www.brookespublishing.com

HISTORY OF CUT SCORES DEVELOPMENT FOR OSEP CHILD OUTCOMES REPORTING IN AEPSi
A Technical Report

ORIGINAL ISSUANCE OF GUIDELINE FROM THE EARLY CHILDHOOD OUTCOMES (ECO) CENTER

ECO issued a paper on July 5, 2006, entitled "Recommendation of the Early Childhood Outcomes (ECO) Center for Determining Age Expected Functioning and the Points of the ECO Rating Scale." This was emailed to the publishers/developers of the online assessment systems on July 20, 2006 after the paper was approved by the Office of Special Education Programs (OSEP) in the U.S. Department of Education. The paper summarized the three child outcomes (positive social-emotional skills, acquisition and use of knowledge and skills, and use of appropriate behaviors to meet needs) and clarified the expansion from the original three categories to four categories in April 2006. This was subsequently expanded that fall to the currently used five categories—a) percentage of children who did not improve functioning, b) percentage of children who improved functioning but not sufficient to move nearer to functioning comparable to same-aged peers, c) percentage of children who improved functioning to a level nearer to same-aged peers but did not reach it (improved developmental trajectory, i.e., rate of growth), d) percentage of children who improved functioning to reach a level comparable to same-aged peers (i.e., gap closers), and e) percentage of children who maintained functioning at a level comparable to same-aged peers. ECO offered these guidelines in part because it was recognized that many states appreciate the value of and prefer to use a criterion- or curriculum-based measure rather than one that provides age equivalents or percentile scores (and even if a tool does have age equivalents or percentiles, cutoffs are still needed to help states determine age-expected functioning).

Publishers were asked to work with ECO to develop guidance that would allow states to use their instruments reliably to classify children's functioning as age-expected or not. They were advised they could approach the task either empirically or conceptually. The publishers were asked to age band the items of their tests for the identification of expected scoring patterns for children at different ages. ECO stated in this paper that "whether assessment publishers approach age banding empirically or conceptually and regardless of whether the instrument is normed or curriculum-based, the publishers (and the states) need guidance from the ECO Center on where to draw the line for what is within and outside the bounds of age expected behavior."

The ECO Center convened two groups of advisors, the Child Technical Work Group and the Implementation Work Group, and after conference calls and in-person meetings, the recommendation emerged to provide a single guideline for assessment publishers rather than having individually determined state guidelines. ECO recommended using the normal curve (i.e., functioning with regard to the three outcomes is normally distributed) and 10% as the estimate of the percentage of the general population ages birth to 5 years considered to have a disability or a delay. Using the normal curve to anchor typical functioning, the ECO Center recommended considering children functioning in the bottom 10% of an outcome area to be

considered functioning below age expectation and all others to be within the range of their same-aged peers.

ECO also issued a table showing the recommended percentage of the population that would fall within each rating on their 7-point scale, the Child Outcomes Summary Form. This table also showed the corresponding upper and lower bounds on the standard deviations for each rating, developmental quotients for assessments with a mean score of 100, and t-scores around a mean of 50. The scale was developed to be sensitive to change among the lowest functioning children and not at all sensitive to change within age expectations (except a bit for ratings 6 and 7). The boundaries were deliberately set to have different percentages of children at each rating. This was done to reflect the fact that there are more atypical children with milder delays than severe delays. An extract from a portion of the table follows:

ECO Rating	Cumulative proportion of population that is this rating or above
7	0.8413
6	0.9032
5	0.9332
4	0.9608
3	0.9803
2	0.9949
1	

The recommendation and the detail in the table were shared with publishers as a guideline by which to identify functioning comparable to same-aged peers, with the understanding that the precision in the table would exceed the level of precision available through many of the assessment tools and processes in place.

Brookes Publishing subsequently participated in the conference call held by ECO on September 12, 2006 with publishers and developers to help them further in this identification of age-expected functioning.

ANALYSIS BY EMRG (Early Intervention Management and Research Group, the AEPS authors)

Over the period from September 2006 through June 2007, the authors/developers of the AEPS undertook the process of identifying cut scores within the AEPS raw scores to identify age-expected functioning. EMRG elected to take on the additional work to do this empirically rather than conceptually. Using data gathered in previous studies conducted by EMRG members, and working with their statisticians familiar with IRT and Rasch modeling, EMRG undertook to:

1. Re-establish and cross-check OSEP cut scores with previously determined OSEP cut scores for Time 1 (now referred to as Near Entry) for 6 sub-databases by Outcome, Level (of the AEPS Test, birth to 3 and 3 to 6), and in 3-month intervals.

2. Establish additional cut scores for Time 2 (now referred to as Near Exit) by Outcome, by Level, and by 3-month intervals, and to establish layers for making decisions concerning the five categories (commonly referred to as baskets).

3. Develop a large set of tables of cut scores for OSEP for both time periods for six sub-databases by Outcome, by Level, and in 3-month intervals that could be turned into an algorithm.

Explanation of Analysis Steps

AEPS Outcome Raw Scores for each of the three child outcomes for each level of the AEPS Test (birth to 3 and 3 to 6) were analyzed using Winsteps 3.6.1 to generate Rasch measures.

A regression-informed line was used to re-establish age-expected functioning cut measures (i.e., a benchmark and a cut score) utilizing the ECO criterion (from the table mentioned above) at 3-month intervals.

The Rasch age-expected functioning cut measures were transformed back to the AEPS scale.
The transformed AEPS scale scores were used to create an OSEP benchmark (0.9032) and an additional cut score (0.9608) necessary for OSEP near-exit reporting categories. Plotting raw scores and age, for each outcome, in scatter plots, three layers were identified in the data—an OSEP benchmark, and one and two layers below the benchmark.

OSEP Categories Decision Tree

In this manner the EMRG work group was able to generate empirically supported classification decision tables, valid and usable for OSEP reporting. When there was limited variability between the benchmark and the cut score, programming was developed in AEPSi so that the scoring notes of A (assistance provided) and B (behavior interfered) could be used to help with the classification of children at near exit.

These decision tables, which show all the possible combinations of Near-Entry and Near-Exit scores, were then used by the AEPSi programmers to enable automatic generation in AEPSi of the a, b, c, d, and e ratings (the five categories) on the three OSEP child outcomes.

For the Level II tables (the 3 to 6 years AEPS Test), test scores from 2,115 children across 16 states were evaluated. For the Level I tables (birth to 3 years), test scores from 1,163 children across 17 states were analyzed.

REANALYSIS ENCOURAGED BY ECO

Two years later, after having the opportunity to study the child outcomes data submitted by the states to OSEP, ECO recommended that the original assumptions be reanalyzed and data and procedures re-analyzed to determine if the reporting was capturing trends and patterns in the data as well as possible. Overall, and across all online publishers' tools, far more children were being reported as falling in category e (% of children who maintained functioning at a level comparable to same-aged peers) than had been expected for children receiving services through Part B, Section 619, and Part C.

ECO further offered to work with the publishers to assist in these re-analyses. Brookes Publishing began this process in the summer of 2009, when ECO, EMRG, and Brookes entered into a mutual nondisclosure agreement (mutual NDA) to protect the confidentiality and privacy of the data and the online programming. In July, 2010, ECO provided Brookes and EMRG with its report, "Assessment, Evaluation, and Programming System for Infants and Children Algorithm, Analysis Final Report." This analysis, while similar to the one conducted in 2007 to arrive at the original cut scores, used newer data and was able to be drawn from a larger population of typically developing children than the data available 2 years earlier. The steps of the study are described below.

Analytic Process

The cut points were developed for total scores for each of the three OSEP child outcomes. Total scores for each of the three outcomes were computed using the crosswalks (between the three outcomes and AEPS Test items) developed by EMRG. The cut points delineate a range of functioning from no immediate foundational skills to all skills considered age appropriate. Empirical validation was completed by replicating analyses across multiple sets of data for children who were identified as typically developing, at risk, or developmentally delayed. Initial clinical validation was done using a review of the skills associated with each age band.

The analysis to create the cut points included several steps. The first step was to clean and compile the data sets used to estimate typical development and the data sets used to validate the cut points. Next ECO scaled the instruments using Rasch modeling and checked for misfit items. Once misfit items were removed, observed means were plotted and smoothed. Smoothed estimates of typical functioning, mean of observed standard deviations (SDs) across age bands, and Rasch item parameters were used to develop the cut scores. The cut scores were applied to several different samples of children with developmental delay and typically developing children as one way of checking the validity of the cut scores. Preliminary content validation of the cut point that demarcates typical development was completed using information from the Rasch scaling and expert review.

Findings

The results are a new set of cut points that can be applied to the *birth to 3 years* and the *3 to 6 years* AEPS Tests to transform scores into the five child outcomes categories of a through e.

Step 1: Cleaning and Compiling Data sets

To estimate the score that would be considered age appropriate functioning for each age band, the sample was restricted to children who were identified as typically developing. The original plan was to create a data set that was 5% to 10% children with disabilities, but the inclusion of children with disabilities in the sample resulted in mean scores by age band that were far below what would be expected. The data used for the *birth to 3 years* analysis included typically developing infants and toddlers combined across three data sets including the data shared by EMRG that was used to develop an earlier set of cut scores (from here on referred to as the Calibration data set) and the first and second batch of de-identified AEPSi data supplied to ECO by Brookes. The data set used for the analysis of the *3 to 6 years* data included typically developing children from the second de-identified AEPSi data sent to ECO.

The following analyses included only children who were identified as typically developing.

Step 2: Computing Rasch Models

Once the final data sets were compiled, the three OSEP outcomes for *birth to 3 years* and *3 to 6 years* were analyzed using Rasch methods to determine if all of the items fit within the scale and to calculate the item parameters. The items were mapped to the three outcomes using the crosswalk provided by EMRG. Items with fit statistics above 2 were removed from the calculation of the total score (for more information in interpreting fit statistics see http://www.winsteps.com/winman/index.htm?diagnosingmisfit.htm). All misfit items were on the *birth to 3 years* assessment. (The range of ability seen in infants and toddlers across the skills associated with each outcome makes calibration of items challenging.) There were no misfit items on the assessment for the *3 to 6 years* assessment.

The five criteria removed from the birth to 3 crosswalk were: quiets to a familiar voice (Social Communication C1.5), uses sensory examination with objects (Cognitive F1.4), uses three-word agent-action-object utterances (Social Communication D3.4), says nursery rhymes (Cognitive G6.2) and swallows liquids (Adaptive A1.4). (This revised crosswalk for *birth to 3 years* has replaced the formerly posted crosswalk on the Brookes web sites and has been supplied to ECO for posting on their web site.)

Step 3: Smoothing Observed Means

Samples of children identified as typically developing assessed on the *birth to 3 years* or on the *3 to 6 years* level of the AEPS Test were included in the final data sets used to estimate the cut score for typically developing children. The children included in the *birth to 3 years* sample included typically developing infants and toddlers combined across three data sets (*n* = 571). The children included in the *3 to 6 years* sample were typically developing children from the second de-identified AEPSi data sent to ECO by Brookes (*n* = 1307).

Total scores were computed for each outcome. When items were missing, mean substitution was used to compute the total score. If more than 70% of items within an outcome were missing, a total score was not computed for that outcome. The SPSS syntax used to compute these total scores was supplied to Brookes by ECO. The distributions of total score on each of the outcomes were inspected to determine the quality of information they would provide for estimating expected performance within an age band. The distribution and progression of mean scores across age bands for the *birth to 3 years* assessment and the *3 to 6 years* assessment were computed.

In some instances, data manipulation was required to assist in the creation of the mean scores that were used to set the cut scores. To set the mean scores, the entire range of possible scores for each outcome was considered. Assuming that the top of the possible score range was the expected location of children in the highest age band, the mean for the oldest children was set around that point; then the observed progression of means and the top point were used to develop the rest of the points on the line.

Step 4: Development of Cut Scores

The next step was to compute the upper and lower bounds of each of the 9 points on the 9-point scale. The 9-point scale is an extension of the 7-point scale used on the Child Outcomes Summary Form (COSF). The 7-point scale is appropriate for children with delays and disabilities but needs more specificity within the age appropriate range to be useful for programs serving typically developing children. The 9-point scale expands the 7-point scale by including two new categories and new definitions for points 7, 8, and 9. All children scoring 6 or above in an outcome area are considered "typically developing" (i.e., not showing developmental delays). The extension of the 7-point scale acknowledges that there are meaningful distinctions in performance among children who are typically developing and that these differences are important because they are associated with the likelihood of success later in school. Children scoring a 9 on the scale are considered to be demonstrating the level and quality of skills needed to succeed across all their current and future settings and situations including kindergarten. Children scoring an 8 are showing fewer skills and as such are at some risk for later school difficulties. Children scoring a 7 are showing even fewer skills and are at high risk for later school difficulties. Children scoring a 6 are typically developing but experiencing challenges in one or more of their current settings and situations and are at high risk for later school difficulties.

For programs not using the 9-point scale, the 9-point scale is transformed into the 7-point scale (that is, the scale for ECO's Child Outcomes Summary Form [COSF] ratings), by collapsing ratings of 7, 8, and 9 together.

(ECO and Brookes agreed that this analysis would initially be done with 9 points and subsequently transformed to 7 points because of interest by some states in having a 9-point rating to enable a closer look at children at risk and typically developing.)

To compute the cut scores for the 9-point scale, several pieces of information about the observed distribution of scores for typically developing children were used: the mean, the standard deviation in raw score units, and the standard error in Rasch measure units. The standard deviations varied greatly by age band. To control for the effect of this potential sampling artifact, the standard deviation was averaged across age bands to represent the population standard deviation. The Rasch standard errors did not vary widely across age bands so they were not averaged across age bands. The development of the cut scores had two main steps:

Step 1: Set the upper bound for a rating of 5 (i.e., the distinction between typically developing and not typically developing) based on Rasch standard error units below the mean.

Step 2: Set the additional cut points based on standard deviation units from the mean.

This two step process was employed because step 1 allowed a line to be drawn between what is considered age-appropriate/typically developing functioning and below. This line was set using the Rasch parameters that theoretically should be less sample specific. Standard deviation units were used to set intervals around the initial cut point because these gave more information about the distribution of raw scores.

Birth to 3 Years Analysis

The mean scores used were smoothed means. The SDs were averaged across age bands for each outcome. The average SDs were: Outcome 1 = 21; Outcome 2 = 35; Outcome 3 = 47. SD units were used to spread the points on the 9-point scale because this allows the spread of ability across the 9 points in a predictable way. Because of differences in the location of the mean item difficulty and the mean ability of the sample across the three outcomes, it was necessary to vary the SD units used to create the intervals across the three outcomes that yielded groups by rating points that meet expectations about population functioning. The intervals were set lowest for Outcome 2 starting at 1.6 SDs below the mean, Outcome 1 was set the next lowest starting at 1.5 SDs below the mean, and Outcome 3 was set the highest starting at 1.2 SDs below the mean.

Step 5: Validation of Cut Points

Birth to 3 Years

The next step was to program these new cut scores into the SPSS syntax used to compute a 9-point rating for each child. The validation for the *birth to 3 years* analysis was done using 4 separate data sets.

1. One data set collected on children participating in one selection of Part C programs (*n* = 223). The data points in the file were all collected at exit. This data set included children tested between June 30, 2008 and July 1, 2009.
2. Another data set collected on children participating in another selection of Part C programs (*n* = 4,653). All children in this data set had assessments finalized between July 1, 2008 and June 30, 2009.
3. Another data set (de-identified) from AEPSi supplied by Brookes (*n* = 546). This data set included children assessed between February 14, 2008 and November 23, 2009.
4. The sample of typically developing children used to develop the cut scores (*n* = 571).

Results for Birth to 3 Years

The distribution followed expectations about the status of children with disabilities for the first and second data sets mentioned above. These expectations are that very few children will score at the bottom of the scale and that most children with disabilities will score between 4 and 7. The third data set above had a large number of children scoring at 1; this could be related to the quality of the data in that data set or to the nature of the population with whom the AEPSi is being used. The fourth data set showed a high number of children scoring at 9, which would be expected.

3 to 6 years

Three data sets provided by Brookes were used to test the new cut points for the *3 to 6 years*. Note that the patterning of data for typically developing children in the first AEPSi data set is unusual and may be the result of mislabeling children with developmental delays as typically developing.

1. The first AEPSi data set (DD *n* = 12,788; TD *n* = 3,661)
2. The data used for the original calibration (DD *n* = 1,048; TD *n* = 1,140)
3. The second AEPSi data set (DD *n* = 5,778; TD *n* = 1,333)

Results for 3 to 6 years

The results for cut points with the Calibration data set and the AEPSi second data set are consistent with expectations for children with disabilities and typically developing children. Typically developing children were seen to score higher than children with developmental delays, and a reasonable proportion of children with developmental delays scored at or above typical development.

Step 6: Content Validation of the Cut Point that Demarcates Typical Development

Next, a descriptive look at what the cut points define as typical development was undertaken. The purpose of this activity was to determine if the cut points set match what is known about typical development. There were three possible results. The first is that the cut points are too low, meaning that children who are not functioning at a level similar to same-aged peers are being scored as if they are. The second is that the cut points could be too high. This would mean that children who were functioning at a level similar to same-aged peers would be scored as if they are functioning at a level below same-aged peers. The third (and most problematic) option is that the behaviors themselves do not align in an order that is consistent with what is known about child development. This would lead to high error rates in identifying children functioning at a level similar to same-aged peers.

Item parameters from the Rasch analysis were used to complete this content validation. The items within each outcome within each level of the instrument were scaled using a partial credit model. This model estimates the ability of children likely to pass at each response level; these item parameters are often called thresholds. For the AEPS, two thresholds were computed for each item. The first threshold is the ability at which a child is more likely to score a 1 than a 0. The second threshold is the ability at which a child is more likely to score a 2 than a 1. In AEPS, these scoring options are defined in the following way:

0 = Child does not meet the criterion

1 = Child inconsistently meets criterion

2 = Child consistently meets the criterion

The age content validation was done by ordering the items within each outcome from easiest to hardest based on the threshold that defined the ability at which a child is more likely to score a 2 than a 1 (that is, the ability at which a child is likely to consistently meet the criterion). To determine the age at which a child would be likely to perform the behavior consistently, the ability score of the upper bound of a 9-point rating scale score of 5 was used for each age band. Conceptually the age bands on the items represent the skills that children in that age band would be expected to consistently perform (90% of children in the age band would be expected to have mastered the skill). The age bands in this analysis were created using the information in the cut score tables above. The sample used to create the age bands is the same as the sample used to create the cut scores.

The results of this study have been incorporated into the automated calculations in AEPSi to generate the child outcomes ratings in the OSEP reports in the online system.

ONGOING WORK

EMRG and Brookes continue to cooperate with the ECO Center to examine refinements in the ongoing development of the OSEP child outcomes reporting process and measurement system. Parties using AEPSi who would be interested in participating in this process are welcome to inquire with Brookes Publishing.

PSYCHOMETRIC STUDIES OF THE AEPS TEST

Reports of the psychometric properties of the AEPS Test are unchanged by the OSEP child outcomes reporting process and the above-reported analyses to assist in the generation of the five categories in OSEP reporting.

Psychometric studies of the AEPS have been ongoing since the mid 1980s, and numerous reports have been published. AEPS has very high ratings across all the usual measures of a test's psychometric strengths. These are reported in Appendix A from AEPS Volume 1, which is also reproduced for download at www.aepsinteractive.com. Studies have looked at interobserver agreement, test–retest reliability, concurrent validity, sensitivity, specificity, and treatment validity. Replication studies have reported the findings.

For more information about AEPSi or this report, please contact any one of the following staff at Brookes Publishing:

Melissa Behm, Executive Vice President (410-337-9580, ext. 144; mbehm@brookespublishing.com)
Heather Shrestha, Editorial Director (410-337-9580, ext. 102; hshrestha@brookespublishing.com)
Monica Belle, Web and Database Developer (410-337-9580, ext. 189; mbelle@brookespublishing.com)