

Published in final edited form as:

*Acad Pediatr.* 2013 ; 13(6): 577–586. doi:10.1016/j.acap.2013.07.001.

## Evidence-Based Milestones for Surveillance of Cognitive, Language, and Motor Development

R. Christopher Sheldrick, PhD and Ellen C. Perrin, MD

Division of Developmental-Behavioral Pediatrics, Floating Hospital for Children, Tufts Medical Center, Boston, Mass

### Abstract

**Objective**—Fewer than half of the nation's pediatricians conduct systematic surveillance of young children's development. Because time and cost are among the barriers, our objective was to create a brief set of parent-report questions about cognitive, motor, and language milestones that is freely available and can be administered and scored quickly.

**Methods**—A team of experts developed candidate items after reviewing existing instruments and prior research. We selected final items based on statistical fit to a graded item response model developed and replicated in separate samples enrolled from primary care settings ( $n = 469$  and  $308$ , respectively). We then developed a 10-item form for each visit on the pediatric periodicity schedule. Combining our initial samples with 395 families enrolled from referral clinics, we tested these forms' concurrent validity with respect to the ASQ-3 and parent reports of developmental diagnoses.

**Results**—A final set of 54 items displayed adequate fit to our statistical model regardless of race/ethnicity, education level, and child gender. Beginning at 4 months, scores on 10-item forms displayed consistent associations with the ASQ-3, and all but the 60-month form detected parents' reports of developmental delay with adequate sensitivity and specificity.

**Conclusions**—The Milestones is one element of the Survey of Well-being of Young Children (SWYC), a brief but comprehensive screening instrument for children under 5 years. The Milestones is a set of evidence-based items with individual normative data that are appropriate for pediatric surveillance. In addition, the scoring of 10-item Milestones forms may provide many advantages of a first level developmental screening instrument.

### Keywords

behavioral; emotional; pediatrics; screening; social

---

Copyright © 2013 by Academic Pediatric Association

Address correspondence to R. Christopher Sheldrick, PhD, 800 Washington St, Box 854, Boston, MA 02111 (rsheldrick@tuftsmedicalcenter.org).

The authors declare that they have no conflict of interest.

**Supplementary Data:** Supplementary data related to this article can be found online at <http://dx.doi.org/10.1016/j.acap.2013.07.001>.  
References 18–26 are cited in the Appendix.

Surveillance and screening of young children's development is central to pediatric care. According to the Centers for Disease Control and Prevention (CDC), as many as 1 in 6 children have a developmental disorder, including 0.5% who have autism, 0.7% who have intellectual disabilities, 7.7% who have learning disabilities, and 3.7% who have developmental delays.<sup>1</sup> A separate study suggests that more than 12% of children under 3 years of age are likely to be eligible for the services provided under the Individuals With Disabilities Education Act (IDEA).<sup>2</sup>

To identify children with these disabilities, the American Academy of Pediatrics (AAP) recommends that every well-child visit include developmental surveillance. This process often includes documentation of parents' reports using a checklist of age-appropriate developmental milestones. Such checklists are recommended by Bright Futures and the CDC, but to our knowledge, no systematic research has evaluated the reliability and validity of each individual question, nor of their collective utility in detecting developmental delays and disorders.

To address the need for a checklist of developmental milestones, we developed a new set of validated questions for use in ongoing surveillance. Known as the Milestones, this instrument is part of a newly developed, comprehensive, and freely available screening tool for children under 5 years of age, the Survey of Well-being of Young Children (SWYC; see [www.theSWYC.org](http://www.theSWYC.org)). In addition to the Milestones, the SWYC includes 3 other components that assess social/emotional functioning,<sup>3,4</sup> autism,<sup>5</sup> and family risk factors.

Feasibility was a central goal in developing the Milestones checklists. Therefore, questions were designed for parents to report directly about their child's accomplishments, ultimately using a computer or a telephone. Questions are short and do not require specified materials or pictures; reading level is low. Only 10 items are recommended at each age, thus facilitating quick administration and scoring. Below we report on the development and the initial validation of the Milestones checklists.

## Methods

### Overview

On the basis of a review of the literature and existing screening instruments, we created 174 new questions about developmental milestones for children through 5 years of age. We collected data from parents about their child's attainment of these milestones. We developed a standard procedure to select items and set clinical thresholds, and we used this procedure to develop a 10-item form for each age on the pediatric periodicity schedule. The concurrent validity and accuracy of these forms were tested using a sample of families enrolled from primary care settings, referral clinics, and neonatal intensive care (NICU) follow-up clinics. All procedures were approved by the Institutional Review Board at Tufts Medical Center.

### Item Development

Our goal was to write items that could be easily and efficiently answered by parents from a range of educational and cultural backgrounds in the context of a pediatric waiting room. Thus we sought to write questions that were short, easy to read, salient to parents, and

appropriate for children through 5 years of age. We began by identifying constructs common across many existing sets of developmental tests and screening instruments; we created additional items where necessary, including indicators of fine motor, gross motor, cognitive, and language skills.

A panel of 11 experts including psychologists, primary care pediatricians, developmental-behavioral pediatricians, occupational therapists, and 8 parents of young children reviewed this initial list of items, providing feedback on clarity, reading level, and relevance. In total, this process resulted in a list of 174 new questions. Questions were screened for Flesch-Kincaid reading level and rewritten if reading levels were over grade 6. The average reading level for the final items was grade 2.7. For each item, response options were “not yet,” “somewhat,” and “very much.”

### Study Samples

Participants were parents of children from birth through 5.5 years recruited from 7 urban practices and community health centers, 7 suburban practice groups, 2 developmental-behavioral assessment clinics, 2 NICU follow-up clinics, 2 child psychiatry clinics, 2 occupational therapy clinics, and 1 speech and language clinic.

We enrolled 4 separate samples: for scale construction and initial validation, 469 families from pediatric primary care practice groups (the primary care sample); 223 families from referral clinics (the referral clinic sample), and 172 families from NICU follow-up clinics (the NICU sample). For replication of item-level parameters, we enrolled 308 families from pediatric primary care practice groups (the replication sample).

### Procedures

Parents were enrolled in 1 of 2 ways. In settings with high patient volumes (including all primary care and some referral clinics), research assistants approached parents in waiting rooms, described the study, and asked them if they would be interested in participating. In clinics with lower patient volumes (including NICU follow-up, developmental assessment, and psychiatric clinics), eligible parents were identified from clinic records. Physicians mailed letters to these parents describing the study and stating that unless they called a dedicated opt-out voice mail number, a research assistant would contact them.

The enrollment process is depicted in Figure 1. Among families identified as eligible in waiting rooms, 87% enrolled, and complete data were obtained from 78% of those enrolled (or 68% of eligible parents). Among families identified from medical records, 59% enrolled in the study, and complete data were obtained from 71% of those enrolled (or 42% of eligible parents).

### Assessments

Parents who enrolled were given a packet of questionnaires to complete in the office or mail back. Packets included age-appropriate questions about cognitive, language, and motor milestones (approximately 120 questions in samples enrolled for scale construction and validation, and selected questions in the replication sample, as described below), the ASQ-3,

demographic information, and questions about family risk factors and developmental problems. Questions about developmental problems began, “To the best of your knowledge, does your child have...” and included yes/no responses for “developmental delay” and “autism or a pervasive developmental disorder.” The ASQ-3 is a well-validated parent-administered instrument<sup>6,7</sup> that assesses children's development in fine motor, gross motor, communication, problem-solving, and personal-social domains. Following published guidelines, a child is considered at risk if his or her score in any domain falls below a threshold set at 2 standard deviations below the mean of the standardization sample. The ASQ-3 assesses children up to 66 months of age through the use of 21 age-specific forms, each with 30 primary multiple choice questions and 6 to 10 open-ended questions.

### Statistical Analysis

Four sets of analyses were conducted using Stata version 12<sup>8</sup> and Mplus version 6.11<sup>9</sup> to support selection of final questions, creation and scoring of age-specific forms, and assessment of concurrent validity and accuracy. Age was calculated on the basis of expected delivery date if the child was under 2 years of age and was born 3 or more weeks before the expected date of delivery.

Following previous assessments of developmental landmarks, our analyses centered on the use of a graded item response theory (IRT) model. In descriptive terms, the graded IRT model assumes that the probability of attaining a “not yet,” “somewhat,” or “very much” response to each milestone depends on the child's developmental age, which is not directly observable (ie, is latent) but will be inferred (ie, predicted) from the model. Item characteristic curves (ICCs) describe the probability of each response to each milestone as a function of developmental age. Example ICCs for a single Milestones question are depicted in Figure 2. When children are at younger developmental ages, parents are very likely to respond “not at all” to this question, but this probability drops as children develop. Conversely, parents are unlikely to respond “very much” when their children are at younger developmental ages, but this probability rises with development. In contrast, the probability of “somewhat” responses begins low, rises to a peak, and then descends at older developmental ages. Further description of ICC parameters is included in the Appendix, as are details of how we estimated our model in Mplus and checked model assumptions (Appendices 1 and 2).

### Selection of Final Questions

We used 3 criteria to choose final items from our 174 candidate items: response rate, item fit, and differential item functioning. For response rates, we eliminated items with 1% of missing data in our primary care sample. For item fit, we analyzed how well the graded IRT model for each item fit our data using Hosmer-Lemeshow tests for multinomial data<sup>10</sup> based on estimates of developmental age calculated in Mplus. Lack of fit does not imply that the construct underlying a given milestone is unimportant, only that the question used to assess it is unreliable as administered and modeled (Appendix 3). For differential item functioning (DIF), an item displays DIF if responses differ between 2 groups after controlling for underlying traits. For example, if children have identical levels of gross motor development but parents from 2 populations (eg, varying by race/ethnicity or socioeconomic status) differ

in their likelihood of endorsing an item such as “my child can kick a ball,” then that item can be said to display DIF. We used logistic regression techniques to test for DIF based on parent education (high school or less vs more than high school), race/ethnicity (white/non-Hispanic vs not), and child gender. Further detail regarding DIF tests is provided in Appendix 4.

To analyze item fit on a normative sample, only participants enrolled from primary care sites who had full-term births (≥ 37 weeks) were included. For an initial pass, all 174 items were tested for response rate and item fit in the primary care sample. Items with adequate fit were administered to and retested in the replication sample and, for additional statistical power, in the combined primary care and replication sample. Items that displayed poor fit or DIF were excluded, and a latent variable model was run with the remaining items to reestimate developmental age for each participant.

### Creation and Scoring of Age-Specific Forms

We developed and applied a standard procedure to create and score 10-item forms for each visit on the pediatric periodicity schedule through age 5 years (ie, 2, 4, 6, 9, 12, 15, 18, 24, 30, 36, 48, and 60 months). Each form included items that varied in level of difficulty. Because a primary goal of surveillance is to detect children with developmental delays, we chose items beginning at 80% of the lower limit of the age range of the form. For example, the least difficult item on the 18 month form has a median age of achievement (ie, scoring “very much”) equivalent to 14.9 months, which is approximately 80% of 18 months.

For use in clinical settings, we developed 2 scoring approaches. The first scoring approach requires a computer, while the second approach is amenable to hand scoring. Both approaches offer estimates of developmental status, which is an estimate of delay based on the ratio of developmental age and chronological age. Calculating the natural logarithm of this ratio provides a continuous estimate of developmental status centered at zero, with negative scores therefore indicating delay.

$$\text{developmental status} = \ln \left( \frac{\text{developmental age}}{\text{chronological age}} \right)$$

We set the clinical threshold for developmental status at  $-0.1625$ , which is equivalent to 15% delay ( $\ln(0.85) = -0.1625$ ). Because states that use percentage delay to determine eligibility for IDEA services use thresholds ranging from 20% to 50% delay,<sup>11</sup> this threshold will enhance sensitivity, as is appropriate for a brief screening instrument.

Both scoring approaches yield continuous scores, to which clinical thresholds can be applied to determine positive/negative screening status. The computer-based scoring approach is based on maximum likelihood estimation and estimates developmental status directly based on the child's age, parents' responses, and ICC parameters. The hand-scoring approach requires summing responses and consulting a growth chart to determine whether the clinical threshold is exceeded. Both scoring approaches use cutoff scores based on a 15% delay.

Further detail regarding computer-based and hand-scoring approaches is provided in Appendices 5 and 6.

### Assessment of Concurrent Validity and Accuracy

To assess concurrent validity, we calculated point-biserial correlations between binary (positive/negative) scores on the ASQ-3 and the continuous Milestones scores, including computer-based and hand-scoring methods. Because standard methods for calculating standard errors associated with point-biserial correlations are often biased,<sup>12</sup> we estimated 95% confidence intervals using estimates across 250 bootstrap samples with replacement.

To assess accuracy in detecting developmental delays and disorders, we analyzed the ability of 2 different screening instruments to predict parents' reports of diagnoses of developmental disorders, including autism and developmental delay. First, we applied clinical thresholds to computer-based and hand-scored Milestones scores and computed area under the ROC curve (AUC), sensitivity, and specificity with respect to parents' reports of diagnoses of developmental disorders. Second, we calculated the same statistics for the ASQ-3 with respect to the same criterion. AUC was included as a summary index of accuracy that includes both sensitivity and specificity. For all analyses of concurrent validity and accuracy, we pooled data from across all 4 samples (including primary care, replication, NICU, and referral clinic samples) to maximize statistical power.

## Results

Sample characteristics are displayed in Table 1. Across all 4 samples, the majority of respondents were mothers (82% to 91%). Demographics were similar across the primary care, referral clinic, and NICU follow-up samples. Most respondents had completed college (53% to 56% across samples), but family incomes varied, with 30% to 35% across samples making less than \$50,000 per year. These samples were also diverse with respect to race and ethnicity, with 29% to 33% reporting minority race or Hispanic ethnicity (for comparison, 23.9% of Massachusetts residents and 36.3% of US residents reported minority race or Hispanic ethnicity in the 2010 census). In contrast, the replication sample included a larger proportion of families of lower income and education. In this sample, only about a third of respondents had completed college, 50% of families reported income under 50,000 per year, and 51% reported minority race and/or Hispanic ethnicity.

### Selection of Final Questions

Of the 174 items tested in the original primary care sample, 96 displayed adequate fit and were administered to the replication sample. Of these, 54 items displayed adequate fit and lack of evidence of DIF. Table 2 presents model parameters and expected age in months at which 25%, 50%, and 75% of parents report "very much" for these milestones based on the full sample of participants enrolled from primary care sites who had full-term births (37 weeks). The 42 items eliminated in the final stage of this process were eliminated for a variety of reasons. Some may represent type 1 errors, while others may reflect constructs that can be reliably assessed using different administration methods. Thus, these items are presented in Appendix 7, as they may be useful for the development of future assessments.

## Creation and Scoring of Age-Specific Forms

Table 2 displays assignment of items to forms. Twelve age-specific forms were created using the method described above, with one exception. Because of the comparatively small number of items appropriate for older children, the least difficult items for the 60-month form begin at age 34 months rather than 48. Thus, the items on the 60-month form are much easier overall than are the items on other forms. Developmental status and residual total score were calculated as described above, and a clinical threshold equivalent to a 15% delay was applied to yield a binary screening result. In our study, 21.8% (95% confidence interval, 18.4 to 25.6) of children enrolled from primary care settings scored positive on the Milestones and 19.0% (95% confidence interval, 15.6 to 22.4) scored positive on the ASQ-3. In comparison, a recent study found that 15.8% of a community sample of children scored positive on the ASQ-3,<sup>13</sup> suggesting that our sample may have higher rates of developmental problems.

## Assessment of Concurrent Validity and Accuracy

With the exception of the 2-month form, point-biserial correlations between the ASQ-3 and continuous Milestones scores based on computer-based scoring and hand scoring were moderate to large for all forms (ranging from 0.40 to 0.70, and all 95% confidence intervals >0.25) (Figure 3). Results for hand scoring closely approximated results for computer-based scoring.

Figure 4 displays AUC, sensitivity, and specificity for the Milestones (both computer-based and hand-scoring methods) and the ASQ-3. Analyses of 2-month and 4-month forms were omitted because reports of developmental delays were rare at these ages (0 and 1 reports, respectively). Confidence intervals for sensitivity are large through 18 months because of comparatively low numbers of children with delays (4 to 9 per age group). AUC fell above 70% for all Milestones forms except at 9 and 60 months, and for all ASQ-3 forms except at 30 months. Results for sensitivity followed a similar pattern, except ASQ-3 forms also fell below 70% at 18 months. Results for specificity all fell above 70%, with the exception of the 60-month Milestones form and the 9-month ASQ-3 form.

## Discussion

The Milestones is a set of 54 parent-report questions focusing on children's motor, cognitive, and language development through 5 years of age. All SWYC Milestones forms are available with scoring instructions online ([www.theSWYC.org](http://www.theSWYC.org)). Individual items demonstrated expected relationships to age in 2 primary care samples, and their functioning was not strongly influenced by race/ethnicity, parent education, or child gender. There are 2 potential uses of these validated 10-item forms: surveillance and first-level screening. The individual items included on each of the 12 forms are valid for ongoing developmental surveillance.

Our results offer initial support for the validity of 9 of the 12 Milestones forms as first-level screening instruments. For the 3 remaining forms, we suspect that different solutions are needed. The 60-month form would benefit from the addition of more difficult items



appropriate for older children, assuming that such items met similar standards of reliability as reported here. We hypothesize that the 2-month form requires greater precision in the measurement of age associated with each milestone, because for newborns, differences in developmental status are apparent over days rather than months. We therefore recommend caution in interpreting results of the 2- and 60-month forms. In contrast, we have no reason to believe that the validity of the 9-month form is any different than that of other Milestones forms. The items met the same standard of reliability, the confidence intervals include acceptable values for sensitivity and specificity, and multiple testing often results in isolated negative findings. Notably, despite the strong record of research supporting its validity, the 30-month form of the ASQ-3 also demonstrated lower sensitivity than expected in this study. We strongly recommend (and plan to conduct) further study of all Milestones forms in a variety of different geographic, cultural, and socioeconomic populations.

We note several limitations shared by our studies of other components of the SWYC.<sup>3-5</sup> We enrolled a convenience sample of English-speaking parents who brought their children to pediatricians' offices or to developmental specialists in the greater Boston area. Our enrollment sites included both primary and secondary care sites with varied practice models, and clienteles with varied socioeconomic status and ethnic backgrounds. Our findings generalize best to similar practice arrangements. In addition, we were not able to enroll all parents who sought pediatric care for their children. In-person recruitment from primary care sites was more successful than our reliance on mailed and telephone contact with parents of children seen in specialty settings. Similar to screening rates in previous studies,<sup>14,15</sup> 68% of eligible parents identified from waiting rooms enrolled and completed study materials. Among families identified through medical records, 42% enrolled and completed study materials. In both cases, reasons for nonresponse could not be determined, and it is possible that factors such as parents' difficulty with literacy and/or English played a role. In addition, limited resources precluded conducting structured diagnostic interviews, and we therefore utilized a validated screening instrument and parents' reports of existing diagnoses as our criterion variables. Further research on the Milestones and the other components of the SWYC is in process that includes structured clinical assessments as the criterion.

Other limitations are unique to the Milestones checklists. For example, to ensure that all items are reliable, we adopted stringent selection criteria that included multiple tests of item fit and differential item functioning. This procedure may have led us to exclude some items that assess important constructs but do not appear to be reliably assessed by parent report (eg, joint attention). In addition, although overall estimates of developmental status from 10-item Milestones forms show promise for identifying children who may benefit from further evaluation, we advise cautious interpretation<sup>16</sup> because a single summary estimate of developmental age may obscure differences in the sequence and rate of development of fine motor, gross motor, language and cognitive skills. Because questions that tap different dimensions are found across ages on the 12 Milestones forms, repeated administrations for each child will offer the opportunity to further investigate the dimensionality of the Milestones checklists, both in longitudinal research and in clinical practice.

The AAP recommends surveillance of children's development at every well-child visit as well as periodic formal developmental screening.<sup>17</sup> The SWYC Milestones offers a set of



evidence-based items with individual normative data that can be administered repeatedly as children grow. Especially if it is used as part of the more comprehensive SWYC, it can be thought of as a surveillance tool. In addition, by offering a scoring procedure for 10-item forms, the Milestones provides many of the advantages of a first level developmental screening instrument. We recommend that a score above the threshold should be followed up by further investigation—by interviewing a parent, observing the child, and/or administering a more detailed validated developmental screening instrument like the ASQ-3. We present initial results supporting reliability and validity, and further research is clearly needed, including replication of initial findings, collection of normative data from larger representative samples, and study of implementation. Extensions amenable to study include implementation in electronic health records and use in settings beyond primary care pediatrics, such as early childhood education and home-visiting programs.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

Research support for the development and validation of the Milestones was provided by The Commonwealth Fund and NIH grant KM1CA156726. We are indebted to The Commonwealth Fund for its support of the larger project of which this is a part. In particular, Dr Ed Schor's vision and imagination motivated us to create a brief public domain instrument to guide surveillance in pediatric primary care settings. Dr Melanie Wall offered expert advice regarding latent variable modeling. We are grateful also for the assistance of Dr Robin Adair, Dr Liz McGowan, and a large number of pediatricians and other child health specialists who allowed us to enroll parents through their offices; to Emily Neger, Shela Merchant, Kate Mattern, Jennifer LaMotte, and Robyn Della Giustina for help with data collection and management; and to the Tufts Medical Center CTSI for ongoing support.

Primary care sites included Harvard Vanguard Medical Associates, Medical Associates Pediatrics, MGH Revere Health Care Center, Pediatric Healthcare Associates, Porter Pediatrics, Southborough Medical Group, Floating Hospital General Pediatrics, Wilmington Pediatrics, Quality Kids Care, Codman Square Health Center, Westwood-Mansfield Pediatric Associates, and Dr Babu Pediatrics PC.

Referral clinics included Floating Hospital Center specialty clinics, including the Center for Children With Special Needs, NICU Follow-up, Child Psychiatry, and International Adoption; OTA Watertown, South Shore Therapies, MGH Child Psychiatry, UMass NICU Follow-up and DBP clinics; Boston Medical Center's Developmental Assessment Clinic; and Harvard Vanguard Speech/Language Pathology.

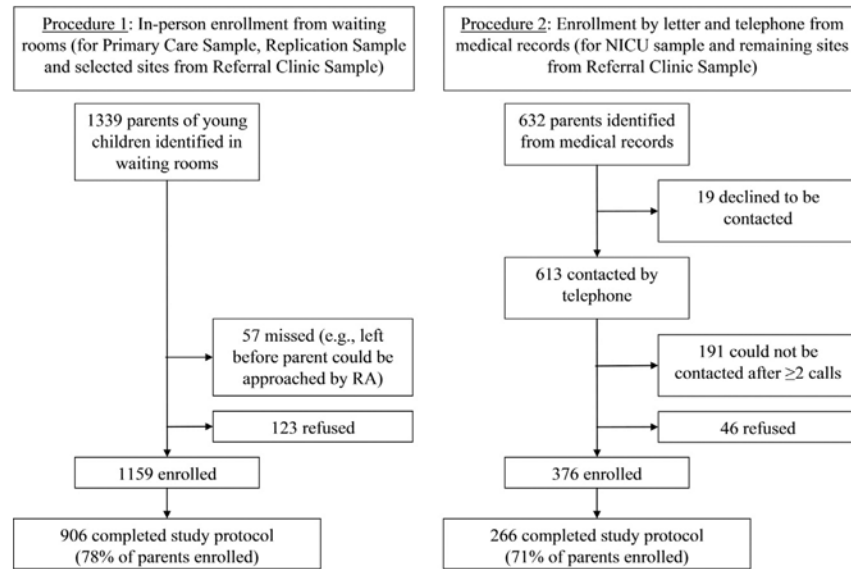
## References

1. Boyle CA, Boulet S, Schieve LA, et al. Trends in the prevalence of developmental disabilities in US children, 1997–2008. *Pediatrics*. 2011; 127:1034–1042. [PubMed: 21606152]
2. Rosenberg S, Zhang D, Robinson C. Prevalence of developmental delays and participation in early intervention services for young children. *Pediatrics*. 2008; 121:e1503–e1509. [PubMed: 18504295]
3. Sheldrick RC, Henson BS, Neger EN, et al. The Baby Pediatric Symptom Checklist (BPSC): development and initial validation of a new social–emotional screening instrument. *Acad Pediatr*. 2013; 13:72–78. [PubMed: 23092547]
4. Sheldrick RC, Henson BS, Merchant S, et al. The Preschool Pediatric Symptom Checklist (PPSC): development and initial validation of a new social–emotional screening instrument. *Acad Pediatr*. 2012; 12:456–467. [PubMed: 22921494]
5. Smith N, Sheldrick RC, Perrin EC. An abbreviated screening instrument for autism spectrum disorders. *Infant Mental Health*. 2012; 34:149–155.
6. Limbos MM, Joyce DP. Comparison of the ASQ and PEDS in screening for developmental delay in primary care settings. *J Dev Behav Pediatr*. 2011; 32:499–511. [PubMed: 21760526]

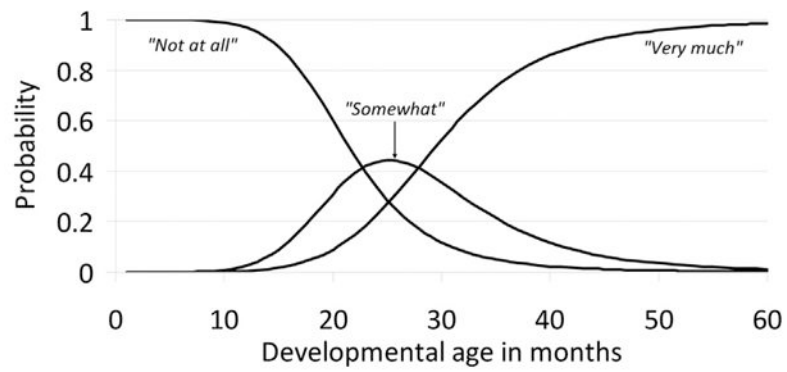
7. Squires, J.; Twombly, E.; Bricker, D., et al. ASQ-3 Ages and Stages Questionnaires User's Guide. 3rd. Lane County, Ore: Brookes Publishing; 2009.
8. StataCorp. College Station. Tex: StataCorp LP; 2011. Stata Statistical Software: Release 12.
9. Muthén, LK.; Muthén, BO. Mplus, Version 6.11. Los Angeles, Calif: Muthén & Muthén; 2011.
10. Fagerland MW, Hosmer DW. A generalized Hosmer-Lemeshow goodness-of-fit test for multinomial logistic regression models. *Stata J.* 2006; 12:447–453.
11. Ringwalt, S. [Accessed January 4, 2013] Summary table of states' and territories' definitions of criteria for IDEA Part C eligibility. 2012. compAvailable at: [http://www.nectac.org/~pdfs/topics/earlyid/partc\\_elig\\_table.pdf](http://www.nectac.org/~pdfs/topics/earlyid/partc_elig_table.pdf)
12. Harris DJ, Kolen MJ. Bootstrap and traditional standard errors of the point-biserial. *Educ Psychol Meas.* 1988; 48:43–51.
13. Thomas SA, Cotton W, Pan X, et al. Comparison of systematic developmental surveillance with standardized developmental screening in primary care. *Clin Pediatr.* 2012; 51:154–159.
14. Hix-Small H, Marks K, Squires J, et al. Impact of implementing developmental screening at 12 and 24 months in a pediatric practice. *Pediatrics.* 2007; 120:381–389. [PubMed: 17671065]
15. Schonwald A, Huntington N, Chan E, et al. Routine developmental screening implemented in urban primary care settings: more evidence of feasibility and effectiveness. *Pediatrics.* 2009; 123:660–668. [PubMed: 19171635]
16. Andersson L. Appropriate and inappropriate interpretation and use of test scores in early intervention. *J Early Interv.* 2004; 27:55–68.
17. 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; 118:405–420. [PubMed: 16818591]
18. Rutter M. Age as an ambiguous variable in developmental research: some epidemiological considerations from developmental psychopathology. *Int J Behav Dev.* 1989; 12:1–34.
19. Drachler ML, Marshall T, Leite JC. A continuous-scale measure of child development for population-based epidemiological surveys: a preliminary study using Item Response Theory for the Denver Test. *Paediatr Perinat Epidemiol.* 2007; 21:138–153. [PubMed: 17302643]
20. Embretson, SE.; Reise, SP. Item Response Theory for Psychologists. New York, NY: Psychology Press; 2000.
21. Mair, P.; Reise, SP.; Bentler, PM. IRT Goodness-of-Fit Using Approaches From Logistic Regression. Los Angeles, Calif: Department of Statistics, UCLA; 2008. Available at: <http://escholarship.org/uc/item/1m46j62q> [Accessed December 1, 2012]
22. French AW, Miller TR. Logistic regression and its use in detecting differential item functioning in polytomous items. *J Educ Measure.* 1996; 33:315–332.
23. Kim SH, Cohen AS, Alagoz C, et al. DIF detection and effect size measures for polytomously scored items. *J Educ Measure.* 2007; 44:93–116.
24. Camilli G, Congdon P. Application of a method of estimating DIF for polytomous test items. *J Educ Behav Stat.* 1999; 24:323–341.
25. Crane PK, Gibbons LE, Jolley L, et al. Differential item functioning analysis with ordinal logistic regression techniques. *Med Care.* 2006; 44(11 suppl 3):S115–S123. [PubMed: 17060818]
26. Cheung YB, Gladstone M, Maleta K, et al. Comparison of four statistical approaches to score child development: a study of Malawian children. *Trop Med Int Health.* 2008; 13:987–993. [PubMed: 18554248]

### What's New

The Milestones, a component of the Survey of Well-being of Young Children (SWYC), is a brief surveillance instrument of cognitive, motor, and language development. It is free, designed for use in primary care, easy to read and score, and may be incorporated into electronic medical record templates.

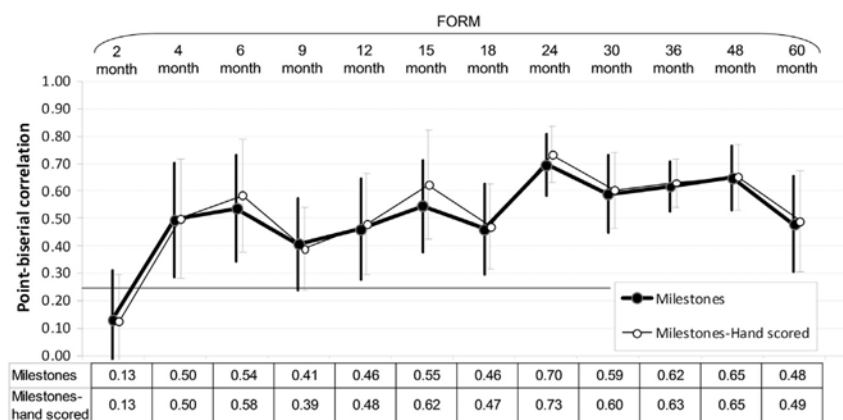


**Figure 1.**  
Enrollment.



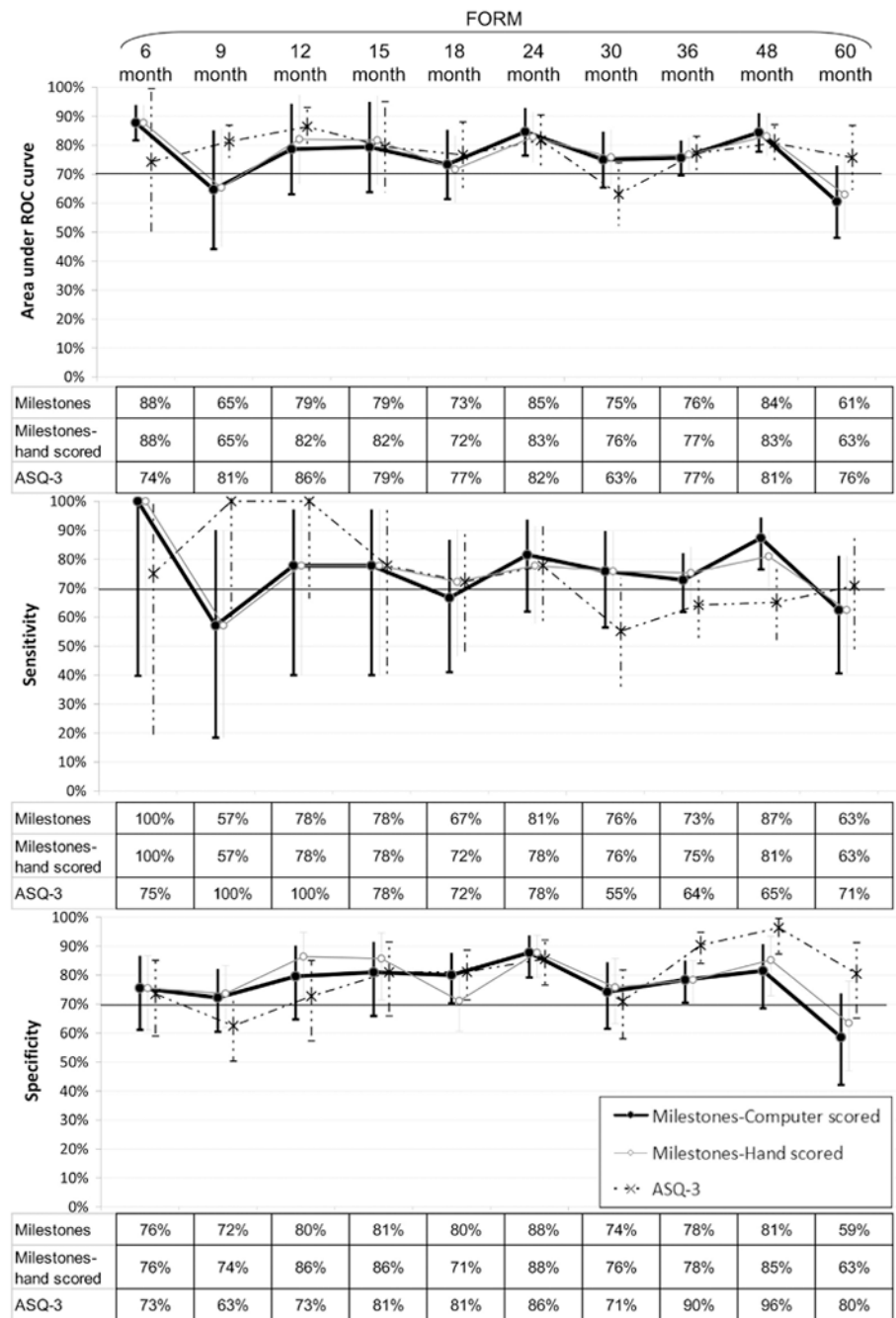
**Figure 2.**

Item characteristics curves for the question, "Does your child name at least one color?"



**Figure 3.**  
Point biserial correlations with ASQ-3.





**Figure 4.**

Accuracy of Milestones electronic scoring, Milestones hand scoring, and ASQ-3 in detecting parents' reports of developmental delays and disorders.





























## Sample Characteristics

*Acad Pediatr.* Author manuscript; available in PMC 2014 November 01.

Characteristic	Sample							
	PC (n = 469)		Referral Clinic (n = 223)		NICU Follow-up (n = 172)		Replication PC (n = 308)	
	n	%	n	%	n	%	n	%
Family/parent characteristics								
Public health insurance	91	19	53	24	50	29	96	31
Mother completed forms	397	85	195	87	157	91	253	82
Parent education								
Less than high school	21	4	7	3	11	6	21	7
High school diploma	112	24	49	22	45	26	111	36
Some college	76	16	42	19	21	12	60	19
College diploma	153	33	71	32	52	30	63	20
Advanced degree	102	22	53	24	39	23	43	14
Not indicated	5	1	1	0	4	2	10	3
Family income								
<\$20,000	81	17	35	16	35	20	99	32
\$20,000–49,999	65	14	32	14	26	15	55	18
\$50,000–99,999	129	28	73	33	57	33	61	20
\$100,000	176	38	73	33	49	28	78	25
Not indicated	18	4	10	4	5	3	15	5

PC = primary care.

**Table 2**  
**Final Item Parameters, Ages of Attainment, and Assignment to Forms**

Milestones item	ICC parameters [ $\beta$ , $-\alpha_1$ , $-\alpha_2$ ]	Age when percent passing = 25%, 50%, 75%		Assigned to form:
		0 months	12 months	
1. Makes sounds that let you know he or she is happy or upset	2.321,-0.89,0.442	 [0.8,1.2,1.9]		2-month form
2. Seems happy to see you	5.995,-0.299,1.478	 [1.1,1.3,1.5]		
3. Follows a moving toy with his or her eyes	3.982,-2,1.444	 [1.1,1.4,1.9]		
4. Turns head to find the person who is talking	3.324,-1.32,2.033	 [1.3,1.8,2.6]		
5. Holds head steady when being pulled up to a sitting position	2.874,-0.371,2.454	 [1.6,2.3,3.4]		4-month form
6. Brings hands together	2.284,0.099,2.081	 [1.5,2.5,4]		
7. Laughs	3.723,1.121,3.478	 [1.9,2.5,3.4]		
8. Keeps head steady when held in a sitting position	2.936,0.107,3.027	 [1.9,2.8,4.1]		
9. Makes sounds like "ga," "ma," or "ba"	2.794,2.144,3.677	 [2.5,3.7,5.5]		6-month form
10. Looks when you call his or her name	2.862,1.634,4.257	 [3.4,4.6,5]		
11. Rolls over	4.238,4.265,6.305	 [3.4,4.4,5.7]		
12. Passes a toy from one hand to the other	5.406,6.093,8.124	 [3.7,4.5,5.5]		
13. Looks for you or another caregiver when upset	2.16,1.79,3.266	 [2.7,4.5,7.5]		9-month form
14. Holds two objects and bangs them together	6.193,8.978,10.719	 [4.7,5.6,6.7]		
15. Holds up arms to be picked up	5.783,8.378,10.976	 [5.5,6.7,8.1]		
16. Gets to a sitting position by him or herself	8.923,16.613,17.779	 [6.5,7.3,8.3]		
17. Picks up food and eats it	6.304,11.001,12.574	 [6.2,7.3,8.7]		12-month form
18. Pulls up to standing	6.905,12.476,13.83	 [6.3,7.4,8.7]		
19. Plays games like "peek-a-boo" or "pat-a-cake"	5.933,10.209,12.45	 [6.8,8.2,9.8]		
20. Calls you "mama" or "dada" or similar name	3.755,6.896,8.368	 [6.9,9.3,12.4]		
21. Looks around when you say things like "Where's your bottle?" or "Where's your blanket?"	3.153,5.48,7.258	 [7.1,10,14.2]		15-month form
22. Copies sounds that you make	2.733,3.8,6.557	 [7.4,11,16.5]		
23. Walks across a room without help	9.663,23.319,24.067	 [10.8,12.1,13.5]		
24. Follows directions - like "Come here" or "Give me the ball"	4.202,8.29,10.842	 [10.2,13.2,17.1]		
25. Runs	10.163,26.071,27.468	 [13.4,14.9,16.6]		18-month form
26. Walks up stairs with help	6.657,16.557,18.256	 [13.2,15.5,18.3]		
27. Kicks a ball	6.352,15.927,17.642	 [13.5,16.1,19.1]		
<b>Note.</b> 12 forms, one for each recommended pediatric visit through 5 years, each has 10 items, with some items common to multiple forms. Parameters define an <i>item characteristic curve</i> (ICC) for each item that indicates the age in months when 25%, 50%, and 75% of children are expected to pass with response="very much"		<b>Percent of children passing (response="very much")</b> 25% 50% 75%  <b>Age in months at 25<sup>th</sup>, 50<sup>th</sup>, &amp; 75<sup>th</sup> %tile</b> [6.3, 7.4, 8.7]		

