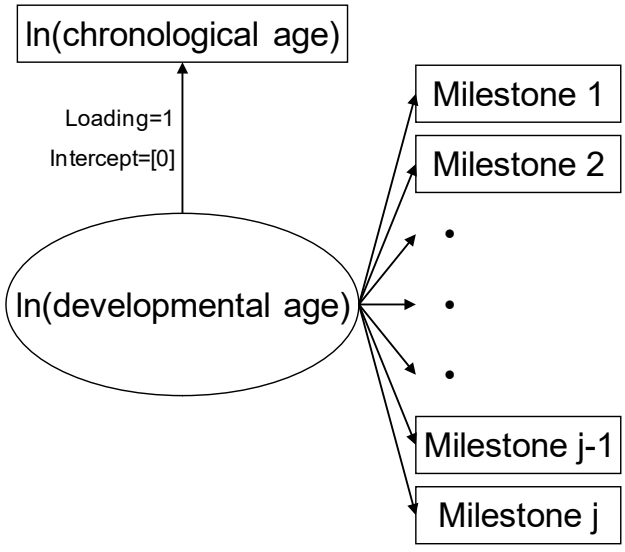
Appendix

A. Item Characteristic Curve (ICC) parameters: ICC equations define probability curves for each response to each milestone. Parameters determine the characteristics of these curves. For example, different milestones are attained at different developmental ages. Therefore, the equation for each milestone includes two parameters (α_1 and α_2) that define the median developmental age at which each response occurs. An additional discrimination parameter (β) describes the slope (or the rate at which the probability of achievement increases with developmental age) of the ICC. A milestone that most children master within a short age-range will have high discrimination, thus offering more information that is relevant over a short time period. As an example, infants typically develop the ability to sit up on their own well before their first birthday. Within a short time-period, mastery of this skill offers information about a child's development, but the utility of knowing that a child has passed diminishes quickly as the vast majority of children achieve this milestone. In contrast, a milestone achieved by children over a wider age-range will have lower discrimination, offering less information but over a longer time period. For example, children begin drawing with a crayon over a comparatively large age range. Failure to pass this milestone may become increasingly concerning with age, but evaluation of this milestone in isolation says little about a child's overall development.

B. Estimation in Mplus: Figure 5 depicts our model of developmental milestones as estimated in Mplus. The latent variable representing developmental age is modeled as the driver of each milestone question and the natural logarithm of c*hronological age*, which was corrected for prematurity (for children ages 2 years and under, age was calculated using expected date of delivery if child was born 3 or more weeks premature). Although it may seem counterintuitive to model developmental age as a driver of chronological age, it is worth noting that chronological age has

Figure 5. Latent variable model



been conceptualized as an ambiguous variable that serves as an indicator for various developmental processes, both psychological and biological.[18] Following this logic, we conceptualized chronological age as an indicator (not a cause or a result) of developmental age. Following past analyses,[19] ln(age) was used to improve model fit. The loading on c*hronological age* was constrained to 1 with an intercept of 0. Thus, the latent variable shares the same units and scale as the natural logarithm of chronological age and can be conceptualized as the natural logarithm of developmental age. For each milestone, this model yields an estimate of slope (β_j) as well as two threshold estimates (α_j), the first reflecting the threshold between "not yet" and "somewhat" responses, and the second reflecting the threshold between "somewhat" and "very much" responses. The model was run with maximum likelihood estimation with robust standard errors. Maximum likelihood uses full information available for each item to estimate model parameters, thus accounting for missing data resulting from the fact that each parent was asked to complete a subset of Milestones questions, as appropriate for the child's age.

We evaluated two key model assumptions. First, Rasch IRT models hold discrimination parameters constant across items. To evaluate the importance of variable discrimination parameters, we re-estimated our model in Mplus using the same data but constraining each β to 1. As hypothesized, a chi-square difference test revealed that the proposed model displayed a much better fit to the data than the Rasch model ($p < .001$).

Second, we evaluated whether data were sufficiently unidimensional for IRT modeling. As an initial test, we conducted an exploratory factor analysis of our final items in Mplus. Eigenvalues of this matrix suggest that a single factor explains most of the variance (Eigenvalue = 35.00), while the second factor explains far less (Eigenvalue = 3.94). Because eigenvalues offer evidence of dimensionality but are not a definitive test,[20] we conducted a further analysis. Specifically, we developed a 2-factor model based on a priori distinctions between motor and non-motor items. We then constrained the correlation between the two factors to one in order to estimate a 1-factor model. A difference test demonstrated that the 2-factor model displayed better fit to the data than did the 1-factor model ($p<.001$). Results of these tests are consistent with the hypothesis that although data are sufficiently unidimensional to support IRT analyses, additional

dimensions of developmental functioning may be evident. Cautions regarding interpretation of developmental age are included in the discussion.

C. Tests of item fit: Our model assumes that each ICC follows a parametric curve between younger ages when the probability of attainment is close to 0 and older ages when the probability of attainment is close to 1. For the model to be estimated appropriately, each item should display acceptable fit to this model. Many causes for lack of fit are possible. For example, unclear wording may introduce random error. If a significant proportion of parents fail to identify attainment of a given milestone (e.g., accurate observation of joint attention often required training), the observed probability will never reach 100% as predicted by the logistic regression curve. Conversely, if a significant proportion of parents report the attainment of a milestone in the absence of mastery (e.g., interpreting non-social smiles as social), the observed probability will begin above 0%. None of these possible causes implies that the construct underlying a given milestone is unimportant, only that the question used to assess it is unreliable as administered

To test item fit using Hosmer-Lemeshow tests for multinomial data,[10] participants' *developmental ages* were estimated in Mplus using the latent variable model described above. The Hosmer-Lemeshow test is conceptually similar to a standard method for testing item fit in IRT,[20] and tests of goodness-of-fit for logistic regression have been used previously to evaluate item fit in IRT models.[21] The test works as follows. Each participant's expected response to a given milestone is calculated as a function of *developmental age* using the model described above. Participants are sorted into 10 groups according to *developmental age*. For each group, the average observed response and the average expected response is calculated, yielding a 10 x 2 table. Ideally, expected proportions match observed proportions. A Pearson chi-square statistic tests lack of fit between observed and expected values. Note that 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.

D. Tests of differential item functioning (DIF): Logistic regression techniques have been recommended for testing DIF,[22] and they yield results that are in general agreement with other IRT modeling techniques.[23, 24] In particular, we followed a method previously used to evaluate

DIF in the Mini Mental State Exam.[25] Specifically, we calculated three ordinal logistic regression models to predict attainment of each milestone that included as independent variables: (1) *developmental age* as estimated from the latent variable model; (2) *developmental age* plus a main effect for the confounding variable, and (3) *developmental age*, the confounding variable plus an interaction term. Items were considered to display uniform DIF if the proportional change in the beta coefficient between models 1 and 2 exceeded 10%. Items were considered to display non-uniform DIF if model 3 offered better fit to the data than model 2 (as determined by whether the product of -2 and the difference in log-likelihoods was significant at the $p<.05$ level on a chi-square distribution with 1 degree of freedom).

E. Computer-based scoring. While estimation of *developmental age* is optimally conducted using maximum likelihood methods, we opted to use "brute force" estimation techniques that are more easily reproduced using readily available software (e.g., Microsoft Excel). The procedure works as follows. Each milestone question is associated with three possible ICC's that reflect the probability of each response as a function of age. ICC's for "0", "1" and "2" responses can be calculated using the following equations:

$$\text{If response = "0": } p_{ij} = 1 - \frac{e^{\alpha_{j1}+\beta_j \cdot \ln(developmental\_age_i)}}{1+e^{\alpha_{j1}+\beta_j \cdot \ln(developmental\_age_i)}}$$

$$\text{If response = "1": } p_{ij} = \frac{e^{\alpha_{j1}+\beta_j \cdot \ln(developmental\_age_i)}}{1+e^{\alpha_{j1}+\beta_j \cdot \ln(developmental\_age_i)}} - \frac{e^{\alpha_{j2}+\beta_j \cdot \ln(developmental\_age_i)}}{1+e^{\alpha_{j2}+\beta_j \cdot \ln(developmental\_age_i)}}$$

$$\text{If response = "2": } p_{ij} = \frac{e^{\alpha_{j2}+\beta_j \cdot \ln(developmental\_age_i)}}{1+e^{\alpha_{j2}+\beta_j \cdot \ln(developmental\_age_i)}}$$

Where p is the probability that child i achieves milestone j, and α_{j1}, α_{j2}, β_j are parameters from Table 2. For each participant's 10 responses, we calculated this probability for each quarter-month (or week) from birth to 72 months, thus yielding 10 ICCs. These 10 ICCs are summarized into a single probability curve by calculating their product to estimate the likelihood of the pattern of responses at each age from birth to 72 months. The child's *developmental age* is the point at which the particular pattern of responses is most likely to occur as indicated by the maximum value of this curve. One drawback to this procedure is that participants who fail all

milestones will be estimated with *developmental age* = 0 while participants who pass all milestones will be estimated with *developmental age* = 72 months.[26] To correct for this, maximum values for each form were set just above the maximum attainable developmental age resulting from at least one imperfect response, while minimum values were set in a similar fashion. An example of how to use this method to score a SWYC form in excel is provided on the SWYC website (www.theSWYC.org).

F. Hand-scoring. The computer-based scoring algorithms described above are not feasible for hand scoring. Therefore, we sought to provide a way to score Milestones forms without access to a computer. We started by calculating the average score a child would be expected to receive if he or she displayed a 15% delay, i.e., scored at the clinical threshold (hereafter, the "expected score"). Because children develop over time, these thresholds vary with age. A scoring table of these thresholds is available on the SWYC website (www.theSWYC.org).

To hand-score the Milestones, users should:

- score each response (0 for "not yet", 1 for "somewhat" and 2 for "very much"),
- sum responses across items (i.e., hereafter, the "sum score"), and
- consult a scoring table to determine whether a child of a given age exceeds the clinical threshold.

To estimate reliability, continuous scores were created by subtracting the *expected score* from the *sum score*. Note that these continuous scores are most accurate for children scoring at or near the clinical threshold.

G. Excluded items. Of the 174 items tested in the *primary care sample*, 96 displayed adequate fit and were administered to the *replication sample*. Of these, 42 were excluded because of DIF or lack of item fit. These items are listed in Table 3.

Table 3. Items discarded because of poor fit and/or differential item functioning

	approximate median age of "very much" response	Evidence of poor item fit	Evidence of non-uniform DIF	Evidence of uniform DIF
Smiles at a person he or she knows	1.3	X	X	X
Turns head towards a sound	1.5	X		X
Puts some weight on his or her legs while being held	2.1	X	X	X
Makes sounds when looking at people or toys	2.2		X	X
Puts toys in his or her mouth	3.3		X	X
Reaches for toys or other things	3.4	X		X
Reaches for people he or she knows	5.5	X		X
Sits for a while without support	5.8		X	X
Crawls or scoots	6.8	X		X
Puts sounds together - like "baba" or "dadada"	7.5	X		X
Uses thumb and finger to pick up a raisin or cheerio	7.9	X		X
Takes 1 or 2 steps while holding on to something	8.8	X		X
Looks when you point to something	9.5	X		X
Puts things inside a box or other container	10.5		X	X
Helps turn the pages of a book	11.4	X		X
Waves hello or good-bye	11.5	X	X	X
Points to things with one finger	11.9	X	X	X
Plays simple games with you - like rolling a ball back and forth	12.2	X	X	X
Says 2 or more words other than "mama" or "dada"	13.0		X	X
Brings things to show you	13.2		X	X
Tries to do things that you do - like wiping the table or sweeping	14.9	X		X
Points to at least 3 body parts - like eyes, ears, or tummy	15.2	X	X	X
Understands Yes/No questions - like "Do you want milk?	15.3	X		X
Asks for "more" of something	16.3	X		X
Uses at least 10 words that you recognize - like "dog" or "goggie" for dog	16.6		X	X
Uses a spoon or fork to eat	17.4	X		X
Points to objects in a picture when you name them - like a horse or a ball	17.8	X	X	X
Understands words like "Wait" or "Stop"	18.0		X	X
Plays make believe - like pretending to feed a doll or to be a pirate	21.1		X	X
Tells you about something that's happened	27.2		X	X
Tells you his or her age when asked	30.2	X		X
Answers questions about what things are used for - like crayons or stoves	30.3	X		X

Names at least 3 shapes - like circle, square, or triangle	31.8	X	X	X
Goes to the bathroom alone (even if you need to wipe or help with buttons)	36.2		X	X
Stays dry all day without diapers	36.9	X	X	X
Tells you what group something belongs to - like "an apple is... a fruit" or "a dog is... an animal"	39.0	X		X
Understands that "tomorrow" means something that hasn't happened yet	39.2	X	X	X
Cuts paper in 2 pieces with scissors	41.1	X		X
Dresses and undresses without help	44.8		X	X
Says the name of the town or city where you live	46.3	X		X
Puts shoes on the correct feet	48.6		X	X
Writes at least 5 letters or numbers	50.6	X	X	X

References

18. Rutter M. (1989). Age as an ambiguous variable in developmental research: Some epidemiological considerations from developmental psychopathology. *International Journal of Behavioral Development*, 12(1):1-34.
19. Drachler ML, Marshall T, Leite JC (2007). A continuous-scale measure of child development for population-based epidemiological surveys: A preliminary study using Item Response Theory for the Denver Test. *Paediatric and Perinatal Epidemiology, 21*(2):138-153.
20. Embretson SE, Reise SP. (2000). *Item response theory for psychologists*. New York: Psychology Press.
21. Mair P, Reise SP, Bentler PM. (2008). IRT goodness-of-fit using approaches from logistic regression. UC Los Angeles: Department of Statistics, UCLA. Retrieved from http://escholarship.org/uc/item/1m46j62q on December 1, 2012.
22. French AW, Miller TR. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement, 33*(3):315-332.
23. Kim SH, Cohen AS, Alagoz C, Kim S. (2007). DIF detection and effect size measures for polytomously scored items. *Journal of Educational Measurement, 44*(2):93-116.
24. Camilli G, Congdon P.(1999). Application of a method of estimating DIF for polytomous test items. *Journal of Educational and Behavioral Statistics, 24*(4):323-341.
25. Crane PK, Gibbons LE, Jolley L, van Belle G. (2006). Differential item functioning analysis with ordinal logistic regression techniques. *Medical Care, 449* (11 Suppl 3): S115-123.
26. Cheung YB, Gladstone M, Maleta K, Duan X, Ashorn P (2008). Comparison of four statistical approaches to score child development: A study of Malawian children. *Tropical Medicine and International Health, 13*(8): 987-993.