Supplemental Information

IRT: GRADED RESPONSE MODEL

In a graded response model, a single question (j) with 3 response options yields 1 discrimination parameter (β_j) and 2 thresholds (α_{j1} and α_{j2}). At any given age, these parameters can be used to determine the proportion of children who are expected to pass each milestone by using the following equations:

For response = "1 : Somewhat" or greater

$$: p_{ij} = 1 - \frac{e^{\alpha_{j1} + \beta_j \cdot \ln(\text{developmental\_age}_i)}}{1 + e^{\alpha_{j1} + \beta_j \cdot \ln(\text{developmental\_age}_i)}},$$

For response = "2 : Very much"

$$: p_{ij} = \frac{e^{\alpha_{j2} + \beta_j \cdot \ln(\text{developmental\_age}_i)}}{1 + e^{\alpha_{j2} + \beta_j \cdot \ln(\text{developmental\_age}_i)}}.$$

Note that each of these 2 equations suggests a different definition of the word "pass." Whereas the first equation requires a parent to report that the child at least somewhat displays the given skill, the second equation requires the parent to report that the child very much displays the given skill.

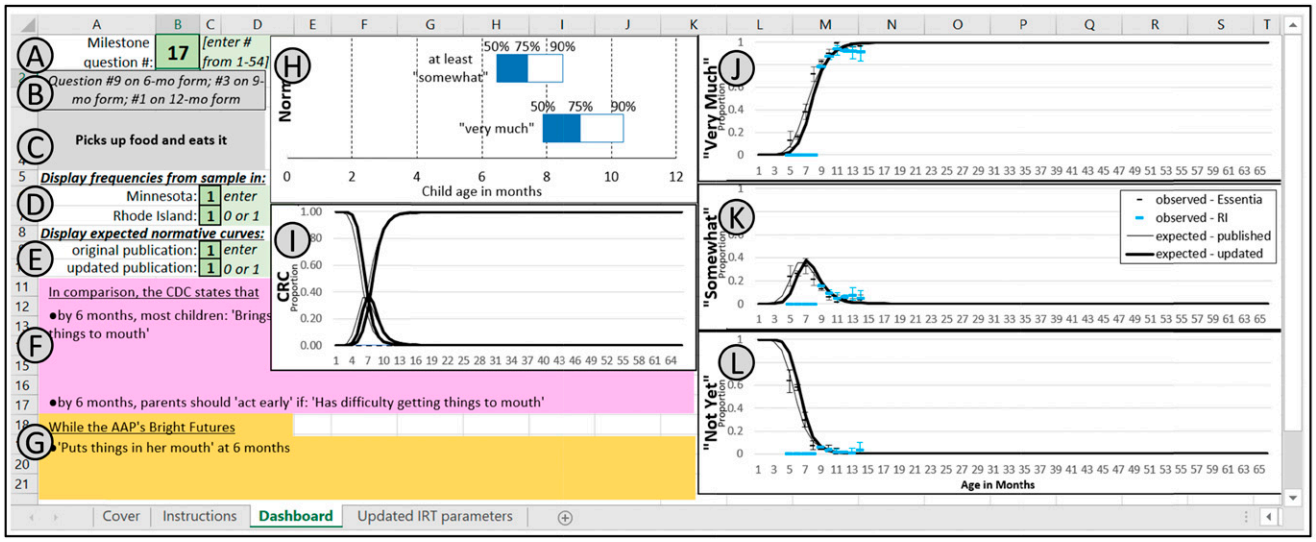
TESTS OF DIF

Following previous publications, we used a multistep approach that included a multiple-group IRT analysis to support a test of statistical significance for each item, followed by an evaluation of the magnitude of DIF.16 In a multiple-group analysis, item-level parameters that define normative curves for each item response (ie, slope and threshold parameters) are constrained to be equal between samples. Results are then compared to a series of separate models that systematically relax constraints for each item, 1 at a time, thus allowing item-level parameters to vary between samples. Evidence that the latter model displays better fit to the data (on the basis of Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and likelihood ratio tests) suggests the presence of DIF between samples. However, because large samples offer substantial statistical power to detect even small (and clinically nonsignificant) differences in item functioning, we included a second step in which we estimated the magnitude of DIF for each item identified in step 1. Following previous guidelines, our estimate of magnitude was defined as the sum of the absolute differences in item characteristic curves weighted by the latent variable scores for each person who completed that item by using a threshold of 0.18 for clinical significance.17 Finally, we estimated the effect of DIF on screening algorithms by comparing estimates of each child's DQ calculated on the basis of norms from the combined population with a DQ based on state-specific norms for items that displayed DIF. DQ is the scoring algorithm for the SWYC's developmental screen.14

CALCULATION OF DQ

To calculate the DQ, developmental age is first estimated on the basis of parents' responses to each question. Specifically, an item response curve is generated for each question answered by using the equations above. Each of these curves represents the changing likelihood of a given response from birth to 72 months of age. Multiplying the curves then offers an estimate of the likelihood of observing the entire pattern of responses reported at each age from birth to 72 months. The maximum value of this curve represents the age at which the pattern of responses is most likely to be observed, and we use this maximum value as our estimate of each child's developmental age. For children who received the maximum score (all answers were "2: very much"), developmental age was set just above the maximum attainable developmental age resulting from at least 1 imperfect response on the set of questions their parents completed. Developmental age for children who received minimum scores (all answers were "0: not at all") was set in a similar fashion. An advantage of the item response curve method is that developmental status is estimated on the basis of all available milestones, even if not every question is answered.

To achieve model convergence, responses for 2 items were recoded as binary by combining "not at all" and "somewhat" responses. Analyses revealed statistically significant DIF between states with respect to 25 questions, but effect sizes fell below our specified threshold of 0.18 for all but 8 items. Most differences were identified in comparison with our Massachusetts sample, which was the smallest and was collected through a research study and not through standard clinical care. In both our Rhode Island and Minnesota samples, we calculated a DQ on the basis of

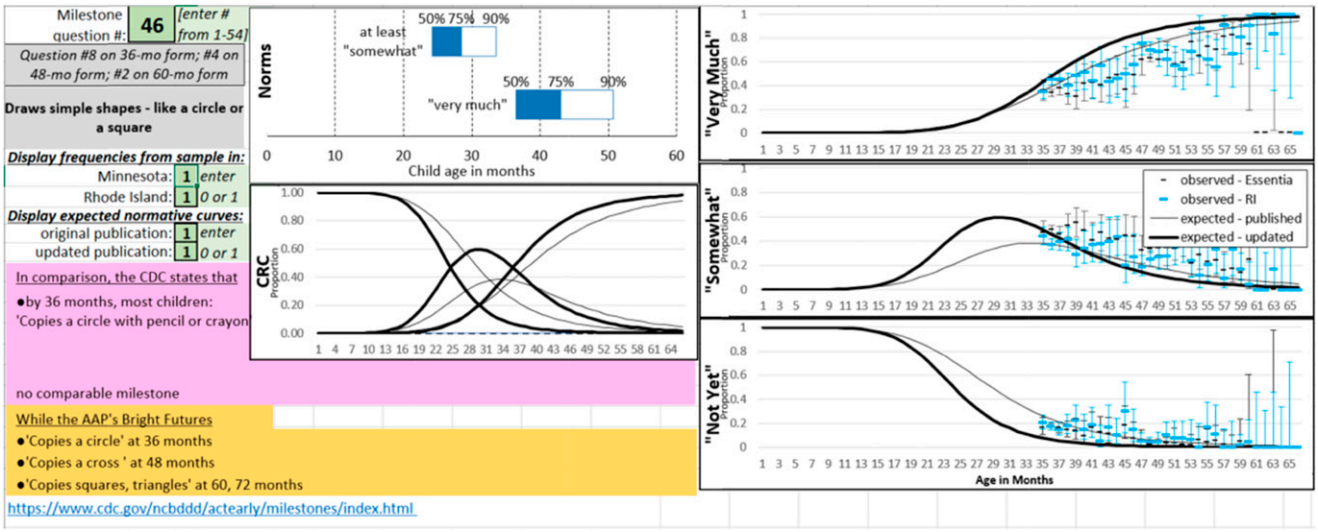


SUPPLEMENTAL FIGURE 2

Annotated screenshot of the Microsoft Excel workbook "SWYC Normative Data and Visualization Tool." To inspect normative data for an individual developmental milestone, the user should complete items in bold. A, The user should select the SWYC Milestone number (number refers to original SWYC Milestone publication). B, Information regarding which SWYC forms include this question. C, Text of SWYC question. D, The user should select data sets from which to display observed frequencies. E, The user should select which parameters to use for calculating interclass correlations. Output includes: F, Text of CDC milestones with similar constructs. G, Text of *Bright Futures* milestones with similar constructs. H, Norms: ages at which 50%, 75%, and 90% of parents are expected to report that their children (1) at least somewhat pass the milestone and (2) very much pass the milestone. I, Category response curves (CRCs) display the expected probability at each child age that a parent will report "not at all," "somewhat," and "very Much." J–L, Calibration: for each response option, these figures are used to compare expected values expressed as item characteristic curves (CRCs) with observed values, which represent that actual proportion of parents who chose each response at each age. AAP, American Academy of Pediatrics; RI, Rhode Island.

summary scores across the 3 samples, and then we recalculated the DQ using sample-specific parameters for these 8 items. The correlations between the alternative scores were 0.999 in the Minnesota sample and 0.996 in the Rhode Island sample, and the means were identical to the second decimal value in both samples. In addition, positive and negative classifications of screening scores on the SWYC's developmental screen remained the same for all children in both samples.

Normative data for all SWYC developmental questions are presented in a Microsoft Excel workbook entitled "SWYC Normative Data and Visualization Tool," which we offer in Supplemental Fig 2. Note that this tool was created by using Microsoft Excel 2016. Backward compatibility with previous versions cannot be assured.



SUPPLEMENTAL FIGURE 3
Example of norms for a later developmental milestone. AAP, American Academy of Pediatrics; CRC, category response curve; RI, Rhode Island.

SUPPLEMENTAL TABLE 4 Regression Results: Main Effects and Interactions With Child Age

	MN Sample		RI Sample		MA Sample	
	β (SE)	*P*	β (SE)	*P*	β (SE)	*P*
Independent variables						
Positive behavioral screen results	−.0022 (0.0071)	.753	.0005 (0.007)	.947	.0158 (0.0242)	.512
Cumulative SDoH						
0	Reference	Reference	Reference	Reference	Reference	Reference
1	−.0089 (0.0075)	.236	−.0062 (0.0084)	.463	−.065 (0.0275)	.018
2+	−.0074 (0.0112)	.509	.147 (0.0471)	.002	−.0854 (0.0375)	.023
Medicaid	.0294 (0.008)	<.001	.0396 (0.0077)	<.001	−.0128 (0.031)	.679
Premature birth	−.1611 (0.0126)	<.001	−.0567 (0.0213)	.008	−.0704 (0.0356)	.048
Child sex: male	−.016 (0.007)	.022	−.031 (0.0066)	<.001	−.0307 (0.0229)	.18
Child race						
White	Reference	Reference	Reference	Reference	Reference	Reference
Asian American	.066 (0.0311)	.034	.0146 (0.0156)	.349	−.0624 (0.0566)	.27
African American	.0851 (0.0199)	<.001	.0304 (0.0124)	.014	.0542 (0.0411)	.187
Other	.0438 (0.0191)	.022	.0303 (0.0141)	.031	.0024 (0.0335)	.943
Hispanic ethnicity	−.012 (0.0232)	.605	−.0061 (0.0123)	.619	−.0255 (0.0338)	.451
Interactions with age						
Positive behavioral screen results	−.0332 (0.0036)	<.001	−.0353 (0.0044)	<.001	−.0487 (0.0103)	<.001
Cumulative SDoH						
0	Reference	Reference	Reference	Reference	Reference	Reference
1	−.0109 (0.0036)	.003	−.0156 (0.0051)	.002	.0099 (0.011)	.37
2+	−.0166 (0.0052)	.001	−.0887 (0.0243)	<.001	.004 (0.0146)	.785
Medicaid	−.0261 (0.0036)	<.001	−.0292 (0.0048)	<.001	−.005 (0.0132)	.704
Premature birth	.0336 (0.0057)	<.001	−.0187 (0.0267)	.484	.0137 (0.0142)	.336
Child sex: male	−.021 (0.0031)	<.001	−.0256 (0.0041)	<.001	−.0167 (0.0091)	.067
Child race						
White	Reference	Reference	Reference	Reference	Reference	Reference
Asian American	−.0266 (0.016)	.096	−.0166 (0.01)	.098	.0129 (0.0212)	.542
African American	−.0399 (0.0085)	<.001	−.0285 (0.0071)	<.001	−.03 (0.0168)	.075
Other	−.0307 (0.008)	<.001	−.0221 (0.0082)	.007	−.0092 (0.0139)	.506
Hispanic ethnicity	.0038 (0.0134)	.777	−.0161 (0.0074)	.03	.985 (−0.0268)	.985

"Positive behavioral screen results" refers to the BPSC for children younger than 18 mo and the PPSC for all other children. MA, Massachusetts; MN, Minnesota; RI, Rhode Island.

SUPPLEMENTAL TABLE 5 Regression Results: Main Effects and Interactions With Child Age by Year

	MN Sample		RI Sample		MA Sample	
	β (SE)	*P*	β (SE)	*P*	β (SE)	*P*
Independent variables						
Positive behavioral screen results	−.006 (0.009)	.498	.004 (0.008)	.663	.047 (0.059)	.429
Cumulative SDoH: 0	−.008 (0.01)	.401	−.003 (0.011)	.782	.015 (0.061)	.809
Cumulative SDoH: 1	.003 (0.014)	.835	.018 (0.023)	.44	−.06 (0.093)	.521
Cumulative SDoH: 2+	.051 (0.009)	<.001	.053 (0.009)	<.001	.128 (0.073)	.077
Medicaid	−.164 (0.015)	<.001	−.121 (0.017)	<.001	−.005 (0.081)	.954
Premature birth	−.013 (0.008)	.118	−.017 (0.008)	.038	−.029 (0.059)	.625
Child sex: male	.069 (0.035)	.05	.029 (0.019)	.129	.06 (0.154)	.696
Child race: white	.089 (0.025)	<.001	.036 (0.015)	.019	−.025 (0.096)	.796
Child race: Asian American	.098 (0.029)	.001	.048 (0.017)	.005	−.107 (0.085)	.205
Child race: African American	−.018 (0.029)	.545	−.003 (0.015)	.856	.005 (0.088)	.953
Child race: other	−.017 (0.01)	.104	−.034 (0.011)	.003	−.063 (0.063)	.315
Interactions with age, y						
Positive behavioral screen results						
1	−.017 (0.01)	.104	−.034 (0.011)	.003	−.063 (0.063)	.315
2	−.073 (0.015)	<.001	−.077 (0.011)	<.001	−.177 (0.067)	.008
3	−.097 (0.016)	<.001	−.106 (0.021)	<.001	−.184 (0.067)	.006
4	−.13 (0.016)	<.001	−.145 (0.03)	<.001	−.207 (0.067)	.002
Cumulative SDoH: 1						
1	−.007 (0.011)	.512	−.013 (0.015)	.356	−.108 (0.065)	.099
2	−.039 (0.015)	.01	−.035 (0.014)	.015	−.016 (0.068)	.817
3	−.046 (0.016)	.005	−.051 (0.025)	.042	−.086 (0.07)	.218
4	−.022 (0.016)	.166	−.022 (0.033)	.512	−.02 (0.068)	.77
Cumulative SDoH: 2+						
1	−.03 (0.015)	.054	−.035 (0.032)	.276	−.061 (0.099)	.539
2	−.058 (0.022)	.008	−.084 (0.031)	.007	.033 (0.101)	.742
3	−.083 (0.024)	<.001	−.147 (0.049)	.003	−.001 (0.103)	.989
4	−.04 (0.021)	.061	−.109 (0.056)	.049	−.019 (0.101)	.853
Medicaid						
1	−.054 (0.01)	<.001	−.05 (0.012)	<.001	−.157 (0.077)	.042
2	−.093 (0.014)	<.001	−.076 (0.012)	<.001	−.156 (0.079)	.049
3	−.097 (0.016)	<.001	−.092 (0.023)	<.001	−.147 (0.081)	.07
4	−.11 (0.015)	<.001	−.076 (0.031)	.014	−.163 (0.083)	.049
Premature birth						
1	.04 (0.014)	.005	.043 (0.022)	.048	−.064 (0.086)	.457
2	.064 (0.021)	.002	.036 (0.021)	.081	−.043 (0.089)	.63
3	.104 (0.024)	<.001	−.049 (0.034)	.147	.004 (0.091)	.967
4	.128 (0.024)	<.001	.106 (0.047)	.025	−.024 (0.089)	.784
Child sex: male						
1	−.019 (0.008)	.018	−.044 (0.011)	<.001	.003 (0.062)	.957
2	−.08 (0.012)	<.001	−.076 (0.011)	<.001	−.083 (0.063)	.191
3	−.069 (0.013)	<.001	−.068 (0.019)	<.001	−.057 (0.064)	.371
4	−.066 (0.013)	<.001	−.055 (0.026)	.034	−.038 (0.064)	.547
Child race: Asian American						
1	−.038 (0.038)	.32	−.047 (0.025)	.058	−.132 (0.161)	.413
2	−.034 (0.052)	.51	−.039 (0.025)	.112	−.08 (0.166)	.632
3	−.085 (0.057)	.132	−.022 (0.051)	.664	−.069 (0.166)	.678
4	−.105 (0.07)	.133	−.102 (0.068)	.134	−.074 (0.162)	.649
Child race: African American						
1	−.038 (0.026)	.154	−.021 (0.021)	.306	.064 (0.102)	.533
2	−.084 (0.038)	.028	−.075 (0.02)	<.001	.028 (0.105)	.793
3	−.189 (0.036)	<.001	−.077 (0.031)	.012	−.045 (0.106)	.67
4	−.116 (0.036)	.001	−.03 (0.041)	.462	−.029 (0.107)	.785
Child race: other						
1	−.084 (0.03)	.006	−.037 (0.024)	.116	.129 (0.089)	.146
2	−.13 (0.045)	.004	−.085 (0.023)	<.001	.1 (0.091)	.272
3	−.125 (0.049)	.011	−.091 (0.038)	.017	.048 (0.091)	.6
4	−.086 (0.047)	.067	−.021 (0.046)	.65	.084 (0.094)	.374

"Positive behavioral screen results" refers to the BPSC for children younger than 18 mo and the PPSC for all other children. MA, Massachusetts; MN, Minnesota; RI, Rhode Island.