Investigation of test characteristics of two screening tools in comparison with a gold standard assessment to detect developmental delay at 36 months: A pilot study

Lisa Currie MSc[1], Linda Dodds PhD[1,2,3], Sarah Shea MD[3], Gordon Flowerdew PhD[1], Jennifer McLean MD[3], Robin Walker MD[3], Michael Vincer MD[3]

L Currie, L Dodds, S Shea, et al. Investigation of test characteristics of two screening tools in comparison with a gold standard assessment to detect developmental delay at 36 months: A pilot study. Paediatr Child Health 2012;17(10):549-552.

BACKGROUND: The ability of the Rourke Baby Record (Rourke) and the Nipissing District Developmental Screen (NDDS) to detect developmental delay is not known.
OBJECTIVE: To determine the test characteristics of the Rourke and NDDS compared with the Bayley Scales of Infant and Toddler Development III for detecting developmental delay in high-risk children.
METHODS: Three-year-olds were recruited from the IWK Health Centre (Halifax, Nova Scotia). Two cut-points were evaluated (one and two or more areas of concern) from the Rourke and NDDS, and were compared with a score of ≤85 on the Bayley Scales of Infant and Toddler Development III.
RESULTS: The majority (67.7%) of the 31 participants reported no concern. At one area of concern, sensitivity was 75% for both the Rourke and NDDS. When two areas of concern were noted, specificity was 93% for the Rourke and 96% for NDDS.
CONCLUSIONS: Both the Rourke and the NDDS appear to be reasonably sensitive and specific, but further investigation is warranted.

Key Words: *Developmental delay; Developmental screening; Developmental surveillance*

L'exploration des caractéristiques de deux outils de dépistage par rapport à une norme de référence de l'évaluation pour déceler le retard de développement à 36 mois : un projet pilote

HISTORIQUE : On ne sait pas à quel point le Relevé postnatal Rourke (RPR) et le *Nipissing District Developmental Screen* (NDDS) peuvent déceler le retard de développement.
OBJECTIF : Déterminer les caractéristiques du RPR et du NDDS par rapport aux échelles de Bayley pour le développement des nourrissons et des tout-petits III (BSITD-III) afin de déceler le retard de développement chez les enfants à haut risque.
MÉTHODOLOGIE : Les chercheurs ont recruté des enfants de trois ans au *IWK Health Centre* (Halifax, Nouvelle-Écosse). Ils ont évalué deux seuils (1 secteur d'inquiétude et au moins 2 secteurs d'inquiétude) du RPR et du NDDS et les ont comparés à un résultat de 85 ou moins aux BSITD-III.
RÉSULTATS : La majorité des 31 participants n'ont déclaré aucune inquiétude (67,7 %). Pour un secteur d'inquiétude, la sensibilité s'élevait à 75 % à la fois dans le RPR et dans le NDDS. Lorsqu'on relevait deux secteurs d'inquiétude, la spécificité s'élevait à 93 % pour le RPR et à 96 % pour le NDDS.
CONCLUSIONS : Tant le RPR que le NDDS semblent être raisonnablement sensibles et spécifiques, mais des explorations plus approfondies s'imposent.

Developmental delay, as measured by the Bayley Scales of Infant Development III (BSITD-III), refers to a limitation in gross motor, fine motor, cognitive, language or personal-social skills in comparison with norm-referenced performance (1). Early detection of delay can result in earlier intervention to improve outcomes (2). Conditions that are known to increase risk of developmental delay include prematurity, low birth weight and neurological injury at birth (3). Family physicians, paediatricians and/or neonatologists typically have contact with high-risk infants for medical follow-up, and are often the health care providers involved with the identification of developmental delay.

The BSITD-III is a well-documented assessment tool to detect developmental delay but is costly and lengthy to administer, and not the choice for surveillance or screening. The Rourke Baby Record (Rourke) is a relatively recent, comprehensive, evidence-based, integrated primary care practice tool, which includes a section designed to support developmental surveillance (4). As a surveillance tool, its ability to identify developmental delay is dependent on assessment over time. This differs from a screening tool, which provides a brief, one-time assessment to assist in the identification of delay (5). The Rourke is used by many physicians and is endorsed by both the College of Family Physicians of Canada and the Canadian Paediatric Society. However, the capability of the Rourke to identify developmental concern is largely undetermined. The Nipissing District Developmental Screen (NDDS) is similar to the Rourke in that it is widely used but has minimal literature to support its ability to detect delayed development (6). An ideal tool needs to be not only time efficient for a medical practice but also sensitive enough to detect deficits in development.

The objective of the present pilot study was to test participant recruitment procedures and methodology to determine the feasibility of completion of a full-scale, adequately powered observational study to determine the test characteristics of the Rourke and NDDS. As well, the present project aimed to provide preliminary findings of the test characteristics of the Rourke and the NDDS compared with the BSITD-III in the detection of delay in high-risk children.

METHODS

High-risk children enrolled in the Perinatal Follow-up Program at the IWK Health Centre (Halifax, Nova Soctia) and scheduled for a 36-month follow-up visit were recruited for the present study

[1]*Department of Community Health and Epidemiology;* [2]*Perinatal Epidemiology Research Unit, Departments of Obstetrics & Gynaecology and Pediatrics;* [3]*Department of Pediatrics, Dalhousie University, Halifax, Nova Scotia*
Correspondence and reprints: Dr Linda Dodds, Perinatal Epidemiology Research Unit, IWK Health Centre, 5980 University Avenue, Halifax, Nova Scotia B3H 4N1. Telephone 902-470-7191, fax 902-470-7190, e-mail l.dodds@dal.ca
Accepted for publication May 23, 2012

TABLE 1
Rourke Baby Record questions pertaining to development for three-year-old children

Understands two-and-three step directions (eg, "Pick up your hat and shoes and put them in the closet")
Uses sentences with five or more words
Walks up stairs using a handrail
Twists lids off jars and turns doorknobs
Shares some of the time
Plays make-believe games with actions and words (eg, pretending to cook a meal, fix a car)
Turn pages one at a time
Listens to music or stories for 5 min to10 min
No parent/caregiver concerns

between November 2, 2010, and June 5, 2011. The study was conducted at 36 months because of the routine administration of the BSITD-III at this age among the high-risk population served by the Perinatal Follow-up Clinic program. Criteria for participants to be considered high risk were the following: a gestational age of ≤31 weeks; or ≤1500 g at birth; or neurological injury at or immediately following birth. Exclusion criteria for participation in the study included children who were non-English speaking and those with major sensory or physical impairment because they would not be able to complete the BSITD-III.

Questions pertaining to development for three-year-old children from the Rourke (Table 1) were administered to the participants' parent over the telephone approximately one week before their appointment in the follow-up clinic. The NDDS was purchased and the questionnaire was given to the parents to complete at the beginning of their appointment at the follow-up clinic, but before the commencement of the BSITD-III. The BSITD-III was administered by the Perinatal Follow-up Program staff as per Perinatal Follow-up Program protocol and as outlined in the BSITD-III administration manual (7).

A 'No' response, indicating that the child did not demonstrate the skill, was termed a 'flag', to maintain consistency with previous literature (6). In the present analysis, the cut-off values indicating need for further assessment for developmental delay on both the Rourke and the NDDS were defined as one or more flagged items. Thresholds of both one or more, and two or more flags were compared with the BSITD-III. As well, the Rourke score with the omission of the open-ended question pertaining to parental concerns was assessed. A score of ≤85 (one SD below the mean) in any domain on the BSITD-III served as the gold standard cut-off for normal performance and was used to indicate developmental delay. A score of ≤70 (two SDs below the mean) on the BSITD-III was also assessed to determine whether the test characteristics differed when using a definition of more impaired delay.

Separate analyses were completed for the Rourke compared with the BSITD-III and the NDDS compared with the BSITD-III. The sensitivity, specificity, likelihood ratios, positive predictive value, negative predictive value, false positive rates, false negative rates, likelihood ratios and accuracy were calculated for the Rourke and NDDS compared with the BSITD-III. Research ethics approval was obtained from the IWK Research Ethics Board.

RESULTS

Of the 64 children eligible to participate during the study period, 31 completed all components of the study (eg, the Rourke, NDDS and BSITD-III). There were 14 (22%) parents who declined participation in the study. One parent/child who had agreed to participate and completed the Rourke assessment did not attend the scheduled appointment (and therefore, did not have the NDDS or BSITD-III assessments completed). There were 11 potential participants who could not be reached by telephone, and seven families whose telephone number was no longer working.

TABLE 2
Descriptive characteristics of participants

Chronological age at assessment, months	
Mean ± SD	37.6±1.8
Median	38.1
Range	33.2–41.0
Gestational age, weeks	
Mean ± SD	31.6±4.6
Median	30.6
Range	26.2–41.6
Birth weight, g	
Mean ± SD	1637.19±878.21
Median	1350.0
Range	690–4481
Twin gestation	13 (41.94)
Male sex	17 (54.84)
Prematurity (birth before 31 weeks' gestation)	17 (54.84)
Prematurity (birth before 37 weeks' gestation)	25 (80.65)
Medical complications at birth	
Neurological	13 (41.94)
Cardiac	7 (22.58)
Respiratory	19 (61.29)
Fetal malnutrition	8 (25.81)

Data presented as n (%) unless otherwise indicated

Developmental delay, defined as scoring ≤85 on any BSITD-III subscale, was observed in 12.9% of children in this sample (95% CI 3.6% to 29.8%). Table 2 describes demographic and medical characteristics of the study participants. The mean age of participants at the time of administration of the BSITD-III was 37 months and six days. The average gestational age at delivery of participants was 31 weeks and six days and the mean birth weight was 1637 g. Participants experienced a range of medical conditions at birth, including neurological events, such as intraventricular hemorrhage or seizures (41.9%), cardiac complications, such as patent ductus arteriosus or cardiac arrest (23.3%), and respiratory events, most frequently idiopathic respiratory distress syndrome requiring oxygen assistance (61.3%). Other complications included septicemia, necrotizing enterocolitis and inguinal hernia.

Of the 31 respondents, 21 (67.7%) indicated that they had no areas of concern on the Rourke. There were five parents with one area of concern (16.1%), and five parents with two or more areas of concern. There were 22 (70.0%) participants who indicated that they had no areas of concern on the NDDS. There were five parents with one area of concern (16.1%) and four parents (12.9%) with two or more areas of concern. There were two (6.5%) participants who scored below one SD of the mean and two (6.5%) who scored below two SDs of the mean on composite scores of the BSITD-III.

Table 3 shows the results of the Rourke and NDDS compared with the BSITD-III score, in which delay was marked by a score ≤85. Using one flagged area of concern as the cut-point on the Rourke, sensitivity was 75%, specificity was 74%, positive predictive value was 30%, negative predictive value was 95% and the likelihood ratio of a positive test was 2.9. Overall accuracy, defined as the number of true positive and true negative results

TABLE 3
Test properties of the Rourke Baby Record and the Nipissing District Developmental Screen (NDDS) compared with the Bayley Scales of Infant Development III (BSITD-III) (≤85)* at one and two flag† cut-points

	Sens	Spec	PPV	NPV	False positive	False negative	LR+	Acc
Rourke								
One flag	75	74	30	95	26	25	2.9	74
Two flags	75	93	60	96	7	25	10.1	90
NDDS								
One flag	75	78	33	95	22	25	3.3	77
Two flags	75	96	75	96	4	25	20.3	94

*Data presented as %, except for the likelihood ratio of a positive test (LR+). *A score of ≤85 (one SD below the mean) in any domain on the BSITD-III served as the gold standard cut-off for normal performance and was used to indicate developmental delay. †A 'No' response, indicating that the child did not demonstrate the skill, was termed a 'flag'. Acc Accuracy; NPV Negative predictive value; PPV Positive predictive value; Sens Sensitivity, Spec Specificity*

TABLE 4
Test properties of the Rourke Baby Record and the Nipissing District Developmental Screen (NDDS) compared with the Bayley Scales of Infant Development III (BSITD-III) (≤70)* at one and two flag† cut-points

	Sens	Spec	PPV	NPV	False positive	False negative	LR+	Acc
Rourke								
One flag	100	72	20	100	28	0	3.6	74
Two flags	100	90	40	100	10	0	9.7	90
NDDS								
One flag	100	76	22	100	24	0	4.1	77
Two flags	100	93	50	100	7	0	14.5	94

*Data presented as %, except for the likelihood ratio of a positive test (LR+). *A score of ≤70 (two SDs below the mean) on the BSITD-III was assessed to determine whether the test characteristics differed when using a more impaired definition of delay. †A 'No' response, indicating that the child did not demonstrate the skill, was termed a 'flag'. Acc Accuracy; NPV Negative predictive value; PPV Positive predictive value; Sens Sensitivity, Spec Specificity*

divided by the total number of participants, was found to be 74%. Similar results were found with omission of the last question on the Rourke, an open-ended question regarding parental concerns (results not shown). When using two flagged areas of concern as the cut-point on the Rourke in comparison with the BSITD-III (one SD below mean), sensitivity was 75%, specificity was 93%, positive predictive value was 60%, negative predictive value was 96%, the likelihood ratio of a positive test rose to 10.1 and overall accuracy was 90%.

The results of the comparison of the NDDS using one flagged area of concern as the cut-point compared with the BSITD-III at one SD below the mean showed that the sensitivity was 75%, specificity was 78%, positive predictive value was 33%, negative predictive value was 95%, and likelihood ratio of a positive test was 3.3; overall accuracy was 77%. On comparison of the NDDS using two flagged areas of concern as the cut-point with the BSITD-III (one SD below the mean), sensitivity was 75%, specificity was 96%, positive predictive value was 75%, negative predictive value was 96%, and likelihood ratio of a positive test was 20.3; overall accuracy was found to be 94%.

The results of the Rourke and NDDS compared with performance on the BSITD-III at two SDs below the mean (≤70) are shown in Table 4. Using one flagged area of concern as the cut-point on the Rourke, sensitivity was 100%, specificity was 72%, positive predictive value was 20%, negative predictive value was 100% and the likelihood ratio of a positive test was 3.6. When the Rourke was assessed using the two flagged areas of concern compared with the BSITD-III score at two SDs below the mean, sensitivity was 100%, specificity was 90%, positive predictive value was 40%, negative predictive value was 100% and the likelihood ratio of a positive test was 9.7; overall accuracy was 90%.

For the NDDS using one flagged area of concern as the cut-point, in comparison with the BSITD-III score with two SDs below the mean, sensitivity was 100%, specificity was 76%, positive predictive value was 22%, negative predictive value was 100% and the likelihood ratio of a positive test was 4.1. For the NDDS using two flagged areas of concern as the cut-point in comparison with the BSITD-III score with two SDs below the mean, sensitivity was 100%, specificity was 93%, positive predictive value was 50%, negative predictive value was 100% and the likelihood ratio of a positive test was 14.5.

DISCUSSION

These preliminary results suggest that the Rourke and the NDDS tests have reasonable sensitivity and specificity and excellent negative predictive value when tested in relation with the BSITD-III among a sample of high-risk children at 36 months of age. Specificity and likelihood ratios were improved when the cut-off was set at two flags. The likelihood ratios found using a one flag cut-point were all <5, indicating small changes between the pretest probability and the post-test probability of developmental delay (8). When using a two flag cut-point, the likelihood ratios were all >10, suggesting large changes between the pretest probability and the post-test probability of developmental delay. Investigation of these tests as screens for developmental delay had not previously been conducted; therefore, these findings suggest that these tools may be appropriate screening tools for developmental delay with relatively few false negatives. The majority of children who tested positive for developmental delay on the BSITD-III also received a positive result on one of the screens.

The Rourke is designed to serve as a surveillance measure, not as a screening tool. Furthermore, its use integrates other components, such as physical examination, as well as parental report on other areas. In the present pilot study, the Rourke was used as a screening tool in that it assessed the child's development at a specific time point. Future studies of the Rourke's ability to detect delay should include information collected from the Rourke as a surveillance measure to more accurately depict the appropriate administration methods of the measure.

The 'gold standard' assessment for developmental delay was the BSITD-III. Anderson et al (9) suggested that the BSITD-III may underestimate developmental delay, and that the reference values used to indicate normal development are not representative of true performance in the general population. When they compared two samples of children, one cohort at elevated risk of delay and one control, the reference values provided by the BSITD-III detected no true difference in performance; however, the actual performance between the two groups was suggestive of delay. In consideration of this finding, participants in the present study may have had delay that, theoretically, may have been identified by the screening tool but not by the BSITD-III.

There were five participants who had a flagged area of concern on either the Rourke or the NDDS but not with both screening

tools. On further analysis of these questions, of the three participants who flagged positive on the Rourke, only two pertained to the child's ability to speak in sentences of five or more words and one indicated a general area of parental concern on question 9. A similar question regarding language production is contained in the NDDS but asks if the child can speak two-to-five word sentences. This discrepancy highlights the importance of appropriately defining developmental milestones in the screening assessments. In terms of flagged areas of concern noted on the NDDS but not the Rourke, two participants indicated concern regarding their child's ability to toss a ball and get dressed with assistance. These two skills do not have a similar counterpart on the Rourke, further suggesting how performance on a developmental screening test can differ depending on how a domain of child development is assessed.

Several limitations of the present study should be noted. The sample of children in the present study was limited to the high-risk children participating in the follow-up program visits, which reduces the representativeness of the results. We were unable to contact 11 (17%) of the potential participants by telephone despite multiple attempts at different times of the day or evening. It is not known whether the participants who we did not reach would have had different responses and outcomes or whether they came to their scheduled clinic visit as planned. Research suggests that high-risk children who regularly attend appointments have lower rates of developmental delay than children who miss appointments (10).

Increasing emphasis is being placed on the 18-month visit with the primary care provider because it is the final visit involving immunizations before the start of school (11). Future research on the ability of the Rourke and NDDS to assess developmental milestones should be conducted at the 18-month immunization visit.

CONCLUSION

The present study provides preliminary evidence that the modified Rourke and the NDDS both perform reasonably well at screening for developmental delay in a high-risk population when compared with the BSITD-III. Furthermore, screening properties of the tools are improved when the criteria for developmental delay extended to two flagged areas of concern on the screening tests. Further larger and sufficiently powered studies should examine the effect of the screening tools on a more diverse population by confirming the results in comparison with children at low risk of developmental delay. With the increased routine use of the Rourke and the NDDS, it is important to demonstrate their abilities to screen for developmental delay.

ACKNOWLEDGEMENTS: The authors thank the staff of the Perinatal Follow-up Program for their assistance and dedication to the project's recruitment and data collection procedures.

REFERENCES

1. Koseck K. Review and evaluation of psychometric properties of the revised Bayley Scales of Infant Development. Pediatr Phys Ther 1999;11:198-204.
2. Leib SA, Benfield DG, Guidubaldi J. Effects of early intervention and stimulation on the preterm infant. Pediatrics 1980;66:83-90.
3. Doig KB, Macias MM, Saylor CF, Craver JR, Ingram PE. The child development inventory: A developmental outcome measure for follow-up of the high-risk infant. J Pediatr 1999;135:358-62.
4. Rourke L, Godwin M, Rourke J, Pearce S, Bean J. The Rourke infant/child maintenance guide: Do doctors use it, do they find it useful, and does using it improve their well-baby visit records? BMC Fam Pract 2009;10:10-28.
5. American Academy of Pediatrics. Identifying infants and young children with developmental disorders in the medical home: An algorithm for developmental surveillance and screening. Pediatr 2006;118:405-20.
6. Dahinten VS, Ford L. Validation of the Nipissing District Developmental Screen for use with infants and toddlers – working paper. <www.effectivepractice.org/site/ywd_effectivepractice/assets/pdf/2c_British_Columbia_Validation_Study.pdf> (Accessed February 6, 2012).
7. Bayley N. Bayley Scales of Infant and Toddler Development, Administration Manual, 3rd edn. San Antonio: Harcourt Assessment Inc, 2006.
8. Jaeschke R, Guyatt GH, Sackett DL; Evidence-Based Medicine Working Group. Users' guides to the medical literature – III. How to use an article about a diagnostic test. JAMA 1994;271:703-7.
9. Anderson PJ, De Luca CR, Hutchinson E, Roberts G, Doyle LW; Victorian Infant Collaborative Group. Underestimation of developmental delay by the new Bayley-III scale. Arch Pediatr Adolesc Med 2010;164:352-6.
10. Tin W, Fritz S, Wariyar U, Hey E. Outcome of very preterm birth: Children reviewed with ease at 2 years differ from those followed up with difficulty. Arch Dis Child Fetal Neonatal Ed 1998;79:F83-7.
11. Williams R, Clinton J. Getting it right at 18 months: In support of an enhanced well-baby visit. Paediatr Child Health 2011;16:647-50.