds.ca>). There are 13 age specific versions of the NDDS, ranging from 1 month to six years (with the number of items ranging from 4 to 22, respectively). The questionnaires address vision, hearing, speech-language, gross motor, fine motor, cognitive, and self-help skills. Responses are scored as a simple yes/no checklist, and when the tool was originally developed, the selection of one or more ‘no’ responses (i.e., the child does not do the behavior in question) indicated the need for further assessment and/or referral. This is known as the one-flag rule. (The two-flag rule requires a minimum of two ‘no’ responses for referral.) The NDDS is published in French, Spanish, and Chinese, and is being translated into Vietnamese.

The NDDS was developed in Ontario during the mid-90s to meet the demand for a comprehensive but time efficient tool that addressed local needs. Items were drawn from a variety of standardized and non-standardized measures. In spite of its increasingly widespread use throughout Canada, and its introduction to the United States in 2001, the NDDS was only recently evaluated, and the scoring rules revised, as part of the 2001 – 2002 Ontario Healthy Babies/Healthy Children (HBHC) evaluation. (Nagy, Ryan, & Robinson, 2002). Preliminary results looked promising. The HBHC study (N = 238) yielded an overall inter-rater agreement rate of 71% between the responses of the parent and non-parent caregiver (i.e., day care staff), and showed that the results of the NDDS were stable between 12 months and 18 months for 65% of the sample. The study also yielded overall agreement rates of 78% when using the original one-flag rule, and 93% when using the newly recommended two-flag rule, to compare the results of the 12-month NDDS with the 12-month Ages and Stages Questionnaire (ASQ, Squires, Potter, & Bricker, 1999).² (Sensitivity and specificity rates were not reported for the NDDS, but our calculations with the cross-tabulated data produced sensitivity and specificity rates of 83% and 95%, respectively, when using the two-flag rule. See Appendix B). However, because the ASQ was also completed by the parent, and because it was used as the sole criterion measure for establishing concurrent validity of the NDDS, the HBHC validation study was limited by its reliance on single-source data. Moreover, given the exclusive focus on the 12-month NDDS, there was a need for further evaluative work. The purpose of this study was to evaluate the concurrent validity of the NDDS for children 4, 18, and 24 months of age, by comparing the NDDS results to the results of standardized child assessments, using the two different scoring rules.

² The ASQ is a measure of early childhood development that is widely used in both research and clinical practice. It has been shown to have fairly strong psychometric properties in terms of overall concurrent validity, although validation work conducted by the ASQ authors showed that sensitivity rates varied by age of the subsample, ranging from 51% among 4-month-olds to 90% among 36-month-olds (Squires et al., 1999). A more recent study of ex-premature infants who were 12, 18, 24 or 48 months old (corrected age), yielded the lowest sensitivity rates among the 18-month-olds (50%) when compared with 12- and 48-month-olds (100%), but the sample sizes of the sub-groups were small (Skellern, Rogers & O’Callaghan, 2001).

METHODS

Sample

Approximately forty children at each of the ages 4, 18, and 24 months were recruited from the Chilliwack area in the Eastern Fraser Valley. Participants were recruited through newspaper advertising and by direct advertising at the public health immunization clinic, a community Family Place centre, and at child health fairs (the three primary venues for the developmental screening program). Recruitment posters were also distributed to physicians' offices, day care centres, preschools, supermarkets, recreation centres, and the library. We sought to obtain a sample that represented the general population and included a range of socioeconomic and biological risk factors, but eligibility criteria required that the responding parent and child speak English. Two children were excluded from the sample in situations where illness or fatigue were believed to have interfered with completion of the assessment, for a final sample size of 118. The mother was the respondent for 111 (94%) of the children in the study.

The sample showed considerable diversity in family socioeconomic characteristics, and 7.6% of the children had been born prematurely according to the caregiver report (although their gestational ages ranged only from 34 to 36.5 weeks). Demographic characteristics are presented in Table 1. Twenty-seven percent of the children lived in families whose household income was less than \$30,000—this is the approximate “low income cut-off” (LICO) for a family of four living in small urban or sub-urban areas in Canada; but another 36% ($n = 43$) lived in families with incomes of \$60,000 or more. Fourteen percent of the children were born to mothers who had less than high school education and 14% were born to mothers who began childbearing during their adolescence (prior to age 20). Most of the children lived in two-parent families (82%) or blended or extended families (10%); only 8% lived in lone-parent families.

Table 1. Characteristics of the Sample

Characteristic	Total N = 118		4 months n = 38	18 months n = 40	24 months n = 40	χ^2
	n	%	%	%	%	
Sex						
Male	54	46%	45%	50%	43%	.477
Female	64	54%	55%	50%	57%	
Teen Mother	17	14%	18%	13%	13%	.732
Lone Parent	9	8%	3%	10%	10%	1.990
Mother's Education						
< High school	17	14%	21%	8%	15%	2.920
= High school	16	14%	8%	20%	13%	
> High school	85	72%	71%	72%	72%	
Income < LICO	32	27%	24%	33%	25%	.904

Note:

LICO = low income cut-off, defined by Statistics Canada.

All chi-square statistics were non-significant at $p > .05$.

These demographic statistics indicate that the sample consisted of a higher proportion of lower income families than is representative of the Chilliwack area population (which is not altogether unexpected given that the sample is limited to families with young children), but was comprised of fewer lone-parent families than expected (BC Statistics).³ There were no statistically significant differences in the demographic characteristics of the three age groups, but it is unlikely that such differences would have been found given the small size of the subgroups.

Measures and Data Collection

We conducted direct assessments of the child's overall development and learning using the Mental Development Index (MDI) of the *Bayley Scales of Infant Development-II* (BSID-II, Bayley, 1993). The BSID-II was selected for use in this study because of its history, psychometric properties,⁴ and appropriateness for children 4 to 24 months of age. It is one of the mostly widely used measures of development for infants and toddlers in both research and clinical practice. The MDI yields standard scores with a mean of 100 and standard deviation (SD) of 15.

Most of the standardized assessments ($n = 101$, 86%) were administered immediately following the parent-completed screening; the remaining 14% took place within two weeks of the initial screening ($M = 6.6$ days, $SD = 4.0$ days). The assessments were conducted at a centralized location (either the public health unit or the community Family Place centre). The child assessments were administered by graduate student research assistants (from the fields of psychology and nursing) who were trained and supervised by one of the members of the research team who is a licensed psychologist. The direct assessments took 20 to 45 minutes depending upon the age of the child. Parents also completed a short demographic questionnaire.

Data Analysis

We calculated the sensitivity and specificity, the positive and negative predictive values, and the over- and under-referral rates for the NDDS based on cross-tabulations showing the agreement between the results of the parent-completed screening test and the BSID-II standardized assessments. The likelihood ratios of positive and negative tests were also calculated. The likelihood ratio of a positive test quantifies how much more likely it is that a developmental delay will be identified in a child who actually is delayed compared with a child who is not, and is particularly useful in determining where to set the cut-off point in a screening test. Likelihood ratios are becoming an increasingly popular means of evaluating and describing the utility of screening tests (Greenhalgh, 1997). See Table 2 for a full set of definitions.

³ Seventeen percent of Chilliwack-Kent families were identified as lone-parent families in the 2001 Canada Census (BC Statistics, 2001).

⁴ Validation studies of the Mental Development Index (MDI) of the BSID-II have yielded median reliability coefficients of .88, .92, and .92 for children ages 4, 18, and 24 months respectively, with a median reliability of .88 among all ages and a test-retest reliability of .87. The stability coefficient for the MDI has been found to be .83 for the 1 to 12 month age range and .91 for the 24 to 36 month range (Bayley, 1993).

Table 2. Definition of Screening Test Characteristics

Sensitivity	The ability of the test to correctly identify children with developmental delays (true positives).
Specificity	The ability of the test to correctly identify children without developmental delays (true negatives).
Total Agreement	The overall accuracy of a screening test. The proportion of all screening tests that correctly classified the child as either a true positive or a true negative.
Positive Predictive Value	The probability that a child who screens positive actually has a developmental delay.
Negative Predictive Value	The probability that a child who screens negative does not have a developmental delay.
Over-referral	The proportion of all children who screen positive but do not have a developmental delay (the proportion of false positives in a population).
Under-referral	The proportion of all children who screen negative when they actually have a developmental delay.
Likelihood ratio—positive	The odds or likelihood that a positive screening result will be found in a child who is delayed, compared with a child who is not delayed.
Likelihood ratio—negative	The odds or likelihood that a negative screening result will be found in a child who is not delayed, compared with a child who is delayed.

One of the difficulties in evaluating and comparing the validity of screening tests is that there is no universal agreement on what constitutes a developmental delay (Frankenburg, 2002). Even when well established standardized assessments such as the BSID-II are used as the “gold standard” criterion measure, there is variability in the cut-off point used to define the presence of a developmental delay. For example, Squires and colleagues (1999) chose 1.5 standard deviations (SD) below the mean of the standardized assessments (Revised Gessell and the BSID-II) as the cut-off point in their validation work of the ASQ, with the justification that -1.5 SD is the eligibility criteria commonly established for early intervention programs in the United States. Skellern et al. (2001), however, set their cut-off point at -1.0 SD in order to capture children with mild and moderate delays, as well as those with more severe developmental delays, arguing that early identification and intervention among the former can yield significant benefits. Lenarski, Singer and Peters (2001) also used a -1.0 SD cut-off on the Differential Ability Scales in their validation study of the Early Screening Profiles. Other researchers studying the Denver Developmental Screening Test-II (Denver-II) have defined scores below -2.0 SD on the BSID-II as significant delays and scores between -1.0 and -2.0 SDs as mild delays—and calculated agreement rates using both values (e.g., Leslie, Gordon, Ganger, & Gist, 2002). In a validation study of the Brigance Infant and Toddler Screen, Glascoe (2002) categorized children as delayed if they scored below the 10th percentile on the assessment measures.

With normally distributed scores on the standardized assessment, the use of -1.0 , -1.5 and -2.0 SDs as cut-off points should yield identification of the lowest scores for 16%, 7%, and 2.3% of a population, respectively (Glass & Hopkins, 1996). The issue, then, in selecting a cut-off point for standardized assessment scores, is to decide the percentage of children (with varying degrees of

delay) that the community wishes to identify. Evidence from the National Longitudinal Survey of Children and Youth has suggested that almost 29% of Canadian children under 12 years of age are experiencing cognitive, academic, or social-emotional-behavioural problems (Willms, 2002). Epidemiological evidence developed on a different set of criteria has suggested that about 18% of American children under 18 years of age are experiencing developmental delays or chronic conditions that leave them at risk for poor development (Newacheck et al., 1998). Finally, Law, Boyle, Harris, Harkness, and Nye (2000) have estimated that the prevalence of speech and language delays (alone) is approximately 6% in the United States; this would suggest that a -2.0 SD cut-off is inadequate.

In order to illustrate the effects of varying cut-off points, we performed the calculations and assessed agreement rates using three different cut-off points on the BSID-II: -1.0 , -1.5 , and -2.0 SDs, and both scoring rules for the NDDS. As recommended by the test authors, the BSID-II scores were corrected for prematurity, and we ensured that the mothers of all prematurely born children in our sample completed the adjusted-age appropriate version of the NDDS (the NDDS does not, otherwise, correct for prematurity.) We did not conduct psychometric analyses by age group due to small size of the subsamples. (For example, if there is only one child identified as delayed by the standardized assessment, the sensitivity rate could only be 0% or 100%, and we concluded that presenting such rates would be misleading.)

FINDINGS

Table 3 presents the results of the screening tests and the standardized assessments. Formulas and an illustration of calculations are provided in Table 4. Thirty-nine of the 118 children (33%) were identified with developmental concerns when applying the one-flag rule to the NDDS results, compared with only 21 children (8%) when the two-flag rule was applied. Setting the cut-off point for the BSID-II at 1.0, 1.5, and 2.0 standard deviations below the mean resulted in 16, 6, and 3 children being identified as “delayed”, respectively. This represented 13.6%, 5.1%, and 2.5% of the sample, fairly close to the 16%, 7%, and 2.3% expected at these points. Figure 1 presents the distribution of the BSID-II MDI scores and indicates that the scores were approximately normally distributed. The mean score of 99.6 and SD of 12.7 also suggest that the scores found in this sample are fairly representative of the larger population, but with somewhat less variability.

We did not find any statistically significant differences in the BSID-II scores by age group, gender, or an interaction of age-by-gender, but the non-significant results⁵ could be due to the small size of the subgroups. The trend, however, does suggest that the younger infants were the least likely to be identified with delays. The same pattern was seen with the NDDS when the one-flag rule was applied, but not the two-flag rule.

⁵ Obtained through two-way analysis of variance (ANOVA)

Table 3. Children Identified as “Delayed” by Screening and Assessments by Age and Gender

	Age Group in Months			Gender		Total N = 118 (%)
	4 n = 38 (%)	18 n = 40 (%)	24 n = 40 (%)	M n = 54 (%)	F n = 64 (%)	
NDDS Results						
1+ flags	8 (21)	18 (45)	13 (33)	26 (48)	13 (20)	39 (33.1)
2+ flags	6 (16)	10 (25)	5 (13)	16 (30)	5 (8)	21 (17.8)
BSID-II Results						
-1.0 SD	2 (5)	6 (15)	8 (20)	9 (17)	7 (11)	16 (13.6)
-1.5 SD	1 (3)	2 (5)	3 (8)	3 (6)	3 (5)	6 (5.1)
-2.0 SD	0 (0)	2 (5)	1 (3)	2 (4)	1 (2)	3 (2.5)

Table 4. Illustration of Calculations using the Two-Flag NDDS Rule and -1.5 SD Cut-off for the BSID-II (N = 118)

Screening Results	Standardized Assessment Results		Total
	Delayed	Not Delayed	
≥ 2 Flags	A True Positive 3	B False Positive 18	21
< 2 Flags	C False Negative 3	D True Negative 94	97
	6	112	118

Sensitivity = A / A + C	3 / 6 = 50.0%
Specificity = D / B + D	94/112 = 83.9%
Total Agreement = A + D / A + B + C + D	97/118 = 82.2%
Positive Predictive Value = A / A + B	3/21 = 14.3%
Negative Predictive Value = D / C + D	94 / 97 = 96.9%
Over-referral = B / A + B + C + D	18/118 = 15.3%
Under-referral = C / A + B + C + D	3/118 = 2.5%
Likelihood ratio—positive = Sensitivity / (1 – specificity)	.50 / .161 = 3.11
Likelihood ratio—negative = (1 – sensitivity) / specificity	.50 / .839 = .60

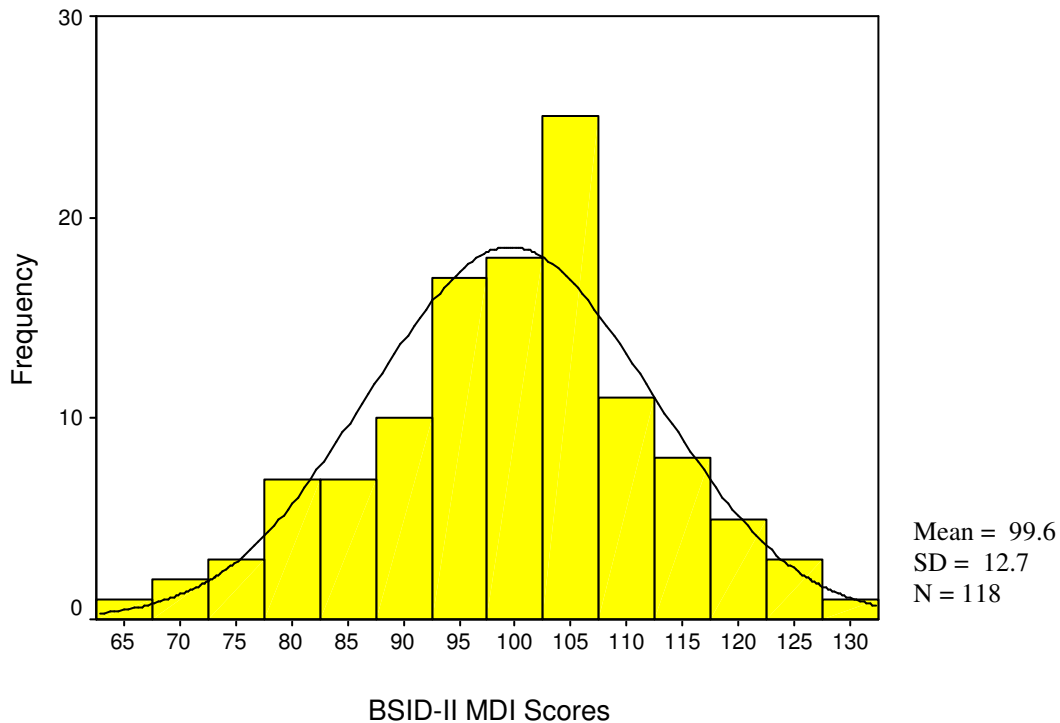
Figure 1. Distribution of BSID-II Mental Development Index Scores

Table 5 presents the psychometric results obtained when comparing the results of the NDDS to the results of the BSID-II. (See Appendix C for details of the cross-tabulations.) Sensitivity rates ranged from 44% to 100% depending upon the cut-off points used for the screening test and the standardized assessment; specificity rates varied less, ranging from 68% to 86%. At the lowest sensitivity, the under-referral rate was 7.6%, which means 1 /13 of the children in the community would fail to be identified and referred when they are experiencing developmental delays. Meisels (1989) has suggested that both sensitivity and specificity rates should be at least 80%, although others have suggested that 70% or 75% is satisfactory (Glascoe, 2002; Barnes, as cited in Meisels, 1989). Using these criteria, a satisfactory level of sensitivity was obtained only when the BSID-II cut-off was set at -2.0 SD (which is intended to identify only the most severely delayed children, or 2.5% of the population). In contrast, 'good' specificity rates were obtained with each of the 3 different BSID-II cut-off points when using the two-flag rule, and marginally acceptable specificity rates were found with the one-flag rule. The likelihood ratios further suggest that using the two-flag rule rather than the one-flag rule improves the likelihood that a developmental delay will be identified in a child who actually is delayed compared with a child who is not. The positive predictive values were low, particularly when using the more conservative cut-off points on the BSID-II. However, the positive and negative predictive values are influenced more by the prevalence of delays in the population than by the specificity of the screening test, and so they speak more to the 'cost' of screening than its accuracy.

Table 5. NDDS Psychometric Validation Results (N =118)

	NDDS One-Flag Rule			NDDS Two-Flag Rule		
	-1.0 SD	-1.5 SD	-2.0 SD	-1.0 SD	-1.5 SD	-2.0 SD
% Sensitivity	56.3	50.0	100.0	43.8	50.0	100.0
% Specificity	70.6	67.9	68.7	86.3	83.9	84.3
% PPV	23.1	7.7	7.7	33.3	14.3	14.3
% NPV	91.1	96.2	100.0	90.7	96.9	100.0
% Over-Referral	25.4	30.5	30.5	11.9	15.3	15.3
% Under-Referral	5.9	2.5	0.0	7.6	2.5	0.0
LR-Positive	1.9	1.6	3.2	3.2	3.1	6.4
LR-Negative	.62	.74	---	.65	.60	--

Note: PPV – positive predictive value; NPV – negative predictive value; LR-Positive – likelihood ratio of a positive test; LR-Negative – likelihood ratio of a negative test.

Gender specific calculations were also conducted, although we recommend that the results be interpreted with caution due to the small size of the subsamples (see Appendix D). Higher sensitivity rates were obtained among the males under both NDDS scoring rules, when using a –1.0 and –1.5 SD cut-off on the BSID-II. Lower specificity rates were obtained among the males for each of the different cut-off criteria, although the levels were problematic only with the one-flag rule.

DISCUSSION

Our findings illustrate the importance of the selected cut-off point when assessing evidence about the psychometric validity of an instrument and when designating the cut-off point to be used for a screening instrument. Sensitivity rates for the NDDS were highest when using the most stringent cut-off for the standardized assessment (–2.0 SDs) but, contrary to expectations, sensitivity was impacted little by the change from the one-flag rule to two-flag rule for the screening tool. The opposite results were obtained for specificity rates.

The sensitivity rates obtained in this study, using standardized child assessments as the criterion measure and the less stringent cut-off points (-1.0 and -1.5 SD), were lower than that obtained by Nagy et al. (2002) when they compared NDDS results with the results of the parent-completed ASQ. This is not unexpected, though, when one considers the potential ‘interaction’ effects of comparing one parent-reported screening tool with another. For example, validation studies of the 4-, 12- and 24-month ASQ with the Revised Gesell and the BSID-II (using a –1.5 SD cut-off) have yielded sensitivity rates of 51%, 85%, and 80%, respectively (Squires et al., 1999). Applying the 83% sensitivity rate obtained by Nagy and colleagues (when comparing the 12-month NDDS to the ASQ), to the sensitivity rates obtained for the ASQ, suggests that the sensitivity of the NDDS might be as low as 42% at 4 months (83% x 51%) and 66% at 24

months ($83\% \times 80\% = 66\%$) when using a standardized assessment as the criterion measure.⁶ From this viewpoint, our results are not inconsistent with those obtained Nagy et al. in the Ontario HBHC evaluation.

A Question of Perspective: Evaluating the Utility of the NDDS

The results of this study suggest that the NDDS is effective at identifying children with severe delays, but is less effective at identifying children with milder delays. This may be viewed as problematic given the recent research literature which indicates that children with mild to moderate delays who are identified early via screening or other means may be particularly well-positioned to derive benefits from referral to early childhood programs such as high quality child care or Mother Goose, even when they do not qualify for early intervention services (Glascoe, 2001). Moreover, children with severe delays are not the population typically targeted by universal screening programs as they are often readily identified prior to school entrance by other means including informal observation, clinical judgment, and parental concerns (Law et al., 2000). Thus, a conservative perspective would suggest that there is reason to question the value of the NDDS as a universal screening tool given that it identified less than half of the children with mild to moderate delays ($6/13 = 46\%$ with the one-flag rule, and $4/13 = 31\%$ with the two-flag rule). Such a perspective would also lead to questions about the consequences of providing false reassurance to parents of children who could benefit from an early childhood programs.

There is an alternative, less conservative perspective that can be taken. This perspective would suggest that (a) screening with the NDDS may result in earlier identification and intervention for those children who are identified and (b) the 31% or 46% of children with mild to moderate delays who were identified with the NDDS might not otherwise have been identified – and indeed, our data provide some support for this. One of the items on the demographic questionnaire asked parents whether or not their child had any developmental delays, or learning or behavioural problems that had been previously identified by a health professional. A review of our data indicates that, of the 3 children who showed severe delays (scores below -2.0 SDs on the BSID-II) and were correctly identified by the NDDS, only one (33%) was reported to have been previously identified with a developmental problem. This suggests that even among those children with severe delays, universal screening with the NDDS may provide for earlier identification. Moreover, of the 6 children who scored between -1.0 and -2.0 SDs on the BSID (mild to moderate delays) and were correctly identified by the NDDS (using the one-flag rule), only one (17%) had been previously identified with a developmental problem.

There are other perspectives that may be considered, beyond those that are concerned solely with the psychometric properties of a screening tool. It could be suggested that the selection and evaluation of a screening tool should also take into account the objectives of the agency or community and the context of the screening program as there may be other important goals to be met through the screening program in addition to the identification of children requiring further assessment and intervention. Practitioners involved in the developmental screening program under study report that they believe that the screening program serves as a vehicle for accessing parents and a focal point for engaging families in educational and support activities related to child development. It is speculated that participating in parent-completed screening may have

⁶ Validity results for the 18-month ASQ were not reported in the 1999 ASQ User's Guide (Squires et al.). Sensitivity results for the 16- and 20-month ASQs were 73% and 65%, respectively.

positive effects on parents' knowledge about child development and the ways in which they support their child's development. If the user-friendly format of the NDDS, as compared to other screening tools, makes parents more willing to participate, this may be another characteristic of the tool to be considered in its overall assessment. We acknowledge that critics of this perspective might argue that, rather than appropriately educating parents about child development, screening instruments with a low sensitivity may provide false reassurance to parents. The counterargument, of course, is that a screening tool should never be used to discount a parent's concerns – a screening tool, by definition, is not diagnostic.

Finally, in evaluating the utility of the NDDS, we suggest that it is important to differentiate between what can be expected of the NDDS tool alone (e.g., if it is mailed out and completed by the parent in isolation) compared with the results that might be expected if the NDDS is completed with the support of a health care professional or child development specialist, and if the NDDS is just one part in a continuum of universal screening strategies within a larger primary prevention approach to facilitating healthy childhood development.⁷

The Possible Influence of Age on Sensitivity Rates

There is some research evidence suggesting that the sensitivity of developmental screening tools may be lower among infants and toddlers than among preschool children, and thus, higher sensitivity rates may have been obtained if children older than 2 years had been included in the sample.⁸ Although the sensitivity of the earlier version of the Denver Developmental Screening Test (DDST) was shown to be generally poor across all age groups, risk categories, and outcome measures (41%), the studies involving younger age samples yielded the lowest sensitivity rates (Frankenburg, Camp, & Van Natta, 1971; Meisels, 1989).⁹ Squires and colleagues' (1999) validation work with the ASQ similarly yielded the lowest sensitivity among 4-month-olds (51%), and another drop at 20 months (65%). Skellern et al.'s (2001) examination of the ASQ among children from 12 to 48 months found the lowest sensitivity among 18-month-olds (50%) but their subsample sizes were small and their study did not include children under 12 months of age. Although there are generally tradeoffs between sensitivity and specificity with changes in the cut-off point, there is less evidence suggesting a relationship between age and specificity.

There are a myriad of reasons why screening tests may be prone to lower sensitivity at the younger ages. First of all, the various domains of development that are encompassed by screening tests (gross motor, fine motor, language, cognitive, and adaptive behaviour) are more interrelated and complex during the infant and toddler period. Secondly, it has been argued that “developmental screening tests generally do not have enough items to identify a variety of problems with equal precision for all the constructs they measure” (Frankenburg, 1994, p. 586).

⁷ We acknowledge that there are other screening tools that were designed to be used in various contexts, including mail-outs (e.g., the ASQ), and that some of these may have higher reported validity at some ages. However, having demonstrated the influence of the cut-off point used to identify ‘true delays’, we caution against comparing the psychometric findings of various studies without considering the operational definitions and data collection methods used in the validation studies.

⁸ We did not calculate sensitivity and specificity by age group. Nonetheless, our results do not suggest that the low sensitivity rates originated among the 4-month-olds because, of the 9 false-negative cases identified using -1.0 SD BSID-II cut-off and the two-flag rule, only 1 child was 4 months old.

⁹ Validation work of the revised Denver-II indicates that overall sensitivity improved to 83%, but results were not reported by age group (Glascoe et al., 1992).

Furthermore, the younger age screens/versions are typically comprised of fewer items (e.g., the NDDS has 4, 13, 17, and 22 items on the 1 month, 4 month, 18 month, and 60 month versions, respectively). This problem may be compounded at younger ages due to the subtlety with which most developmental delays present during infancy.

Study Limitations

These results were obtained within the context of a research study – and although the research assistants were available to answer questions asked by the parents, they did not review the NDDS results with parents after its completion as is typically done by public health nurses in the Chilliwack community. Nonetheless, we expect that this practice would affect specificity results more than sensitivity, as it is the “no” responses that would likely elicit more discussion than the “yes” responses. For example, anecdotal evidence from the Chilliwack practitioners suggest that parents sometimes report a “no” because the child has not had the opportunity to try a particular item (e.g., to pick up a cheerio), and the practitioner would explore this by asking about equivalent activities. It is possible, however, that parents were providing what they perceived to be socially desirable responses as it is clear by the design of the tool (and the tear-off educational section in particular) that the listed items are milestones expected of the normally developing child. This would have affected sensitivity rates rather than specificity.

A single criterion measure was used in this study. Children may have delays in one or several domains and, ideally, a concurrent battery would include measures of cognitive, motor, and social-emotional development, activities of daily living, and discrete measures of both receptive and expressive language. It is, however, difficult to find reliable and valid measures that differentiate development by such domains for children under 2 years of age. It is also difficult to complete a comprehensive direct assessment of a young child within the constraints (time and financial) of a research study, since such assessments typically require administration of multiple measures across several sessions. Of the measures available for use with infants and toddlers, the BSID-II is one of the most popular in both research and clinical practice making it a ‘gold-standard’ criterion measure. In addition, the MDI of the BSID-II has strong reliability and encompasses a breadth of key items spanning different developmental domains, making it an attractive solution to the assessment quandary that accompanies this type of research. Moreover, if the BSID-II had captured a significantly more narrow range of developmental domains than the NDDS, we would expect this to have influenced specificity rates more than sensitivity rates. Finally, we acknowledge that the sample was limited to children 2 years of age and younger, and that the sample size was relatively small.

Further Research

This analysis represents the first phase of a larger 5-year evaluation study of the NDDS-based screening program in the Eastern Fraser Valley. The next phase of our study will involve an evaluation of the concurrent validity of the NDDS among children who are 3 to 3 ½ years of age, and an evaluation of the predictive validity of early screening with the NDDS (at 4, 18, and 24 months) for predicting developmental delays at 3 to 3 ½ years of age. However, we will also continue to collect data among 4-, 18-, and 24-month-olds so that we may conduct age and gender specific analyses. In addition to our investigation of the NDDS tool, we will be assessing the impact of the system on parents and on the community’s ability to foster child development, and measuring changes in the overall school readiness of children in the community.

CONCLUSION

A strength of this study was its use of multiple data sources. Although it might be argued that the majority of the parents were completing the screening tool within the artificial environment of a research study, the participants were all members of the local community in which there has been considerable promotion of developmental screening, and the research team followed a carefully designed protocol with several built-in checks and balances. Also, the study was conducted in the community settings where service provision typically occurs.

We have presented two possible interpretations of our findings, one that emphasizes the potential benefits of the tool, the other that emphasizes the psychometric limitations of the tool. If the NDDS is used, we recommend that it be used with caution, with clearly considered objectives for its use, and with an understanding of the benefits and consequences of different cut-off criteria. For example, using the one-flag rule rather than the two-flag rule may facilitate the identification of children with mild to moderate delays. When selecting a screening tool or evaluating its adequacy, the benefits of early identification, the impact of failing to identify children with delays, the cost of conducting follow-up assessments for children who do not require intervention, and the availability of resources in that particular community for screening, assessment, and intervention must all be considered.

Given that this study was focused on the psychometric properties of the NDDS, we did not address its cost effectiveness. A basic tenet of screening is that, for greater cost effectiveness, higher risk populations be targeted. However, two key arguments for universal developmental screening are (a) that there are no singular determinants of developmental delays among children, and (b) that targeted screening may not have its intended effect if those targeted feel stigmatized. This leads us back to the quest for a universal screening tool that is psychometrically valid, cost-effective, and user-friendly – but with some uncertainty as to whether all three are possible. What may be more important is that the user understand the inherent tradeoffs and have appropriate expectations for universal screening.

To communities that are in the process of implementing a developmental screening program, we would advise that it is important, when reading reports of the psychometric validity of various screening tests, that the characteristics and quality of the studies be considered. In their systematic review of the screening literature for speech and language delays, Law et al. (2000) found an inverse relationship between the quality of the study and its sensitivity ($r = -.48$) and likelihood ratio ($r = -.34$) particularly in studies that used community samples rather than clinical or high-risk samples.¹⁰ Practicalities must also be considered, as there would be little to be gained through the use of a tool with strong validity if it yielded a low participation rate due to participant burden or other constraints.

To our knowledge, this is the first external evaluation of the concurrent validity of the NDDS using standardized child assessments as the criterion. We acknowledge the need for further psychometric evaluation of the NDDS, with children older than 2 years of age, and with larger samples. We also recognize the importance of additional research into the effects of participating in screening, drawing on other methodological approaches.

¹⁰ Quality was assessed in terms of methodological rigor (i.e., threats to internal or external validity) and the adequacy of the report that would allow replicability (Law et al., 2000).

REFERENCES

- American Academy of Pediatrics Committee on Children with Disabilities (2001). Developmental surveillance and screening of infants and young children. *Pediatrics*, 108(1), 192-196.
- Bayley, N. (1993). *Bayley Scales of Development- Second Edition (BSID-II)*. San Antonio, TX: The Psychological Corporation.
- British Columbia Statistics (2001). 2001 Chilliwack-Kent District Profile. /www.bcstats.gov.bc.ca/data/cen01/profiles/PED_11.pdf [retrieved July 20, 2004].
- Frankenburg, W. K. (1994). Preventing developmental delays: Is developmental screening sufficient? *Pediatrics*, 93(4), 586-593.
- Frankenburg, W. K. (2002). Developmental surveillance and screening of infants and young children. *Pediatrics*, 109, 144-145.
- Frankenburg, W. K., Camp, B. W., & Van Natta, P. A. (1971). Validity of the Denver Developmental Screening Test. *Child Development*, 42, 475-485.
- Glascoe, F. P. (1999). Using parents' concerns to detect and address developmental and behavioral problems. *Journal of the Society for Pediatric Nursing*, 4, 24-35.
- Glascoe, F. P. (2001). Are overreferrals on developmental screening tests really a problem? *Archives of Pediatric & Adolescent Medicine*, 155, 54-59.
- Glascoe, F. P. (2002). The Brigance Infant and Toddler Screen: Standardization and validation. *Developmental and Behavioral Pediatrics* 23, 145-150.
- Glascoe, F. P. (2001). Are overreferrals on developmental screening tests really a problem? *Archives of Pediatric and Adolescent Medicine*, 155, 54-59.
- Glascoe, F. P. (2002). Two views of developmental testing. *Pediatrics*, 109(6), 1181-1182.
- Glascoe, F. P., Byrne, K. E., Ashford, L. G., Johnson, K. L., Chang, B., & Strickland, B. (1992). Accuracy of the Denver-II in developmental screening. *Pediatrics*, 89, 1221-1225.
- Glascoe, F. P., & Dworkin, P. H. (1995). The role of parents in the detection of developmental and behavioral problems. *Pediatrics*, 95, 829-836.
- Glass, G. V., & Hopkins, K. D. (1996). *Statistical methods in education and psychology* (3rd Ed.). Toronto, ON: Allyn and Bacon.
- Greenhalgh, T. (1997). How to read a paper: Papers that report diagnostic or screening tests. *British Medical Journal*, 315(7107), 540-543.

- Hertzman, C., McLean, S. A., Kohen, D., Dunn, J., & Evans, T. (2002). *Early development in Vancouver: Report of the Community Asset Mapping Project (CAMP)*. Vancouver: University of British Columbia, Human Early Learning Partnership.
- Hills, T. W. (1987). *Screening for school entry*. Urbana, IL. (ERIC Document Reproduction Services No. ED281607).
- Janus, M., & Offord, D. R. (2000). *Early Development Instrument: A guide*. Hamilton, ON: McMaster University, The Canadian Center for Studies of Children at Risk.
- King, T. M., & Glascoe, F. P. (2003). Developmental surveillance of infants and young children in pediatric primary care. *Current Opinion in Pediatrics*, 15(6), 624-629.
- King, E. H., Logsdon, D. A., & Schroeder, S. R. (1992). Risk factors for developmental delay among infants and toddlers. *Children's Health Care*, 21, 39-52.
- Kohen, D., Dahinten, V. S., & McIntosh, C. (2003, April). *Mechanisms of neighbourhood effects on Canadian preschoolers*. Symposium paper presented at the 2003 Biennial Meeting of the Society for Research in Child Development, Tampa FL.
- Law, J., Boyle, J., Harris, F., & Harkness, A., & Nye, C. (2000). The feasibility of universal screening for primary speech and language delay: findings from a systematic review of the literature. *Developmental Medicine & Child Neurology*, 42, 190-200.
- Lenarski, S., Singer, M., & Peters, M. (2001). Utility of the Early Screening Profiles in identifying preschoolers at risk for cognitive delays. *Psychology in the Schools*, 38(1), 17-24.
- Leslie, L. K., Gordon, J. N., Ganger, W., & Gist, K. (2002) Developmental delay in young children in child welfare by initial placement type. *Infant Mental Health Journal*, 23, 496-516.
- Meisels, S. J. (1989). Can developmental screening tests identify children who are developmentally at risk? *Pediatrics*, 83, 578-585.
- Meisels, S. J., & Atkins-Burnett, S. (1994). *Developmental screening in early Childhood: A guide* (4 ed.). Washington: The National Association for the Education of Young Children.
- Meisels, S. J., & Atkins-Burnett, S. (2000). The elements of early childhood assessment. In J. P. M. Shonkoff, Samuel J. (Ed.), *Handbook of early childhood intervention* (2nd ed.) (pp. 231-257). Cambridge University Press.
- Montgomery, M. L. (1999). Use of the Child Development Inventory to screen high-risk populations. *Clinical Pediatrics*, 38, 535-539.

- Nagy, P., Ryan, B., & Robinson, R. (2002, December). Nipissing Instrument Validation Report, 2001-2002. In *Evaluation of the Healthy Babies, Healthy Children Program*. Unpublished Report of the Early Years and Healthy Child Development Branch, Ontario Ministry of Community, Family and Children's Services.
- Newacheck, P. W., Strickland, B., Shonkoff, J. P., Perrin, J. P., McPherson, M., McManus, M., Lanver, C., Fox, H., & Arangoss, P. (1996). An epidemiologic profile of children with special health care needs. *Pediatrics*, *102*, 117-123.
- Skellern, C., Rogers, Y., & O'Callaghan, M. J. (2001). A parent-completed developmental questionnaire: Follow up of ex-premature infants. *Journal of Paediatrics & Child Health*, *37*, 125-129.
- Squires, J. (1996). Parent-completed developmental questionnaires: A low-cost strategy for child-find and screening. *Infants and Young Children*, *9*(1), 16-28.
- Squires, J., Potter, L., & Bricker, D. (1999). *The ASQ user's guide: Second edition*. Baltimore: Paul H. Brookes.
- Willms, J. D. (2002). Research findings bearing on Canadian social policy. In J. D. Willms (Ed.), *Vulnerable children. Findings from Canada's National Longitudinal Survey of Children and Youth*, pp. 331-358. Edmonton, AB: University of Alberta Press.

APPENDIX A

Research Questions

1. What is the concurrent validity (sensitivity and specificity) of the Nipissing District Developmental Screen (NDDS), when compared with results from a battery of direct child assessments, for the identification of developmental delays among children aged 4, 18, 24, 36, and 60 months?
2. What is the stability over time of the NDDS when comparing parent-ratings obtained at 4, 18, 24, 36, and 60 months? (This will also allow us to answer the question “what is the utility of ongoing screening beyond 24 months of age?”)
3. Is participation in a parent-completed screening system associated with parents’ perceptions of (a) increased knowledge about child development; (b) changes in parenting behaviours and activities to foster child development; and (c) increased confidence in parenting?
4. What is the predictive validity of early screening with the NDDS (at 4, 18, and 24 months) for predicting developmental delays at 36 and 60 months?
5. What are the most effective means of mobilizing a community to develop and implement a universal and integrated developmental screening system?
6. Is the implementation of a universal and integrated developmental screening system associated with changes in the community’s capacity for early identification and intervention with young children?
7. Are there changes in children’s school readiness at the population or community level as measured by the Early Development Instrument (administered in 2002 and to be repeated in 2006) that may be associated with the implementation of a universal and integrated developmental screening system?

APPENDIX B

**Results of the Ontario HBHC Instrument Validation Study:
Cross-Tabulations of Agreement Between the 12-Month NDDS and the ASQ (N=226)¹**

One-Flag Rule

<i>NDDS Results</i>	<i>ASQ Results</i>		Total
	<i>Delayed</i>	<i>Not Delayed</i>	
Risk (≥ 1 Flags)	22	48	70
Typical Development	1	155	156
	23	203	226

Two-Flag Rule

<i>NDDS Results</i>	<i>ASQ Results</i>		Total
	<i>Delayed</i>	<i>Not Delayed</i>	
Risk (≥ 2 Flags)	19	11	30
Typical Development	4	192	196
	23	203	226

*Note:*¹. Using a cut-off of 1.5 standard deviations below the mean of the ASQ

Agreement Between the 12-Month NDDS and the ASQ

	<i>% Sensitivity</i>	<i>% Specificity</i>	<i>% Total Agreement</i>	<i>% PPV</i>	<i>% Over-Referral</i>	<i>% Under-Referral</i>
NDDS 1Flag	95.7%	76.4%	78.3%	31.4%	21.2%	0.5%
NDDS 2 Flags	82.6%	94.6%	93.4%	63.3%	4.9%	1.8%

Note: Sensitivity and specificity rates were calculated by the investigators of the present study, based on data reported in the unpublished Nipissing Instrument Validation Report, 2001-2002 (Nagy, Ryan, & Robinson, 2002). The Nipissing Instrument Validation Report presented the cross-tabulated data, but reported only the total agreement rates.

APPENDIX C

Cross-Tabulations of Agreement Between the Nipissing District Developmental Screen and the BSID-II Assessment Using Various Cut-off Points (N = 118)

A. Using the One-Flag Rule on the NDDS to Indicate Developmental Risk

BSID-II: - 1.0 SD Cut-off

<i>Screening Results</i>	<i>Standardized Assessment Results</i>		Total
	<i>Delayed</i>	<i>Not Delayed</i>	
Risk (≥ 1 Flags)	9	30	39
Typical Development	7	72	79
	16	102	118

BSID-II: - 1.5 SD

<i>Screening Results</i>	<i>Standardized Assessment Results</i>		Total
	<i>Delayed</i>	<i>Not Delayed</i>	
Risk (≥ 1 Flags)	3	36	39
Typical Development	3	76	79
	6	112	118

BSID-II: - 2.0 SD

<i>Screening Results</i>	<i>Standardized Assessment Results</i>		Total
	<i>Delayed</i>	<i>Not Delayed</i>	
Risk (≥ 1 Flags)	3	36	39
Typical Development	0	79	79
	3	115	118

B. Using the Two-Flag Rule on the NDDS to Indicate Developmental Risk

BSID-II: - 1.0 SD

<i>Screening Results</i>	<i>Standardized Assessment Results</i>		Total
	<i>Delayed</i>	<i>Not Delayed</i>	
Risk (≥ 2 <i>Flags</i>)	7	14	21
Typical Development	9	88	97
	16	102	118

BSID-II: - 1.5 SD

<i>Screening Results</i>	<i>Standardized Assessment Results</i>		Total
	<i>Delayed</i>	<i>Not Delayed</i>	
Risk (≥ 2 <i>Flags</i>)	3	18	21
Typical Development	3	94	97
	6	112	118

BSID-II: - 2.0 SD

<i>Screening Results</i>	<i>Standardized Assessment Results</i>		Total
	<i>Delayed</i>	<i>Not Delayed</i>	
Risk (≥ 2 <i>Flags</i>)	3	18	21
Typical Development	0	97	97
	3	115	118

APPENDIX D

NDDS Psychometric Validation Results by Gender, Using the One-Flag Rule

	NDDS One-Flag Rule					
	-1.0 SD		-1.5 SD		-2.0 SD	
	M	F	M	F	M	F
% Sensitivity	66.7	42.9	66.7	33.3	100.0	100.0
% Specificity	55.6	82.5	52.9	80.3	53.8	81.0
% Total Agreement	57.4	78.1	53.7	78.1	55.6	81.3
% PPV	23.1	23.1	7.7	7.7	7.7	7.7
% NPV	89.3	92.2	96.4	96.1	100.0	100.0
% Over-Referral	37.0	15.6	44.4	18.8	51.8	18.8
% Under-Referral	5.6	6.3	1.9	3.1	0	0
LR-Positive	1.5	2.5	1.4	1.7	2.2	5.3
LR-Negative	0.6	0.7	0.6	0.8	---	---

Note: Males: $n = 54$, Females: $n = 64$.

NDDS Psychometric Validation Results by Gender, Using the Two-Flag Rule

	NDDS Two-Flag Rule					
	-1.0 SD		-1.5 SD		-2.0 SD	
	M	F	M	F	M	F
% Sensitivity	55.6	28.6	66.7	33.3	100.0	100.0
% Specificity	75.6	94.7	72.5	93.4	73.1	93.7
% Total Agreement	72.2	87.5	72.2	90.6	74.1	93.8
% PPV	31.3	40.0	12.5	20.0	12.5	20.0
% NPV	89.5	91.5	97.4	96.6	100.0	100.0
% Over-Referral	20.4	4.7	25.9	6.3	25.9	6.3
% Under-Referral	7.4	7.8	1.9	3.1	0	0
LR-Positive	2.3	5.4	2.4	5.1	3.71	15.75
LR-Negative	0.6	0.8	0.5	0.7	---	---

Note: Males: $n = 54$, Females: $n = 64$.