4. CONCURRENT CALIBRATIONS AND EQUATING DESIGN FOR THE BSF-R IN THE 2-YEAR NATIONAL DATA COLLECTION

A shortened and streamlined version of the Bayley Scales of Infant Development, Second Edition (BSID-II) called the Bayley Short Form—Research Edition (BSF-R) was specially developed to assess child developmental status in the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B). One of the major justifications for using the Bayley scales in the ECLS-B was that they would produce results comparable with other child development studies that also report results for the BSID-II. Care was taken in selecting BSF-R item subsets so that ECLS-B results would be as consistent as possible with BSID-II. In particular, Mental Development Index (MDI) and Psychomotor Development Index (PDI) scores, developed by the publisher from a nationally representative standardization sample of children collected in 1991–92, could then also be reported in the ECLS-B.

For test construction purposes, the BSID-II standardization dataset was obtained from The Psychological Corporation, and all 178 mental and 111 motor items for infants between 1 and 42 months of age were calibrated using an Item Response Theory (IRT) two-parameter logistic (2-PL) model, a response model that was specifically chosen to highlight the more discriminating BSID-II items, which could then be identified for use in the ECLS-B. Although the initial item calibrations were undertaken by Westat, they are referred to here as *publisher calibrations* since these are based on the publisher's standardization dataset.

There are no weights on the standardization dataset. The standardization sample is representative of the national infant population and is considered to be self-weighting. The standardization dataset is comprised of the standardization sample and additional observations. None of these observations have case weights.

By contrast, the ECLS-B dataset is a stratified cluster sample based on unequal selection probabilities. Sample weights are used with the ECLS-B dataset so that it will then be representative of the national infant population in 2001. Thus, through calibration, scaling, scoring, and analysis and throughout this report, the standardization data were unweighted and the ECLS-B data were weighted.

4.1 Results of the BSF-R Adaptive Testing Strategy

Both the full BSID-II and BSF-R short forms were designed to be administered as adaptive tests. A core item set, appropriate for children in the target age group, was administered first. The raw score total for this core item set was then used to determine if additional basal or ceiling item sets should also have been administered. BSF-R adaptive tests followed procedures of administration similar to those used in BSID-II. The BSF-R diverged from the BSID-II primarily in its use of shortened core, basal, and ceiling item sets. The BSF-R was composed of shorter tests that were not designed to be strictly parallel tests.

Moreover, the BSF-R was specially adapted for home administration as part of household survey interviews conducted in the ECLS-B. The BSF-R was completed by interviewers guided by a standard schedule of task administrations, involving the structured presentation of stimulus material intended to elicit child responses. Still other items were scored based on observed child behavior occurring at any moment during assessment. Additionally, one or more items were scored from each task administration or observation. The first three sections of table 4-1 report the number of items in each of the basal, core, and ceiling item sets, followed by the number of task administrations and observations completed by interviewers before recording item responses. The sum of task administrations and observations does not equal the total number of items because in several instances more than one item was scored from a single task administration or observation. The difference between the sum of task administrations plus observations, and the total number of items, is shown in the last section of the table. This difference represents the number of items that cannot be considered entirely independent items.[1]

Although tests with a different number of items and other minor adaptations do not satisfy the rigorous requirements for test equating, tests based on the same item pool can often be calibrated on a common scale metric. Tests from the same item pool then yield unbiased ability estimates with the same central tendency but different standard errors. IRT procedures used in BSF-R design and development offered the prospect of producing comparable scores sharing the same scale metric used by the publisher. This metric was used to report BSF-R results, including model-based estimates of BSID-II raw scores and developmental index scores.

[1] Table 4-1 summarizes an exceedingly complex observational setting. This is because some item responses are recorded based on prior observation while requiring a separate administration on another occasion. It is even possible for a mental item to be scored from a motor administration. Except for item counts in the first section of the table, the numbers in other sections should be considered approximations rather than represent a full accounting of the situation encountered on each occasion.

Table 4-1. Number of BSF-R items, administrations, observations, and dependencies, by BSF-R scale, round of data collection, and item set: 2001–02 and 2003–04

	9-month data collection			2-year data collection		
Characteristic and item set	Total	Mental scale	Motor scale	Total	Mental scale	Motor scale
Number of items						
Total	66	31	35	69	37	32
Basal	21	9	12	18	9	9
Core	26	13	13	36	19	17
Ceiling	19	9	10	15	9	6
Number of administrations						
Total	34	17	17	43	25	18
Basal	6	2	4	7	4	3
Core	18	9	9	26	14	12
Ceiling	10	6	4	10	7	3
Number of observations						
Total	15	9	6	8	6	2
Basal	9	5	4	4	3	1
Core	2	2	0	3	2	1
Ceiling	4	2	2	1	1	0
Number of dependencies						
Total	17	5	12	18	6	12
Basal	6	2	4	7	2	5
Core	6	2	4	7	3	4
Ceiling	5	1	4	4	1	3

NOTE: Item count: The number of items that are scored in each item set. Administrations: The number of task administrations. Each administration is defined as the structured presentation of the stimulus materials to obtain the child's response(s). Thus, multiple cores can be obtained from the same administration. Observations: Observations are items that do not require the structured presentation of stimulus materials but are scored by direct observation of the child's spontaneous behavior. Dependencies: A dependent item is redundant with another item and does not provide unique information about the child's ability and therefore does not increase construct representation. Item dependencies may also exacerbate any construct-irrelevant factors that may be associated with an item (e.g., prior familiarity with the item). The number of dependencies is calculated as the difference between the total item count and the sum of the number of administrations and observations.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

The BSF-R was designed so that most children received only the core item set, while other children received both the core item set plus either the basal or ceiling item set, depending on basal and ceiling decision rules based on the raw score total obtained with the core item set. Basal and ceiling decision rules were defined so that children within a standard deviation to either side of the age group mean ability estimate received only the core item set. In ideal circumstances, this would imply that about 68 percent of the children received only the core item set. Below a certain minimum core item set raw score value, another 16 percent were expected to receive the additional complement of basal items. Above a certain maximum core item set raw score value, another 16 percent were expected to receive the additional complement of ceiling items.

These percentages were expected to vary depending on the actual raw score values obtained with the core item sets. Basal and ceiling item sets contained 6 to 11 items that were specially selected to cover the child population well into the tails of each ability distribution. Table 4-2 shows how many children received the core, basal, and ceiling item sets on the BSF-R mental and motor tests. This shows that the item sets performed more or less as expected at 2 years. However, at 9 months, both the mental and motor distributions shifted upward toward the higher levels of ability, resulting in a greater use of both mental and motor ceiling items. While this resulted in a certain amount of inefficiency—in the sense that more items needed to be administered during field work—in principle, the ceiling item sets handled appropriately this need for a test with more difficult items.

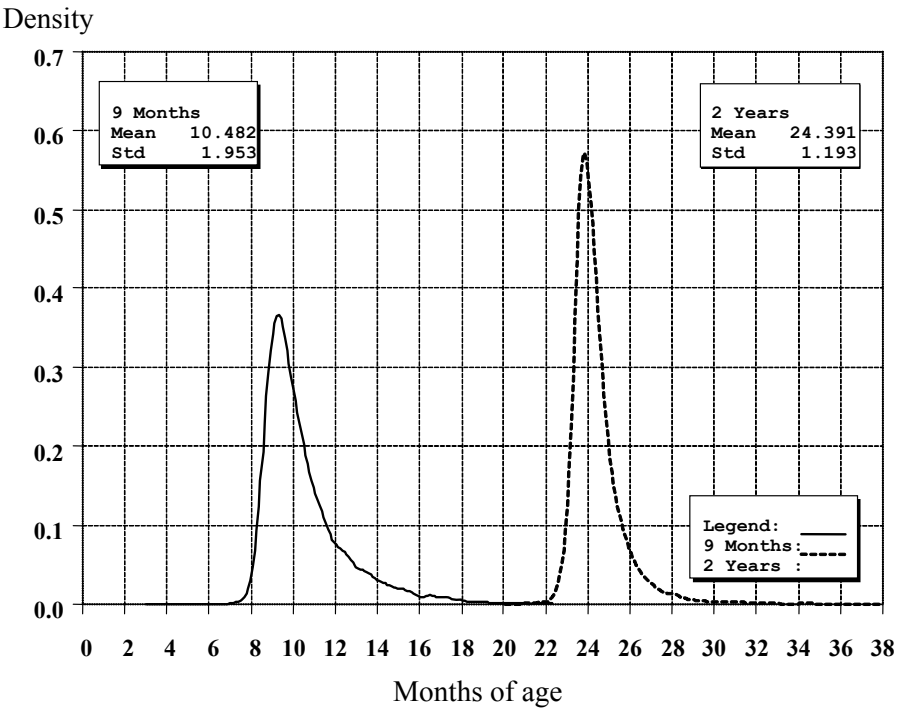
Table 4-2. Number and percentage of children, and mean ability estimates, by BSF-R scale, item set, and round of data collection: 2001–02 and 2003–04

BSF-R scale and item set	9-month data collection			2-year data collection		
	Number of children	Percent	Mean	Number of children	Percent	Mean
Mental scale						
Total	10,200	100.0	-0.982	8,900	100.0	4.306
Basal	250	2.4	-2.624	1,100	12.2	2.249
Core	5,650	55.4	-1.466	6,150	69.1	4.221
Ceiling	4,300	42.2	-0.982	1,650	18.7	5.967
Motor scale						
Total	10,200	100.0	-0.942	8,850	100.0	2.889
Basal	450	4.5	-3.074	1,150	12.8	1.779
Core	6,300	62.0	-1.592	7,150	80.7	2.966
Ceiling	3,400	33.4	0.553	600	6.5	4.114

NOTE: Detail may not sum to total because of rounding. Sample sizes have been rounded to the nearest 50.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Figure 4-1 shows why 9-month ability distributions shifted upward, resulting in the administration of fewer basal and more ceiling item sets than initially expected. This was because the age distribution at 9 months was shifted to the right and was highly skewed. Instead of being assessed at an average age of 9.5 months as initially expected, the first wave of infants were assessed at an average age of 10.5 months. For every additional month of age, mental ability estimates are expected to rise by fully 0.5 population standard deviations. In that case, only 7 percent required the basal and 31 percent required the ceiling items. However, it was not just that the age distribution was shifted to the right but also that it was highly skewed. In this case, even fewer infants were expected to receive the basal and even more received the ceiling item sets. Notice that for the second wave of assessments the age distribution was much closer to the expected 24.5 months of age, and the distribution was much less skewed. Thus, the age distribution of the ECLS-B sample at 9 months and 2 years was to a considerable extent responsible for the distribution of children who received the core, basal, and ceiling item sets.

Figure 4-1. Kernel density estimation for age distributions of children in the 9-month and 2-year ECLS-B data collections: 2001–02 and 2003–04



NOTE: Kernel density estimation obtained with weighted ECLS-B sample observations. A kernel density plot is a nonparametric representation of density that has been smoothed (e.g., by using a Gaussian function). Std = standard deviation.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Because mental and motor development is explosive during infancy, infant age and development are closely related. This age-development relationship can be exploited during item calibration and scoring to improve the accuracy of item parameters and ability estimates. Observations were first clustered by age group, and the mean and standard deviation—representing the ability distribution in each age group—were used to condition group member ability estimates. The gains in precision obtained with multiple group IRT[2] are thought to be slight but help to ensure consistency when individual observations are scored.

Multiple group IRT (Bock and Zimowski 1997) was applied to ECLS-B item calibrations using Bilog-MG (Zimowski et al. 1997) and in-house software. The first set of software represents an industry standard and was useful for assessing the precision and accuracy of results. In-house software provided better graphics for visual inspection of item fit, together with almost unlimited flexibility during test equating and analysis. The two sets of software use multiple group IRT and produce results that are essentially identical. In multiple group IRT, item parameter values are estimated simultaneously together with the latent group ability distributions.

4.2 Examining the Potential of BSF-R Item Sets

An examination of the psychometric properties of BSF-R instruments began with an assessment of the potential of BSF-R item subsets before these were actually used in the ECLS-B. Among the 2,939 observations in the publisher's dataset between 1 and 42 months of age, 1,700 comprise the standardization sample, including a subset of 900 standardization observations between 8 and 30 months of age (see table 2-1). These age groups coincide most closely with the range of ages found in the ECLS-B. To examine the potential of BSF-R item sets, observations were compared after they were first scored with the full complement of BSID-II items and then again scored using BSF-R item subsets. While in principle this includes both the 9-month and 2-year item subsets, in practice each child was scored on the BSF-R items for which there were valid responses in the standardization dataset.[3]

[2] For further information about multigroup IRT, please refer to Bock, R.D., and Zimowski, M.F. (1997). Multiple Group IRT. In W.J. van der Linden and R.K. Hambleton, (Eds.), *Handbook of Modern Item Response Theory.* New York: Springer-Verlag.

[3] This implies that some children probably were scored on the 9-month item sets, others on the 2-year item sets, and still others on some items from both sets. As shown in table 4-8, the publisher standardization sample had insufficient number of observations to conduct separate analyses at 9 months and 2 years.

Having already calibrated all of the BSID-II items and scored all of the standardization sample observations with the full set of publisher item calibrations, the same observations were scored a second time using only the BSF-R item subsets. Publisher item calibrations were used on both occasions. Consequently, any differences encountered between the two sets of scores would have reflected a bias introduced by using the BSF-R item subsets. Results presented in table 4-3 permit a comparison of central tendencies, standard errors, and two measures of residual goodness of fit when the standardization sample was scored twice with different item subsets from publisher item calibrations.

When mean ability estimates obtained with the full BSID-II were compared with means obtained using BSF-R item subsets, these were expected to yield unbiased estimates of average ability. Mean expected a posteriori (EAP) ability estimates reported in table 4-3 were virtually identical on the mental scale and within a 10th of a population standard deviation on the motor scale (see section 5.2 for a more detailed description of the EAP). This was expected since the same standardization sample observations and publisher item calibrations were used to obtain both sets of means. Nevertheless, the results supported expectations. The central tendencies obtained with BSF-R item subsets faithfully reproduced those obtained using the full complement of BSID-II items.

Table 4-3. Descriptive statistics for 900 standardization sample observations scored with publisher calibrations using both the full BSID-II and BSF-R item sets: 2001–02 and 2003–04

Scale	Full BSID-II	BSF-R item subset[1]
Mental		
Mean EAP ability estimate	2.511	2.512
Mean EAP standard error	0.282	0.345
Information-weighted mean square residual goodness of fit—Infit	0.954	0.915
Outlier-sensitive mean square residual goodness of fit—Outfit	0.964	0.931
Motor		
Mean EAP ability estimate	1.648	1.634
Mean EAP standard error	0.349	0.399
Information-weighted mean square residual goodness of fit—Infit	0.935	0.891
Outlier-sensitive mean square residual goodness of fit—Outfit	0.884	0.835

[1] Includes both 9-month and 2-year item subsets.

NOTE: Publisher standardization dataset observations 8 through 30 months of age that corresponding most closely with age groups found in the ECLS-B sample. EAP = expected a posteriori.

SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

For both the mental and motor scales, BSF-R instruments yielded a somewhat larger standard error when compared with the corresponding standard errors obtained from scores using BSID-II item sets. The average standard error for the 900 standardization sample observations on the BSF-R mental subset was 0.345 population standard deviation, which compared with 0.282 for the full BSID-II mental set. The average standard error on the BSF-R motor item set was 0.399, which compared with 0.349 for the full BSID-II motor set. Indeed, the somewhat larger standard errors obtained with the BSF-R instruments were expected since these were obtained using smaller item subsets. Although no single child would ever be administered all 178 mental items or all 111 motor items, BSID-II core, basal, and ceiling sets were invariably larger than those included in BSF-R instruments.

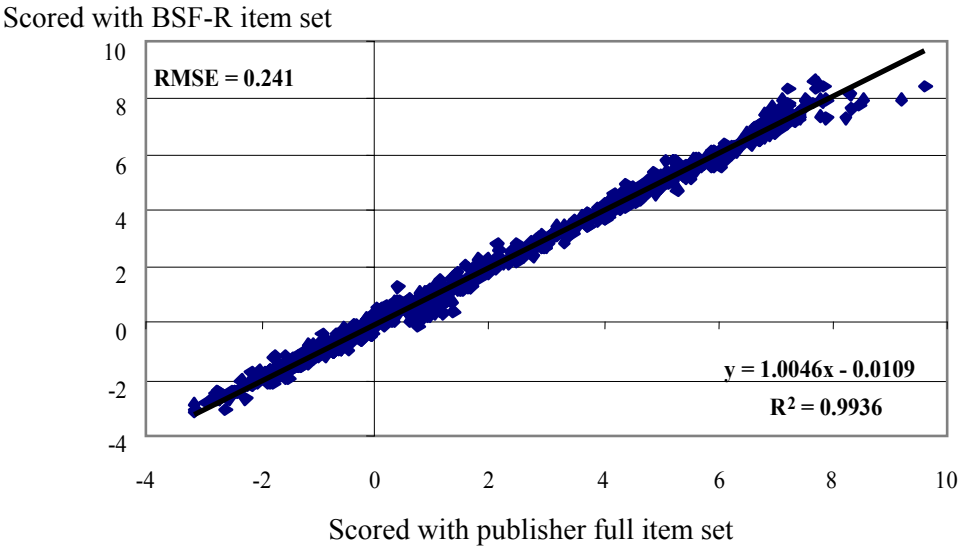
These comparisons were reassuring in the sense that somewhat larger standard errors for the BSF-R short forms had been expected. The BSF-R tests were never intended to be strictly parallel tests yielding virtually identical ability estimates and similar standard errors found in BSID-II. Instead, using publisher item parameters, the BSF-R item subsets were expected to perform like τ-equivalent tests, producing essentially identical ability estimates but somewhat larger standard errors.[4]

This same analysis can be repeated across the entire ability range for standardization sample observations between the ages of 8 and 30 months. Linear relationships between observations scored with the full BSID-II and scored again with the shorter BSF-R item subsets are shown in figures 4-2 and 4-3. To the extent that it was possible for BSF-R item subsets to produce results that were identical to those produced with the full BSID-II, the EAP ability estimates would have aligned themselves precisely along a straight line having an origin of zero and slope of unity.

Indeed, the two figures show that the central tendency of the relationship between the two sets of ability estimates had an origin very close to zero and a slope very close to unity. The r^2 coefficients reported in the figures are also close to unity, suggesting that the relationship between the two sets of scores was nearly perfect. However, the ability range between 8 and 30 months of age is so large that the r^2 statistics may be somewhat misleading. A better measure of the imperfection in measurement is provided by the root mean squared error (RMSE) reported in each of the figures. These values show that the expected error of estimation obtained with the reduced item subsets was approximately one-quarter of a population standard deviation (RMSE = 0.241 for the BSF-R mental and RMSE = 0.224 for the motor). The average error is exceedingly small, suggesting that, under clinical conditions, the BSF-R item subsets were capable of predicting BSID-II ability estimates with considerable precision across a broad range of ability.

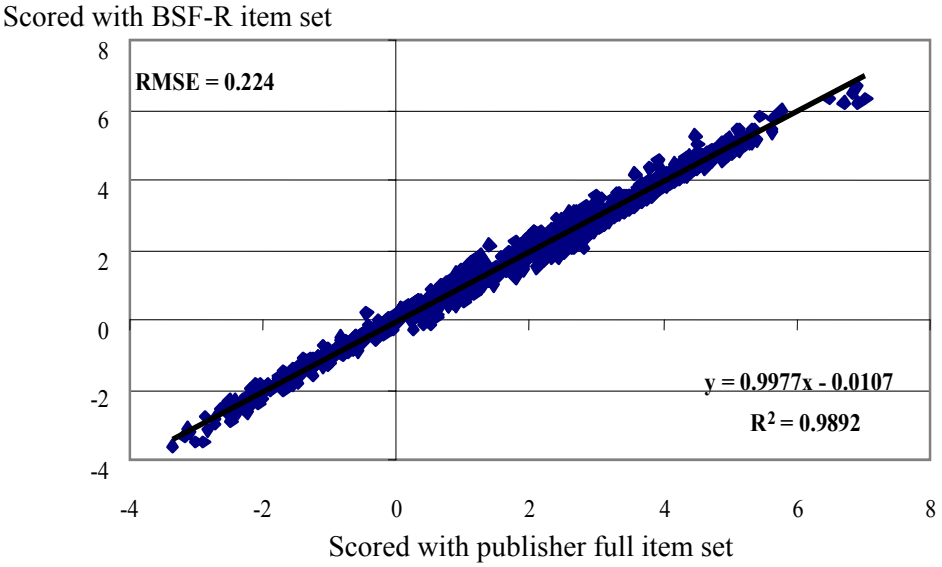
[4] The τ-equivalent refers to measurements that have the same true scores but possibly different standard errors (Lord and Novick 1968).

Figure 4-2. Expected a posteriori ability estimates, using fixed publisher item calibrations throughout; standardization sample observations scored first with the full BSID-II mental (*x* axis) and then with the BSF-R mental item set (*y* axis): 1993



NOTE: RMSE = root mean squared error; R^2 = proportion of variance in the data explained by the regression equation.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993.

Figure 4-3. Expected a posteriori ability estimates, using fixed publisher item calibrations throughout; standardization sample observations scored first with the full BSID-II motor (*x* axis) and then with the BSF-R motor item set (*y* axis): 1993



NOTE: RMSE = root mean squared error; R^2 = proportion of variance in the data explained by the regression equation.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993.

Measurement accuracy was also assessed with person-fit analyses of individual response vectors. Person-fit indices showed to what extent a response pattern was considered typical. One would ordinarily expect an examinee to obtain correct responses to easy items, provide correct and incorrect responses to items in the vicinity of his or her ability, and obtain incorrect responses to hard items. Fit statistics measure the extent to which a response pattern contains surprisingly correct responses to difficult items or surprisingly incorrect responses to easy items. As such, fit statistics show the extent to which the data are found to be appropriate for the IRT model. To assess this issue, average person fit statistics are also reported in table 4-3.

Both outlier sensitive (Outfit) and information-weighted (Infit) mean square statistics are reported in table 4-3. Outfit is based on the sum of squared residuals normalized by the variance around its expectation. The disadvantage of this statistic is that it is quite sensitive to unexpected responses to items that are much too easy or much too difficult. Infit is an information-weighted measure that gives less weight to remote items in determining on the magnitude of the fit statistic (Linacre and Wright 1994). The expected value for the mean square residual on both of these indices is 1.0. Departures from expectation are represented by values noticeably above or below unity, where large values represent excessive noise and small values represent insufficient stochastic variation needed for useful measurement. For reasonably large samples, fit statistics greater than 1.1 indicate departures from expected response patterns that require further attention (Smith, Schumacker, and Bush 1998).

Applying these same criteria, all of the Infit statistics reported in table 4-3 are slightly less than unity. An Infit index of 0.9 implies that there is 10 percent less randomness than expected among item responses that closely matched the respondent's ability level. In this case, individual responses were too predictable, fit the response model too closely, and provided redundant information when these observations were scored with IRT. In the case of the Bayley, this may have resulted from coding several item responses from a single task administration. The assessor effectively behaved as if he or she were imputing responses rather than recording behavior observed after independent trials. Outfit statistics for both motor item sets were satisfactory but again showed evidence of redundant information contained in the item responses.[5]

In general, Infit values in excess of the criterion value of 1.1 are a more serious problem than Outfit values in excess of the same criterion value. This is because high Infit values show that the data fail

[5] Although it can be safely assumed that standardization sample observations followed publisher recommendations in applying basal and ceiling item sets, no effort was made here to assure that responses to BSF-R items follow basal and ceiling rules prescribed for BSF-R administration.

to fit the response model at the point where they are most needed to estimate a person's level of ability. High Outfit values are easier to manage because in a worst case scenario suspect responses could be replaced with missing values without serious impact on ability estimates.

Table 4-4. Number and percentage distribution of 900 standardization sample observations scored with publisher calibrations using both the full BSID-II and combined 9-month and 2-year BSF-R item sets, by level of outfit and scale: 1993

		Full BSID-II		BSF-R subsets[1]	
Fit	Level of outfit	Number	Percent	Number	Percent
Mental scale					
Total	$0 \leq y < \infty$	900	100.0	900	100.0
Excellent	$0 \leq y < 1$	618	68.7	663	73.7
Acceptable	$1 \leq y < 3$	263	29.2	200	22.2
Problematic	$3 \leq y < 5$	19	2.1	29	3.2
Unacceptable	$5 \leq y < \infty$	0	#	8	0.9
Motor scale					
Total	$0 \leq y < \infty$	900	100.0	900	100.0
Excellent	$0 \leq y < 1$	655	72.8	668	74.2
Acceptable	$1 \leq y < 3$	236	26.2	225	25.0
Problematic	$3 \leq y < 5$	8	0.9	4	0.4
Unacceptable	$5 \leq y < \infty$	1	0.1	3	0.3

Rounds to zero.
[1] Includes both 9-month and 2-year item subsets. Outfit = outlier-sensitive mean squared residual goodness of fit.
NOTE: Standardization dataset observations 8 through 30 months of age. Detail may not sum to total because of rounding.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993.

Fit statistics are generated for each observation, which implies that observations should be examined individually. Table 4-4 reports sample frequencies for the Outfit statistic broken down into four categories. Outfit < 1 represents model fit that is considered extremely good but may suggest redundant information. This is followed by a category where 1 < Outfit < 3, which may show some evidence of misfit but is usually considered satisfactory. Outside this range, 3 < Outfit < 5 would be considered problematic and elicit attention to individual cases. At the extreme, where Outfit > 5, model fit is considered unacceptable, and individual cases should probably not be given score values. The table shows that the 900 observations between 8 and 30 months of age were generally well represented by publisher item calibrations, with very few observations that might be considered problematic. The full complement of BSID-II standardization sample observations essentially escape being labeled unacceptable, with the possible exception of a single observation on the motor scale. Generally speaking, subsets of BSF-R items

performed almost as well but revealed a handful of observations on both the mental and motor scales that probably should not have been given score values.

4.3 BSF-R Compatibility

The performance of BSF-R instruments used in the ECLS-B fieldwork were then considered. Assuming that IRT parameter invariance properties hold, then it would have been possible to score ECLS-B observations directly using publisher item calibrations.[6] This possibility was examined using ECLS-B longitudinal data collected at 9 months and 2 years of age. As reported in table 4-5, the ECLS-B sample consisted of approximately 10,215 children assessed at two points in time. This includes 10,197 children assessed on the mental scale and 10,163 on the motor scale during the 9-month data collection and 8,912 assessed on the mental scale and 8,824 assessed on the motor scale during the 2-year data collection. The difference between the total assessed at 9 months and 2 years is largely a reflection of the 1,359 children who were not assessed at the 2-year data collection, since only 8 children who had not been assessed at 9 months were assessed at 2 years. Frequency counts for completed assessments showed that nearly all who took the mental assessment also took the motor assessment. Due to delays in scheduling fieldwork, comparatively few children were assessed prior to 9 months or prior to 2 years of age, whereas many more were assessed as many as several months beyond the expected ages.

Given that the full set of BSID-II items had already been calibrated using the publisher standardization dataset, the possibility of using publisher item parameters to score ECLS-B observations was logically considered. Assuming that the ECLS-B data were to fit the publisher IRT model, all of the resulting EAP ability estimates could then be easily reported on the scale metric used by the publisher. This would obviate any need for an independent set of ECLS-B item calibrations or any kind of scale equating. Instead, BSF-R IRT ability estimates obtained with publisher calibrations could be used to calculate publisher IRT true scores, each of which would provide a model-based estimate of the BSID-II number-right raw score. Model-based estimates of raw scores or developmental index scores reported in the ECLS-B would then be directly compared with BSID-II results reported elsewhere.

[6] Item difficulty parameter estimates from separate calibrations align themselves along a straight line, indicating that a simple transformation of origin and scale is all that is needed to place one set of items on the same scale metric as the other item set. This constitutes a rigorous test of IRT parameter invariance properties.

Table 4-5. Cross-classification of number of children assessed/not assessed at 9-month and 2-year round of data collection, by BSF-R scale: 2001–02 and 2003–04

	2 years		
9 months	Not assessed[1]	Assessed	Total
Mental scale			
Total	1,300	8,900	10,200
Not assessed	#	#	#
Assessed	1,300	8,900	10,200
Motor scale			
Total	1,350	8,850	10,200
Not assessed	#	50	50
Assessed	1,350	8,800	10,150

Rounds to zero.
[1] Not assessed includes children who failed to complete two-thirds of the core item set and by decision of NCES were not scored. Detail may not sum to total due to rounding. Sample sizes have been rounded to the nearest 50.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

For this to work properly, it was first necessary to demonstrate that ECLS-B data conformed to the publisher IRT model. Implicitly, this would test the hypothesis that ECLS-B instruments, interviewers, and administration procedures, when used in a household interview setting, resulted in recorded responses that were consistent with those obtained with the full BSID-II when this was used in a clinical setting. Evidence supporting this hypothesis would provide a strong argument to support the validity of ECLS-B measures since these could then be shown to produce τ-equivalent results essentially identical to those obtained under clinical conditions. If the evidence *failed* to support this hypothesis, this finding would imply that ECLS-B instruments, interviewers, and procedures produced results that were inconsistent with those of the full BSID-II when used under clinical conditions. However, in this eventuality, it would still be possible to calibrate BSF-R items on a common-scale metric consistent with the BSID-II.

For this first experiment, only ECLS-B data collected using BSF-R instruments, interviewers, and administration procedures were used. ECLS-B data were scored directly using publisher item calibrations. Experiment results are reported in table 4-6. These results show that ECLS-B mental item responses are fairly inconsistent in the vicinity of the child's ability and very inconsistent on items far removed from the child's ability. With mean Infit indices ranging from 1.364 at 9 months to 1.513 at 2 years, this implies that, *from the BSID-II perspective*, there was considerable noise in the vicinity of child ability. Interpreted literally, this index shows that there is anywhere from 36 percent to 50 percent more random noise in ECLS-B data than would be expected had the data actually conformed to publisher item

calibrations. This finding was of critical importance because it showed that ECLS-B data failed to fit the response model at the point where they were most needed to estimate child ability.

Table 4-6. Mean fit indices for ECLS-B observations scored directly with publisher calibrations, by BSF-R scale and round of data collection: 2001–02 and 2003–04

ECLS-B subsample	Mean Squared Residual Fit Index	Mean fit values	
		BSF-R mental item subset	BSF-R motor item subset
9 months	Information-weighted mean squared residual goodness of fit—Infit	1.364	1.096
	Outlier-sensitive mean squared residual goodness of fit—Outfit	2.045	1.267
2 years	Information-weighted mean squared residual goodness of fit—Infit	1.513	1.081
	Outlier-sensitive mean squared residual goodness of fit—Outfit	2.204	1.125

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

There seems to be an even larger problem with items that were either too easy or too difficult for the child, resulting in misjudgments when these item responses were recorded. Here, Outfit values slightly in excess of 2.0 that might not cause great concern with individual observations were considered unacceptably large for a group mean.

Both Infit and Outfit indices were more satisfactory in the case of the motor scale, however, the Infit means were close to the critical value of 1.1. Outfit indices for the motor scale were somewhat larger. Collectively, these indices showed that ECLS-B data do not fit the publisher response models as well as one would have liked. These findings suggest that the ECLS-B data required their own set of item calibrations and an appropriate equating design so that test results could be reported on the publisher scale metric.

Outfit frequencies for the ECLS-B sample shown in table 4-7 confirm what has already been stated regarding mean fit indices. Although model fit on the mental test was either excellent or acceptable for the majority of observations when these were scored using publisher calibrations, model fit was either problematic or unacceptable for large numbers of other observations. Motor fit was generally satisfactory at 2 years but not entirely satisfactory for an appreciable number of observations at 9 months.

Table 4-7. Number and percentage distribution of ECLS-B sample observations scored with publisher calibrations, using combined 9-month and 2-year BSF-R item sets, by level of outfit and scale: 2001–02 and 2003–04

		BSF-R mental scale		BSF-R motor scale	
Fit	Level of outfit	Number	Percent	Number	Percent
9 months					
Total	$0 \leq y < \infty$	10,200	100.0	10,200	100.0
Excellent	$0 \leq y < 1$	4,100	39.9	5,950	58.3
Acceptable	$1 \leq y < 3$	4,050	39.9	3,350	33.1
Problematic	$3 \leq y < 5$	1,200	11.9	600	6.1
Unacceptable	$5 \leq y < \infty$	850	8.3	250	2.6
2 years					
Total	$0 \leq y < \infty$	8,950	100.0	8,900	100.0
Excellent	$0 \leq y < 1$	1,600	17.7	4,050	45.7
Acceptable	$1 \leq y < 3$	5,200	58.1	4,750	53.4
Problematic	$3 \leq y < 5$	1,500	16.6	50	0.8
Unacceptable	$5 \leq y < \infty$	700	7.7	#	0.1

Rounds to zero.
NOTE: Outfit = outlier-sensitive mean squared residual goodness of fit. Detail may not sum to total because of rounding. Sample sizes have been rounded to the nearest 50.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

These findings are understandable given the unique set of circumstances encountered in ECLS-B fieldwork. It was reasonable to expect that a household interview setting would be different from a clinical setting. In a household interview, space sufficient for an assessment is often lacking, lighting is frequently inadequate for careful observation, and there is little opportunity to regulate the conditions in which to conduct an assessment. Nor was it expected that laymen interviewers with limited experience in child development would be able to replicate clinical measures in the absence of the specialized knowledge and expertise expected of trained clinicians. Although it had already been demonstrated that BSF-R item subsets were capable of providing unbiased estimates of measurement outcomes obtained with the full BSID-II, it was expected that the ECLS-B experience would yield somewhat different results due to fieldwork conditions and adaptations introduced to simplify administration of some of the items. However, even in these circumstances, it was possible to calibrate ECLS-B item subsets on the publisher scale metric.

4.4 BSF-R Conditioning

The task of calibrating the BSF-R item subsets on a common scale metric using ECLS-B data was then considered. Table 4-8 reports frequency counts for both the ECLS-B sample and publisher standardization dataset, broken down by months of age. The ECLS-B sample contained large numbers of assessments at two points in time. This age breakdown highlights ECLS-B observations intended for assessments at 9 months and 2 years of age. Due to the usual complexity of scheduling interviews, the age distribution on both occasions was highly skewed. Although some children in the sample were easily located and promptly interviewed, many others could only be interviewed after a series of scheduling delays. Thus, the age distribution at 9 months became skewed and waned at about the same age where 2-year assessments began. The challenge was to find a satisfactory means of placing scores for all these children on a common scale metric. A consistent scale metric is required for the longitudinal analysis of ECLS-B data.

With age distributions such as these, there was little opportunity to use item linkages between the 9-month and 2-year BSF-R tests to establish a common vertical scale. In fact, there were only two common items linking the two BSF-R mental tests and only a slightly more expressive number of eight item linkages between the two motor scales. In any case, these were fairly atypical items that served as ceiling items at 9 months and basal items at 2 years. With relatively few items and smaller numbers of respondents for these items, there was little opportunity to develop an equating design based on common item linkages between BSF-R tests at 9 months and 2 years.

By contrast, the strength of the publisher dataset lay not so much in the number of standardization sample and other observations in this dataset, but rather with the strategic positioning of these observations over such an extensive range of infant ability found between 1 and 42 months of age.[7] This design assures the largest possible number of observations linking adjacent age item sets. In fact, there is an average of 633 (± 235 observations) for each mental item, ranging from a minimum of 257 to a maximum of 1,130 observations used to calibrate the publisher mental scale. There is an average of 564 (± 228 observations) for each motor item, ranging from a minimum of 174 to a maximum of 1,031 observations used to calibrate the publisher motor scale. For both the mental and motor scale, these observations provide a solid string of items calibrated across the widest possible range of infant development, assured by the extraordinary age variation found between 1 and 42 months of age.

[7] The publisher's standardization sample contained 100 observations for each of 17 selected age groups. The 1,700 standardization sample observations are complemented by an additional 1,239 observations of other infants. The higher percentage of basal items administered to this second group suggests that perhaps 4.5 percent of these observations show some evidence of deficient ability. The standardization sample and other observations in the combined sample of 2,939 observations were used to calibrate publisher item sets, affording the largest possible number of item responses linking adjacent age item sets.

Table 4-8. Frequency count for ECLS-B longitudinal sample and publisher standardization dataset, by test and months of age: 2001–02 and 2003–04

Months of age	Mental scale					Motor scale				
	ECLS-B		Publisher			ECLS-B		Publisher		
	9-month	2-years	Stdz	Other	Total	9-month	2-years	Stdz	Other	Total
Total	10,200	8,900	1,700	1,250	22,050	10,150	8,850	1,700	1,250	21,950
1	†	†	100	50	150	†	†	100	50	150
2	†	†	100	50	150	†	†	100	50	150
3	†	†	100	50	150	†	†	100	50	150
4	#	†	100	#	150	#	†	100	#	150
5	#	†	100	#	100	#	†	100	#	100
6	50	†	100	#	150	50	†	100	#	150
7	150	†	†	†	150	150	†	†	†	150
8	850	†	100	50	950	850	†	100	50	950
9	3,400	†	†	†	3,400	3,350	†	†	†	3,350
10	2,650	†	100	#	2,800	2,650	†	100	#	2,800
11	1,300	†	†	†	1,300	1,300	†	†	†	1,300
12	700	†	100	150	950	700	†	100	150	950
13	450	†	†	†	450	450	†	†	†	450
14	250	†	†	†	250	250	†	†	†	250
15	150	1	100	100	350	150	#	100	100	350
16	100	†	†	†	100	100	†	†	†	100
17	100	†	†	†	100	100	†	†	†	100
18	50	#	100	50	200	50	#	100	50	200
19	#	#	†	†	#	#	#	†	†	#
20	#	50	†	†	50	#	50	†	†	50
21	#	250	100	100	450	#	250	100	100	450
22	#	450	†	†	450	#	400	†	†	450
23	†	1,450	†	†	1,450	†	1,450	†	†	1,450
24	†	4,150	100	200	4,500	†	4,150	100	200	4,450
25	†	1,600	†	†	1,600	†	1,600	†	†	1,600
26	†	550	†	†	550	†	550	†	†	550
27	†	200	100	50	350	†	200	100	50	350
28	†	100	†	†	100	†	100	†	†	100
29	†	50	†	†	50	†	50	†	†	50
30	†	#	100	100	250	†	#	100	100	250
31	†	#	†	†	#	†	#	†	†	#
32	†	#	†	†	#	†	#	†	†	#
33	†	#	†	†	#	†	#	†	†	#
34	†	#	†	†	#	†	#	†	†	#
36	†	#	100	150	250	†	#	100	150	250
37	†	†	†	†	†	†	#	†	†	#
38	†	#	†	†	#	†	#	†	†	#
42	†	†	100	50	150	†	†	100	50	150

† Not applicable.
# Rounds to zero.
NOTE: Stdz: Publisher standardization sample. Other: Nonstandardization sample observations included in the publisher dataset. BSID ages for ECLS-B observations rounded to nearest whole number. Detail may not sum to total because of rounding. Sample sizes have been rounded to the nearest 50.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04; Publisher BSID-II standardization dataset, The Psychological Corporation, 1993.

Given the ECLS-B design and available publisher data, the best approach to build a comparable vertical scale metric to link the 9-month and 2-year samples was to use the publisher standardization dataset as a bridge linking the two data collections. Frequency counts for the publisher standardization sample are also reported in table 4-8. The standardization sample had fewer observations than the ECLS-B sample, but these were strategically positioned at regular intervals all the way from 1 to 42 months of age. Collectively, across all age groups, the standardization sample contained information on the full complement of 178 mental and 111 motor items. By contrast, designed for use at 9 months and 2 years, BSF-R instruments were based on item subsets that include 66 mental and 59 motor items. Each of these items provided a linkage between BSF-R short forms and the full BSID-II. These linkages were used to establish a consistent scale metric between BSF-R tests at 9 months and 2 years of age and to establish an ECLS-B scale metric that was consistent with publisher documentation.

Scale equating in the ECLS-B was approached in the context of a Non-Equivalent groups with Anchor Test (NEAT) design, having both internal and external anchor items (von Davier and von Davier 2004). The NEAT design envisions two populations *P* and *Q*, each represented by samples of examinees that take two different tests. The sample from population *P* takes test *Y*, while the sample from population *Q* takes test *X*. Each of these tests contains a subset of common items *V*. This formulation was appropriate in the present context, where the publisher standardization sample was drawn from *P* and the ECLS-B longitudinal sample was drawn from *Q*. The challenge was to identify items in *V* that act as internal anchor items.[8] In the first experiment reported earlier, using publisher item parameter calibrations, all 66 mental and all 59 motor items were effectively placed in *V*, with no remaining items in *Y* or *X*. The *V* item parameters remained fixed, effectively making these the strongest possible anchor items. This experiment revealed that ECLS-B data were substantially inconsistent with publisher item calibrations.

The second experiment was based on concurrent item calibrations obtained using both ECLS-B and publisher data in a single run. This new design is shown in table 4-9, where all common items were placed in *V* and any remaining items in *Y*.[9] Items from *Y*, *V*, and *X* can be calibrated simultaneously by coding item responses that were not observed and remained missing by design as "not

[8] Internal anchor items are items internal to the test that serve to set the scale metric. External anchor item belong to an external test being used to set the scale metric.

[9] The V item set is a subset of the X item set at this stage, awaiting subsequent analysis, whereupon some of the V items will be transferred to X (with no corresponding publisher item in the Y set). The NEAT design anticipates the second stage of analysis, when some V items will have been transferred to X.

presented." Several features of this design should be noted. The first feature is that standardization sample observations were calibrated concurrently with ECLS-B observations, yielding a new set of item parameters. These are referred to as ECLS-B item calibrations in order to distinguish them from the original set of publisher item calibrations. The second feature is that BSID-II items not administered in ECLS-B (*Y* in the table) had item parameters that remain fixed so that they effectively acted as external anchor items. These items were positioned across the full range of ability and not just at the extremes of the scale. Parameters for these items remained unchanged during item calibration. The third feature of this design is that parameters for BSF-R *V* items were allowed to float until they found their positions in parameter space relative to the *Y* item parameters that remained fixed.

Table 4-9. First Non-Equivalent groups with Anchor Test (NEAT) design, by item sets: 2001–02 and 2003–04

Population		NEAT item sets			Total
		Y	*V*	*X*	
Mental					
P	Publisher	112	66	†	178
Q	ECLS-B	†	66	0	66
Motor					
P	Publisher	52	59	†	111
Q	ECLS-B	†	59	0	59

† Not applicable.
NOTE: NEAT = Non-Equivalent groups with Anchor Test design; *Y* = external anchor items with fixed item parameters; *V* = internal conditioned items; *X* = other BSF-R items; *P* = publisher standardization dataset; *Q* = ECLS-B sample.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04; Publisher BSID-II standardization dataset, The Psychological Corporation, 1993.

Standardization dataset observations drive the equating design since they were primarily scored on strategically positioned BSID-II items whose parameters remained fixed but also on *V* items whose parameters were allowed to float. Collectively, the standardization dataset observations acted as a set of Bayesian priors on the BSF-R item parameters, coaxing these parameters into positions that were

consistent with the fixed set of parameters in *Y*.[10] Admittedly, this was a weak equating design because there were no internal anchor items in the ECLS-B that could act as knots to fix the scale. Instead, calibration relied exclusively on the conditioning provided by publisher standardization dataset observations acting as a stabilizing counterweight. The publisher test effectively played the role of an external anchor test.

This design used common item linkages and Bayesian priors with the full set of BSID-II items to calibrate ECLS-B items in a single run using all 22,391 observations from both the ECLS-B and publisher's standardization datasets. The full set of BSID-II mental items was calibrated using 2,938 publisher standardization dataset observations (13 percent of the total), together with 19,117 (87 percent) observations from the ECLS-B sample. This may overstate the importance of the publisher standardization dataset in one sense and understate it in another. If publisher observations *over the age range covered by ECLS-B sample* were considered, then there were only 1,724 publisher observations of comparable age, which is about 8 percent as large as the total number of ECLS-B observations. On the other hand, BSID-II observations had proportionally more weight in the tails of the ECLS-B ability distributions, where there were relatively few ECLS-B items and where the standardization dataset observations were only needed to help calibrate the BSF-R basal and ceiling item sets. While the numbers of observations involved in calibrating the full set of BSID-II motor items were slightly different, the proportions involved were virtually identical. While publisher observations represented 13 percent of the total combined sample, only 8 percent of those observations were of comparable age.

Full BSID-II item sets with 178 mental items and all 111 motor items were used in the concurrent calibration, including many items that were not present in any of the ECLS-B short forms.[11] There were numerous item linkages relating ECLS-B short forms to the backbone of BSID-II items with

---

[10] Bayesian priors are probability distributions that are used to condition poorly fitting parameters during estimation. The distributions impose a penalty on improbable parameter values. For the analyses described here, instead of imposing Bayesian priors on individual IRT item parameters, well-conditioned publisher standardization dataset observations were added to the ECLS-B sample during item calibration to accomplish this same purpose. In this role, standardization dataset observations condition the full set of data.

An alternative approach to using observations from the standardization data set would have been to estimate a fully Bayes model with informative prior distributions. When prior distributions are based on publisher item parameter estimates, progressively stronger priors yield parameter estimates that look increasingly more like those obtained with the publisher's dataset. However, the goal of this calibration was to identify a subset of ECLS-B items that were consistent with their publisher counterpart items so that a consistent scale metric could be obtained. This goal was better accomplished with the data augmentation approach described here. Publisher-ECLS-B comparisons subsequently identified a subset of ECLS-B items that were consistent with the corresponding publisher items. These items then became the only direct link between ECLS-B and the publisher in the second stage of this analysis, when item parameters were again estimated, effectively setting the scale.

[11] Conceivably, some of the BSID-II items were too easy for the ECLS-B population and could possibly have been left out of the concurrent calibrations. However, there was no harm done by including these very easy items in the calibrations since their item parameters remained fixed. Items in the extremes of each scale play a much more limited role as external anchor items.

fixed parameters. By using the full set of BSID-II items, it was possible to see how ECLS-B items line up with publisher items across nearly 20 population standard deviations of ability between 1 and 42 months of age. Nor did this exhaust the benefits of this design. It was also possible to separately score the standardization sample observations alternately using either publisher or ECLS-B item calibrations. By scoring the same observations twice with different sets of item parameters, the resulting scale score distributions can be compared, showing the extent to which ECLS-B item parameters replicated the results obtained with the full set of publisher items.

Fit indices for ECLS-B observations scored with the new set of ECLS-B item parameters are reported in table 4-10. All fit indices fell well below the critical value of 1.1. The data fit the IRT model exceptionally well. All indices fell below unity, reflecting the redundancy of information found in assessor-imputed item responses. In general, ECLS-B data were consistent with the item response model. In particular, Outfit indices show that the problem with inconsistent responses far removed from the child's ability level had now been resolved. However, a certain amount of redundant information became apparent throughout, but was most expressive in the motor scale at 9 months. This redundancy was no great cause for concern, because it does not affect scoring, but Infit < 1 implied that IRT standard errors would be somewhat underestimated.

Table 4-10 shows that ECLS-B data fit the response model obtained with concurrent calibration using this first equating design. Sample frequencies reported in table 4-11 confirm this, where virtually all of the observations on both occasions exhibit person fit that was either excellent or acceptable. The improvement in the mental scale on both occasions was most striking. On the motor scale, person fit improved substantially at both 9 months and 2 years. There seems to be little question that the concurrent item calibrations succeeded in producing response models that were consistent with the ECLS-B data.

Table 4-10. Mean fit indices for ECLS-B observations scored after concurrent calibration, by BSF-R scale and round of data collection: 2001–02 and 2003–04

ECLS-B subsample	Mean squared residual fit index	BSF-R mental item subset	BSF-R motor item subset
9 months	Information-weighted mean squared residual goodness of fit—Infit	0.961	0.859
	Outlier-sensitive mean squared residual goodness of fit—Outfit	0.943	0.837
2 years	Information-weighted mean squared residual goodness of fit—Infit	0.939	0.919
	Outlier-sensitive mean squared residual goodness of fit—Outfit	0.911	0.899

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

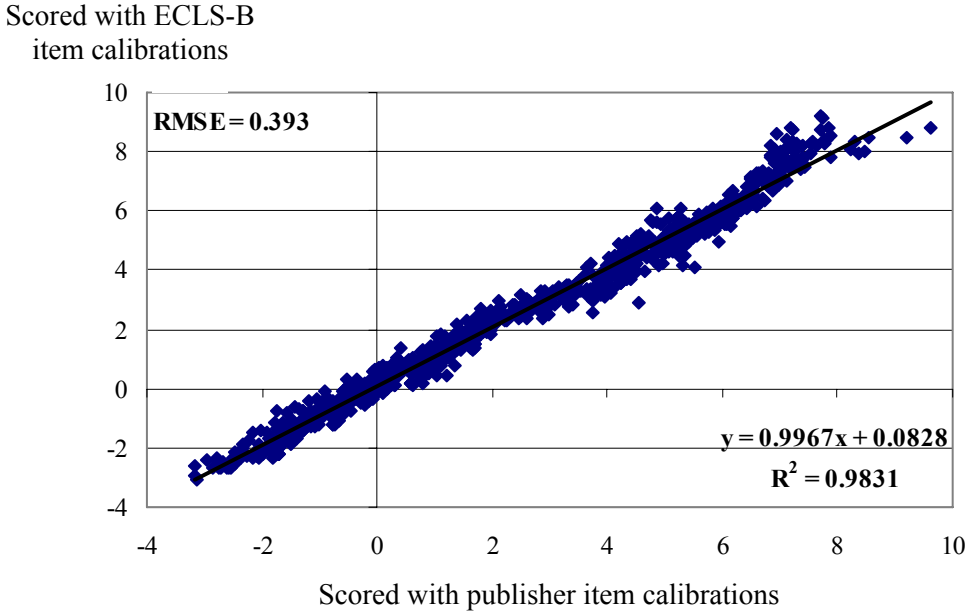
Table 4-11. Number and percentage of ECLS-B sample children, by level of fit for the BSF-R scales, after concurrent calibration, by BSF-R scale and round of data collection: 2001–02 and 2003–04

		BSF-R mental subset		BSF-R motor subset	
Fit	Level of outfit	Number	Percent	Number	Percent
9 months					
Total	$0 \leq y < \infty$	10,200	100.0	10,150	100.0
Excellent	$0 \leq y < 1$	6,800	66.8	7,700	75.6
Acceptable	$1 \leq y < 3$	3,350	32.7	2,100	20.8
Problematic	$3 \leq y < 5$	50	0.5	250	2.6
Unacceptable	$5 \leq y < \infty$	#	#	100	1.0
2 years					
Total	$0 \leq y < \infty$	8,900	100.0	8,850	100.0
Excellent	$0 \leq y < 1$	5,950	66.9	5,800	65.8
Acceptable	$1 \leq y < 3$	2,950	33.0	3,000	34.2
Problematic	$3 \leq y < 5$	#	0.1	#	0.1
Unacceptable	$5 \leq y < \infty$	#	#	#	#

Rounds to zero.
NOTE: Frequencies may differ slightly from table 4-8 due to weighting and rounding considerations. Outfit = outlier-sensitive mean squared residual goodness of fit. Detail may not sum to total because of rounding. Sample sizes have been rounded to the nearest 50.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

To further examine the quality of item calibrations, it was necessary to consider how the BSF-R scoring performed in relation to the full BSID-II. Figures 4-4 and 4-5 show how the two item sets compared when scoring the same set of 900 standardization sample observations described earlier. The quality of fit was not quite as tight as it was in the previous set of figures, when publisher item parameters were used with both item sets, but it was still respectable. Both r^2 coefficients were quite high, although again this was largely a reflection of the enormous range of ability. A better measure of fit was provided by the root mean squared residuals shown at the upper left of the figures, each expressed in population standard deviation units. Generally speaking, the mental scores were accurate to within RMSE = 0.393 of a population standard deviation, while motor scores are accurate to within RMSE = 0.345, when using the full BSID-II as a standard for comparison. In practice, the BSF-R scales were not altogether as precise as publisher item parameters originally had suggested.

Figure 4-4. Expected a posteriori ability estimates for standardization sample observations scored first with publisher item calibrations (full BSID-II mental items) and then scored with ECLS-B item calibrations (BSF-R) following concurrent item calibration: 2001–02 and 2003–04

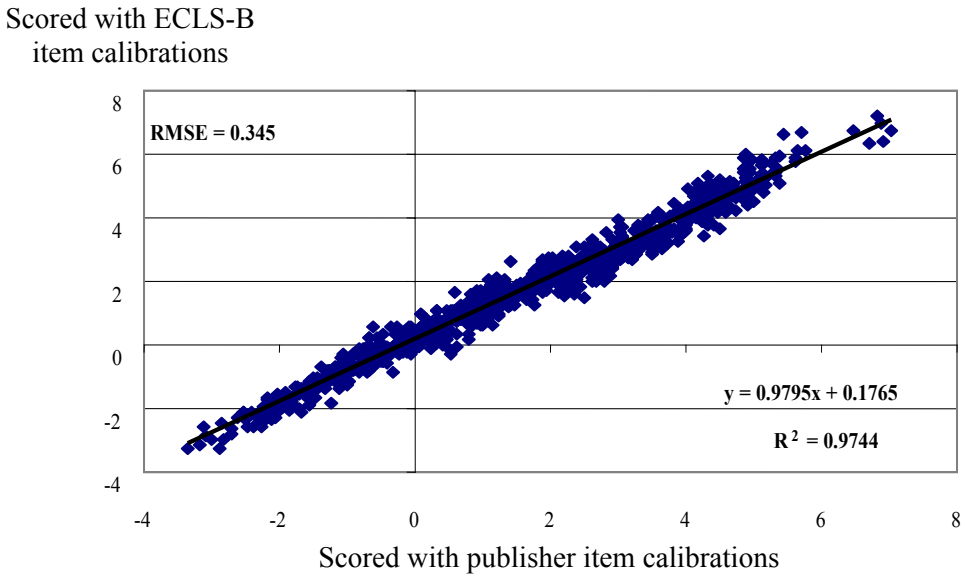


NOTE: RMSE = root mean squared error; R^2 = proportion of variance in the data explained by the regression equation.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

To the extent that BSF-R item subsets produced results similar to those obtained with the full BSID-II, expected a posteriori (EAP) ability estimates aligned themselves closely with a straight line

having an origin of zero and slope of unity.[12] Regression lines in both figures passed close to the scale origin, which coincided with the average ability of 12-month-old infants, and again both slope coefficients were close to unity. Although the BSF-R instruments do not provide anything like τ-equivalent tests, because BSF-R and BSID-II item parameters are often inconsistent, they can still be calibrated on the publisher scale metric.

Figure 4-5. Expected a posteriori ability estimates for standardization sample observations scored first with publisher item calibrations (full BSID-II motor items) and then scored with ECLS-B item calibrations (BSF-R) following concurrent item calibration: 2001–02 and 2003–04



NOTE: RMSE = root mean squared error; R^2 = proportion of variance in the data explained by the regression equation.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

4.5 BSF-R Equating Design

The new set of ECLS-B item parameters was compared with publisher item parameters so that a set of BSF-R items was identified to serve as internal anchor items. Differential item function (DIF) analysis was used to identify BSF-R items that were inconsistent with publisher item parameters. An item

[12] IRT produces a likelihood function for the response vector at each ability level, $L(X|\theta)$. However, the objective in testing is to obtain an estimate of the probability of an ability given the person's response vector, $P(\theta|X)$. This is known was the expected a posteriori (EAP) probability. Bayes' theorem is used to obtain $P(\theta|X)$, based on the relationship:

$$P(\theta|X) \propto L(X|\theta)\, P(\theta).$$

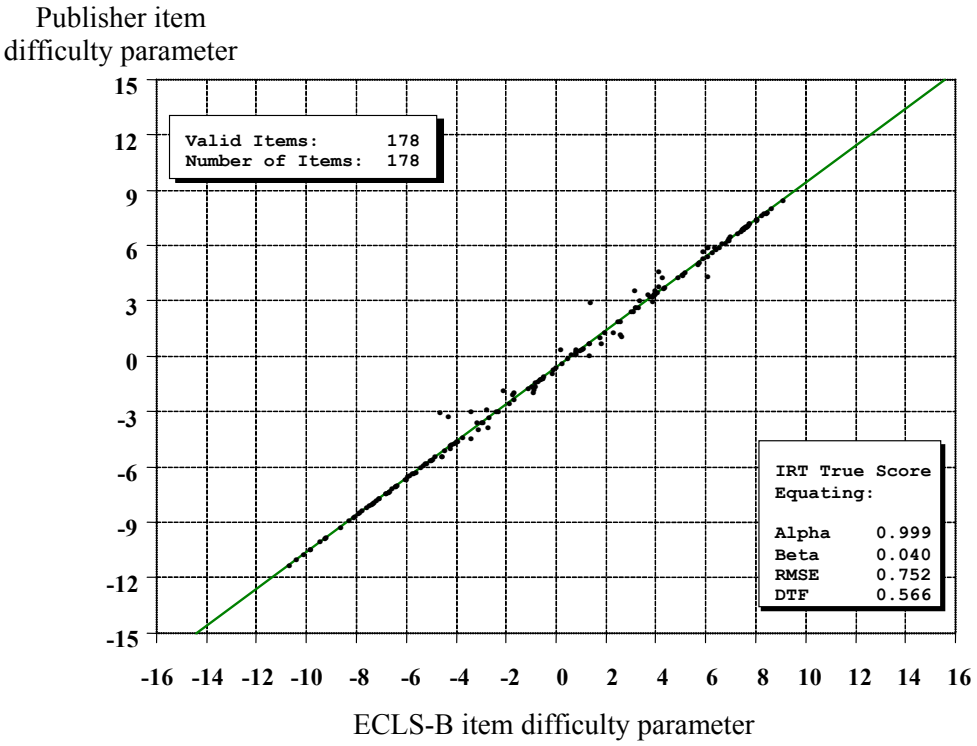
Maximum likelihood is found at the point where this function peaks, also known as the EAP ability estimate. In this sense, the EAP is simply the best available estimate of the person's ability.

has been said to exhibit DIF "if individuals of the same ability, but from different groups, do not have the same probability of getting the item right" (Hambleton, Swaminathan, and Rogers 1991, p. 110). In the present context, DIF was used in a somewhat different sense to investigate differences between observers and settings rather than differences between population subgroups.

DIF analysis was used to compare BSF-R instruments used by laymen as part of a household survey interview with use of the full BSID-II by trained professionals in a clinical setting. One looks for DIF affecting individuals of the same ability in two populations P and Q, respectively represented by ECLS-B and the publisher standardization samples. At issue was whether any of the BSF-R items behaved substantially differently in the ECLS-B than they did in BSID-II. The item might still be used in scaling and scoring, but it would play no further role in setting the scale metric. Where this was found to be true, the item was not used to equate the BSF-R with the BSID-II. DIF analysis was used to identify inconsistent items and a subset of highly consistent items that could serve as internal anchor items.

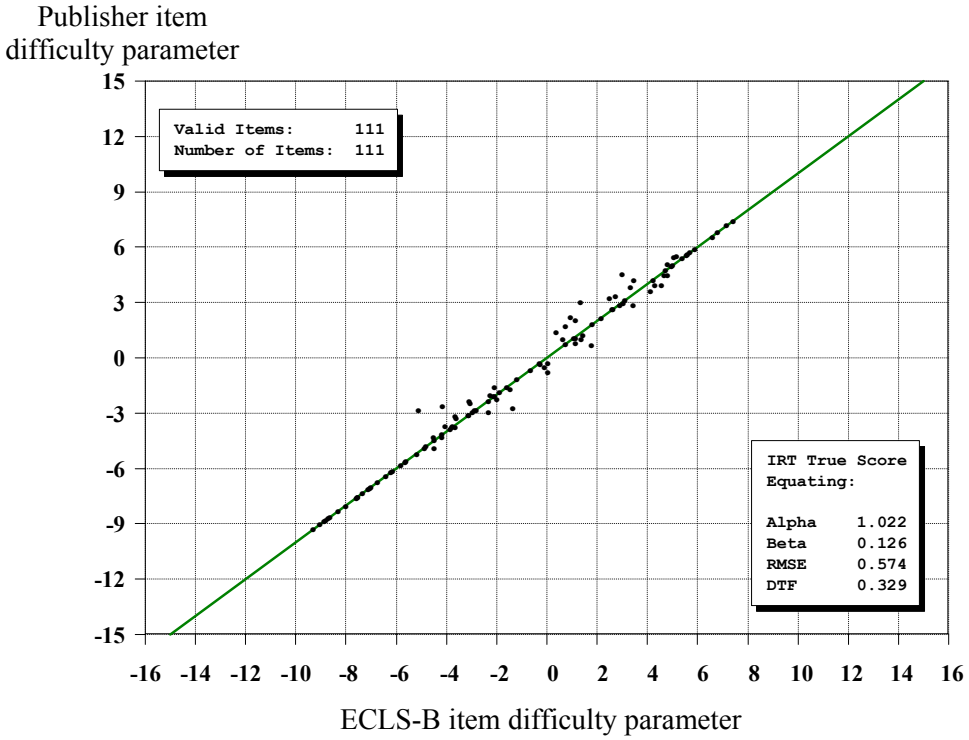
A conceptual grasp of differential item functioning in this context was provided by plotting publisher and ECLS-B item difficulty parameters b_j along perpendicular axes, as shown in figures 4-6 and 4-7. Units of measurement shown in the figures represent population standard deviation units, where publisher observations 12 months of age form the $N(0,1)$ reference population that defines the graph origin and scale. In IRT, item difficulty and person proficiency parameters share a common scale. Bearing in mind this scale metric, there were several instances where the item difficulty parameters of BSF-R items diverge from those of publisher items. This judgment was made empirically to make sure a sufficient number of items would be maintained. These items behaved very differently in the two settings, so much so that they could be considered to be entirely different items lacking any counterpart found among publisher items.

Figure 4-6. ECLS-B mental item difficulty parameters b_j on the x axis plotted against the corresponding publisher difficulty parameter on the y axis after concurrent calibration: 2001–02 and 2003–04



NOTE: Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Alpha = linear transformation of scale; Beta = linear transformation of origin; RMSE = root mean squared error; DTF = differential test functioning.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Figure 4-7. ECLS-B motor item difficulty parameters b_j on the x axis plotted against the corresponding publisher difficulty parameter on the y axis after concurrent calibration: 2001–02 and 2003–04



NOTE: Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Alpha = linear transformation of scale; Beta = linear transformation of origin; RMSE = root mean squared error; DTF = differential test functioning.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

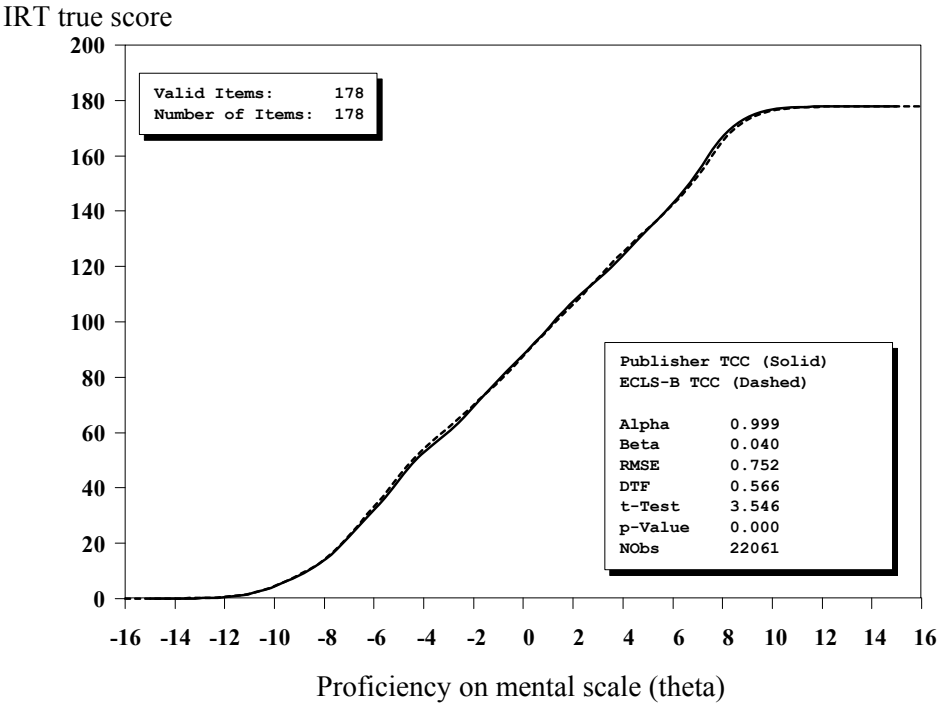
Items showing little or no evidence of DIF have item difficulty parameters b_j located along the diagonal line extending from the lower left to upper right of each figure. Obvious examples include difficulty parameters for *Y* items from the equating design, whose fixed parameters remain the same on both axes. These items are clearly visible lying precisely on the diagonal, especially toward the extremes of the figures, where no BSF-R items are located. At these extremes, BSID-II items appropriate for 1 month of age appear on the diagonal at the lower left of each figure, while BSID-II items appropriate for 42 months appear at the upper right. These item parameter difficulty values appear precisely along the diagonal because item parameters remained unchanged during concurrent calibration. These item parameters convey a clear sense of the central tendency of each scale, which coincides with a 45°-angle line.

Moving closer to the center of each figure, a mixture of BSF-R *V* and BSID-II *Y* items is encountered, where some points lie close to the diagonal while others are farther removed. Here, too, one finds interstitial *Y* items with fixed parameters lying exactly along the diagonal in the midst of other ECLS-B items. Somewhat farther away from the diagonal, one finds item difficulty parameters that are still relatively close to the diagonal. These represent BSF-R items that behaved almost exactly like the corresponding publisher items and were thus worth considering as internal anchor items. Far removed from the diagonal are BSF-R items that are highly inconsistent with the corresponding publisher items. These points represent items that appeared to be much harder or easier in the ECLS-B administration than in the standardization dataset. This inconsistency suggested that these items should play no further role in equating. These items should be considered unique to BSF-R as if they had no counterpart in BSID-II.

The equating constants reported in the box at the lower right of each figure were based on IRT true-score equating (Stocking and Lord 1983). This method used test characteristic curves (TCCs) to align a source test such as BSF-R with a target test such as the publisher. Test equating was accomplished by finding a linear transformation of origin and scale that minimized the weighted area between the two TCCs, as shown in figures 4-8 and 4-9. The equating constants obtained with IRT true-score equating represent the linear transformation that best aligns the two tests. This included a transformation of origin (beta) and of scale (alpha). The figures report values for beta close to zero and values for alpha close to unity. This shows that no true-score equating was required after concurrent calibration.

A measure of differential test functioning (DTF) is the DTF index, where smaller values represent the extent to which the ECLS-B and publisher tests measured the same trait and larger values represent the extent to which the pair of tests fail to align. The DTF index is reported in squared units. The root mean square of this value or RMSE represents the average number-right raw score units separating the two TCCs displayed on the y axis in figures 4-8 and 4-9 (*not* the population standard deviation units displayed on the x axis). The RMSE values in the ECLS-B were both found to be relatively small in relation to the 178 mental and 111 motor items in each respective test. Although the ECLS-B and publisher tests were well aligned over an extensive range of ability, there were many items that were not closely aligned on the two tests. In this case, they should play no further role in *test equating*, although they were retained for scoring.

Figure 4-8. Test characteristic curves (TCCs) for BSF-R and BSID-II mental scales after concurrent calibration: 2001–02 and 2003–04

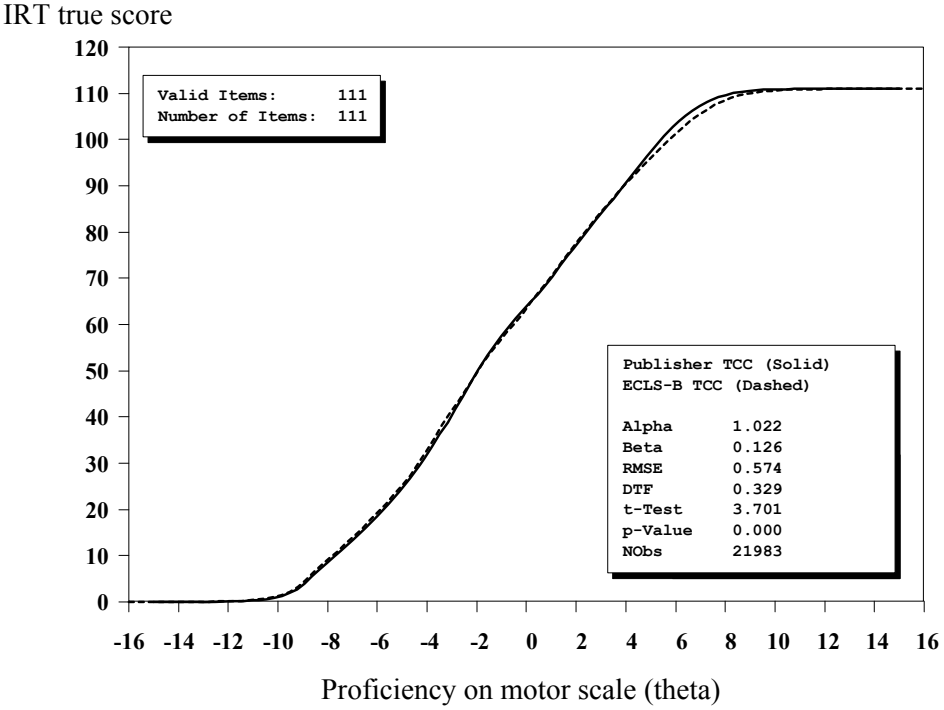


NOTE: Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Alpha = linear transformation of scale; Beta = linear transformation of origin; RMSE = root mean squared error; DTF = differential test functioning; t-Test = statistical test of the difference between the two curves shown in the figure; p-Value = probability of the results of the t-Test; NObs = number of observations.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Concurrent calibration assures that the two tests are generally well aligned. This alignment is seen both in the diagonal lines shown in figures 4-6 and 4-7, and in the close fit between the two TCCs shown in figures 4-8 and 4-9.[13] In these circumstances, the noncompensatory index (NC-DIF) provides an appropriate measure of individual item DIF (Raju, van der Linden, and Fleer 1995). Parametric IRT models were used to calculate the NC-DIF index. NC-DIF indices represent the weighted mean squared distance between item characteristic curves (ICCs) obtained with separate calibrations. The square-root of the NC-DIF index is thus the weighted average distance separating the two ICCs.

[13] T-test results indicate that there is a significant difference between the TCC based on publisher data and the TCC based on ECLS-B data. However, with large samples such as ECLS-B, relatively small differences will almost always be statistically significant. In this case, the magnitude of the difference is relatively small: 0.752 raw score points on a test that includes 178 items. In practice, children would not be administered all 178 items, but rather only a subset of about 35 items. Out of 35 items, a difference of 0.752 raw score points is 2% to either side of the publisher raw score standard.

Figure 4-9. Test characteristic curves (TCCs) for BSF-R and BSID-II motor scales after concurrent calibration: 2001–02 and 2003–04

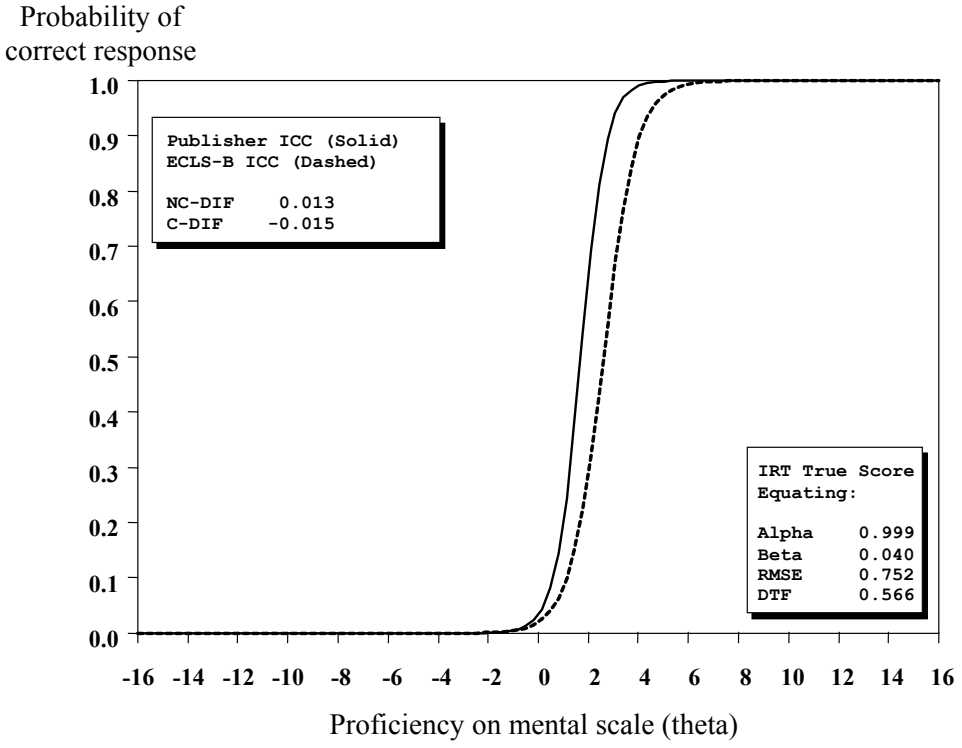


NOTE: Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Alpha = linear transformation of scale; Beta = linear transformation of origin; RMSE = root mean squared error; DTF = differential test functioning; t-Test = statistical test of the difference between the two curves shown in the figure; p-Value = probability of the results of the t-Test; NObs = number of observations.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

An analysis was conducted to identify BSF-R items exhibiting three different levels of DIF. RMSE < 0.02 were used to identify BSF-R items exhibiting low DIF. These BSF-R items were virtually identical to their counterpart items in BSID-II and were used as internal anchor items with fixed parameters identical to the publisher standard. A mid DIF level with RMSE in the range 0.02 < RMSE < 0.08 was used to identify BSF-R items that played a more limited role in scale equating. These items continued to receive conditioning from publisher standardization dataset observations acting as a stabilizing counterweight. Finally, a residual high DIF level was used for all remaining BSF-R items considered to have no counterpart among BSID-II items. The mean RMSE value in this category was 0.12, implying that ECLS-B and publisher item characteristic curves were fully separated by 12 percentage points.

Figure 4-10 shows an example of an item exhibiting appreciable DIF. Although the two ICCs run broadly parallel to one another, the population-weighted mean vertical distance between them is 0.126 or almost 13 percentage points. Although the ECLS-B item was strongly discriminating, with an item discrimination parameter a = 0.856, since it had NC-DIF > 0.08, such items were subsequently disregarded for purposes of scale equating.

Figure 4-10. Item characteristic curves (ICCs) for mental item MEN110 (Names one object) on BSF-R and BSID-II: 2001–02 and 2003–04



NOTE: BSF-R Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Alpha = linear transformation of scale; Beta = linear transformation of origin; RMSE = root mean squared error; DTF = differential test functioning; NC-DIF = noncompensatory DIF; C-DIF = compensatory DIF.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

On this basis, ECLS-B items in each of the three DIF levels were identified. Item frequencies for these levels are reported in table 4-12. The table shows that 10 (15 percent) of the mental and 8 (14 percent) of the motor items show virtually no DIF. In all important respects these BSF-R items were identical to their BSID-II counterparts. Another 29 (44 percent) of the mental and 31 (52 percent) of the motor items exhibit tolerable levels of DIF. These items were in the same general vicinity of their BSID-II counterpart items and thus performed as expected. However, 27 (41 percent) mental and 20 (34

percent) motor items performed much differently under household survey conditions than they would have been expected to perform in BSID-II under clinical conditions.

Table 4-12. Frequency count and percentage of BSF-R item parameters (using ELCS-B combined 9-month and 2-year data) for NEAT design items exhibiting low, medium, and high levels of DIF when compared with BSID-II item parameters (using publisher data): 2001–02 and 2003–04

		Mental scale		Motor scale	
DIF level	NEAT item set	Number	Percent	Number	Percent
Total		66	100.0	59	100.0
Low	V_A	10	15.2	8	13.6
Mid	V_B	29	43.9	31	52.5
High	X	27	40.9	20	33.9

NOTE: NEAT = Non-Equivalent groups with Anchor Test design; V_A = internal anchor items with fixed item parameters; V_B = internal conditioned items; X = other BSF-R items. Detail may not sum to total because of rounding.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Items exhibiting appreciable DIF were examined to see why they may have been sensitive to changes in setting or administrative skill. In a general sense, one could say that many of these items placed excessive demand on interviewer assessors to infer infant intentionality, for example, MEN059—Manipulates Bell, Showing Interest in Detail or Men131—Attends to Story. While it was possible to teach scoring criteria, it was not always possible to teach the interpretation of signs that were sometimes required before a response could be scored. Many ECLS-B interviewers lacked prior experience in child development, and there was no available means to provide them with ready made experience that would enable them to interpret what they observed so as to determine whether a child's response was clearly intentional or not.

The majority of BSF-R items performed in ECLS-B as they would be expected to perform under the best clinical conditions. This too was anticipated since some of the items could be scored objectively, leaving little or no margin for interpretation, for example, MEN089—Puts Six Beads in Box or MEN126—Names Three Objects. The challenge rather was to find a satisfactory procedure for identifying these best performing items.

Results provided by the DIF analysis were used to reformulate the equating design so that BSF-R item calibrations would become more consistent with BSID-II. Those items with appreciable DIF were effectively excluded from equating by reformulating the equating design as shown in table 4-13.

Excluding these items from equating was accomplished within the NEAT framework by transferring BSF-R items with large DIF from item set *V* to *X*. Whereas in the previous concurrent calibration there were no *X* items at all in the design, after this reformulation, there were 27 such items on the mental and another 20 on the motor scale.

Table 4-13. Second Non-Equivalent groups with Anchor Test (NEAT) design: 2001–02 and 2003–04

Mental						
		NEAT item sets				
Population		*Y*	V_A	V_B	*X*	Total
P	Publisher	139	10	29	†	178
Q	ECLS-B	†	10	29	27	66
Motor						
		NEAT item sets				
Population		*Y*	V_A	V_B	*X*	Total
P	Publisher	72	8	31	†	111
Q	ECLS-B	†	8	31	20	59

† Not applicable.
NOTE: NEAT = Non-Equivalent groups with Anchor Test design; *Y* = external anchor items with fixed item parameters; V_A = internal anchor items with fixed item parameters; V_B = internal conditioned items; *X* = other BSF-R items; *P* = publisher standardization dataset; *Q* = ECLS-B sample.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Equating now depended entirely on the remaining 39 (59 percent) mental and 39 (66 percent) motor items exhibiting little or no DIF. A distinction was made between more restricted numbers of virtually identical items that served as internal anchor items V_A and more encompassing numbers of similar but not identical items that continued to receive conditioning in the form of standardization dataset observations acting as the stabilizing counterweight V_B. These were the items shown to be most consistent with their respective counterpart publisher items.[14]

Based on this newly reformulated NEAT design, another concurrent calibration was performed, where BSF-R items V_A acted as internal anchors, BSF-R items V_B acted as a stabilizing counterweight and continued to receive conditioning from publisher observations, while the remaining BSF-R items placed in *X* assumed entirely new identities bearing no relation to publisher items. This implied that there were no remaining standardization dataset observations to effectively act as Bayesian priors on the new set of BSF-R *X* items. Standardization dataset responses to these items were removed to *Y*, where all item parameters were based on publisher calibrations and remained fixed during concurrent

[14] By augmenting the data with observations from the standardization dataset during item parameter estimation, the reliability of the Mental test (measured by the ratio of true-score variance to total variance) increased from .84 to .98, a 17 percent improvement. The reliability of the Motor test increased from .96 to .97, a 1 percent improvement.

calibration. Parameters for the new set of BSF-R *X* items were left free to float and find their positions in parameter space based only upon their relationships with other ECLS-B item responses.[15] At the same time, a more highly consistent set of BSF-R V_B items were coaxed into position by standardization dataset observations acting as a set of Bayesian priors. Item parameters were obtained with ECLS-B observations combined with well-conditioned standardization dataset observations. After recalibration, the new set of BSF-R item parameters coincided more closely with ECLS-B item responses and also adhered more closely to the publisher scale metric.

Fit indices for ECLS-B observations scored after DIF analysis in the second concurrent calibration run are reported in table 4-14. These indices remained essentially unchanged from the previous calibration. The reformulated design has resulted in trivial improvements to person fit on the mental scale at both 9 months and 2 years of age. There was essentially no improvement to person fit on the motor scale, where all four indices were only a fraction higher than previously. Fit statistics again show that there was a certain amount of redundant information among BSF-R item responses in the ECLS-B, a tendency that proved to be somewhat more apparent in the motor test at 9 months.

Table 4-14. Mean fit indices for ECLS-B observations scored after DIF analysis and concurrent calibration, by BSF-R scale and round of data collection: 2001–02 and 2003–04

ECLS-B subsample	Mean squared residual fit index	Mental scale	Motor scale
9 months	Information-weighted mean squared residual goodness of fit—Infit	0.954	0.862
	Outlier-sensitive mean squared residual goodness of fit—Outfit	0.932	0.829
2 years	Information-weighted mean squared residual goodness of fit—Infit	0.940	0.937
	Outlier-sensitive mean squared residual goodness of fit—Outfit	0.909	0.920

SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Sample frequencies reported in table 4-15 support these same conclusions. Mental observations exhibiting problematic fit improved slightly and declined from 171 observations previously to 136 with the current response model. All remaining improvement in fit on the mental scale was trivial, affecting only a handful of observations. On the motor scale, the number of observations with problematic

[15] BSF-R *X* items showing evidence of DIF remain in the scales for scaling and scoring. Item calibrations reveal that BSF-R items fit ECLS-B data appropriately and thus should be considered as part of each scale. Issues of scale content and construct validity provided additional justification for retaining the items in each scale. When scored, these items increase the precision of ability estimates and, ultimately, enhance scale reliabilities. With maximum likelihood estimation, raw scores play no role in IRT scaling and scoring.

fit actually increased at 9 months and 2 years. Observations exhibiting unacceptable fit increased by 23 at 9 months but declined by 2 at 2 years. The reformulated equating design at best produced trivial improvements in model fit at 9 months of age. However, the scale metric was made more consistent with the second recalibration.

Table 4-15. Number and percentage of ECLS-B sample children, by level of fit for ECLS-B observations scored after differential item function analysis and concurrent calibration, by BSF-R scale and round of data collection: 2001–02 and 2003–04

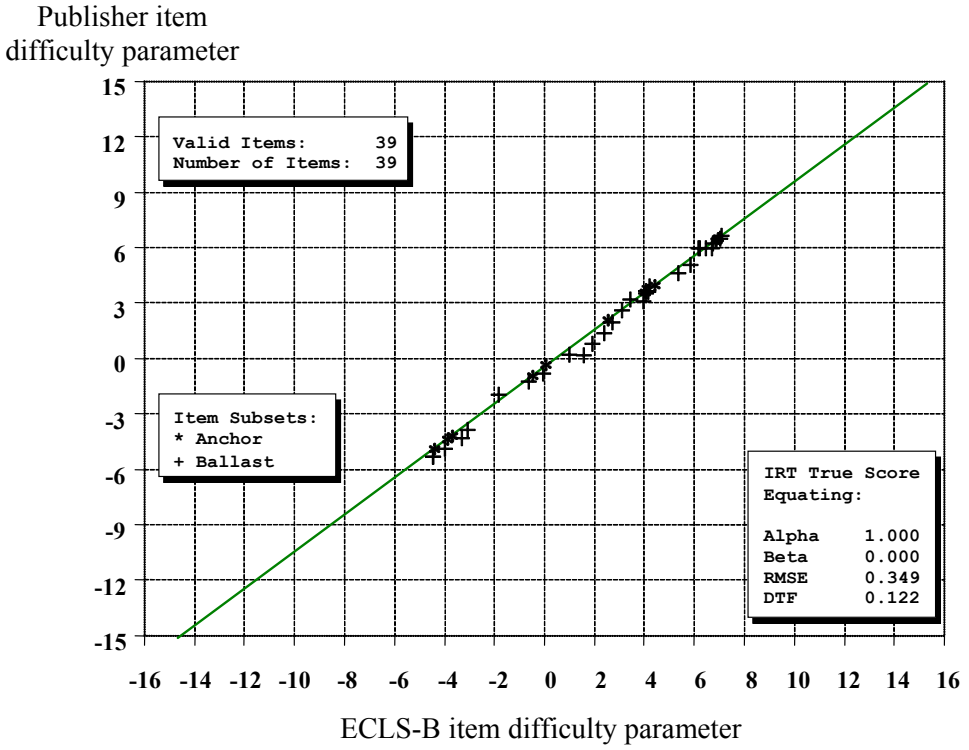
		BSF-R mental scale		BSF-R motor scale	
Fit	Level of outfit	Number	Percent	Number	Percent
9 months					
Total	$0 \leq y < \infty$	10,200	100.0	10,150	100.0
Excellent	$0 \leq y < 1$	6,650	65.2	7,600	74.8
Acceptable	$1 \leq y < 3$	3,500	34.5	2,250	22.0
Problematic	$3 \leq y < 5$	50	0.3	250	2.4
Unacceptable	$5 \leq y < \infty$	#	#	100	0.8
2 years					
Total	$0 \leq y < \infty$	8,900	100.0	8,850	100.0
Excellent	$0 \leq y < 1$	5,900	66.3	5,750	64.7
Acceptable	$1 \leq y < 3$	3,000	33.6	3,100	35.2
Problematic	$3 \leq y < 5$	#	0.1	#	0.1
Unacceptable	$5 \leq y < \infty$	#	#	#	#

Rounds to zero.
NOTE: Frequencies differ slightly from table 4-8 and table 4-11 due to weighting and rounding considerations. Outfit = outlier-sensitive mean squared residual goodness of fit. Detail may not sum to total because of rounding. Sample sizes have been rounded to the nearest 50.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

As a final step in the item calibrations, publisher items *Y*, which were not used in the ECLS-B, were removed from the BSF-R mental and motor scales. This left only the BSF-R items in each scale, consisting of the V_A and V_B items used in equating and the *X* items that were allowed to float freely. Figures 4-11 and 4-12 show how well the common items *V* performed in equating the BSF-R to the publisher standard. In figures 4-11 and 4-12, ECLS-B item difficulty parameters b_j on the *x* axis are once again plotted against the corresponding publisher parameters on the *y* axis. Publisher items *Y* with fixed parameters, that would otherwise appear on the diagonal, have now been removed from each scale, leaving only the BSF-R items. Notice also that one of the V_B items in the 2-year motor basal item set, MOT062—Walks Alone (Basal), had to be eliminated for the scale due to its dependency with

MOT063—Walks Alone with Good Coordination (Basal), scored not only from a single task administration but also from a single observation, leaving a total of 38 common items in the motor scale.

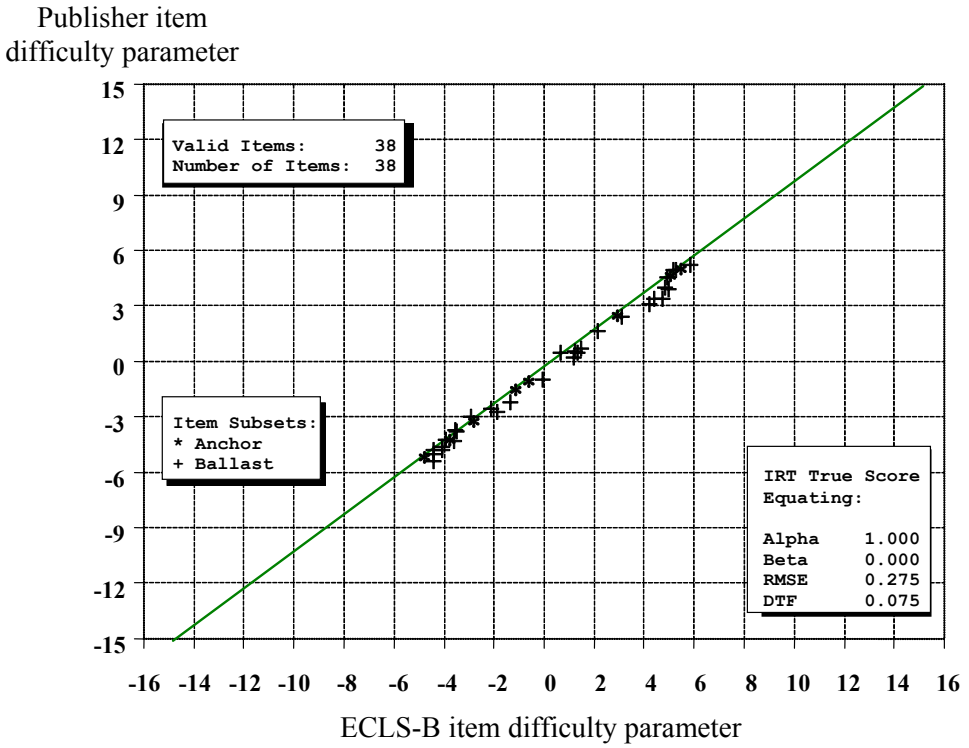
Figure 4-11. ECLS-B mental item difficulty parameters b_j on the x axis plotted against the corresponding publisher difficulty parameter on the y axis after differential item function analysis and concurrent calibration: 2001–02 and 2003–04



NOTE: Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Alpha = linear transformation of scale; Beta = linear transformation of origin; RMSE = root mean squared error; DTF = differential test functioning. Scatter plot of item difficulties for common items. Anchor items with fixed item parameters lie exactly along the diagonal and stabilizing items conditioned on publisher standardization dataset observations lie near the diagonal.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Item difficulty parameters for BSF-R items used in equating now lie very close to the diagonal in each figure. The two figures show the BSF-R items situated at strategic intervals across a broad range of ability, stretching across approximately 12 population standard deviations on the mental and approximately 11 population standard deviations on the motor scales. These ranges roughly coincide with the 9-month and 2-year latent ability distributions to be shown presently in figures 4-17 and 4-18. These BSF-R items are sufficiently close to the diagonal, sufficient in number, and strategically positioned across a broad range of ability to assure that both the mental and motor scales were effectively calibrated on the publisher scale metric.

Figure 4-12. ECLS-B motor item difficulty parameters b_j on the x axis plotted against the corresponding publisher difficulty parameter on the y axis after differential item function analysis and concurrent calibration: 2001–02 and 2003–04

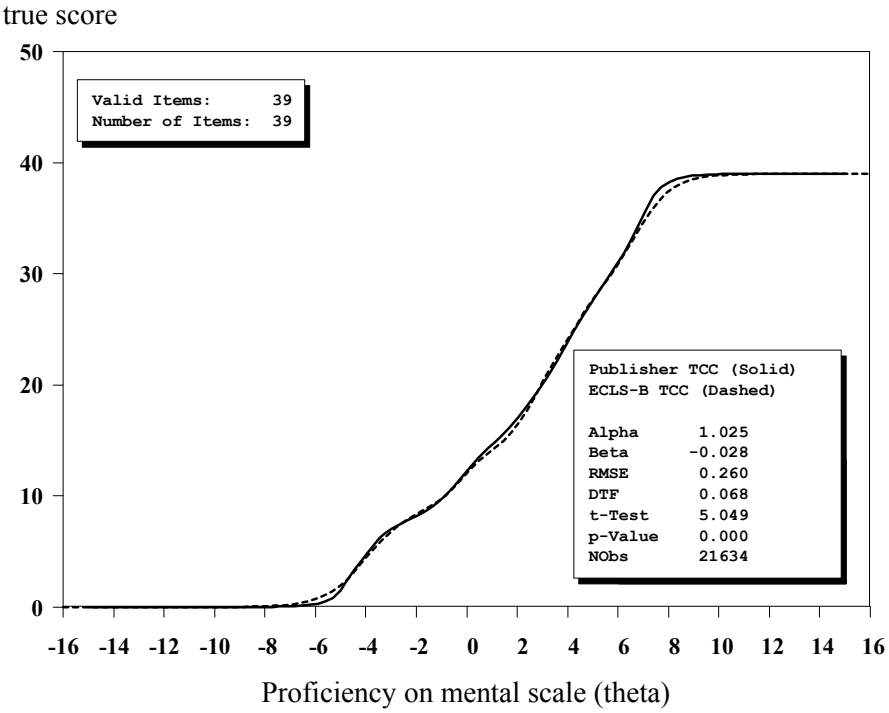


NOTE: Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Alpha = linear transformation of scale; Beta = linear transformation of origin; RMSE = root mean squared error; DTF = differential test functioning. Scatter plot of item difficulties for common items. Anchor items with fixed item parameters lie exactly along the diagonal and stabilizing items conditioned on publisher standardization dataset observations lie near the diagonal.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Figure 4-13 shows the TCC alignment for the 39 common items in the mental scale, while figure 4-14 shows the alignment for the 38 common items in the motor scale.[16] The RMSE representing the average distance between the TCCs on the vertical axis for the two tests in both cases is small, although these values need to be considered in relation to the number of items in each test, which has also declined. Although it could be argued that new set of figures represents an improvement in terms of the closer relationship between each pair of TCCs, this ignores the behavior of BSF-R *X* items.

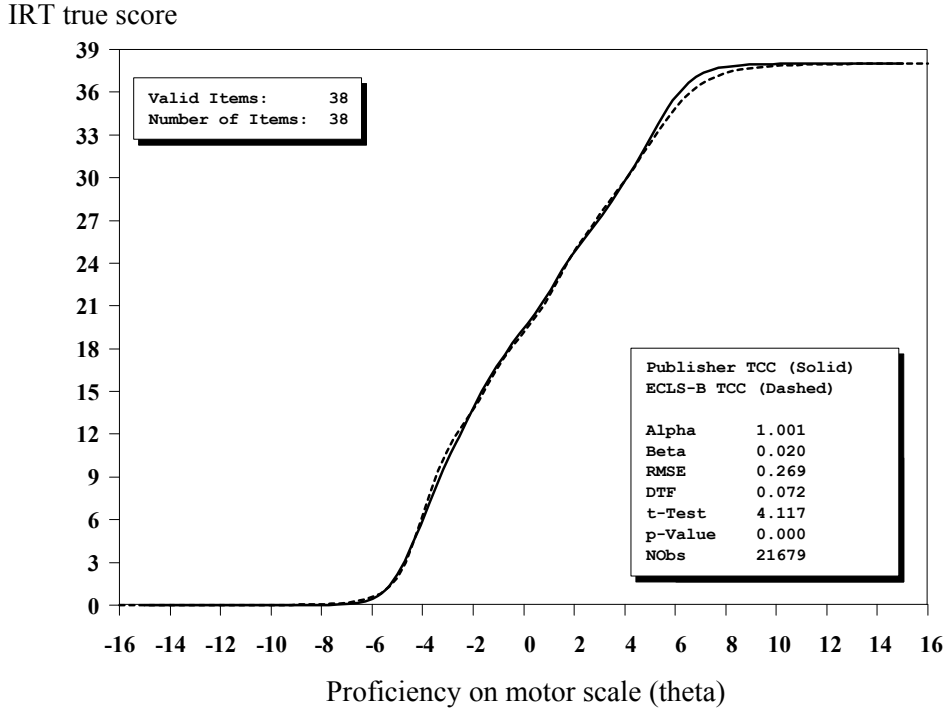
[16] As previously noted, MOT062—Walks Alone (Basal), had to be excluded from the scale due to its dependency with MOT063—Walks Alone with Good Coordination (Basal), leaving a total of 38 common items in the motor scale.

Figure 4-13. Test characteristic curves (TCCs) for BSF-R and BSID-II mental scales after differential item function analysis and concurrent calibration: 2001–02 and 2003–04



NOTE: Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Alpha = linear transformation of scale; Beta = linear transformation of origin; RMSE = root mean squared error; DTF = differential test functioning; t-Test = statistical test of the difference between the two curves shown in the figure; p-Value = probability of the results of the t-Test; NObs = number of observations.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

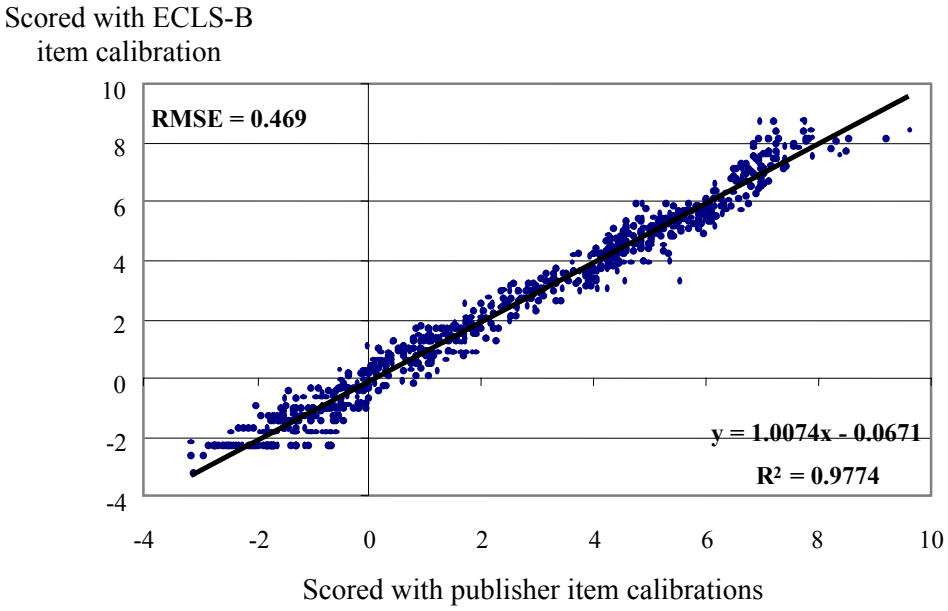
Figure 4-14. Test characteristic curves (TCCs) for BSF-R and BSID-II motor scales after differential item function analysis and concurrent calibration: 2001–02 and 2003–04



NOTE: Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Alpha = linear transformation of scale; Beta = linear transformation of origin; RMSE = root mean squared error; DTF = differential test functioning; t-Test = statistical test of the difference between the two curves shown in the figure; p-Value = probability of the results of the t-Test; NObs = number of observations.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

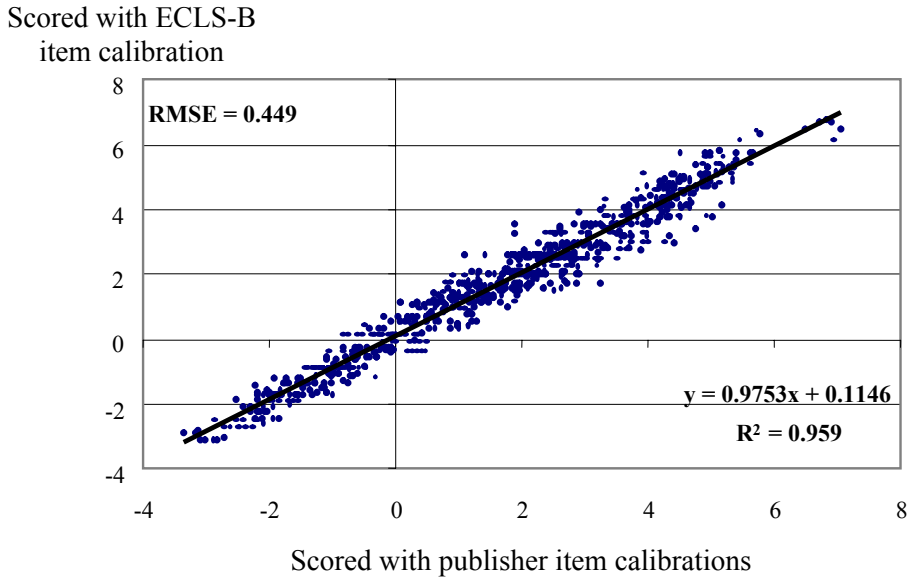
To further examine the quality of item calibrations, consider how BSF-R scoring performed in relation to the full BSID-II on the set of 900 standardization dataset observations. Figures 4-15 and 4-16 show the relationship between the two sets of scores using only the limited number of V_A and V_B BSF-R items when scoring with ECLS-B item parameters. This is because the BSF-R *X* items took on an entirely new identity without parallel in BSID-II. These additional items should have provided some additional precision when scoring ECLS-B observations, but they cannot be used to score publisher observations. The subset of common BSF-R items produced scores that have essentially the same intercept and slope as those produced with publisher item calibrations. Average RMSEs were in the vicinity of 0.45 on both tests. The full set of BSF-R items produced results that were at least as precise and conceivably somewhat more precise.

Figure 4-15. Expected a posteriori ability estimates for standardization sample observations are scored first with the full publisher BSID-II mental items calibrations and then with ECLS-B mental item calibrations (BSF-R) following concurrent item calibration with the second NEAT design: 2001–02 and 2003–04



NOTE: NEAT = Non-Equivalent groups with Anchor Test; RMSE = root mean squared error; R^2 = proportion of variance in the data explained by the regression equation.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Figure 4-16. Expected a posteriori ability estimates for standardization sample observations are scored first with the full publisher BSID-II mental items calibrations and then with ECLS-B motor item calibrations (BSF-R) following concurrent item calibration with the second NEAT design: 2001–02 and 2003–04



NOTE: NEAT = Non-Equivalent groups with Anchor Test; RMSE = root mean squared error; R^2 = proportion of variance in the data explained by the regression equation.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

4.6 BSF-R Developmental Growth

Infant mental and motor development is so explosive in the early years of life that the range of latent ability found in the ECLS-B spans approximately 12 population standard deviations.[17] Rarely in psychometric research is there an opportunity to capture so much variation in mental and physical status. The challenge for the designers of the BSF-R instruments was to measure each child's mental and motor ability accurately across this broad range of ability, using a reduced item set selected from the BSID-II, while still maintaining comparability with the publisher score metric. After the adjustments discussed above, the design effort met these requirements.

[17] The age-specific latent ability distributions in the publisher standardization dataset have standard deviations that are nearly equal to 1, with small tendency for the variation in mental and physical ability to increase as age approaches 42 months.

Figures 4-17 and 4-18 show kernel density[18] estimations for the ECLS-B mental and motor latent ability distributions for the 9-month and 2-year data collections. Publisher calibrations set the scale metric shown in each of the figures, where standardization data observations in cross-section at 12 months of age have a $N(0, 1)$ distribution, with mean $\mu = 0$ and standard deviation $\sigma = 1$. ECLS-B observations in the 9-month data collection were generally younger than 12 months of age, and thus were represented by negative scale values appearing to the left of each figure. ECLS-B observations in the 2-year data collection were all well above 12 months of age, and were thus represented by positive scale values appearing to the right of each figure.

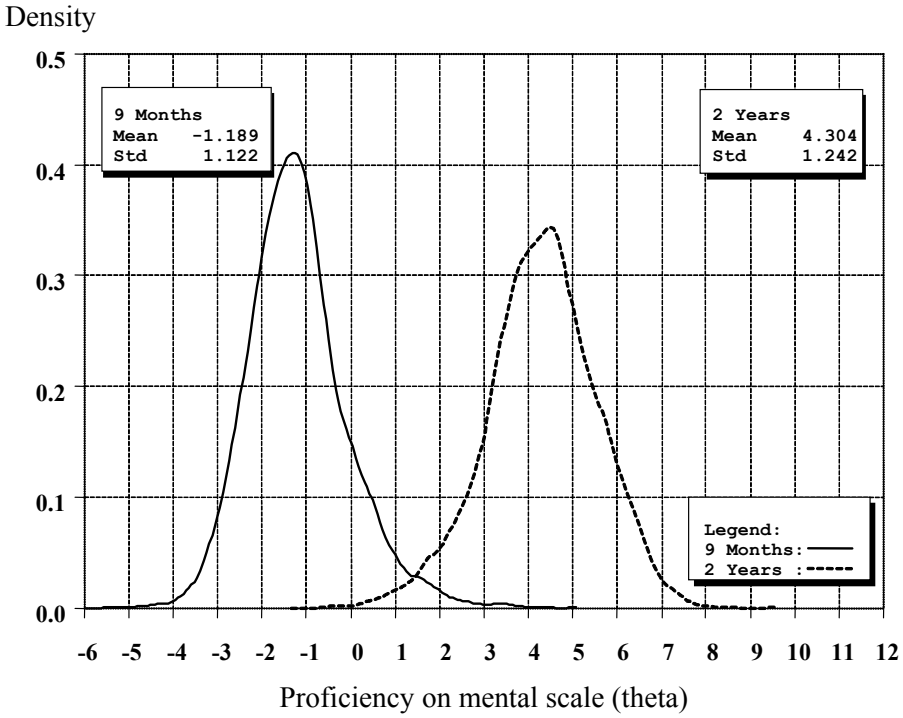
There was considerable variation in mental and motor ability within each cross-section of data, especially for the 9-month data collection. The dispersion of scores is largely a reflection of the distribution of ages in each wave of data. Indeed, the broad range of ages encountered in the ECLS-B implied that mean ability estimates for both the 9-month and 2-year data collections could not reasonably be expected to represent mean ability at exactly 9 or 24 months of age. It was necessary to take this diversity of ages explicitly into account when estimating mean ability at precisely 9 or 24 months.

This could be accomplished by modeling mental and motor scale scores as a function of age at time of assessment. The value of age at precisely 9 months could then be entered into the age-ability equation to obtain a predicted ability score value at precisely 9 months. Indeed, any age value in months could then be entered into the regression equation to predict ability scores across a whole range of ages. In this fashion, a age-ability regression provided a continuous function that could be used to delineate the mean trajectory of ability scores anywhere from say 7 to 28 months, without seriously extrapolating beyond the ages found in the ECLS-B sample, as shown below in figures 4-19 and 4-20.

It should be noted that the ECLS-B is a true longitudinal study, in the sense that the same individuals were assessed at two points in time. Each individual's growth trajectory could thus be summarized in terms of an initial status at exactly 9 months, together with an average monthly growth rate between 9 and 24 months, arriving at a final status at 2 years of age.

[18] A kernel density plot is a non-parametric representation of density that has been smoothed (e.g., by using a Gaussian function).

Figure 4-17. Kernel density estimation for ECLS-B mental latent ability distributions for 9-month and 2-year data collections, in publisher scale metric: 2001–02 and 2003–04

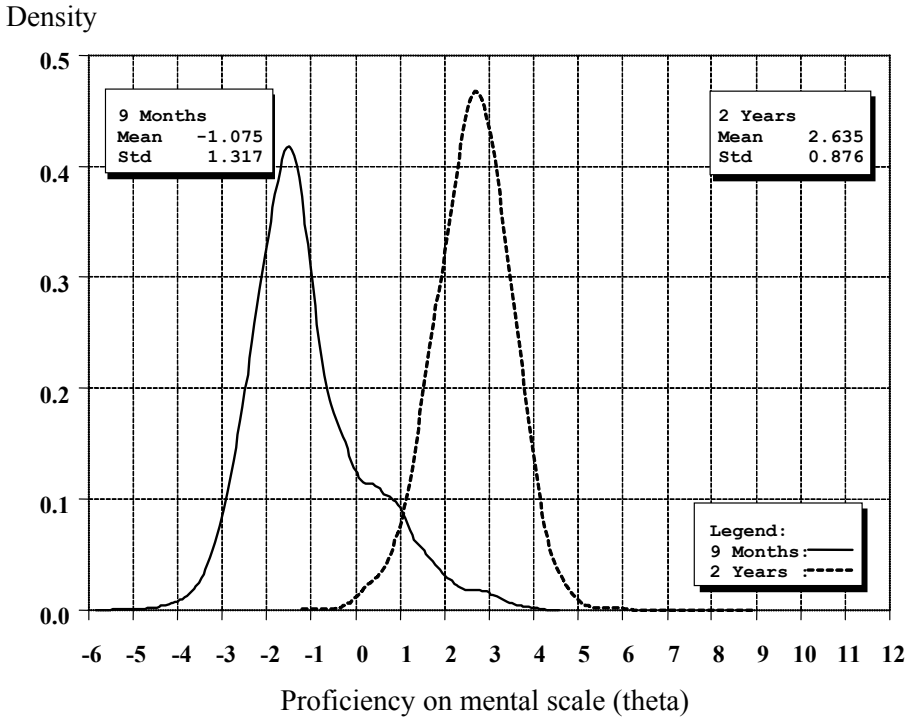


NOTE: Kernel density estimation obtained with weighted ECLS-B sample observations. A kernel density plot is a nonparametric representation of density that has been smoothed (e.g., by using a Gaussian function). Std = standard deviation.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

However, as clearly shown in the publisher standardization sample, mental and motor growth is not linear over these ages, and instead decelerates with advancing age. This deceleration introduces a slight curvature in growth trajectories, where growth effectively slows down as age increases. With observations at only two points in time, it was not possible to quantify this degree of curvature for individual growth trajectories in the ECLS-B, however, it was possible to estimate the degree of curvature for the ECLS-B sample as whole.

ECLS-B initial status was not assessed at precisely 9 months of age but over a range of ages extending roughly between 6 and 19 months. Similarly, final status was not assessed at precisely 2 years but over a range of ages extending roughly between 20 and 30 months. Thus, the ECLS-B sample covers a wide age range extending approximately from 9 to 30 months of age. This broad range of ages made it possible to estimate the overall deceleration in mental and motor growth in ECLS-B.

Figure 4-18. Kernel density estimation for ECLS-B motor latent ability distributions for 9-month and 2-year data collections, in publisher scale metric: 2001–02 and 2003–04



NOTE: Kernel density estimation obtained with weighted ECLS-B sample observations. A kernel density plot is a nonparametric representation of density that has been smoothed (e.g., by using a Gaussian function). Std = standard deviation.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

The essential comparison of interest was provided by national norms, as represented by the publisher standardization dataset, which was collected in 1991–92. The objective was to determine how the overall mental and motor growth of ECLS-B infants in 2001–03 compared with publisher norms that had been obtained one decade earlier. For this purpose, it should be recalled that the publisher dataset is cross-sectional, in the sense that all infants were assessed at only one point in time. The growth trajectory provided by the publisher standardization dataset is based on a series of successive cross-sections obtained for many ages between 1 and 42 months.

A subset of these observations was selected by the publisher to form a nationally representative sample of infants in order to establish national norms. All of the observations in the standardization dataset were used in regressions modeling mental and motor age-ability relationships reported below, and a dummy variable was included in each regression to identify standardization sample

observations. BSID age was used throughout and an age-squared term was used to capture the deceleration in growth.[19]

In the discussion of age-ability regression that follows, decimal BSID months of age rather than chronological age was used throughout. BSID age was obtained from chronological age adjusted for premature birth in ECLS-B. Variables for low birth weight, household socioeconomic status (SES) and race-ethnicity were also included in the ECLS-B regression equations modeling mental and motor status as a function of BSID months of age and a BSID age-squared variable included to capture deceleration in growth in the sample as a whole.

A multilevel analysis was used to model overall mental and motor growth trajectories in relation to age, with time nested within children nested within ECLS-B sample clusters. The multilevel model takes the clustering of the sample design explicitly into account when coefficient standard errors are calculated. The equations for the multilevel, full maximum likelihood model that was estimated are given below.[20] The level-1 model was estimated using 106,450 plausible values, with 5 values per child, usually at two points in time[21]; the level-2 model was estimated using 12,243 children; and the level-3 model was estimated using 159 ECLS-B strata clusters and the 1 publisher group.

Level-1 Model

$$Y = P0 + P1*(AGE9) + P2*(AGE9SQ) + E$$

Level-2 Model

$$P0 = B00 + R0$$
$$P1 = B10 + R1$$
$$P2 = B20$$

Level-3 Model

$$B00 = G000 + G001(PUB) + U00$$
$$B10 = G100 + G101(PUB) + U10$$
$$B20 = G200 + G201(PUB)$$

[19] Decimal values for BSID age and the corresponding BSID age-squared variable were centered on zero at 9 months. Publisher sources assure that there were no premature infants in the standardization sample, and standardization sample observations included in the analysis were identified by a dummy variable in the growth analysis regression. The overall intercept coefficient and standardization dummy variable intercept coefficient in the publisher regression were summed together to provide an estimate of the overall mean standardization sample initial status at precisely 9 months of age. The overall slope coefficient and standardization dummy variable slope coefficient in the publisher regression were summed together to provide an estimate the overall mean standardization sample monthly growth rate.

[20] The outcome variable Y is either the Mental or Motor EAP ability estimate. Age9 is months of age centered on exactly 9 months so that the intercept P0 will represent mean ability at precisely 9 months of age. Age9SQ is the squared term for Age9. PUB is the dummy variable used to identify publisher observations belonging to a single group cluster.

[21] Plausible values were used so that the error variances for R0 and R1 could be obtained. Sample weights were divided by 5 to compensate for the number of plausible values per child.

Sample weights, based on sample selection probabilities, were used in the ECLS-B regression. The publisher standardization dataset is self-weighting and represents the national population in 1991–92. The objective here was not to conduct a complete analysis of all these variables but rather to provide an essential summary of the data showing overall mental and motor growth in relation to BSID age. For this purpose, BSID age and the corresponding BSID age-squared variable were centered on zero at 9 months of age.[22]

Figure 4-19 presents the essential summary of the data, where average mental attainment in the ECLS-B sample is compared with average mental ability estimates obtained with the publisher standardization sample. For the ECLS-B sample, the mental age-ability relationship was estimated to be:

$$\theta_{E,Men} = -1.870 + 0.513x - 0.007x^2,$$

where $\theta_{E,Men}$ is the mental score obtained with ECLS-B calibrations and x is BSID months of age minus 9 months. Standard errors for each coefficient are small due to the large sample of ECLS-B observations at both points in time. The standard error for ECLS-B mental initial status was 0.028; the standard error for linear growth was 0.013; and the standard error for BSID age-squared was 0.001.

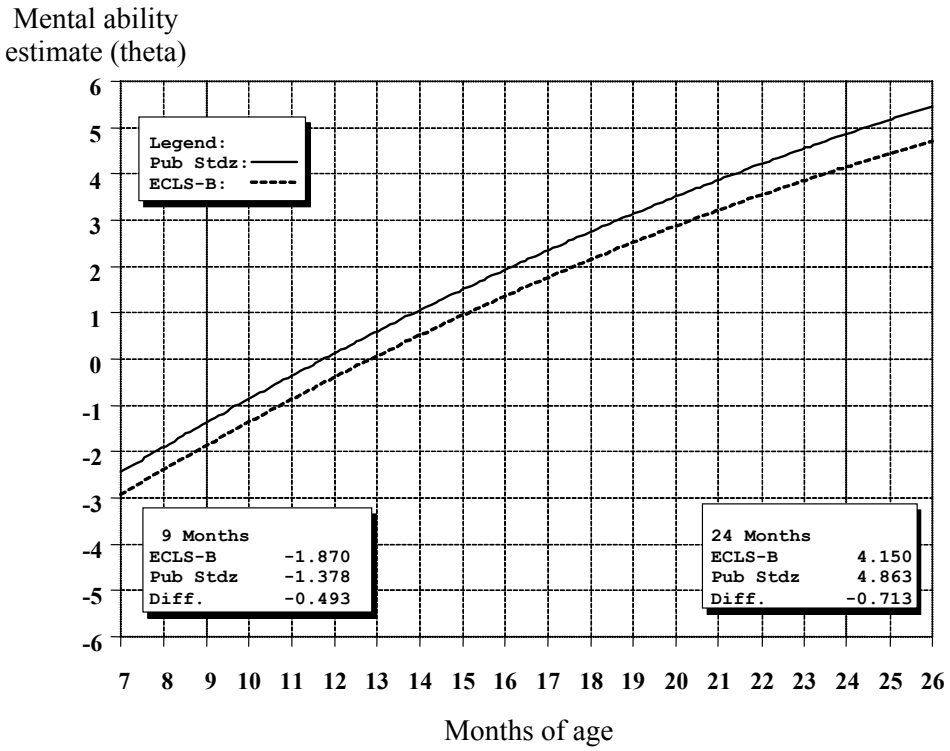
For the publisher standardization sample, the corresponding mental age-ability relationship was estimated to be:

$$\theta_{P,Men} = -1.378 + 0.521x - 0.007\,x^2,$$

where $\theta_{P,Men}$ is the mental score obtained with publisher calibrations and x is once again BSID months of age minus 9 months. Standard errors are quite small due to the large number of age cross-sections assessed between 1 and 42 months of age. The standard error for publisher mental initial status was 0.030; the standard error for linear growth was 0.009; and the standard error for BSID age-squared, to three decimal places, was 0.000.

[22] Birth weight was centered at normal weight in the ECLS-B, while household socioeconomic status and each of a series of race-ethnicity dummy variables were centered at their respective ECLS-B sample means. This allowed the ECLS-B regression intercept coefficient to be interpreted as overall mean ECLS-B initial status at precisely 9 months of age among infants who were both carried to term and showed no deficit in birth weight. The ECLS-B regression slope coefficient should be interpreted as the overall mean ECLS-B monthly growth rate for the same population of infants who were both carried to term and showed no deficit in birth weight.

Figure 4-19. Essential summary of the data showing mental growth in relation to age as estimated in publisher standardization sample and ECLS-B sample data: 2001–02 and 2003–04



NOTE: Multilevel regressions obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. "Pub Stdz" refers to the publisher standardization sample; "ECLS-B" refers to the ECLS-B sample.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Conventional tests of statistical significance showed that the difference between the publisher norm and the ECLS-B initial status at precisely 9 months was statistically significant, whereas the difference in monthly growth rate was not statistically significant. To three decimal places, the age-squared coefficients were identical and negative, representing the deceleration in growth over age as measured by the mental scale.[23]

These relationships showed that mental growth was roughly parallel in both the ECLS-B and publisher standardization samples, although the ECLS-B sample initial status revealed an appreciable

[23] These results are reported in the N(0, 1) metric defined at 12 months of age used in IRT scaling and scoring. To translate these results into the *N*(250, 50) metric found in the ECLS-B public-use data files, apply the following formula:

$$\theta_{\text{Men, N(250, 50)}} = [(\theta_{\text{Men, N(0, 1)}} - \bar{x}) / \sigma_x] \times 50 + 250,$$

where $\bar{x}$ = -1.189 represents the ECLS-B mental sample for the 9-month assessment and σ_x = 1.124 represents the ECLS-B motor sample standard deviation for the 9-month assessment. A similar transformation would place publisher mental results on the same *N*(250, 50) metric used in the ECLS-B.

deficit in relation to the standardization sample initial status at 9 months. The ECLS-B mean initial status was found at -1.870 and the publisher standardization sample mean initial status at -1.378, where both numbers are expressed in population standard deviations in cross-section at 12 months. The deficit in ECLS-B mean initial status was thus estimated to be -1.870 – (-1.378) = -0.493 or about half a population standard deviation below publisher norms established in 1991–92.

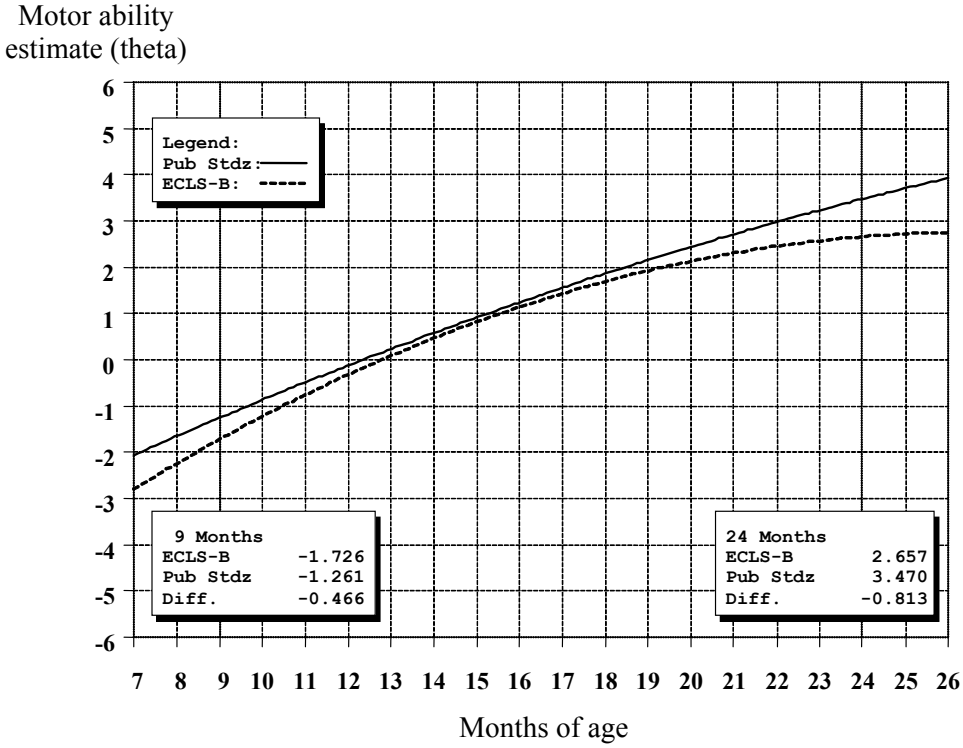
The linear and quadratic components of mental growth were generally similar, and to three decimal places the quadratic terms were identical. However, linear growth in the ECLS-B was found to be somewhat lower than in the publisher standardization sample. Although this growth rate deficit was small, on the order of 0.513 - 0.521 = -0.007 population standard deviations per month, its cumulative effect over 24 - 9 = 15 months would be fairly substantial. In this fashion, the initial ECLS-B deficit of -0.493 at 9 months widened to an estimated deficit of -0.713 population standard deviations at precisely 2 years of age. Not only was the deficit at 2 years statistically significant, it was also relatively large, on the general order of seven-tenths of a population standard deviation. In conducting this analysis, no attempt was made to adjust for demographic differences in the newborn U.S. population. It should be noted that the demographic profile of the infant population changed substantially since 1991–92. During the 1990s, there was substantial immigration resulting in large increases in the Hispanic population. Immigration, combined with high fertility rates among Hispanics, resulted in an increase in the percentage of newborns who were Hispanic. In 1990, 14.5 percent of newborns were Hispanic compared to 23.6 percent in 2001. Not surprisingly, the percentage of newborns with a foreign-born mother also increased from 15.7 percent in 1990 to 22.5 in 2001. Another change was that the percent of births that were preterm rose 12 percent since 1990, from 10.6 percent to 11.9 percent in 2001.

Figure 4-20 presents the essential summary of the data for the motor scale, where the average motor development of the ECLS-B sample is compared with that of the publisher standardization sample. For the ECLS-B sample, the motor age-ability relationship was estimated to be:

$$\theta_{E,Mot} = -1.726 + 0.511x - 0.015\,x^2,$$

where $\theta_{E,Mot}$ is the motor score obtained with ECLS-B calibrations and x is BSID months of age minus 9 months. Once again, standard errors are small due to the large size of the ECLS-B sample at both points in time. The standard error for ECLS-B initial status was 0.028; the standard error for linear growth was 0.012; and the standard error for BSID age-squared was 0.001.

Figure 4-20. Essential summary of the data showing motor growth in relation to age as estimated in publisher standardization sample and ECLS-B sample data: 2001–02 and 2003–04



NOTE: Multilevel regressions obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. "Pub Stdz" refers to the publisher standardization sample; "ECLS-B" refers to the ECLS-B sample.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

For the publisher standardization sample, the corresponding motor age-ability relationship was estimated to be:

$$\theta_{P,Mot} = -1.261 + 0.393x - 0.005\,x^2.$$

Once again, standard errors are small due to the large number of ages assessed in cross-section between 1 and 42 months of age. The standard error for publisher motor initial status was 0.023; the standard error for linear growth was 0.008; and the standard error for BSID age-squared was 0.001.

All three of the differences between the corresponding ECLS-B and publisher coefficients were statistically significant. The deficit in ECLS-B mean initial status relative to the publisher norm at precisely 9 months of age was 11.726 - (-1.261) = -0.466 population standard deviations. ECLS-B linear growth was 0.511 - 0.393 = 0.118 population standard deviations higher than publisher linear growth, but in compensation the ECLS-B coefficient for the BSID age-squared variable was negative and three times as large as the publisher quadratic coefficient. Consequently, both growth trajectories were concave from below, but the ECLS-B curve was more sharply bowed, indicating a higher rate of deceleration in growth. The ECLS-B growth trajectory initially appears to grow more rapidly than the publisher growth trajectory, until the two curves almost intersect in the vicinity of 15 months, after which the ECLS-B curve appears to grow more slowly.[24]

The curvature of the ECLS-B profile appears to be accentuated when this is compared with that of the publisher standardization sample. Despite the broad range of ages found in the ECLS-B sample, this was perhaps insufficient to provide a good estimate of the deceleration in motor growth. However, the ECLS-B sample contained approximately 10,000 observations at 9 months and again at 2 years. For this reason, it is reasonable to assume that initial status at 9 months and the final status at 2 years were accurately estimated. This showed the initial deficit of -0.466 growing to an even larger motor deficit of -0.813 at 2 years or roughly eight-tenths of a population standard deviation. This revealed that the ECLS-B sample of infants in 2001–02 started with an appreciable deficit in initial status at 9 months, growing more slowly on average over the next 15 months, to yield deficit that was almost twice as large at 2 years. The general trend in motor growth resembled that found previously for mental growth,

[24] These results are reported in the N(0, 1) metric defined at 12 months of age used in IRT scaling and scoring. To translate these results into the *N*(250, 50) metric found in the ECLS-B public-use data files, apply the following formula:

$$\theta_{Men,\,N(250,\,50)} = [(\theta_{Men,\,N(0,\,1)} - \overline{x}\,) / \sigma_x] \times 50 + 250,$$

where $\overline{x} = -1.075$ is the ECLS-B motor sample mean for the 9-month assessment and $\sigma_x = 1.323$ is the ECLS-B motor sample standard deviation for the 9-month assessment. A similar transformation would place publisher motor results on the same *N*(250, 50) metric used in the ECLS-B.

although the curvature of motor growth was found to be much more accentuated. Again, presumably part of the deficit in motor growth should be attributed to demographic changes occurring in the U.S. infant population between 1991–92 and 2001–03.

In a more general sense, this analysis showed that the publisher dataset provided growth coefficient precision comparable to that found in the ECLS-B at a fraction of the sample size by assessing many ages in cross-section. This is clearly an efficient design for establishing national norms. On the other hand, the publisher standardization dataset provided no information about the rate of growth for individuals. Only the ECLS-B sample provided information about the growth rates of individual infants. This is an efficient design for assessing the impact of childrearing and other practices on individual growth rates.

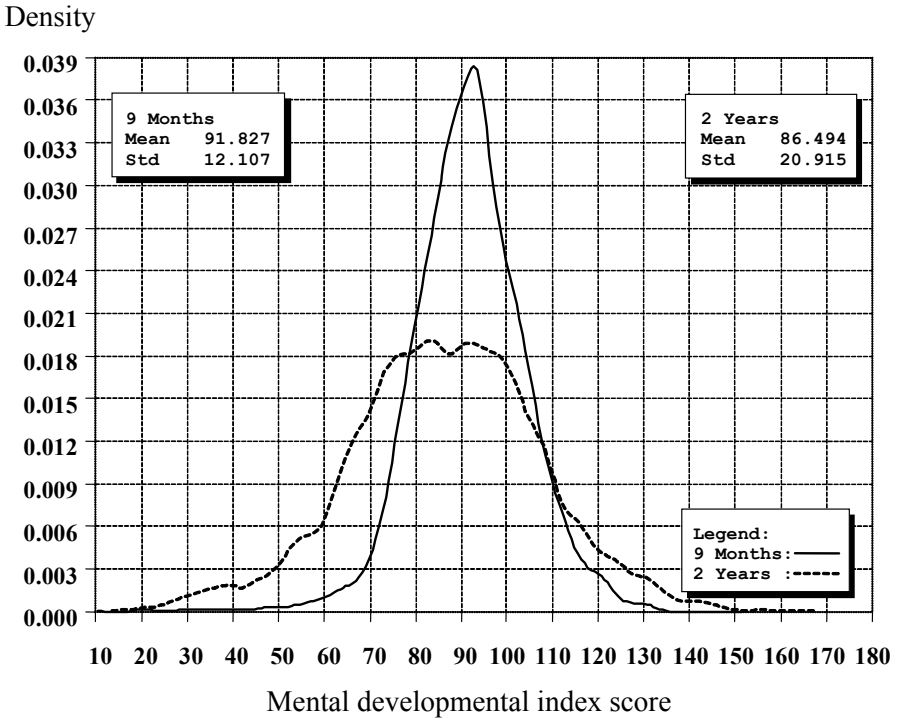
4.7 BSF-R Developmental Indices

Fortunately, it is possible to make simple comparisons in cross-section without the need for age-ability regressions. This possibility is provided by publisher developmental index scores. Developmental index scores are age-normed ability estimates. In publisher documentation, developmental index scores are obtained with raw scores and age, adjusted for prematurity, by using a lookup-table. In the ECLS-B, the same scores were obtained by using IRT, BSID age, and regression estimates.

With the benefit of the equating design, EAP ability estimates obtained in the ECLS-B could be reported on the same scale metric used in publisher item calibrations. Thus, EAP ability estimates obtained in the ECLS-B could be used with publisher item calibrations to obtain an IRT true-score, which was a model-based estimate of the publisher's number-right raw score. IRT true-scores, together with BSID age were then applied to a regression equation to produce developmental index scores for individual observations. Figures 4-21 and 4-22 show weighted kernel density estimations for ECLS-B mental and motor developmental index scores.

The developmental index score represents the child's position in an $N(100, 15)$ norm reference distribution with a mean $\mu = 100$ and standard deviation $\sigma = 15$. The two figures show that the developmental status of ECLS-B infants in the weighted sample at the time of the 9-month assessment

Figure 4-21. Weighted kernel density estimations for mental developmental index score distributions in the ECLS-B 9-month and 2-year data collections: 2001–02 and 2003–04

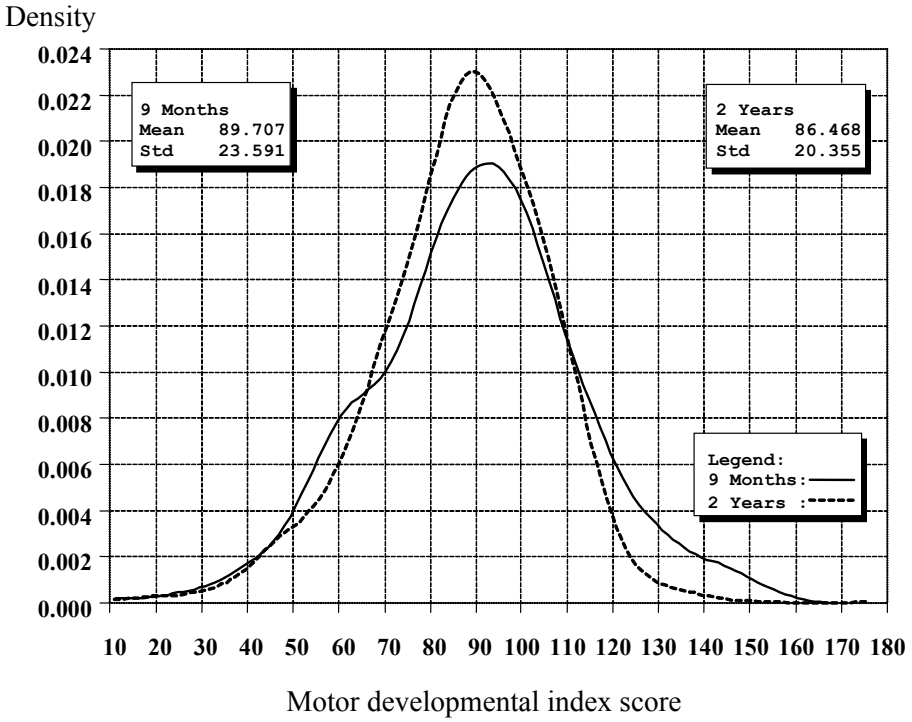


NOTE: Kernel density estimation obtained with weighted ECLS-B sample observations. A kernel density plot is a nonparametric representation of density that has been smoothed (e.g., by using a Gaussian function). Std = standard deviation.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

was lower than the national average on both the mental and motor scales. Moreover, by the time of the 2-year assessment, ECLS-B children had fallen farther behind the national standard provided by the publisher on both the mental and motor scales.

The growth deficit on the mental scale was especially dramatic, showing that the central tendency of the ECLS-B sample for the 2-year assessment is nearly a full population standard deviation below norm. The ECLS-B mental distribution was also more heterogeneous—and, therefore, unequal—for the 2-year assessment than it was for the 9-month assessment, with the dispersion in mental developmental index scores almost doubling in size.

Figure 4-22. Weighted kernel density estimations for motor developmental index score distributions in the ECLS-B 9-month and 2-year data collections: 2001–02 and 2003–04



NOTE: Kernel density estimation obtained with weighted ECLS-B sample observations. A kernel density plot is a nonparametric representation of density that has been smoothed (e.g., by using a Gaussian function). Std = standard deviation.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

On the motor scale, the ECLS-B sample began with an average deficit disadvantage of two-thirds of a population standard deviation in relation to publisher norms at the time of the 9-month assessment. This relative position declined by the time of the 2-year assessment. By the time of the 2-year assessment, the ECLS-B infant population is almost a full population standard deviation below publisher norms. The variation in motor development decreased very slightly in relation to publisher norms, indicating that the population became relatively more homogenous in terms of psychomotor development.

The central tendencies and the trends in central tendencies over time are similar for both the mental and motor scales. The initial deficits at the time of the 9-month data collection were similar on both the mental and motor scales at approximately two-thirds of a population standard deviation. Both deficits increase to nearly a population standard deviation by the time of the 2-year data collection. By 2 years, ECLS-B found much greater diversity in mental and motor status in 2003–04 compared with the publisher standardization sample obtained in 1991–92. It is possible that the ECLS-B sample as a true

probability sample of the U.S. infant population was more inclusive than the publisher norm sample and, therefore, produced lower performance estimates due to broader coverage of the population obtained in the ECLS-B. ECLS-B item responses from 2001–03 consistently reflected performance levels well below those reported by the publisher in 1991–92.

All of these conclusions depend critically on the quality of the ECLS-B equating design. To the extent that scale equating was properly implemented, then these results reflect real differences in the population, and the lower mean developmental scores reflect lower levels of performance across a broad range of developmentally relevant tasks. It is for this reason that the authors of this report have gone to lengths to show that the ECLS-B observations have been scored on the publisher metric. The fact that 60 percent of the mental and 66 percent of the motor items performed the same in the ECLS-B as they do in BSID-II implies that the same scale metric has been maintained throughout. In this case, the mean differences in ability seen among children in the ECLS-B reflect real differences in the infant population, of which the ECLS-B data are representative, rather than artifactual differences that otherwise might be attributable to the BSF-R short form or to fieldwork conditions and procedures used in the ECLS-B.

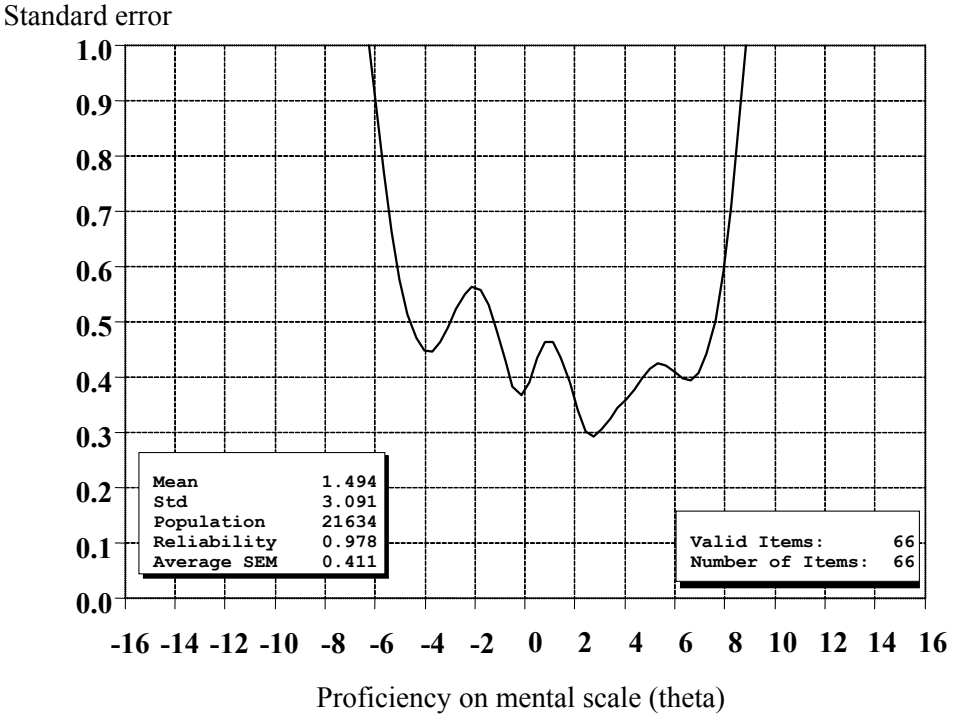
4.8 BSF-R Precision and Reliability

Due to the wide range of age and ability found in the ECLS-B data collections, standard errors of measurement are probably more informative than reliability coefficients as a means for assessing measurement precision. Figures 4-23 and 4-24 show standard errors of measurement at different levels of ability for the BSF-R mental and motor tests used in the ECLS-B. These errors were not adjusted for the redundancy of information observed earlier in BSF-R Infit and Outfit indices because it is not the convention to do so in this type of IRT model. Consequently, standard errors and reliability coefficients reported in the figures and table that follow may appear to be more precise than they actually are. In reviewing these figures, be mindful that the ECLS-B population for the 9-month assessment was centered near $\theta = -1.2$ on the mental and -1.1 on the motor scale. For the 2-year assessment, the ECLS-B population was centered near $\theta = 4.3$ on the mental scale and $\theta = 2.6$ on the motor scale.

The size of standard errors shown in figure 4-23 reveal the limitations in the precision of the BSF-R mental test, where many of the easier items in the 9-month core item set provided little discrimination below $\theta = -1.5$. Mental basal and ceiling item sets were substantially more informative, providing better precision in the tails of the 9-month distribution. Standard errors over .3 in the core item

set indicates a lack of efficiency in basal and ceiling decision rules, implying that some infants failed to receive the required basal or ceiling item sets. The BSF-R mental test generally provided substantially better precision for the 2-year assessment, where precision remained high and standard errors short over most of the latent ability distribution.

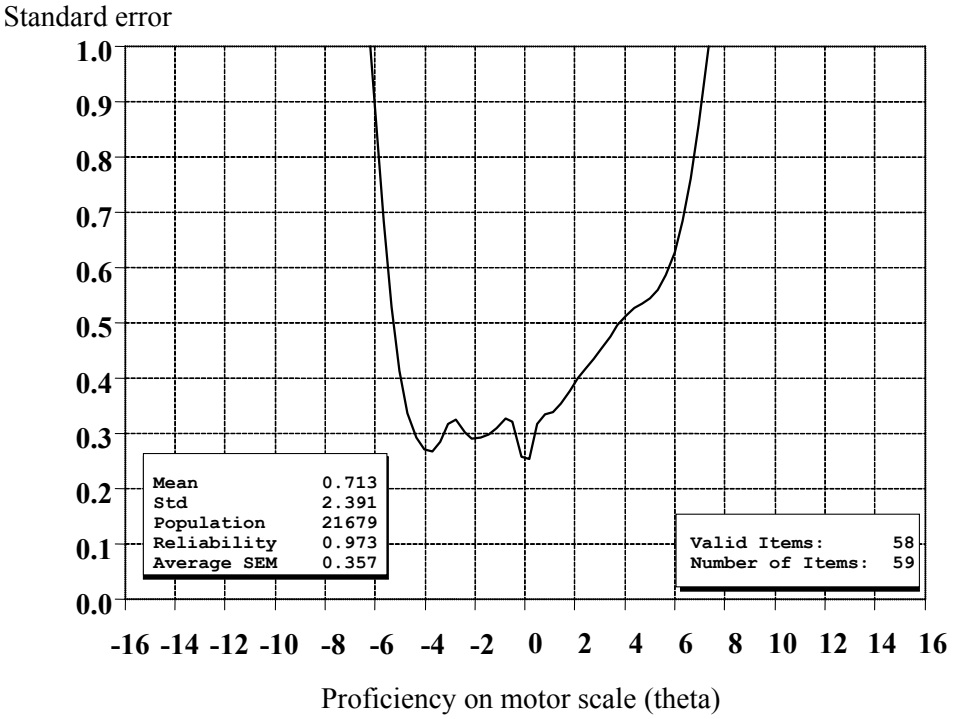
Figure 4-23. Standard errors of measurement for the BSF-R mental test used in the ECLS-B, across all levels of ability: 2001–02 and 2003–04



NOTE: Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Std = standard deviation.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Figure 4-24 shows the precision of the BSF motor test across the entire 9-month ability distribution. The situation was different at 2 years, where BSF motor precision was relatively poor at higher levels of ability. Roughly half of the 2-year assessments obtained good to excellent precision before standard errors began to rise substantially at higher levels of ability. The larger standard errors limit reliability in the BSF-R motor test for the 2-year assessment, making this the least reliable of the BSF-R instruments used in the ECLS-B.

Figure 4-24. Standard errors of measurement for the BSF-R motor test used in the ECLS-B, across all levels of ability: 2001–02 and 2003–04



NOTE: Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Std = standard deviation.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Reliability coefficients reported in the figures were based on normal distributions having the same reported mean and standard deviation. In this sense, these were theoretical reliabilities, based on plausible distributional assumptions rather than directly on ECLS-B observations. From this theoretical perspective, the BSF-R mental was estimated to have an overall reliability of $r_{xx} = 0.978$ and the motor an overall reliability of $r_{xx} = 0.973$. The overall IRT reliability coefficient obtained with ECLS-B

observations was $r_{xx} = 0.975$ for the BSF-R mental.[25] The corresponding figure for the BSF-R motor was 0.969. In both cases, the sample-based estimates coincided almost exactly with reliability coefficients calculated based on distribution assumptions.

All of these coefficients were very high because of the broad range of ages considered in the ECLS-B. Standard errors and reliability coefficients by wave of assessment are also reported in table 4-16. These are more realistic reliability coefficients since the differences in age between assessments is no longer a factor. These coefficients show the somewhat lower reliabilities obtained for the mental at the time of the 9-month assessment and the motor at the time of the 2-year assessment. Standard errors continue to be reported in population standard deviation units. The reference population that sets the scale metric is that of the publisher standardization dataset at 12 months of age.

Table 4-16. Standard errors and reliability coefficients for the 9-month and 2-year BSF-R mental and motor-scales: 2001–02 and 2003–04

Test	Mean standard error[1]	Reliability (internal consistency)
Mental		
Total	0.47	0.98
9 months	0.49	0.81
2 years	0.44	0.88
Motor		
Total	0.38	0.97
9 months	0.33	0.94
2 years	0.45	0.73

[1] Standard errors reported in population standard deviation units, where the 12-month age group is $N(0, 1)$.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

[25] While the information function provides the most comprehensive measure of IRT score reliabilities, it is helpful to provide a single index of test reliability in IRT. For IRT scales, the ratio of the average measurement error variance to total variance can be used for this purpose, after subtracting this value from unity. This yields a measure of true score variance as a proportion of total variance:

$$r_{xx} = \frac{\text{error variance}}{\text{total variance}} \approx 1 - \frac{\sum_{k=1}^{q} \sigma_{e_k}^2 A(X_k)}{\sum_{k=1}^{q} \left(X_k - \overline{X}\right)^2 A(X_k)},$$

where the $A(X_k)$ are normal ordinate weights for points X_k spanning the distribution of ability, with $\sum_{k=1}^{q} A(X_k) = 1$ over q quadrature points.

4.9 BSF-R Assessor Effects

This section describes the estimation of assessor effects for repeated measures of mental and motor development used in the ECLS-B.[26] Section 4.6 considered repeated measures of mental and motor status nested within children, nested within sampling clusters. In that section, a multilevel model was used to assess individual growth among children who—for the purpose of that analysis—remained in the same sampling clusters over the course of the investigation. When children cross contextual boundaries during an investigation, the data no longer have such a neat, nested, hierarchical structure. Instead, the analysis involves cross-classifications of children by social settings that change during the course of investigation.

One such migration occurs when considering assessor effects on measures of mental and motor growth used in the ECLS-B, as shown in exhibit 4-1. The data in the table represent only a small selection of ECLS-B assessments for purposes of illustration, broken down by child, data collection, and interviewer. Each row of the table represents a child, whereas each column represents an interviewer. For brevity, only 22 children are listed. The histories of children 050 and 060 illustrate the change in assessor that occurred from the first data collection to the second. These children shared assessor 1020 at the time of the 9-month data collection but were assigned to different assessors at the time of the 2-year data collection, when child 050 was assessed by interviewer 1019 and child 060 was assessed by interviewer 1121. For the 2-year collection, child 060 joined child 150, when both were assessed by interviewer 1121.

Thus, each child assessment could have been conducted by a different interviewer. For a group of children assessed by the same assessor at the time of the 9-month data collection, some of these children might have been assessed by the same interviewer at the time of the 2-year collection, while others were assessed by yet another interviewer. An ECLS-B interviewer might have assessed all of the same children at both points in time but for the 2-year collection could have assessed additional children who were previously assessed by another interviewer. This resulted in a complex data analysis structure, where lower-level units (repeated developmental measures) were cross-classified by two higher-level units (children and interviewer assessors).

Assessor effects on infant growth can be conceived as deflections upward or downward from each child's individual growth trajectory. In principle, it is possible that measures of mental and motor

[26] A similar estimation of assessor effects in the publisher data could not be made for comparison to the ECLS-B because there is no information about assessors in the standardization dataset.

Exhibit 4-1. Data structure for a selection of ECLS-B assessments, by child, data collection and interviewer assessor: 2001–02 and 2003–04

	9-month interviewer assessor										2-year interviewer assessor							
Child	1017	1019	1020	1021	1023	1108	1111	1166	1167	Child total	1019	1067	1108	1111	1120	1121	Child total	Assessments total
Assessor total	2	1	3	1	1	8	4	1	1	22	2	3	10	4	1	2	22	44
050			1							1	1						1	2
060			1							1						1	1	2
070			1							1	1						1	2
150								1		1						1	1	2
180									1	1				1			1	2
190							1			1				1			1	2
260							1			1				1			1	2
310					1					1					1		1	2
380	1									1		1					1	2
390	1									1		1					1	2
400				1						1		1					1	2
440						1				1			1				1	2
450		1								1			1				1	2
460						1				1			1				1	2
470						1				1			1				1	2
550							1			1			1				1	2
601						1				1			1				1	2
602						1				1			1				1	2
610						1				1			1				1	2
640						1				1			1				1	2
700						1				1			1				1	2
730							1			1				1			1	2

NOTE: Selection of ECLS-B sample observations for an unweighted multilevel regression of age-ability relationships cross-classified by child and interviewer assessor. Each 1 represents a single assessment.

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

growth in the ECLS-B were deflected upward or downward by exposure to different assessors. Assessor effects would be represented by the variance of this deflection. Part of the variation in growth curves that otherwise might have been attributed to individual growth differences among infants would instead be attributed to assessor effects. Hence, it was desirable to estimate the magnitude of assessor effects. Taking assessor effects explicitly into account was expected to reduce the temporal instability in infant outcomes.

A cross-classified random effects model, estimated using full maximum likelihood estimation, was used to examine the magnitude of assessor effects on measures of child developmental status and growth used in the ECLS-B. Unweighted ECLS-B data, rather than weighted data, were used because the purpose of the analyses was to evaluate the quality of the ECLS-B assessment data, rather than to generalize the findings to the population of children. HCM2, part of the HLM 6.0 software package, was used for this purpose. ECLS-B age-ability relationships were modeled in much the same way as described previously in section 4.6 of this report. This new model again included BSID age centered at precisely 9 months of age and a corresponding BSID age-squared variable. Continuous variables representing premature birth and household SES were added to the model, as were dummy variables indicating whether a child was Black or Hispanic. The equations for the 2-level model used for the mental scale are specified below.[27] The level-1 model was estimated using 91,430 plausible values, with 5 values per child, usually at two points in time; the level-2 model was estimated using 9,412 children.

Level-1 Model

$$Y = P0 + P1*(AGE9) + P2*(AGE9SQ) + e$$

Level-2 Model

$$P0 = theta(0) + b00 + c00 + (G01)*PREMONTH + (G02)*SESML + (G03)*HISP$$

$$P1 = theta(1) + b10 + c10 + (G11)*PREMONTH + (G12)*SESML + (G13)*AFAM + (G14)*HISP$$

$$P2 = theta(2)$$

[27] The outcome variable Y is the Mental EAP ability estimate. Age9 is months of age centered on exactly 9 months so that the intercept P0 will represent mean ability at precisely 9 months of age. Age9SQ is the squared term for Age9. PREMONTH is months premature at birth. SESML is a maximum likelihood scale score representing socio-economic status. HISP is a dummy variable representing Hispanic and AFAM is a dummy variable representing African-American children.

The independent variables, which are all characteristics of children, used in this analysis were chosen because they are consistently related to status or growth, or both, in ECLS-B. They were not included in order to assess the impact of these variables, because the objective was to examine the extent to which assessor effects deflect measures of mental and motor growth between 9 months and 2 years of age. Premature birth, SES, and ethnicity were accounted for in the model so that none of these factors would be confounded with assessor effects.

In this analysis, the essential summary of mental growth in relation to age was given by the regression equation:

$$\theta_{E,\,Men} = -1.872 + 0.459x - 0.004x^2,$$

where $\theta_{E,\,Men}$ is child mental scale score and x is BSID age in ECLS-B. The interested reader may want to compare this growth curve with the formula shown earlier: $\theta_{E,Men} = -1.870 + 0.513x - 0.007x^2$. Standard errors for each of these coefficients were once again small due to the large sample of ECLS-B observations at both points in time. The standard error for ECLS-B mental initial status at precisely 9 months of age was 0.018; the standard error for the linear monthly growth rate was 0.005 and the standard error for BSID age-squared to three decimal places was 0.000. The reader may want to compare these values with those reported earlier in section 4.6. Standard errors for the cross-classified mental regression coefficients were in every case smaller than those found in the original age-ability equation reported in section 4.6 of this report.

Part of the variability that had once been attributed to individual child differences could now be attributed to assessor effects. Additionally, some of the within child variation in growth could also now be attributed to assessor effects. The intra-class correlation for mental initial status at precisely 9 months of age, conditional upon growth and the additional child control variables (premature birth, SES and ethnicity), was given by the ratio of the variance for child initial status in relation to total variance:

$$\frac{\tau_{b_{00}}}{\tau_{b_{00}} + \tau_{c_{00}} + \sigma^2} = \frac{0.386}{0.386 + 0.040 + 0.235} = 0.584,$$

or about 58 percent of total variance, where $\tau_{b_{00}}$ is between children true score variance, $\tau_{c_{00}}$ is between assessor variance, and σ^2 is random error variance. This value can be directly interpreted as a reliability coefficient, defined in classical test theory as the ratio of true score variance to total variance. From this perspective, the value of 0.584 represents the reliability of the BSF-R measure of mental status in cross-section at precisely 9 months of age. The complement to this value is 1 - 0.584 = 0.416 or about 42

percent, reflecting relatively high levels of measurement error of one form or another on the BSF-R mental test at 9 months (random error plus assessor effects).[28]

The proportion of random error in relation to total variance on the mental test was:

$$\frac{\sigma^2}{\tau_{b_{00}} + \tau_{c_{00}} + \sigma^2} = \frac{0.235}{0.386 + 0.040 + 0.235} = 0.355,$$

or roughly 36 percent of the total variance in initial status at precisely 9 months of age. Assessor effects on initial status thus appeared to be relatively small partly because random error was so large. On the BSF-R mental test, random error accounts for more than a third of the total variance in initial status at 9 months.

The intra-class correlation for assessor effects on mental initial status is given by the ratio of assessor variance to total variance:

$$\frac{\tau_{c_{00}}}{\tau_{b_{00}} + \tau_{c_{00}} + \sigma^2} = \frac{0.040}{0.386 + 0.040 + 0.235} = 0.061,$$

which represents about 6 percent of total variance in cross-section at precisely 9 months of age. Thus, the impact of assessor effects on the measurement of mental initial status was small, representing only 6 percent of total variance. As a proportion, assessor effects on mental measures represented 0.040 / 0.275 = 0.146 or about 15 percent of total measurement error (assessor effects plus random error). The complement to this figure was 1 - 0.146 = 0.854 or about 85 percent, which represented random error, by far the largest component of total measurement error found in the BSF-R mental. From this perspective, assessor effects on mental initial status appeared to be relatively small if only because random error was so large.

Variance components as usual were reported in squared units of measurement. The corresponding standard deviations for these values were $\sigma_{b_{00}} = \tau_{b_{00}}^{1/2} = \sqrt{0.386}$ = 0.621 for mental true score initial status; $\sigma_{c_{00}} = \tau_{c_{00}}^{1/2} = \sqrt{0.040}$ = 0.200 for assessor effects on mental initial status; and $\sigma = \sqrt{0.235}$ = 0.484 for the random error in initial status. These units were the same as those used in scaling and scoring the BSF-R mental test, expressed in standard deviation units of the 12-month-old cohort found in the publisher standardization dataset.

[28] Five plausible values were used to represent internal inconsistency measurement error at each point in time for each observation.

Turning to consider the impact of assessor effects on mental monthly growth rate, the variance attributed to assessor effects was $\tau_{c_{10}}$ = 0.001, while true score growth rate variance among children was $\tau_{b_{10}}$ = 0.006. With data at only two points in time, the random error component for growth rates could not be estimated. Instead, the importance of assessor effects on growth rates was obtained by considering the variance of assessor effects in relation to the true-score variance of growth rates, which was estimated to be $\tau_{c_{10}} / \tau_{b_{10}}$ = 0.001 / 0.006 = 0.105 or about 11 percent as large as the true-score variance in mental growth rates.

In standard deviation units, this was $\sigma_{c_{10}} = \tau_{c_{10}}{}^{1/2} = \sqrt{0.001}$ = 0.033 or assessor effects on mental growth rates and $\sigma_{b_{10}} = \tau_{b_{10}}{}^{1/2} = \sqrt{0.006}$ = 0.078 for true score variation in growth between children. How should one interpret the size of assessor effects in relation to growth? The average growth rate was estimated to be 0.459 population standard deviations per month, so the expected impact of an assessor effect one standard deviation above average mental growth would be 0.459 + 0.033 = 0.492 units per month, whereas the impact of an assessor effect one standard deviation below average growth would be 0.459 - 0.033 = 0.427 population standard deviations per month. The difference in mental growth rates over these two extremes would be 0.492 - 0.427 = 0.065 of a population standard deviation per month.

A parallel HCM2 analysis was used to obtain an essential summary of motor growth in relation to age in ECLS-B, as represented by the regression equation:

$$\theta_{\text{E, Mot}} = -1.742 + 0.516x - 0.015x^2,$$

where $\theta_{\text{E, Mot}}$ is the child motor scale score and x is once again BSID age. The interested reader may want to compare this growth curve with the equation shown earlier: $\theta_{\text{E,Mot}}$ = -1.726 + 0.511x - 0.015 x^2. Standard errors were again small due to the large size of the ECLS-B sample. The standard error for ECLS-B motor initial status at precisely 9 months of age was 0.022; the standard error for the linear monthly growth rate was 0.005 and the standard error for BSID age-squared to three decimal places was 0.000. Standard errors for the cross-classified random effects model of assessor effects were again smaller in every case than those reported for the original motor regression equation in section 4.6 of this report.

The intra-class correlation for motor initial status at precisely 9 months, conditional upon growth and the additional child control variables (premature birth, SES, and ethnicity), was given by the ratio of variance of child initial status to total variance:

$$\frac{\tau_{b_{00}}}{\tau_{b_{00}} + \tau_{c_{00}} + \sigma^2} = \frac{0.918}{0.918 + 0.053 + 0.155} = 0.815,$$

or about 82 percent of total variance, due to the higher reliability of the BSF-R measure of motor status at precisely 9 months of age. The complement to this value is 1 - 0.815 = 0.185 or about 19 percent, reflecting relatively low levels of measurement error of one form or another on the BSF-R motor test at 9 months (random error plus assessor effects).

The intra-class correlation for assessor effects on the BSF-R motor is given by the ratio of assessor variance to total variance:

$$\frac{\tau_{c_{00}}}{\tau_{b_{00}} + \tau_{c_{00}} + \sigma^2} = \frac{0.053}{0.918 + 0.053 + 0.155} = 0.047,$$

which represents about 5 percent of total variance in cross-section at precisely 9 months of age. Thus, the impact of assessor effects on the measurement of motor initial status was again small, representing only 5 percent of total variance.

As a part of total measurement error (assessor effects plus random error), assessor effects on the motor represented 0.053 / 0.208 = 0.255 or about 26 percent of total measurement error. The complement to this was 1 - 0.255 = 0.745 or about 75 percent, which represented random error found in the BSF-R motor test, again by far the largest component of total measurement error.

The proportion of random error in relation to total variance was:

$$\frac{\sigma^2}{\tau_{b_{00}} + \tau_{c_{00}} + \sigma^2} = \frac{0.155}{0.918 + 0.053 + 0.155} = 0.138,$$

or roughly only 14 percent of total variance. Assessor effects on motor initial status at 9 months appeared to be relatively more substantial if only because random error was so much smaller on the BSF-R motor. On the BSF-R mental, random error represented fully 36 percent of the total variance in initial status whereas on the motor random error represented only 14 percent of the total variance in initial status.

Variance components as usual were reported in squared units of measurement. The corresponding standard deviations for these values are $\sigma_{b_{00}} = \tau_{b_{00}}^{1/2} = \sqrt{0.918}$ = 0.958 for motor true score initial status; $\sigma_{c_{00}} = \tau_{c_{00}}^{1/2} = \sqrt{0.053}$ = 0.230 for assessor effects on initial status; and $\sigma = \sqrt{0.155}$

= 0.394 for the random error in initial status. These units were the same as those used in scaling and scoring the BSF-R motor test, which again were based on the motor standard deviation for the 12-month-old cohort found in the publisher standardization dataset.

Turning to consider the impact of assessor effects on motor monthly growth rate, the variance attributed to assessor effects was $\tau_{c_{10}} = 0.001$, while true score growth rate variance among children was $\tau_{b_{10}} = 0.006$. Although, to three decimal places, these values appeared to be identical to corresponding values reported previously for mental growth, in fact, these values were smaller by just a tiny fraction. The importance of assessor effects on growth rates was obtained by considering the variance of assessor effects in relation to the true-score variance of growth rates, which was estimated to be $\tau_{c_{10}} / \tau_{b_{10}} = 0.001 / 0.006 = 0.114$ or again about 11 percent as large as the true-score variance in motor growth rates.

In standard deviation units, this variation was $\sigma_{c_{10}} = \tau_{c_{10}}^{1/2} = \sqrt{0.001} = 0.025$ for assessor effects on motor growth and $\sigma_{b_{10}} = \tau_{b_{10}}^{1/2} = \sqrt{0.006} = 0.074$ for true score variation in motor growth between children. The importance of assessor effects can again be assessed in relation to average growth. The average motor growth rate was estimated to be 0.516 population standard deviations per month in cross-section, so the expected impact of an assessor effect one standard deviation above average motor growth would be 0.516 + 0.025 = 0.541 per month, while the impact of an assessor effect one standard deviation below average growth would be 0.516 - 0.025 = 0.491 population standard deviations per month. The difference in motor growth rates over these two extremes would be 0.541 –0.491 = 0.050 of a population standard deviation per month. The value of this difference was somewhat smaller for motor than it was for mental growth rates.

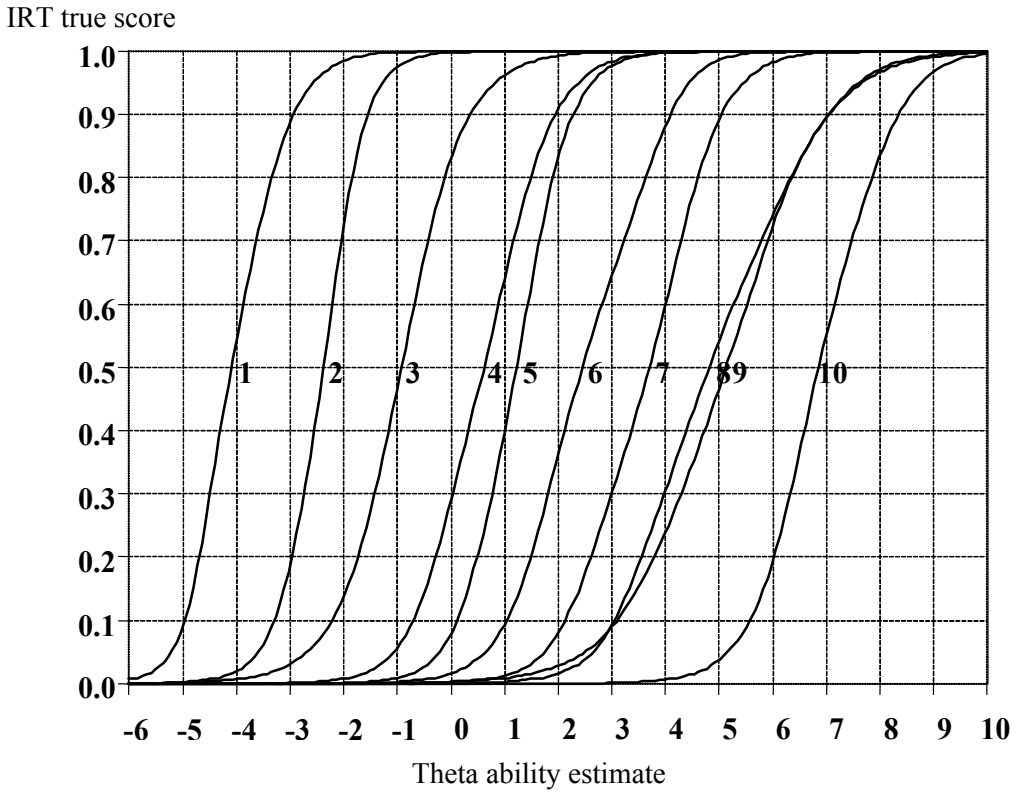
From this analysis, it seems evident that random or internal consistency measurement error was substantially more important than assessor effects on both the BSF-R mental and motor tests used in the ECLS-B. Measurement error in one form or another (random error and assessor effects) accounted for about 42 percent of total variance in initial status on the mental and about 19 percent of the total variance on the motor. The ratios of assessor effects variance to total variance in initial status was about 6 percent for the mental and about 5 percent for the motor. Random or internal consistency error variance represented fully 36 percent of the total variance in initial status on the mental and 14 percent on the motor. For both the BSF-R measures of mental and motor development, internal consistency error proved to be a more important measurement issue than assessor effects in the ECLS-B.

The variance of assessor effects in relation to true-score variance for growth rates was about 11 percent on both the mental and motor tests. On this basis, the impact of assessor effects on growth rates appeared to be moderate in the ECLS-B. However, it should be noted that the potentially important issue of test-retest reliability cannot be addressed in ECLS-B due to the design limitation with data collections at two points in time. In this context, it is perhaps worth noting that test-retest designs are never very reliable since the reliability of growth rates depends fundamentally on, and rises rapidly with, the number of observations obtained for each subject (Bryk and Raudenbush 1987; Willet 1989, 1997). This is particularly unfortunate in a longitudinal study such as the ECLS-B since the impact of existing housing conditions and other contingent social contexts primarily should be assessed in relation to developmental growth rates rather than in relation to developmental status.

4.10 BSF-R Proficiency Level Probabilities

One of the convenient features of IRT is that items and persons share the same scale metric. This implies that persons at any given level of ability can be characterized by items at that same threshold. In the ECLS-B, small clusters of items of similar content at roughly the same level of ability were used to represent developmental milestones for young children. Item clusters containing anywhere from 3 to 7 items were identified so that short subscales were built with publisher item calibrations. In the ECLS-B, 10 such subscales were identified for the mental scale and an additional 10 were also identified for the motor scale, as shown in figures 4-25 and 4-26.

Figure 4-25. Response functions for proficiency level subscales representing 10 developmental milestones on the mental scale: 1993

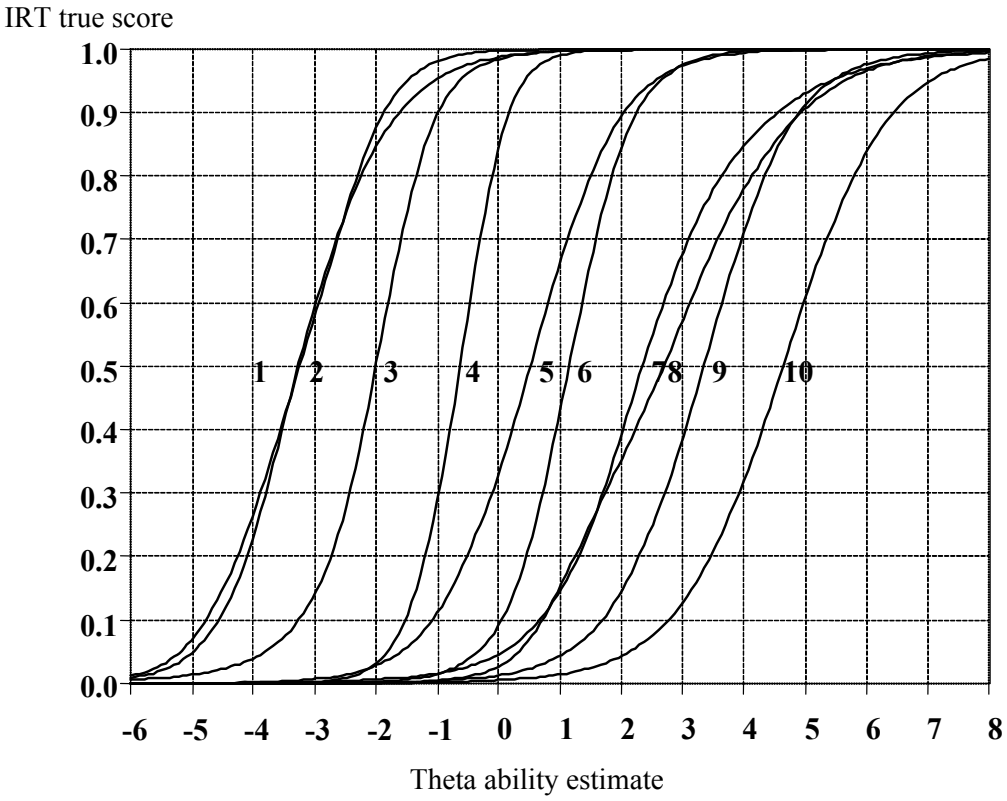


NOTE: Item Response Theory (IRT) item calibrations obtained with unweighted publisher standardization dataset observations. 1 = Explores objects; 2 = Explores purposefully; 3 = Jabbers expressively; 4 = Early problem solving; 5 = Names object; 6 = Receptive vocabulary; 7 = Expressive vocabulary; 8 = Listening/comprehension; 9 = Matching/discrimination; 10 = Early counting/quantitative.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993.

At higher levels of ability, the test curve characteristic (TCC) for these subscales equaled the total number of items in the scale.[29] To produce proficiency level probabilities, subscale true-scores were divided by the total number of items in the scale. This was equivalent to summing the probabilities computed from each of the subscale component items at a given level of ability and dividing this sum by the total number of items in the subscale, this representing the maximum possible score. This produced a response function rising from zero at low levels of ability to unity at high levels of ability. The response function represented the probability of having reached the developmental milestone represented by the items in the subscale. In this way, a proficiency level subscale performed much like a super-item, and the resulting response function looked much like an item characteristic curve.

[29] Both the ICC and the TCC have response functions that have a similar shape. The only difference is that the ICC represents a probability between 0 and 1, whereas the TCC represents the raw score. Dividing these scores by the maximum possible score on the test, the raw score can be interpreted as a probability. In this sense, proficiency level subscales can behave or act as super-items.

Figure 4-26. Response functions for proficiency level subscales representing 10 developmental milestones on the motor scale: 1993



NOTE: Item Response Theory (IRT) item calibrations obtained with unweighted publisher standardization dataset observations. 1 = Eye-hand coordination; 2 = Sitting; 3 = Pre-walking; 4 = Stands alone; 5 = Skillful walking; 6 = Balance; 7 = Fine motor control; 8 = Uses stairs; 9 = Alternating balance; 10 = Motor planning.
SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993.

Publisher item calibrations were used to build the proficiency level subscales so that a consistent scale metric was maintained. In the ECLS-B, BSF-R item responses were used to obtain a mean expected a posteriori (EAP) ability estimate. This ability estimate was then applied to the subscale response function to obtain a proficiency level probability. In this fashion, one ability estimate yielded 10 proficiency probabilities, representing the probability that a child had reached each of the 10 developmental milestones. Mean probabilities on both the mental and motor tests are reported in table 4-17.

Table 4-17. Mean proficiency level probabilities for the 10 proficiency level subscales of the BSF-R mental and motor scales at 9 months and 2 years: 2001–02 and 2003–04

Test	9 months	2 years
Mental scale		
Explores objects	0.989	1.000
Explores purposefully	0.871	1.000
Jabbers expressively	0.415	0.999
Early problem solving	0.111	0.985
Names object	0.048	0.976
Receptive vocabulary	0.015	0.848
Expressive vocabulary	0.003	0.645
Listening/comprehension	0.001	0.373
Matching/discrimination	0.002	0.326
Early counting/quantitative	0.000	0.042
Motor scale		
Eye-hand coordination	0.914	1.000
Sitting	0.894	0.999
Pre-walking	0.719	0.999
Stands alone	0.326	0.998
Skillful walking	0.182	0.928
Balance	0.092	0.897
Fine motor control	0.046	0.563
Uses stairs	0.036	0.489
Alternating balance	0.015	0.310
Motor planning	0.005	0.108

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Based on the weighted ECLS-B sample, table 4-17 reports progressively lower probabilities over a sequence of progressively more difficult developmental milestones in cross-section for both the 9-month and 2-year assessments. The one exception to this pattern is from the 9-month probability for listening/comprehension to the 9-month probability for matching/discrimination. For the 9-month data collection, 41.5 percent of the infants babbled, whereas for the 2-year data collection, virtually all of the children surpassed this developmental milestone. Generally speaking, for the 9-month assessment, an appreciably large number of infants performed successfully on the first five subscales on both the mental and motor scales. For the 2-year assessment, appreciably large numbers of children performed satisfactorily on all but the very last of the subscales.

The ECLS-B sample could be used to explicitly model each of these proficiency level probabilities as a function of age. Since probabilities are non-linear, it would be advisable to represent each probability using normal deviates, *probits* or *logits*. Proceeding in this fashion, it would be possible to use regression equations obtained with the weighted ECLS-B sample to search for the mean age at which each developmental level is attained. From a psychometrics perspective, developmental threshold would occur at the age where the predicted probability is 0.5. Another approach would be to consider substantive thresholds where developmental mastery would occur at the age where the predicted probability would be, for example, 0.67 or 0.8.[30] The ECLS-B proficiency level subscales consisting of items taken from the BSID-II, are reported in table 4-18, along with publisher item parameters.

4.11 BSF-R Differential Test and Item Functioning

BSF-R 9-month and 2-year item sets were examined for evidence of DTF and DIF in the ECLS-B. This involved comparisons of test performance between a focal group (e.g., African American children) and a reference group (e.g., White children), once individuals in the two groups have been matched or *blocked* on their ability estimates. It was not expected that the different subgroups would perform identically on the same test. Rather, children from two different groups, *who were otherwise identical in terms of their overall ability*, should have had the same probability of obtaining correct responses to the set of items. There should have been no relative advantage or disadvantage in obtaining correct responses based on the child's subgroup membership.

[30] If the probability of mastery is set very high—for example 0.9 or 0.95—then the mastery age will drift very far away from the age threshold, which is the age where infants are actually acquiring the skill set represented by the developmental milestone. Mastery probabilities of 0.67 or 0.8 are merely suggested as compromises that will keep the mastery age in the vicinity of the age threshold.

Table 4-18. Proficiency level subscales for the BSF-R mental and motor scales: Items in each proficiency level subscale and their IRT ability and discrimination parameters: 2001–02 and 2003–04

Proficiency level subscale label	BSID-II item number and item label		Item difficulty *(b)*	Item discrimination *(a)*
Explores objects	MEN045	Picks up Cube	-4.813	2.501
	MEN048	Plays with String	-4.796	1.836
	MEN052	Bangs in Play	-3.930	1.158
	MEN053	Reaches for Second Cube	-3.819	1.218
	MEN055	Lifts Inverted Cup	-4.361	1.402
	MEN057	Picks up Cube Deftly	-3.773	1.167
	MEN059	Manipulates Bell, Showing Interest in Detail	-2.963	1.643
Explores purposefully	MEN062	Pulls String Adaptively to Secure Ring	-2.652	1.096
	MEN065	Retains Two of Three Cubes for 3 Seconds	-2.405	1.616
	MEN066	Rings Bell Purposely	-2.393	1.546
	MEN069	Looks at Pictures in Book	-2.192	1.805
Jabbers expressively	MEN076	Jabbers Expressively	-0.749	0.940
	MEN078	Vocalizes Four Different Vowel-Consonant Combinations	-1.114	0.838
	MEN081	Responds to Spoken Request	-1.015	1.233
Early problem solving	MEN089	Puts Six Beads in Box	-0.280	1.521
	MEN095	Puts Nine Cubes in Cup	0.692	0.953
	MEN102	Retrieves toy (Visible Displacements)	1.021	1.099
	MEN104	Uses Rod to Attain Toy	1.012	1.177
Names object	MEN100	Uses Two Different Words Appropriately	0.734	1.316
	MEN101	Shows Shoes, Other Clothing, or Object	0.746	1.569
	MEN106	Uses Word(s) to Make Wants Known	1.613	1.969
	MEN110	Names One Object	1.732	1.186
Receptive vocabulary	MEN099	Points to Two Pictures	1.944	1.066
	MEN108	Points to Three of Doll's Body Parts	1.919	1.228
	MEN122	Points to Five Pictures	3.660	1.452

See note at end of table.

Table 4-18. Proficiency level subscales for the BSF-R mental and motor scales: Items in each proficiency level subscale and their IRT ability and discrimination parameters: 2001–02 and 2003–04—Continued

Proficiency level subscale label	BSID-II item number and item label		Item difficulty *(b)*	Item discrimination *(a)*
Expressive vocabulary	MEN111	Combines Word and Gesture	2.487	1.615
	MEN114	Uses A Two-Word Utterance	3.109	0.896
	MEN121	Uses Pronoun(s)	3.975	1.197
	MEN126	Names Three Objects	4.144	1.283
	MEN133	Names Five Pictures	4.365	1.160
Listening/comprehension	MEN131	Attends to Story	3.511	1.397
	MEN134	Displays Verbal Comprehension	4.459	0.939
	MEN140	Understands Two Prepositions	5.184	0.865
	MEN142	Multiple-Word Utterances Response to Picture Book	6.524	0.932
Matching/discrimination	MEN125	Matches Pictures	3.967	1.003
	MEN128	Matches Three Colors	4.052	0.667
	MEN137	Matches Four Colors	5.194	0.801
	MEN144	Discriminates Pictures I	5.593	1.293
	MEN151	Discriminates Pictures II	6.521	0.794
Early counting/quantitative	MEN141	Understands Concept of One	5.925	1.327
	MEN146	Counts (Number Names)	6.464	1.871
	MEN147	Compares Masses	6.486	1.105
	MEN152	Repeats Three Number Sequences	6.770	0.857
	MEN156	Understands Concept of More	7.520	1.193
	MEN159	Counts (Stable Number order)	6.933	1.276
	MEN164	Counts (Cardinality)	8.039	1.364
Eye-hand coordination	MOT031	Uses Partial Thumb Opposition to Grasp Cube	-3.845	1.267
	MOT032	Attempts to Secure Pellet	-3.741	1.135
	MOT041	Uses Whole Hand to Grasp Pellet	-2.859	0.919
	MOT049	Uses Partial Thumb Opposition to Grasp Pellet	-2.642	0.720

See note at end of table.

Table 4-18. Proficiency level subscales for the BSF-R mental and motor scales: Items in each proficiency level subscale and their IRT ability and discrimination parameters: 2001–02 and 2003–04—Continued

Proficiency level subscale label	BSID-II item number and item label		Item difficulty *(b)*	Item discrimination *(a)*
Sitting	MOT022	Sits with Slight Support for 10 Seconds	-4.469	1.170
	MOT028	Sits Alone Momentarily	-4.142	1.162
	MOT034	Sits Alone for 30 Seconds	-3.195	1.082
	MOT036	Sits Alone Steadily	-3.260	0.974
	MOT043	Moves Forward Using Prewalking Methods	-2.883	1.023
	MOT051	Moves from Sitting to Creeping Position	-2.400	1.520
Pre-walking	MOT044	Supports Weight Momentarily	-2.751	0.625
	MOT045	Pulls to Standing Position	-2.488	1.219
	MOT046	Shifts Weight while Standing	-2.257	1.403
	MOT052	Raises Self to Standing Position	-2.001	1.895
	MOT053	Attempts to Walk	-1.724	1.223
	MOT054	Walks Sideways while Holding on to Furniture	-1.604	1.801
Stands alone	MOT059	Stands up I	-0.490	1.395
	MOT060	Walks with Help	-1.186	1.676
	MOT061	Stands Alone	-0.669	1.734
	MOT062	Walks Alone	-0.344	2.274
Skillful walking	MOT063	Walks Alone with Good Coordination	-0.295	0.917
	MOT067	Walks Backward	0.760	1.088
	MOT071	Walks Sideways	1.020	0.829
Balance	MOT065	Squats Briefly	1.386	1.087
	MOT068	Stands up II	0.993	1.070
	MOT072	Stands on Right Foot with Help	1.006	1.267
	MOT073	Stands on Left Foot with Help	1.213	1.371
Fine motor control	MOT074	Uses Pads of Fingertips to Grasp Pencil	2.037	1.077
	MOT075	Uses Hand to Hold Paper in Place	2.199	0.763
	MOT090	Grasps Pencil at Nearest End	3.227	0.513

See note at end of table.

Table 4-18. Proficiency level subscales for the BSF-R mental and motor scales: Items in each proficiency level subscale and their IRT ability and discrimination parameters: 2001–02 and 2003–04—Continued

Proficiency level subscale label	BSID-II item number and item label		Item difficulty *(b)*	Item discrimination *(a)*
Uses stairs	MOT069	Walks down Stairs with Help	1.058	1.420
	MOT079	Walks up Stairs Alone, Placing Both Feet on Each Step	2.632	0.949
	MOT080	Walks down Stairs Alone, Placing Both Feet on Step	3.104	0.867
	MOT095	Walks up Stairs, Alternating Feet	4.210	0.675
Alternating balance	MOT082	Stands Alone on Right Foot	2.887	0.950
	MOT083	Stands Alone on Left Foot	2.959	0.754
	MOT086	Swings Leg to Kick Ball	3.803	1.194
	MOT089	Walks on Tiptoe for Four Steps	3.584	0.697
Motor planning	MOT088	Laces Three Beads	3.892	0.643
	MOT091	Imitates Hand Movements	4.208	0.830
	MOT093	Manipulates Pencil in Hand	4.524	0.866
	MOT096	Copies Circle	4.506	0.694
	MOT098	Imitates Postures	5.059	0.820
	MOT101	Buttons One Button	5.382	0.787

SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

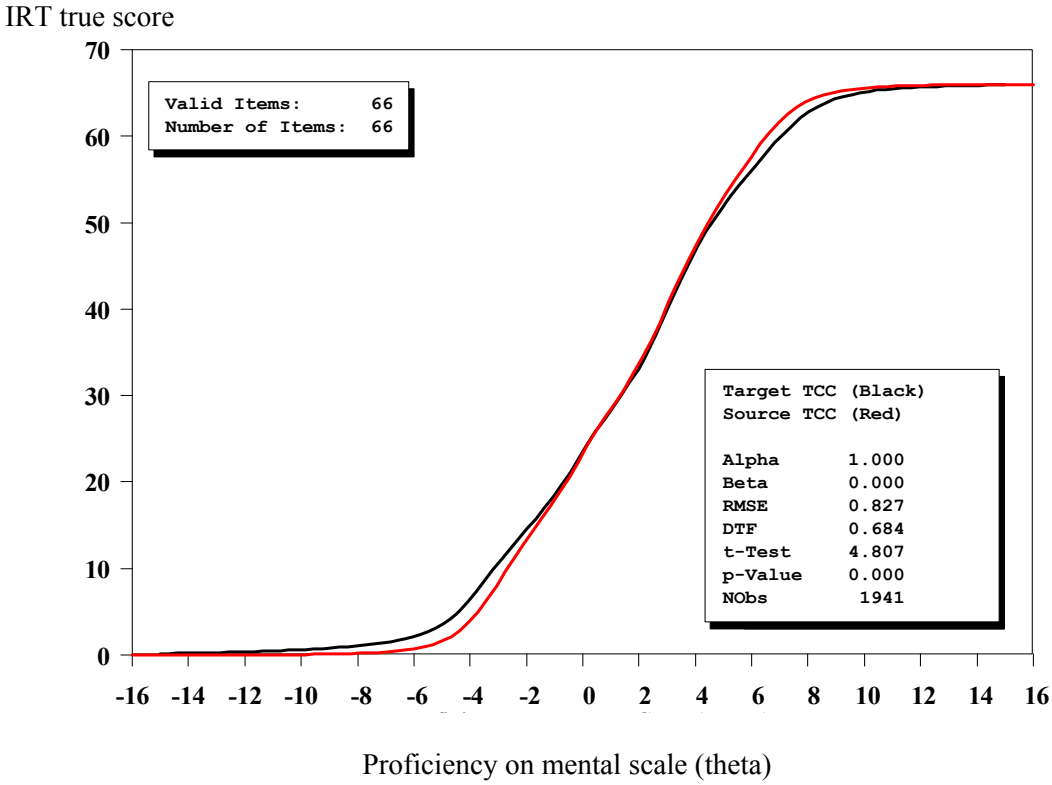
A test is said to exhibit DTF when individuals having the same ability, but from different groups, fail to obtain the same number of correct responses. IRT provides a unified framework for investigating issues of statistical bias at both the test and item levels. A test shows evidence of statistical bias when, at the same level of ability, two groups fail to obtain the same score. DTF is examined in the ECLS-B using parametric IRT procedures developed by Raju, van der Linden, and Fleer (1995).

For this purpose, a series of separate response vector files were created for focal minority groups and reference majority groups using observations obtained at 9 months and 2 years. Each file was then scored separately using identical sets of BSF-R item parameters. The scoring effectively classifies each observation by ability level. As each observation was scored separately in each group, marginal likelihoods were accumulated for each item response across all levels of ability. Once all observations were scored in this fashion, new sets of IRT parameters were fitted to the marginal likelihoods in a single iteration. The new sets of item parameters represented the response characteristics for each respective focal or reference group across all levels of ability.

The issue to be addressed in DTF analysis was whether children at the same level of ability on average obtained the same number-right score on the same test. This issue was examined in IRT by comparing the TCCs for the two groups. The TCC is the sum of the ordinates of the ICCs at each level of ability, $\xi = \sum_{j=1}^{n} P_j(\theta)$. The TCC represents the expected number of correct responses, expressed in raw score metric, equivalent to the number of items that would be answered correctly on a test. Any misalignment of TCCs reveals evidence of DTF. The total number-right score at each level of ability was examined by comparing IRT true-scores for each focal and reference group comparison.

The new sets of item parameter estimates were used for these group comparisons. The TCC for the focal (source) and reference (target) tests were compared across all levels of ability. The weighted sum of squared differences between the source and target test characteristic curves was used as a DTF index. The DTF coefficient quantified the degree of misalignment between the two curves, expressed in squared raw score units. The square root of the DTF coefficient was an RMSE, expressed in raw score units. The magnitude of RMSE values were interpreted bearing in mind the maximum raw score possible or the average raw score on the test in question. These residual measures of dispersion around the target TCC were the DTF statistics most frequently reported in the literature, (Raju, van der Linden, and Fleer 1995) as shown in figure 4-27.

Figure 4-27. DTF analysis, showing mental test characteristic curves for BSF-R Asian focal group and White reference group before equating: 2001–02 and 2003–04



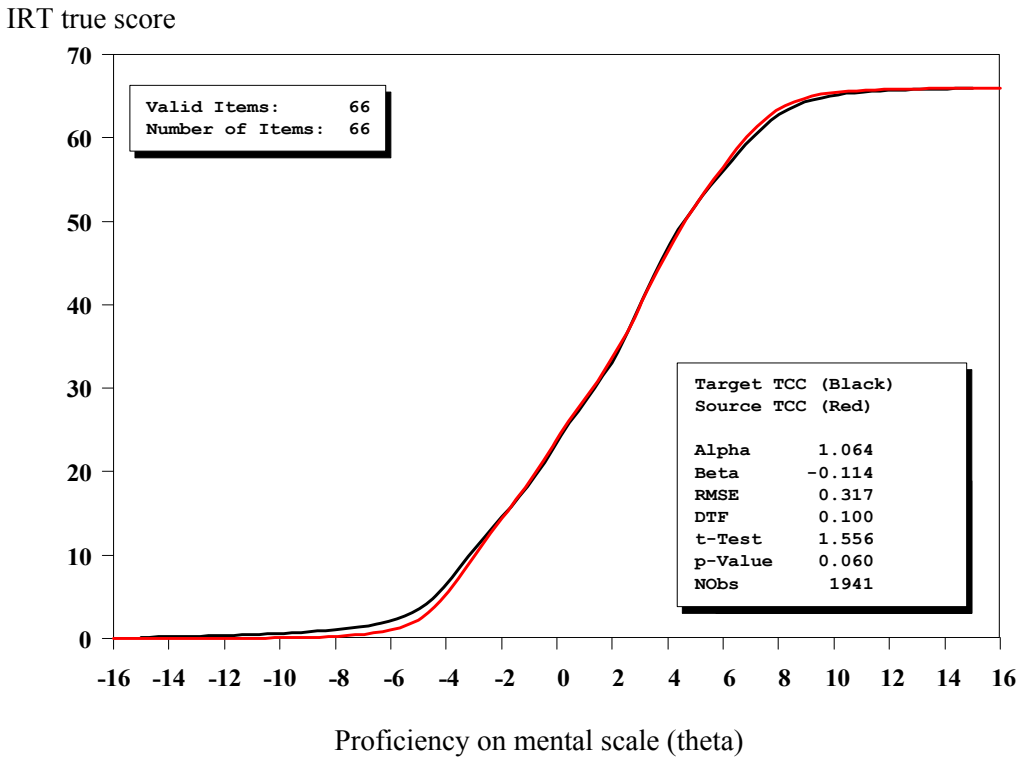
NOTE: Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Alpha = linear transformation of scale; Beta = linear transformation of origin; RMSE = root mean squared error; DTF = differential test functioning; t-Test = statistical test of the difference between the two curves shown in the figure; p-Value = probability of the results of the t-Test; NObs = number of observations.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

However, from a practical point of view, researchers were more concerned about the overall magnitude and direction of statistical bias as this may have affected ability estimates. With large samples such as the ECLS-B, virtually any DTF coefficient was statistically significant. This implied that it was appropriate to generalize from the focal and reference group samples to the same groups in the ECLS-B population and affirmed that *at least some DTF greater than zero* exists when these instruments were used with these subgroups of the population. If some statistical bias existed in the population, then it was often more meaningful to ascertain the overall direction and magnitude of this statistical bias.

Thus, it was also helpful to consider the average overall difference between test scores in the two groups in terms of the population standard deviation units expressed by the IRT scale metric. Estimates of the average overall statistical bias were obtained with IRT true-score equating, which

showed the linear transformation of origin and scale that would be needed to align the source (focal) and target (reference) tests. In the context of DTF analysis, equating constants α (slope) and β (origin) were expressions of the overall statistical bias expected when the assessment instrument was used with the focal group. The overall group effect was represented by the intercept coefficient β, whereas a group by ability interaction effect was represented by the slope coefficient α, as shown in figure 4-28.

Figure 4-28. DTF analysis, showing mental test characteristic curves for BSF-R Asian focal group and White reference group after equating: 2001–02 and 2003–04



NOTE: Item Response Theory (IRT) item calibrations obtained with weighted ECLS-B sample observations and unweighted publisher standardization dataset observations. Alpha = linear transformation of scale; Beta = linear transformation of origin; RMSE = root mean squared error; DTF = differential test functioning; t-Test = statistical test of the difference between the two curves shown in the figure; p-Value = probability of the results of the t-Test; NObs = number of observations.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

Both equating coefficients α and β were reported in population standard deviation units and were, thus, effect-size measures of the average statistical bias of focal group ability estimates relative to reference group ability estimates. Under conditions of perfect test alignment in the two groups and no evidence of DTF, the expectation was to find origin $\beta = 0$ and slope $\alpha = 1$, indicating that no statistical bias was present when comparing the two groups. When $\beta \neq 0$, a group effect was present, demonstrating

that some spurious group-related trait unrelated to the trait purportedly measured by the test was also captured by the assessment instrument. When $\alpha \neq 1$, a group-ability interaction effect unrelated to the test objective was also present. Although these coefficients were reported less frequently in the DTF literature, conceptually they were very useful and easy to understand.

DTF statistics for the ECLS-B are reported in table 4-19 for the three focal and reference group comparisons considered in the ECLS-B. With the large sample size available in the ECLS-B, many DTF and RMSE measures were statistically significant prior to equating, whereas virtually no measure of dispersion between TCCs was statistically significant once α and β were used to relate the focal and reference groups. This demonstrates rather conclusively that a group effect and group by ability interaction effect accounted for virtually all of a distinctly linear form of statistical bias. At risk of little or no simplification, the statistical bias represented by the β coefficient, conveniently expressed in population standard deviation units, was generally sufficient to summarize the overall difference in test performance.

Inspecting the β coefficient values in table 4-19 reveals statistical biases that ranged from minute to small for the nine focal and reference group comparisons considered in this exercise. These comparisons included three race-ethnicity focal groups (African Americans, Hispanics, and Asians), one comparison each for gender, premature and SES focal groups, and three comparisons for maternal-child attachment behavior focal groups as measured and identified by the TAS-45 in the ECLS-B. Small statistical biases affected each minority group on the mental test, ranging from -0.010 population standard deviation bias for the low-SES group down to -0.221 for attachment style D (disorganized) children. Indeed, there was evidence of a modest amount of DIF for both the C and D attachment styles on both ECLS-B instruments. By contrast, there was little, if any, evidence of statistical bias of appreciable magnitude for African Americans or any other minority ethnic group on either ECLS-B test, except possibly a small -0.114 population standard deviation bias for Asians on the mental test.

Table 4-19. BSF-R differential test functioning (DTF) statistics (DTF index, RMSE, Alpha, and Beta) for focal group—reference group comparisons on the mental and motor scales: 2001–02 and 2003–04

Focal—reference group comparison	Statistic	Mental scale	Motor scale
Black — White			
	DTF	0.080	0.057
	RMSE	0.282	0.239
	Alpha	1.004	1.023
	Beta	-0.058	-0.025
Hispanic — White			
	DTF	0.055	0.005
	RMSE	0.235	0.073
	Alpha	0.994	1.001
	Beta	-0.019	0.005
Asian — White			
	DTF	0.684	0.123
	RMSE	0.827	0.351
	Alpha	1.064	1.020
	Beta	-0.114	-0.049
Female — Male			
	DTF	0.040	0.469
	RMSE	0.199	0.685
	Alpha	0.992	1.006
	Beta	0.042	0.140
Low SES — High SES			
	DTF	0.024	0.041
	RMSE	0.156	0.202
	Alpha	0.997	1.014
	Beta	-0.010	-0.034
Premature — Full Term			
	DTF	0.132	0.168
	RMSE	0.363	0.409
	Alpha	1.022	1.020
	Beta	-0.025	-0.051
A — B Attachment Style			
	DTF	0.140	0.064
	RMSE	0.375	0.253
	Alpha	1.002	1.008
	Beta	-0.072	-0.049
C — B Attachment Style			
	DTF	1.058	0.379
	RMSE	1.029	0.616
	Alpha	1.015	1.010
	Beta	-0.216	-0.111

See note at end of table.

Table 4-19. BSF-R differential test functioning (DTF) statistics (DTF index, RMSE, Alpha, and Beta) for focal group—reference group comparisons on the mental and motor scales: 2001–02 and 2003–04—Continued

Focal—reference comparison	Statistic	Mental scale	Motor scale
D — B Attachment Style			
	DTF	1.082	0.472
	RMSE	1.040	0.687
	Alpha	1.016	0.990
	Beta	-0.221	-0.127

NOTE: An item exhibits DIF "if individuals of the same ability, but from different groups, do not have the same probability of getting the item right" (Hambleton, Swaminathan, and Rogers 1991, p. 110). Alpha = linear transformation of scale; Beta = linear transformation of origin; RMSE = root mean squared error; DTF = differential test functioning.
SOURCE: U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.

DTF and RMSE coefficients reported in table 4-19 are primarily reflections of this same, relatively small statistical bias, as this affected the number-right raw score. The largest RMSE value was a 1.040 raw score point disparity at any level of ability for the D—B attachment style comparison. These coefficients revealed nothing about the direction of statistical bias, whereas coefficient β showed that this was negative or zero for virtually all groups except possibly for females on the motor test. Several DTF and RMSE coefficients were statistically significant prior to a linear transformation, whereas rarely were any of the differences statistically significant after a linear transformation of origin and of scale. This showed that the principal difference between focal and reference groups was usually a question of systematic linear statistical bias. However, it is important to emphasize that DTF should be observed and quantified prior to any such linear transformation.

DIF has also been examined in the ECLS-B. DIF identified individual items that showed an unexpectedly large difference in the probability of a correct response when comparing individuals in the focal and reference groups at the same level of ability. For the ECLS-B sample, DIF indices were calculated using the parametric IRT procedures developed by Raju, van der Linden, and Fleer (1995). Table 4-20 summarizes these results, showing all of the items on the BSF-R that exhibited weighted RMSEs of 0.10 or more. Items showing lower levels of DIF have been excluded from the table in order to save space. Like the DTF indices presented previously, these weighted root mean square NC-DIF indices reflect the magnitude of the distance between ICCs but not the direction of bias.

Table 4-20. BSF-R differential item functioning (DIF) for mental and motor items that exhibited weighted root mean squared errors (RMSEs) of 0.10 or more: 2001–02 and 2003–04

		Selected demographic characteristics						Attachment style classifications		
Item	Label	Premature	Low SES	Female	Black	Hispanic	Asian	A_ Avoidant	C_ Ambivalent	D_ Disorganized
MEN058	Retains Two Cubes for 3 Seconds (Basal)	†	†	†	†	†	†	0.112	†	†
MEN099	Points to Two Pictures (Basal)	†	†	†	†	†	†	†	0.117	0.134
MEN101	Shows Shoes, Other Clothing, or Object (Ceiling)	†	†	†	†	†	†	†	0.170	†
MEN102	Retrieves toy (Visible Displacements) (Ceiling)	†	†	†	†	†	0.104	†	†	†
MEN104	Uses Rod to Attain Toy (Ceiling)	†	†	†	†	†	†	†	0.106	†
MEN107	Follows Directions (Doll) (Basal)	†	†	†	†	†	†	†	†	0.116
MEN108	Points to Three of Dolls Body Parts (Basal)	†	†	†	†	†	†	†	0.113	0.139
MEN110	*Names One Object (Basal)*	*†*	*†*	*†*	*†*	*0.105*	*†*	*†*	*†*	*†*
MEN117	Imitates A Two-Word Sentence (Core)	†	†	†	0.131	†	†	†	†	†
MEN122	Points to Five Pictures (Core)	†	†	†		†	†	†	†	0.107
MEN123	Builds Tower of Six Cubes (Core)	0.131	†	†	†	0.116	†	†	0.106	0.140
MEN124	Discriminates Book, Cube and Key (Core)	†	†	†	†	†	†	†	†	0.131
MEN128	Matches Three Colors (Core)	†	†	†	0.132	†	†	†	†	†
MEN131	Attends to Story (Core)	†	†	†	†	†	†	0.141	0.140	0.190
MEN137	Matches Four Colors (Core)	†	†	†	0.140	†	†	†	†	†
MEN141	Understands Concept of One (Core)	†	†	†	†	0.102	0.105	†	†	†
MOT074	Uses Pads of Fingertips to Grasp Pencil (Core)	†	†	0.137	†	0.117	†	†	†	†
MOT075	Uses Hand to Hold Paper in Place (Core)	†	0.109	0.137	†	†	†	†	0.157	0.141
MOT084	Walks Forward on Line (Core)	†	†	†	†	†	†	†	0.107	0.116
MOT090	Grasps Pencil at Nearest End (Core)	†	†	0.144	†	†	†	†	†	†
MOT093	Manipulates Pencil in Hand (Core)	†	†	0.132	†	†	†	†	†	†
MOT102	Stands Alone on Left Foot for 4 Seconds (Ceiling)	†	†	†	†	†	†	†	0.211	†

† Not applicable.

NOTE: An item exhibits DIF "if individuals of the same ability, but from different groups, do not have the same probability of getting the item right" (Hambleton, Swaminathan, and Rogers 1991, p. 110). Items are considered to exhibit expressive DIF when the weighted root mean squared difference between focal and reference group item characteristic curves exceeds 0.10 or 10 percentage points. SES = socioeconomic status.

SOURCE: Publisher BSID-II standardization dataset, The Psychological Corporation, 1993; U.S. Department of Education, National Center for Education Statistics, Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), 9-month and 2-year data collections, 2001–02 and 2003–04.