View Correction

# Comparative Accuracy of Developmental Screening Questionnaires

R. Christopher Sheldrick, PhD[1]; Susan Marakovitz, PhD[2]; Daryl Garfinkel[2] ; et al

≫ Author Affiliations | Article Information

❝ Cite    Ⓒ Permissions    ⬈ Metrics    💬 Comments

▤ RELATED ARTICLES    🖾 FIGURES    ↓ SUPPLEMENTAL CONTENT

## Key Points

**Question** Which screening question simes are most accurate for detecting developmental delays among

Sections      PDF      Share

**Findings** In this diagnostic accuracy study of 1495 families enrolled from primary care settings, trade-offs in sensitivity and specificity were observed among 3 screening tools (Ages and Stages Questionnaire, Third Edition, Parents' Evaluation of Developmental Status, and Survey of Well-being of Young Children: Milestones), but no one questionnaire emerged as superior overall. All questionnaires displayed specificity higher than 70%, but sensitivity exceeded 70% only for the Parents' Evaluation of Developmental Status with respect to severe delays and for the Survey of Well-being of Young Children: Milestones with respect to severe delays among children younger than 42 months.

**Meaning** Results of this study suggest that all 3 developmental screening questionnaires offer modest advantages to pediatric practitioners for detecting developmental delays.

## Abstract

**Importance** Universal developmental screening is widely recommended, yet studies of the accuracy of commonly used questionnaires reveal mixed results, and previous comparisons of these questionnaires are hampered by important methodological differences across studies.

**Objective** To compare the accuracy of 3 developmental screening instruments as standardized tests of developmental status.

**Design, Setting, and Participants** This cross-sectional diagnostic accuracy study recruited consecutive parents in waiting rooms at 10 pediatric primary care offices in eastern Massachusetts between October 1, 2013, and January 31, 2017. Parents were included if they were sufficiently literate in the English or Spanish language to complete a packet of screening questionnaires and if their child was of eligible age. Parents completed all questionnaires in counterbalanced order. Participants who screened positive on any questionnaire plus 10% of those who screened negative on all questionnaires (chosen at random) were invited to complete developmental testing. Analyses were weighted for sampling and nonresponse and were conducted from October 1, 2013, to January 31, 2017.

**Exposures** The 3 screening instruments used were the Ages & Stages Questionnaire, Third Edition (ASQ-3); Parents' Evaluation of Developmental Status (PEDS); and Survey of Well-being of Young Children (SWYC): Milestones.

**Main Outcomes and Measures** Reference tests administered were Bayley Scales of Infant and Toddler Development, Third Edition, for children aged 0 to 42 months, and Differential Ability Scales, Second Edition, for older children. Age-standardized scores were used as indicators of mild (80-89), moderate (70-79), or severe (<70) delays.

primarily female (1325 [88.7%]), with a mean (SD) age of 33.4 (6.3) years. Of the 20.5% to 29.0% of children with a positive score on each questionnaire, 35% to 60% also received a positive score on a second questionnaire, demonstrating moderate co-occurrence. Among younger children (<42 months), the specificity of the ASQ-3 (89.4%; 95% CI, 85.9%-92.1%) and SWYC Milestones (89.0%; 95% CI, 86.1%-91.4%) was higher than that of the PEDS (79.6%; 95% CI, 75.7%-83.1%; *P*<.001 and *P*=.002, respectively), but differences in sensitivity were not statistically significant. Among older children (43-66 months), specificity of the ASQ-3 (92.1%; 95% CI, 85.1%-95.9%) was higher than that of the SWYC Milestones (70.7%; 95% CI, 60.9%-78.8%) and the PEDS (73.7%; 95% CI, 64.3%-81.3%; *P*<.001), but sensitivity to mild delays of the SWYC Milestones (54.8%; 95% CI, 38.1%-70.4%) and of the PEDS (61.8%; 95% CI, 43.1%-77.5%) was higher than that of the ASQ-3 (23.5%; 95% CI, 9.0%-48.8%; *P*=.012 and *P*=.002, respectively). Sensitivity exceeded 70% only with respect to severe delays, with 73.7% (95% CI, 50.1%-88.6%) for the SWYC Milestones among younger children, 78.9% (95% CI, 55.4%-91.9%) for the PEDS among younger children, and 77.8% (95% CI, 41.8%-94.5%) for the PEDS among older children. Attending to parents' concerns was associated with increased sensitivity of all questionnaires.

**Conclusions and Relevance**  This study found that 3 frequently used screening questionnaires offer adequate specificity but modest sensitivity for detecting developmental delays among children aged 9 months to 5 years. The results suggest that trade-offs in sensitivity and specificity occurred among the questionnaires, with no one questionnaire emerging superior overall.

## Introduction

Accurate instruments are widely recognized as essential if universal developmental screening is to fulfill its goals. The value of a questionnaire's results for case conceptualization, decision-making, and ultimately service receipt depends on the questionnaire's ability to yield accurate information.[1] Thus, organizations such as the US Preventive Services Task Force[2] and the Canadian Task Force on Preventive Health Care[3] carefully consider evidence on the screening instruments' sensitivity and specificity when making determinations about their overall effectiveness in improving children's health.

Studies that estimate the sensitivity and specificity of developmental screening questionnaires abound, yet few publications meet consensus reporting guidelines for diagnostic accuracy, such as the revised Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2).[4] For example, a range of evidence is frequently cited to support the Ages & Stages Questionnaire (ASQ) and the Parents' Evaluation of Developmental Status (PEDS), 2 of the most widely used developmental screening questionnaires in pediatrics.[5,6] This body of research includes samples derived from primary care and specialty populations,

evidence (and explicitly citing publishers' manuals and websites), the American Academy of Pediatrics (AAP) consensus statement on developmental screening reports that the ASQ displays a sensitivity range of 0.70 to 0.90 and a specificity range of 0.76 to 0.91, whereas the PEDS displays a sensitivity of 0.96 and a specificity of 0.83.[7] These values are above the 0.70 threshold commonly recommended to represent adequate sensitivity and specificity.[8]

To assess the effectiveness of universal developmental screening in primary care settings, a meta-analysis included only studies that were conducted in low-risk populations and used a standardized diagnostic evaluation.[9] That meta-analysis identified only 4 studies that met the inclusion criteria, all of which assessed the ASQ's accuracy and 1 of which also assessed the accuracy of the PEDS. For the ASQ, the meta-analysis found a median sensitivity of 55.0% (range, 47.1%-66.7%) and a median specificity of 86.0% (range, 38.6%-94.3%); for the PEDS, it found a sensitivity of 41.1% (95% CI, 24.7%-59.3%) and a specificity of 89.3% (95% CI, 85.1%-92.5%).[9] The review also noted a high risk of bias in 3 studies in at least 1 QUADAS-2 domain, and a fourth study displayed unclear risk of bias in 3 QUADAS-2 domains. The relatively small number of studies identified echoed the conclusion of an earlier systematic review that concluded "there are surprisingly few published studies that describe the psychometric characteristics of the developmental screening tests … and even fewer studies that demonstrate their utility and validity in clinical settings."[10,11(p29)] Because studies with different designs, that were conducted with different populations, and that include multiple reference standards scored with varying definitions of developmental delay cannot be effectively compared using quantitative methods, the precise cause of the discrepancy between these systematic reviews and the AAP statement is unclear.

To better inform decisions about developmental screening for young children, we conducted a diagnostic accuracy study with a primary aim of comparing 3 prominent developmental screening instruments: ASQ-3, PEDS, and the Survey of Well-being of Young Children (SWYC): Milestones,[12] a freely available screening instrument that is included in the most recent AAP guidelines for developmental screening.[7] All 3 of these instruments are cited in the Bright Futures guidelines of the AAP.[13] To control for heterogeneity in methods that challenge meta-analyses, we provided direct comparisons in a single study. As a secondary aim, we explored the accuracy of (1) PEDS: Developmental Milestones, a follow-up assessment recommended to increase the predictive value of the PEDS, which was included at the request of its author, and (2) a single question about parent concerns on the SWYC that was recommended by the AAP.[13] We assessed the accuracy of both measures alone and in combination with their parent questionnaire.

## Methods

### Participants

---

English or Spanish language to complete the questionnaires and if their child was of eligible age. Of approximately 3370 families approached (**Figure**), 2597 (77%) offered consent to contact and were eligible, and 1545 (60%) of these families completed a packet of screening instruments. Fifty children with known developmental delays or autism, as reported by parents, were excluded from further analyses. Every child with a positive score on at least 1 questionnaire was offered a comprehensive evaluation, and each child with a negative score on all screening instruments had a 10% chance of selection (**Figure**). Among the 951 families selected, 642 (68%) completed evaluations.

⤢ View Large       ↓ Download

**Figure. Enrollment and Evaluation**


Enrollment and Evaluation

Screening packets included age-appropriate questionnaires that assess risk for developmental delay, behavioral disorders, and autism. Standardized developmental tests were administered during the evaluation.

## Study Procedure

Participants were asked to complete a packet of age-appropriate developmental, behavioral, and autism-specific screening questionnaires in counterbalanced order as well as to answer questions regarding demographic characteristics and race/ethnicity (using National Institutes of Health categories). Parents could choose to complete the questionnaires in the waiting room or at home and then return the forms using a prestamped envelope. Study procedures followed QUADAS-2 recommendations[4] and were approved by the institutional review board at Tufts University School of Medicine. Written informed consent was provided by all participants.

## Developmental Screening Questionnaires

Developmental screening questionnaires included the ASQ-3 (third edition),[14] PEDS,[15] and SWYC Milestones[12] (eAppendix in the **Supplement**). Although research suggests that provision of props and toys may not be necessary for ensuring the accuracy of the ASQ,[16] all parents were provided with materials (eg, blocks, crayons) to facilitate completion of the questionnaire, as recommended in the manual. During the first phase of the study, the ASQ-2 (the second edition of the ASQ) was administered. The SWYC Milestones was administered with the question, "Do you have any concerns about your child's learning or

☰ **Sections**        📄 **PDF**        ⌣ **Share**

## Developmental Assessment

Research assistants double-entered the data using software with automatic scoring. One of our senior investigators (R.C.S.) determined which families would be invited for evaluations on the basis of questionnaire results and a random number generator. Child assessment visits were conducted by one of our trained examiners (including D.G.), supervised by one of our licensed clinicians (S.M.), and videotaped for later review. Bilingual examiners conducted the assessments with Spanish-speaking families. Protocols were adapted for Spanish-speaking children to include tests with demonstrated validity for this population. Examiners and their supervisors were unaware of the screening results. The median (interquartile range [IQR]) time from screening to evaluation was 73 (49-113) days.

## Developmental Status Tests

Reference tests included the Bayley Scales of Infant and Toddler Development, Third Edition, to evaluate language and cognitive development for children from 9 through 42 months of age, and the Differential Ability Scales, Second Edition, for older children. To assess the language development of Spanish-speaking children, we used a published translation of the Differential Ability Scales, Second Edition; a previous translation of the Bayley Scales of Infant and Toddler Development, Second Edition, cognitive scales; and the Spanish edition of the Preschool Language Scale, Fifth Edition. Fine and gross motor development were assessed for all children using the Battelle Developmental Inventory, Second Edition. Scores were categorized as typical (age-standardized scores of ≥90), mild (80-89), moderate (70-79), or severe (<70) delays.

## Statistical Analysis

Using Stata, version 15 (StataCorp LLC), we calculated the proportion of positive scores on each questionnaire and the co-occurrence with other questionnaires. Next, sensitivity and specificity for each questionnaire were analyzed and compared. These analyses were conducted separately for children younger or older than 42 months, because they received different reference tests. Following published recommendations,[17,18] we used generalized estimating equations with logit links to simultaneously estimate true and false positive fractions and their 95% CIs while accounting for clustering by practice. We included covariates and their interactions with questionnaire 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, negative predictive value, positive likelihood ratio, and negative likelihood ratio[19,20] with respect to mild to severe delays. We calculated the

Sections

PDF

Share

dom selection of children with a negative score—ie, planned missing data). Following published recommendations, we addressed unplanned missing data (eg, declining to attend the evaluation) by multiple imputation with chained equations using models that included variables predicting both missingness and outcome variables.[22] These variables included developmental questionnaire scores, parents' concerns, parents' perceptions of screening, and demographic variables (income, educational level, and race/ethnicity). Twenty multiple-imputed data sets were created on the survey-weighted data set. To assess for misspecification, we compared the analyses based on the missing data model with those calculated through complete case analysis.[23]

All tests were 2-tailed, and a type I error rate of 0.05 was used to evaluate statistical significance. Statistical analyses were performed from October 1, 2013, to January 31, 2017.

# Results

In total, 1495 families of children aged 9 months to 5.5 years participated. **Table 1** presents self-reported demographic characteristics. The mean (SD) age of the children at enrollment was 2.6 (1.3) years, 779 (52.1%) were male, and approximately one-third were of nonwhite race and/or Hispanic ethnicity (compared with 30% in the Greater Boston metropolitan area and 39% in the United States).[24] Parent respondents were primarily married (1022 [68.4%]) and female (1325 [88.7%]), with a mean (SD) age of 33.4 (6.3) years. The sample was diverse with respect to socioeconomic status, with 475 parents (31.7%) reporting a high school education or less and 353 (23.4%) reporting a graduate degree.

⊡ View Large    ↓ Download

**Table 1. Baseline Characteristics of Screened Sample**

Baseline Characteristics of Screened Sample

Logistic regressions revealed the differences in nonresponse at each of the 2 points at which selection bias was possible. Parents who did not complete the screening packets (n=1052), compared with those who did (n=1495), were more likely to report younger child age (mean [SD] age, 2.7 [1.4] years vs 2.6 [1.3] years; *P*=.001) as well as nonwhite race (345 [32.8%] vs 395 [26.3%]) and Hispanic ethnicity (268 [17.9%] vs 238 [22.6%]; *P*=.003) (eTable 1 in the **Supplement**). Parents who were offered but declined to complete comprehensive evaluations for their children (n=309) were more likely than those whose children completed evaluations (n=642) to report black race (59 [19.1%] vs 89 [13.9%]; *P*=.02), being un-

Sections          PDF          Share

[SD] age, 32.0 [6.2] years vs 33.5 [6.4] years; *P*=.001) (eTable 2 in the **Supplement**). These variables were included in the models of nonresponse.

**Table 2** presents the proportion of children with a positive score on each questionnaire and co-occurrence with other questionnaires. Among the 20.5% to 29.0% of children with a positive score on 1 question-naire, the proportion of those who also obtained a positive score on a second questionnaire ranged from 35% to 60%. Parents were more likely to score positive on the PEDS (422 [29.0%]) than in response to the single SWYC question about concern (127 [8.8%]). Whereas most parents who reported being very much concerned on the SWYC question also obtained a positive score on each of the 3 primary screening questionnaires (ASQ-3: 11 [78.6%]; PEDS: 14 [100%]; SWYC: 14 [100%]), the converse was not true; only a minority of parents whose children had a positive score on 1 of the 3 primary screening instruments re-ported being even somewhat concerned (ranging from 64 [21.7%] to 108 [25.7%]).

⬒ **View Large**    ↓ **Download**

**Table 2. Proportion Screening Positive and Co-occurrence With Positive Scores on Other Questionnaires**

Proportion Screening Positive and Co-occurrence With Positive Scores on Other Questionnaires

**Table 3** presents estimates of sensitivity and specificity for severe, moderate to severe, and mild to severe (any) delays (see eTables 3 and 4 in the **Supplement** for adjusted and unadjusted estimates of sensitivity by severity level). Point estimates suggest that all 3 questionnaires displayed adequate specificity (ie, ≥0.70).[8] Sensitivity exceeded 70% only with respect to severe delays for the PEDS (78.9%; 95% CI, 55.4%-91.9%) and for the SWYC Milestones (73.7%; 95% CI, 50.1%-88.6%) among younger children (<42 months) and for the PEDS among older children (77.8%; 95% CI, 41.8%-94.5%). Patterns were similar across adjusted and unadjusted analyses. Questionnaire order was not statistically significant. Although the estimate of the ASQ-3's sensitivity was higher than that of the ASQ-2, the difference was not statisti-cally significant. No differences were found between Spanish and English language forms, with the excep-tion of the Spanish version of the ASQ, which was more sensitive than the English version among younger children.

⬒ **View Large**    ↓ **Download**

**Table 3. Sensitivity and Specificity of Primary Screening Instruments[a]**

Sensitivity and Specificity of Primary Screening Instruments

⦙≡
**Sections**

📄
**PDF**

⤙
**Share**

Comparisons between questionnaires revealed that, among younger children (<42 months), the ASQ-3 (89.4%; 95% CI, 85.9%-92.1%) and the SWYC Milestones (89.0%; 95% CI, 86.1%-91.4%) were both more specific than the PEDS (79.6%; 95% CI, 75.7%-83.1%; $P<.001$ and $P=.002$, respectively), but the differences in sensitivity were not statistically significant. Among older children (43-66 months), the SWYC Milestones (54.8%; 95% CI, 38.1%-70.4%) and the PEDS (61.8%; 95% CI, 43.1%-77.5%) were both more sensitive to mild delays compared with the ASQ-3 (23.5%; 95% CI, 9.0%-48.8%; $P=.012$ and $P=.002$, respectively), but the ASQ-3 (92.1%; 95% CI, 85.1%-95.9%) was more specific than both the SWYC Milestones (70.7%; 95% CI, 60.9%-78.8%) and the PEDS (73.7%; 95% CI, 64.3%-81.3%; $P<.001$).

In secondary analyses among younger children (<42 months), requiring a positive score on the SWYC Milestones and a finding of parent concern yielded lower sensitivity to severe delays (57.9%; 95% CI, 35.5%-77.4%) but higher specificity overall (95.8%; 95% CI, 93.7%-97.2%). In contrast, defining a positive result as consisting of either a positive score on the SWYC Milestones or a finding of parent concern yielded higher sensitivity to severe delays (89.5%; 95% CI, 66.1%-97.4%) but lower specificity overall (87.3%; 95% CI, 84.2%-89.8%). Rescreening children with a positive score on the PEDS with the PEDS: Developmental Milestones increased specificity (83.9%; 95% CI, 80.3%-86.9%) but had no effect on sensitivity (78.9%; 95% CI, 55.4%-91.9%). Similar patterns were observed among older children (42-66 months).

Table 4 presents positive predictive value, negative predictive value, positive likelihood ratio, and negative likelihood ratio with respect to any delay. Among children who had a positive score on any of the 3 primary questionnaires, 44.0% to 60.6% had at least a mild delay on the reference tests (ie, positive predictive value), whereas 77.7% to 80.2% of children with a negative screen tested in the typical range (ie, negative predictive value). Because these statistics were associated with base rate (which varied across samples), we also report the likelihood ratios, which are based directly on sensitivity and specificity (not base rate). Positive likelihood ratio ranged from 1.87 (95% CI, 1.24-2.83) to 3.95 (95% CI, 2.86-5.47), indicating that the odds of having a developmental delay were approximately 2 to 4 times higher if a child had a positive screen. Negative likelihood ratio ranged from 0.83 (95% CI, 0.69-1.00) to 0.52 (95% CI, 0.34-0.78), indicating that the odds of having a developmental delay were approximately 20% to 50% lower if a child had a negative score. Diagnostic odds ratio ranged from 2.9 to 6.3, suggesting mild to moderate overall accuracy.

View Large    Download

**Table 4. Positive Predictive Value, Negative Predictive Value, and Likelihood Ratios With Respect to Any Delay[a]**

Sections          PDF          Share

# Discussion

Results of this study suggest that developmental screening questionnaires offer modest advantages to primary care practitioners for detecting developmental delays. Moderate co-occurrence of positive results among screening instruments is consistent with previous findings,[25] as is the finding that high levels of concern are likely to coincide with positive screening scores but that positive screening scores reflect parents' concerns in only a few cases.[26] Inclusion of standardized developmental tests allowed us to extend these findings to address accuracy. Moderately high positive predictive values suggest that a sizable proportion of children with a positive score on the ASQ-3, PEDS, or SWYC Milestones meet the criteria for developmental delay if formally tested. However, although the sensitivity for severe delays approached or exceeded 70%, it fell below this mark for moderate and mild delays. Positive and negative likelihood ratios were also modest.

Results also suggest that sensitivity increases when questionnaire results are considered while attending closely to parent concerns. The PEDS, which exclusively assesses parental concerns, displayed point estimates for sensitivity to severe delays that were higher than the estimates for other questionnaires. Inclusion of a parent's concern when interpreting the SWYC Milestones results increased this instrument's sensitivity. However, achieving this level of sensitivity requires the capacity and motivation among practitioners to closely evaluate children whose parents report being somewhat concerned or who endorse as few as 1 concern as required to obtain a positive score on the PEDS. For many pediatricians, the predictive value of this comparatively low level of concern may fall below the threshold necessary to justify action.[27,28]

Findings of modest accuracy raise questions about the utility of universal developmental screening. Many countries outside the United States do not endorse universal screening.[3] However, questionnaires with modest accuracy may still contribute to clinical care. Given that screening is typically conducted in the context of developmental surveillance (a standard element of a pediatric well-child visit that includes observation of the child), a screening questionnaire's ability to add relevant information to what is typically gathered through the clinical examination is important to increase the accuracy of clinical judgment. Although comparisons to standard pediatric care are outside the scope of the present study, we believe that the fact that the diagnostic odds ratios reported here exceed those documented in a systematic review of the accuracy of standard pediatric surveillance[29] is indirect evidence that screening instruments can provide useful information. Moreover, these questionnaires may offer other advantages beyond their psychometric properties. Investigators have long noted that screening instruments' usefulness depends not only on their accuracy but also on their ability to inform case conceptualization and medical decision-making.[1,30] This idea is consistent with recent research suggesting that screening questionnaires can play

Sections    PDF    Share

This study's results suggest trade-offs among screening questionnaires, but no questionnaire was found to be clearly superior. For example, the PEDS displayed some of the lowest diagnostic odds ratios, yet it had the highest sensitivity to severe delays. The sensitivity of the ASQ-3 fell below 70% for all delay levels, yet its positive predictive value was uniformly high. These findings suggest differences in scoring thresholds, which indicate trade-offs between sensitivity and specificity. Other characteristics (such as the feasibility and face validity of the PEDS, the detailed information on varied domains of development offered by the ASQ-3, and the parallel with the schedule of pediatric visits and comprehensive nature of the SWYC Milestones) may be equally important when choosing a screening instrument.

## Limitations

This study has several limitations. Sample sizes precluded analyses of smaller age groups specific to each screening form, and they yielded relatively large CIs for many estimates; therefore, point estimates were subject to significant sampling variation and should be interpreted with caution. Although the study was designed to generalize to primary care populations, families who reported black race and/or lower socioeconomic status were less likely to follow through on referrals for complete evaluations. This factor limited our ability to address outcomes for these populations. Moreover, the mean child age was slightly older than that recommended in standard AAP guidelines for screening. In addition, the results diverged from the findings in some previous studies. Whether this heterogeneity is best explained by the variations in reference tests or study populations, differences among studies highlight that sensitivity is not a property of a screening questionnaire but rather a description of how that screening instrument performs in a given context, for a given use, and with a given population. In the absence of consistent results across studies, stable psychometric properties of any particular questionnaire should not be assumed.

This study was also limited by the developmental tests that served as reference standards. Questions have been raised about inflated scores for the Bayley Scales of Infant and Toddler Development, Third Edition,[34] which may have affected our results. More generally, lack of perfect reliability among reference standards is known to depress estimates of sensitivity and specificity[35,36]; however, violations of conditional independence (eg, from residual effects of severity after accounting for delay status) can, in turn, inflate such estimates.[17] These factors add a degree of uncertainty to the findings.

## Conclusions

This study's results suggest that developmental screening instruments may offer valuable information to pediatric practitioners, although these findings do not lead to definitive recommendations. As has been argued previously, screening instruments are, at best, one element in a larger system of care.[23] We recommend that future research move beyond evaluating the accuracy of screening instruments to using

Sections          PDF          Share

# Article Information

**Corresponding Author:** R. Christopher Sheldrick, PhD, Department of Health Law, Policy & Management, Boston University School of Public Health, 715 Albany St, Boston, MA 02118 (rshldrck@bu.edu).

# References

1. Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. *Med Decis Making*. 1991;11(2):88-94. doi:10.1177/0272989X9101100203

   PubMed | Google Scholar | Crossref

2. US Preventive Services Task Force. 2015 Procedures manual. https://www.uspreventiveservicestaskforce.org/Page/Name/procedure-manual. Accessed March 30, 2019.

3. Canadian Task Force on Preventive Health Care. Procedure manual. https://canadiantaskforce.ca/wp-content/uploads/2016/12/procedural-manual-en_2014_Archived.pdf . Published March 2014. Accessed March 30, 2019.

4. Whiting PF, Rutjes AW, Westwood ME, et al; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Ann Intern Med*. 2011;155(8):529-536. doi:10.7326/0003-4819-155-8-201110180-00009

   PubMed | Google Scholar | Crossref

5. Arunyanart W, Fenick A, Ukritchon S, et al. Developmental and autism screening: a survey across six states. *Infants Young Child*. 2012;25(3):175-187. doi:10.1097/IYC.0b013e31825a5a42

   Google Scholar | Crossref

6. Radecki L, Sand-Loud N, O'Connor KG, Sharp S, Olson LM. Trends in the use of standardized tools for developmental screening in early childhood: 2002-2009. *Pediatrics*. 2011;128(1):14-19. doi:10.1542/peds.2010-2180

7. Lipkin PH, Macias MM; COUNCIL ON CHILDREN WITH DISABILITIES, SECTION ON DEVELOPMENTAL AND BEHAVIORAL PEDIATRICS. Promoting optimal development: identifying infants and young children with developmental disorders through developmental surveillance and screening. *Pediatrics*. 2020;145(1):e20193449. doi:10.1542/peds.2019-3449
PubMed | Google Scholar | Crossref

8. Council on Children With Disabilities; Section on Developmental Behavioral Pediatrics; Bright Futures Steering Committee; Medical Home Initiatives for Children With Special Needs Project Advisory Committee. Identifying infants and young children with developmental disorders in the medical home: an algorithm for developmental surveillance and screening. *Pediatrics*. 2006;118(1):405-420. doi:10.1542/peds.2006-1231
PubMed | Google Scholar | Crossref

9. Warren R, Kenny M, Fitzpatrick-Lewis D, et al. *Screening and Treatment for Developmental Delay in Early Childhood (Ages 1-4): Systematic Review*. Hamilton, Ontario: McMaster University; 2014.

10. Drotar D, Stancin T, Dworkin PH, Sices L, Wood S. Selecting developmental surveillance and screening tools. *Pediatr Rev*. 2008;29(10):e52-e58. doi:10.1542/pir.29-10-e52
PubMed | Google Scholar | Crossref

11. Drotar D., Stancin T, Dworkin P. Pediatric developmental screening: understanding and selecting screening instruments. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.605.692&rep=rep1&type=pdf.Published online February 26, 2008. Accessed March 30, 2019.

12. Sheldrick RC, Perrin EC. Evidence-based milestones for surveillance of cognitive, language, and motor development. *Acad Pediatr*. 2013;13(6):577-586. doi:10.1016/j.acap.2013.07.001
PubMed | Google Scholar | Crossref

13. Hagan JF, Shaw JS, Duncan PM. *Bright Futures: Guidelines for Health Supervision of Infants, Children, and Adolescents*. 4th ed. Itasca, IL: American Academy of Pediatrics; 2016.

14. Squires J, Twombly E, Bricker D, Potter L. *ASQ-3 Ages and Stages Questionnaires User's Guide*. 3rd ed. Lane County, OR: Brookes Publishing; 2009.

15. Glascoe FP. *Collaborating with Parents: Using Parents' Evaluation of Developmental Status to Detect and Address Developmental and Behavioral Problems*. Nolensville, TN: Ellsworth & Vandermeer Press; 1998.

16. San Antonio MC, Fenick AM, Shabanova V, Leventhal JM, Weitzman CC. Developmental screening using the Ages and Stages Questionnaire: standardized versus real-world conditions. *Infants*

17. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. New York, NY: Oxford University Press; 2003.

18. Leisenring W, Alonzo T, Pepe MS. Comparisons of predictive values of binary medical diagnostic tests for paired designs. *Biometrics*. 2000;56(2):345-351. doi:10.1111/j.0006-341X.2000.00345.x
PubMed | Google Scholar | Crossref

19. Grimes DA, Schulz KF. Refining clinical diagnosis with likelihood ratios. *Lancet*. 2005;365(9469):1500-1505. doi:10.1016/S0140-6736(05)66422-7
PubMed | Google Scholar | Crossref

20. Youngstrom EA, Choukas-Bradley S, Calhoun CD, Jensen-Doss A. Clinical guide to the evidence-based assessment approach to diagnosis and treatment. *Cognit Behav Pract*. 2015;22(1):20-35. doi:10.1016/j.cbpra.2013.12.005
Google Scholar | Crossref

21. Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol*. 2003;56(11):1129-1135. doi:10.1016/S0895-4356(03)00177-X
PubMed | Google Scholar | Crossref

22. Enders CK. *Applied Missing Data Analysis*. New York, NY: Guilford Press; 2010.

23. McIsaac M, Cook RJ. Statistical methods for incomplete data: some results on model misspecification. *Stat Methods Med Res*. 2017;26(1):248-267. doi:10.1177/0962280214544251
PubMed | Google Scholar | Crossref

24. US Census Bureau. State and county quickfacts. https://www.census.gov/quickfacts/fact/table/US/PST045219. Accessed July 15, 2018.

25. Sices L, Stancin T, Kirchner L, Bauchner H. PEDS and ASQ developmental screening tests may not identify the same children. *Pediatrics*. 2009;124(4):e640-e647. doi:10.1542/peds.2008-2628
PubMed | Google Scholar | Crossref

26. Sheldrick RC, Neger EN, Perrin EC. Concerns about development, behavior, and learning among parents seeking pediatric care. *J Dev Behav Pediatr*. 2012;33(2):156-160.
PubMed | Google Scholar | Crossref

27. Sheldrick RC, Garfinkel D. Is a positive developmental-behavioral screening score sufficient to justify referral? A review of evidence and theory. *Acad Pediatr*. 2017;17(5):464-470. doi:10.1016/j.acap.2017.01.016

2015;56(9):936-948. doi:10.1111/jcpp.12442

PubMed | Google Scholar | Crossref

29. Sheldrick RC, Merchant S, Perrin EC. Identification of developmental-behavioral problems in primary care: a systematic review. *Pediatrics*. 2011;128(2):356-363. doi:10.1542/peds.2010-3261
PubMed | Google Scholar | Crossref

30. Balogh EP, Miller BT, Ball JR, eds. *Improving Diagnosis in Health Care*. Washington, DC: National Academies Press; 2015. doi:10.17226/21794

31. Sheldrick RC, Frenette E, Vera JD, et al. What drives detection and diagnosis of autism spectrum disorder? looking under the hood of a multi-stage screening process in early intervention. *J Autism Dev Disord*. 2019;49(6):2304-2319. doi:10.1007/s10803-019-03913-5
PubMed | Google Scholar | Crossref

32. Coker TR, Chacon S, Elliott MN, et al. A parent coach model for well-child care among low-income children: a randomized controlled trial. *Pediatrics*. 2016;137(3):e20153013. doi:10.1542/peds.2015-3013
PubMed | Google Scholar | Crossref

33. Mimila NA, Chung PJ, Elliott MN, et al. Well-child care redesign: a mixed methods analysis of parent experiences in the PARENT trial. *Acad Pediatr*. 2017;17(7):747-754. doi:10.1016/j.acap.2017.02.004
PubMed | Google Scholar | Crossref

34. Aylward GP. Continuing issues with the Bayley-III: where to go from here. *J Dev Behav Pediatr*. 2013;34(9):697-701. doi:10.1097/DBP.0000000000000000
PubMed | Google Scholar | Crossref

35. Omurtag A, Fenton AA. Assessing diagnostic tests: how to correct for the combined effects of interpretation and reference standard. *PLoS One*. 2012;7(12):e52221. doi:10.1371/journal.pone.0052221
PubMed | Google Scholar | Crossref

36. Schmidt FL, Hunter JE. Measurement error in psychological research: lessons from 26 research scenarios. *Psychol Methods*. 1996;1(2):199-223. doi:10.1037/1082-989X.1.2.199
Google Scholar | Crossref