Supplementary Online Content

Sheldrick RC, Marakovitz S, Garfinkel D, Carter AS, Perrin EC. Comparative accuracy of developmental screening questionnaires. *JAMA Pediatr*. Published online February 17, 2020. doi:10.1001/jamapediatrics.2019.6000

This supplementary material has been provided by the authors to give readers additional information about their work.

eTable 1. Demographics of Children Who Completed and Did Not Complete Screening

	Frequency (%) or Mean (SD)		
	Completed screening	Declined screening	p-value
n	1495	1052	
Child			
Age at enrollment	2.6 (1.3)	2.7 (1.4)	0.001
Sex			
Male	779 (52.1)	538 (51.1)	0.76
Female	716 (47.9)	514 (48.9)	
Race			
White	1102 (73.7)	707 (67.2)	0.001
African-American	198 (13.2)	132 (12.5)	0.59
Asian	97 (6.5)	72 (6.8)	0.73
Other	98 (6.6)	141 (13.4)	
Ethnicity			
Hispanic	268 (17.9)	238 (22.6)	0.003
Not Hispanic	1129 (75.5)	673 (64)	

eTable 2. Demographics of Children Who Completed and Did Not Complete Developmental Evaluations

	Frequency (%) or Mean (SD)		
	Completed evaluation	Declined evaluation	p-value
n	642	309	
Child			
Age at enrollment	2.4 (1.3)	2.4 (1.4)	0.96
Age at evaluation	2.6 (1.3)	--	--
Sex			
Male	353 (55)	168 (54.4)	0.68
Female	289 (45)	141 (45.6)	
Race			
White	458 (71.3)	202 (65.4)	0.19
African-American	89 (13.9)	59 (19.1)	0.02
Asian	47 (7.3)	20 (6.5)	0.71
Other	48 (7.5)	28 (9.1)	
Ethnicity			
Hispanic	121 (18.8)	71 (23)	0.09
Not Hispanic	473 (73.7)	210 (68)	
Premature birth	76 (11.8)	35 (11.3)	0.91
Participating parent			
Sex			
Male	64 (10)	43 (13.9)	0.71
Female	578 (90)	266 (86.1)	
Prefers Spanish language	24 (3.7)	13 (4.2)	0.66
Marital Status			
Married	452 (70.4)	181 (58.6)	0.001
Not Married	166 (25.9)	115 (37.2)	
Age, mean years (SD)	33.5 (6.4)	32 (6.2)	0.001
Education			
High School or less	205 (31.9)	138 (44.7)	0.001
Some college	88 (13.7)	49 (15.9)	
College degree	185 (28.8)	65 (21)	
Graduate degree	164 (25.5)	51 (16.5)	
not indicated	0 (0)	6 (1.9)	
Family income, $			
<30,000	84 (13.1)	32 (10.4)	0.23
30,000-49,999	43 (6.7)	22 (7.1)	
50,000-99,999	121 (18.8)	52 (16.8)	
>=100,000	186 (29)	54 (17.5)	
not indicated	208 (32.4)	149 (48.2)	
Public health insurance	143 (22.3)	76 (24.6)	0.36

eTable 3. Sensitivity and Specificity of Primary Screening Instruments by Severity of Delay

	Younger children (0-42 months)			
	Sensitivity			Specificity (no delays)
Primary Aim:	**Severe Delays (4.0%)**	**Moderate Delays (9.5%)**	**Mild Delays (14.5%)**	
SWYC Milestones	73.7% [CI: 50.1%-88.6%]	51.7% [CI: 39.1%-64.0%]	33.1% [CI: 21.7%-47.0%]	89.0% [CI: 86.1%-91.4%]
ASQ-3	60.0% [CI: 29.7%-84.2%]	50.0% [CI: 30.2%-69.8%]	23.1% [CI: 9.7%-45.5%]	89.4% [CI: 85.9%-92.1%]
PEDS	78.9% [CI: 55.4%-91.9%]	54.0% [CI: 41.6%-65.8%]	28.0% [CI: 17.9%-40.9%]	79.6% [CI: 75.7%-83.1%]
Secondary Aim:				
PEDS:DM	60.8% [CI: 49.6%-71.0%]	81.0% [CI: 66.2%-90.2%]	67.2% [CI: 40.0%-86.3%]	42.7% [CI: 30.2%-56.2%]
PEDS AND PEDS:DM	78.9% [CI: 55.4%-91.9%]	47.6% [CI: 35.6%-59.9%]	22.7% [CI: 14.1%-34.4%]	83.9% [CI: 80.3%-86.9%]
SWYC: somewhat concerned	66.7% [CI: 33.2%-89.0%]	32.5% [CI: 19.8%-48.3%]	22.4% [CI: 12.6%-36.8%]	95.5% [CI: 93.1%-97.1%]
SWYC: very concerned	11.1% [CI: 1.5%-50.2%]	2.5% [CI: 0.3%-15.9%]	1.5% [CI: 0.2%-10.2%]	97.4% [CI: 92.1%-99.2%]
SWYC Milestones OR somewhat concerned	89.5% [CI: 66.1%-97.4%]	56.7% [CI: 43.9%-68.6%]	39.9% [CI: 26.7%-54.9%]	87.3% [CI: 84.2%-89.8%]
SWYC Milestones AND somewhat concerned	57.9% [CI: 35.5%-77.4%]	31.7% [CI: 21.2%-44.4%]	19.3% [CI: 11.7%-30.2%]	95.8% [CI: 93.7%-97.2%]
	Older children (43-66 months)			
	Sensitivity			Specificity (no delays)
Primary Aim:	**Severe Delays (8.3%)**	**Moderate Delays (10.2%)**	**Mild Delays (15.7%)**	
SWYC Milestones	44.4% [CI: 17.5%-75.1%]	36.3% [CI: 11.0%-72.3%]	53.8% [CI: 38.2%-68.8%]	70.7% [CI: 60.9%-78.8%]
ASQ-3	50.0% [CI: 12.2%-87.8%]	33.3% [CI: 8.3%-73.5%]	20.5% [CI: 10.5%-36.1%]	92.1% [CI: 85.1%-95.9%]
PEDS	77.8% [CI: 41.8%-94.5%]	28.2% [CI: 8.5%-62.4%]	66.7% [CI: 50.5%-79.7%]	73.7% [CI: 64.3%-81.3%]
Secondary Aim:				
PEDS:DM	87.4% [CI: 78.3%-93.1%]	86.9% [CI: 79.4%-92.0%]	85.7% [CI: 56.7%-96.5%]	13.1% [CI: 6.7%-23.9%]
PEDS AND PEDS:DM	77.8% [CI: 41.8%-94.5%]	35.9% [CI: 12.4%-69.1%]	61.5% [CI: 45.5%-75.4%]	78.8% [CI: 70.2%-85.4%]
SWYC: somewhat concerned	80.0% [CI: 30.4%-97.3%]	32.1% [CI: 8.4%-71.0%]	31.3% [CI: 17.6%-49.2%]	93.6% [CI: 87.2%-96.9%]
SWYC: very concerned	40.0% [CI: 9.8%-80.3%]	4.6% [CI: 0.5%-31.4%]	2.0% [CI: 0.6%-6.2%]	99.2% [CI: 94.4%-99.9%]
SWYC Milestones OR somewhat concerned	55.6% [CI: 24.9%-82.5%]	50.3% [CI: 16.3%-84%]	59% [CI: 43%-73.2%]	70.0% [CI: 60.1%-78.3%]
SWYC Milestones AND somewhat concerned	44.4% [CI: 17.5%-75.1%]	50.3% [CI: 16.3%-84%]	28.2% [CI: 16.3%-44.3%]	87.9% [CI: 81%-92.5%]

Note. ASQ = Ages & Stages Questionnaire; PEDS = Parents' Evaluation of Developmental Status; SWYC = Survey of Wellbeing of Young Children

eTable 4. Frequencies and Unadjusted Estimates of Sensitivity and Specificity Among Referred Children

	Younger children (9-42 months)			
	Positive score (Sensitivity)			Negative Score (Specificity) n=302 (62.7%)
Primary Aim:	**Severe Delays; n=21 (4.4%)**	**Moderate Delays; n=69 (14.3%)**	**Mild Delays; n=90 (18.7%)**	
SWYC Milestones	n=15 (71.4%)	n=35 (50.7%)	n=37 (41.1%)	n=236 (78.1%)
ASQ-3	n=14 (66.7%)	n=35 (50.7%)	n=40 (44.4%)	n=229 (75.8%)
PEDS	n=16 (76.2%)	n=36 (52.2%)	n=34 (37.8%)	n=184 (60.9%)
Secondary Aim:				
PEDS:DM	n=21 (100%)	n=57 (82.6%)	n=67 (74.4%)	n=113 (37.4%)
PEDS AND PEDS:DM	n=16 (76.2%)	n=32 (46.4%)	n=28 (31.1%)	n=207 (68.5%)
SWYC: somewhat concerned	n=14 (66.7%)	n=22 (31.9%)	n=24 (26.7%)	n=274 (90.7%)
SWYC: very concerned	n=2 (9.5%)	n=2 (2.9%)	n=2 (2.2%)	n=302 (100%)
SWYC Milestones OR somewhat concerned	n=18 (85.7%)	n=38 (55.1%)	n=44 (48.9%)	n=226 (74.8%)
SWYC Milestones AND somewhat concerned	n=12 (57.1%)	n=23 (33.3%)	n=23 (25.6%)	n=274 (90.7%)
	Older children (43-66 months)			
	Positive score (Sensitivity)			Negative Score (Specificity) n=89 (57.8%)
Primary Aim:	**Severe Delays; n=10 (6.5%)**	**Moderate Delays; n=15 (9.7%)**	**Mild Delays; n=40 (26%)**	
SWYC Milestones	n=5 (50%)	n=9 (60%)	n=21 (52.5%)	n=43 (48.3%)
ASQ-3	n=6 (60%)	n=7 (46.7%)	n=8 (20%)	n=73 (82%)
PEDS	n=8 (80%)	n=7 (46.7%)	n=27 (67.5%)	n=47 (52.8%)
Secondary Aim:				
PEDS:DM	n=10 (100%)	n=15 (100%)	n=34 (85%)	n=22 (24.7%)
PEDS AND PEDS:DM	n=8 (80%)	n=7 (46.7%)	n=25 (62.5%)	n=55 (61.8%)
SWYC: somewhat concerned	n=8 (80%)	n=7 (46.7%)	n=13 (32.5%)	n=78 (87.6%)
SWYC: very concerned	n=4 (40%)	n=0 (0%)	n=0 (0%)	n=88 (98.9%)
SWYC Milestones OR somewhat concerned	n=6 (60%)	n=11 (73.3%)	n=23 (57.5%)	n=42 (47.2%)
SWYC Milestones AND somewhat concerned	n=5 (50%)	n=6 (40%)	n=11 (27.5%)	n=70 (78.7%)

Note. ASQ = Ages & Stages Questionnaire; PEDS = Parents' Evaluation of Developmental Status; SWYC = Survey of Wellbeing of Young Children

eAppendix. Additional Detail Regarding Protocol

Introduction

Increasing numbers of states and professional organizations are recommending and even mandating systematic screening for developmental delays among young children in pediatric settings. Unfortunately, there is a weak evidence base from which pediatricians are forced to make decisions to inform such screening. To address the considerable gaps in the research literature on screening, we aimed to conduct a large, diagnostic accuracy study to compare 3 sets of developmental-behavioral screening instruments for children under 5 years of age (the Ages and Stages Questionnaire, the Parent's Evaluation of Developmental Status, and our own instrument, the Survey of Wellbeing on Young Children). The importance of accuracy for a screening instrument is widely acknowledged. If even a modest proportion of pediatricians adopted a more sensitive screening instrument, a large number of additional children with developmental-behavioral disorders would be identified and thus have the opportunity to receive appropriate interventions. A corresponding increase in specificity would result in a dramatic reduction in the number of false positive cases, thereby alleviating burden on primary care pediatricians, specialists, and parents.

For this reason, studies that estimate sensitivity and specificity of developmental screening questionnaires abound. We cite two primary sources of evidence regarding the sensitivity and specificity of the ASQ and the PEDS: the systematic review conducted to support recommendations of the Canadian Task Force on Preventive Health Care (which cites four studies), and the American Academy of Pediatrics (AAP) consensus statement on developmental screening (which cites the ASQ and PEDS manuals). For the SWYC Milestones, we rely on the original published study. Further detail regarding the population, age-group, and reference standards included in these studies are provided in the table below.

Methods

Developmental assessment

Child assessment visits were conducted by trained examiners, supervised by a licensed psychologist, and videotaped for later review. Specifically, 12 research staff administered developmental testing over the course of the study. The same two clinicians (one supervising clinical psychologist and one bilingual developmental-behavioral pediatrician) provided supervision regarding developmental testing throughout the study. All testing was videotaped unless the participant requested to opt-out (which did not otherwise affect eligibility). The tapes were reviewed by the study clinicians to ensure adherence to protocol and to advise research assistants regarding administration and scoring to ensure fidelity.

Analyses

Sampling strategy. Our design takes advantage of the fact that children who score negative on any given screening instrument actually represent two different populations: (1) those who score positive on at least one other screening instrument for development, behavior or autism, and (2) those who score negative on all screening instruments. By inviting 100% children in group #1 for evaluations, our design oversamples children who are most likely to represent false negatives for any single screening instruments (i.e., those who screen positive on a different screening instrument). Moreover, these children are the most important ones to evaluate when comparing the accuracy of screening instruments because they populate the discrepant cells (i.e., instances of disagreement reflecting positive scores on one test but negative on another) in analyses of dependent samples. The decision to sample 10% of children who screen negative applies only to group #2—those who score negative on ***all*** screening instruments. Sampling of this group of children can influence overall estimates of accuracy, but in a way that influences estimates for all screeners equivalently (e.g., these children are either true negatives or false negatives on every screener). Thus, comparisons among screeners are unaffected. The *a priori* power estimates we conducted suggested that 10% would be sufficient, and this sampling is reflected by our inverse probability weights and therefore the confidence intervals we report in this paper. Therefore, the confidence intervals we report should be generalizable to pediatric populations.

Statistical Analyses. To estimate and compare sensitivity and specificity for each questionnaire we used generalized estimating equations (GEEs) with logit links to simultaneously estimate true and false positive fractions and their 95% confidence intervals. We included covariates and their interactions with screener-type to account for administration in Spanish and for use of an earlier edition of the ASQ. To account for severity, we separately

assessed sensitivity to mild, moderate, and severe delays and then calculated specificity among children with no evidence of delay. From these statistics, we also calculated positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (LR+) and negative likelihood ratio (LR-) with respect to mild-to-severe delays. The method of using GEEs to estimate the effects of covariates on sensitivity and specificity was developed by Pepe and colleagues in a series of papers[1,2] and detailed in a published book.[3] We chose this approach because it is, to our knowledge, the only regression-based method that supports estimation and comparison of key indicators of diagnostic accuracy based on dependent samples, as employed in this study. As the authors note, "Surprisingly, statistical methods for comparing tests with regard to [diagnostic accuracy] parameters have not been available for the most common study design in which each test is applied to each study individual. [Therefore,] we propose a statistic for comparing the predictive values of two diagnostic tests using this paired study design. The proposed statistic is a score statistic derived from a marginal regression model and bears some relation to McNemar's statistic. As McNemar's statistic can be used to compare sensitivities and specificities of diagnostic tests, parameters that condition on disease status, our statistic can be considered as an analog of McNemar's test for the problem of comparing predictive values, parameters that condition on test outcome."[1]

Table. Further detail regarding validation studies for ASQ, PEDS, and SWYC

	Population			Screener		Reference standard	
Reference	**sample**	**location**	**age**	**type**	**rule**	**type**	**clinical threshold**
Rydz et al 2006	primary care	US	18 mo	ASQ	<2 SD on ≥1 domain	Battelle Develop-mental Inventory	not reported
Steenis et al., 2015	convenience	Nether-lands	2-42 mo	ASQ-III	<2 SD	Bayley-III NL	2 SD
Limbos et al., 2011	primary care	Canada	12-60 mo	ASQ-II, PEDS	<2 SD on ASQ; ≥1 predictive concern on PEDS	Bayley-III, WISC-3, PLS-4 + VABS-2	<10th %tile
Gollenberg et al., 2009	convenience sample enrolled in longitudinal study	US	24 mo	ASQ-II	<2 SD	Bayley-II	< 1 SD & < 2 SD
PEDS manual	Primary sample: Self-selected parents who used PEDS Online plus parents who completed PEDS during well-child visits	US	M = 35 mo (0-6 years)	PEDS	Predictive concern	"eligible for services via diagnostic testing"	varied by state criteria: "e.g., two 25% delays, 1 ½ sd below the mean, etc"
ASQ-III manual	"Identified group" from EI/ECSE programs and a "typical group" from child care centers, preschool programs, and internet ads	US	2-60 months	ASQ-3	<2 SD below mean in one or more areas	Battelle Developmental Inventory "in over 90% of cases"	<75 on any scale or subscale

Sheldrick & Perrin, 2013	primary care	US	2-66 months	SWYC, ASQ-III	<2 SD	Parent-report of developmental delay	N/A

Based on estimates of sensitivity, specificity, and prevalence reported in Table 2, additional parameters were estimated as follows:

$$PPV = \frac{sensitivity * prevalence}{sensitivity * prevalence + (1 - specificity) * (1 - prevalence)}$$

$$NPV = \frac{specificity * (1 - prevalence)}{specificity * (1 - prevalence) + (1 - sensitivity) * prevalence}$$

$$LR_{+} = \frac{sensitivity}{1 - specificity}$$

$$LR_{-} = \frac{1 - sensitivity}{specificity}$$

$$DOR = \frac{LR_{+}}{LR_{-}}$$

References

1. Leisenring W, Alono T, Pepe MS. Comparisons of predictive values of binary medical diagnostic tests for paired designs. *Biometrics.* 2000;56(2):345-351.
2. Moskowitz CS, Pepe MS. Comparing the predictive values of diagnostic tests: sample size and analysis for paired study designs. *Clinical trials*. 2006;3: 272-279.
3. Pepe MS. *The statistical evaluation of medical tests for classification and prediction*. Oxford; 2003.