



# Conducting a Developmental Assessment in Young Children

Glen P. Aylward

Published online: 30 June 2020

© National Register of Health Service Psychologists 2020

## Abstract

Developmental assessment of infants and toddlers presents a host of clinical challenges, many of which are unique to this age range. These include determining whether the child has a delay or a deficit, selecting the proper assessment tools, and accurately interpreting the findings. This process has four components: administration of structured items, direct observations, caregiver report, and history.

## Clinical Vignette

Jenny is a 2-year, 1-month-old toddler whose parents report continuing concerns about her development. Her pediatrician has responded with a request for a detailed developmental assessment. Language is a major area of worry, although motor skills are delayed, and Jenny is described as “clumsy.” Jenny was born at 25 weeks gestational age. She spent her first 130 days after birth in the hospital. She had numerous health issues, including respiratory distress syndrome, intraventricular hemorrhage (IVH) Grade IV, and bronchopulmonary dysplasia (BPD). Jenny had Apgar scores of 3, 5, and 7 at 1, 5, and 10 minutes, respectively. An early intervention specialist screened Jenny at 18 months but found her to be ineligible for services because the magnitude of delay was not enough to warrant intervention. Sharing the parents’ concerns, the pediatrician is now asking for more detailed assessment, observation, and measurement of the toddler’s current functioning and development. What tests should you give to obtain critical information regarding Jenny’s developmental status? Will the test findings provide insight with regard to what interventions are needed? What developmental problems might you expect, given Jenny’s medical history?

## Clinical Challenges

### Prevalence of Developmental Problems

Approximately 15% of the pediatric population has developmental problems. Of this subset, 45% have speech or language issues, 38% display developmental delay in other domains such as motor or adaptive skills, and 17% have autism or other disabilities (Feldman, 2020).

The prevalence rate of developmental problems is higher in those born preterm. In fact, of children born at 25 weeks gestational age, only 5% do **not** have developmental concerns (Berry et al., 2017; Hoekstra, Ferrara, Couser, Payne, & Connett, 2004). These concerns range from cognitive to motor to neuropsychological function (particularly executive functioning). There is essentially an inverse relationship between gestational age and developmental disabilities: the younger the gestational age, the greater the likelihood of developmental problems.

In 2016, a total of 750,000 children between the ages of 3 and 5 received special education services. The increase in early assessment and service use is endorsed and encouraged by the American Academy of Pediatrics. Once identified, these children could then be placed in intervention programs, which have been shown to improve developmental outcomes. Most pediatricians now provide developmental screening services and interpretation of the results (Lipkin et al., 2020).

### Delay Versus Disorder

A critical issue in developmental assessment of an infant is the differentiation between a delay and a disorder. A delay suggests that the child is demonstrating the proper developmental sequence but at a slower pace. Emerging skills are present, but mastery of a task or specific developmental acquisition has not occurred or is not occurring as quickly as expected. By virtue of being called a delay, it is assumed that the child will eventually progress and “catch up.” In contrast, a deficit or disorder suggests that the toddler’s development is atypical, and there is a low probability that the developmental skill will be mastered in the future. A disorder is typically associated with developmental impairment—either mental, physical, or both.

## Developmental Assessment Components

Developmental assessment consists of four components that the examiner should incorporate: (a) inclusion of the caregiver in the evaluation process (including their reports of behavior and milestones), (b) direct observation of the child's behaviors, (c) administration of structured developmental assessment instruments, and (d) consideration of the child's and family's history (Aylward, 2020).

There are multiple layers of issues in the assessment of infants and toddlers. One is determining what tests to use. Should it be a screening test or an actual assessment? If the referral reason is general, one would start with a Level 1, parent-completed screening measure (e.g., Ages and Stages Questionnaire-3; Bricker & Squires, 2009) or the Parents' Evaluation of Developmental Status; Glascoe, 1998). If the child was referred to you to determine if there is a more encompassing developmental problem (e.g., a generalized developmental delay), then a psychologist-administered, hands-on, Level II standardized assessment involving the four components listed previously is warranted.

A broad-band assessment such as the Bayley-4 would be appropriate, followed by a more specific narrow-band testing instrument such as the Modified Checklist for Autism in Toddlers (Robins, Fein, & Barton, 2009) in the case of suspected autism spectrum disorder (ASD), or the Preschool Language Scales-5 (Zimmerman, Steiner, & Pond, 2011) or the Clinical Evaluation of Language Fundamentals® Preschool-2 (CELF Preschool-2; Semel, Wiig, & Secord, 2004) for children identified as having language development concerns.

### Inclusion of the Caregiver in the Evaluation Process and Use of Caregiver Report

There are multiple reasons to utilize caregiver report and have the caregiver participate in the assessment process. First, it reduces anxiety in the parent and child. Second, caregiver report augments data that are collected and provides verification of ambiguous findings during the assessment. Parents are better able to describe day-to-day behaviors of the child, versus what is displayed in a 60- or 90-minute session. Third, testing time is reduced, as are missing scores. Finally, this approach is compatible with the "authentic" or "naturalistic" approach to developmental assessment. It is good practice to routinely elicit parental concerns about their child's behavior. Moreover, interviewing caregivers helps to identify risk and protective factors in the child's environment, social milieu, and medical background.

### Direct Observation of Behaviors and Milestones

Nonetheless, as part of the overall clinical evaluation, examiners should regularly observe a child's qualitative behavior

during testing. Areas of importance include (a) vision (aberrant eye position [esotropia, exotropia], uncoordinated eye movements); (b) excessive tone/decreased tone that is obvious in execution of motor tasks; (c) asymmetries in arm or leg use; (d) poor motor modulation (reaching, letting go of objects); (e) strongly established hand preference at an early age (< 12 months); (f) very short attention span, excessively high activity level, or increased impulsivity for age; and (g) emotional dysregulation.

Many "red flags" are not the result of maturational lags or delays, but are due to deficits or deviance in development or neurological impairment. Keen observational skills tend to differentiate a clinician from a technician.

Patterns of dysfunction are more concerning than an individual "abnormal" sign. You should also consider the functional impact or significance of an abnormal or questionable finding. More specifically, how does the finding affect the child's development? For example, tightness of the lower extremities in a 24-month-old child is less concerning if she is able to walk unassisted or run, versus being unable to ambulate because she stands on her tiptoes constantly; walks in an unstable, awkward manner; or is unable to run. The tightness may reflect mild to moderate cerebral palsy or a developmental coordination disorder.

With regard to milestones in general terms, initially (up to approximately 9–12 months of age) emphasis is placed on motor development. More specifically, cortical suppression or inhibition of automatic reflexes that are mediated by lower brain centers (e.g., hands fisted, asymmetric tonic neck posturing ["fencing posture"], plantar grasp [foot grasp]) causes these reflexes to disappear. As these early reflexes are inhibited, voluntary motor behaviors become possible such as intentional grasping, transferring hand to hand, and walking. A disorder is suggested if these reflexes persist. The next type of milestone involves further development and refinement of motor skills and tone such as head control, rolling over, crawling, cruising around furniture, walking, and running (12–18 months). By 18 to 24 months, language development and its associated social aspects, as well as the combined use of gestures and words, is the main area of focus.

### Developmental Assessment Tools That Are Directly Administered

There is no "gold standard" for infant developmental tests. Any developmental test used should be considered a "reference standard." This is because there is a high degree of variability and no absolute, definitive values or cutoffs in developmental assessment that would be similar to those found in lab analyses. Tests are not valid or invalid in many cases. Rather, it is how the test is used and interpreted that makes test scores valid or invalid for formulating certain assumptions

or diagnoses (Aylward & Zhu, 2019). Proper test use also includes awareness of a test's strengths and weaknesses. Although the use of standardized psychological tests with age-based norms is helpful, developmental tests that simply score items in a dichotomous "yes/no" fashion may not be useful in the clarification of the delay/disorder distinction. Having a score for "emergent" skills, (i.e., a 3-option scoring system) helps to some degree. For example, on the Bayley-4 item that requires the child to jump with both feet off the floor, the child could (a) show mastery by having both feet in the air simultaneously (a 2-point response); (b) have only one foot off the floor at a time with the other providing support (1 point); or (c) not be able to show any jumping motion whatsoever (score of 0). A dichotomous scoring format would combine the last two options, when in actuality they are quite different (emergent versus not existent). The emergent score could be considered a delay, while the absence of the skill has a higher probability of being a deficit.

The types of items included in a selected developmental test are important. Many infant tests are heavily weighted with canalized, sensorimotor items (e.g., reaching, picking up an item, bringing hand to mouth). These common skills are "prewired"—remain intact in all but the most impaired children—and tend to be resistant to negative biologic influences (insult to the central nervous system). These simple behaviors are not indicative of later levels of function, yet they are prevalent in many infant tests. This leads to confusion regarding a delay or disorder (and results in poor prediction) because of the sensitivity and specificity of these canalized behaviors. Essentially the items are not failed unless there are extreme, negative circumstances (e.g., a moderate to severe perinatal event such as asphyxia). As a result, sensitivity is reduced and specificity would be inflated because most children would pass the item. The difficulty in distinguishing between a delay and a disorder is exacerbated with the use of screening tests because they contain a limited sampling of items and therefore are restricted in their ability to identify patterns of developmental concerns. In fact, this may explain why the early intervention examiner in the opening vignette did not consider Jenny in need of intervention at 18 months.

More complex behaviors of infants that are indicative of true cognitive or intellectual ability evolve over time. The challenge for the practitioner administering assessments to young children is to identify early components or precursors of these more complex behaviors. For example, being selectively attentive to what is going on in the examination room, and habituation are early indicators of higher-order cognitive skills or executive function (Bayley & Aylward, 2019b). Rather than tapping isolated skills, the practitioner needs to identify early behavioral indicators that require greater integration of neural networks. These indicators reflect the ability of the overall brain system to function in an efficient, organized, and cohesive manner. Clinicians are often faced with

tests that are excessively redundant. Other tests are overly inclusive regarding the breadth of developmental skills they assess. Unfortunately, length, redundancy, and resultant disinterest and fatigue in the child compromise findings and raise questions regarding a test's validity.

### Problems in Test Selection and Interpretation

Testing infants and toddlers requires different skills than those needed for testing older children. These differences include being flexible regarding the order of administration that best suits the young child (nonlinear administration), being attuned to nonverbal behaviors that might be indicative of the toddler's frustration or disinterest, and determining when a break is necessary. Along these lines, there are unique interpretive issues as well. Test selection and interpretation will also have an impact on the determination of a delay or a disorder. This in turn, will directly affect intervention eligibility.

The most frequent method used to "quantify" the magnitude of a developmental delay (and perhaps differentiate it from a disorder) is "percent delay." This is computed as the ratio of chronological age (or corrected age in the case of premature infants)  $\times$  100. Assuming a child's chronological age is 24 months and the language score is at an 18-month level, this equates to a 25% delay. However, ratios are not as simple as they may initially seem because (a) ratios are not comparable across infancy and toddlerhood, (b) the standard deviation does not remain constant, (c) confidence intervals vary greatly, and (d) the velocity of change of a developmental construct is not consistent across ages. Moreover, the mental age estimate (numerator) is totally dependent on the test used (Aylward, 2020). Despite these shortcomings, percent delay cutoffs are used routinely, suggesting a degree of preciseness that simply does not exist. Standard deviations below the mean (e.g.,  $-1.0$ ,  $-1.5$ ,  $-2.0$  SD) are more accurate.

A related problem of developmental assessment is age equivalents. Although this concept is helpful when explaining findings to caregivers, it is also inaccurate and subject to misinterpretation. A 4-year-old who is functioning at a 2-year-old level is not the same as a 2-year-old functioning age appropriately. Age equivalents are particularly vulnerable to misinterpretation in tests with steep *item gradients* (where a minor change in raw scores translates into major alterations in the age equivalent). Standard scores, percentiles, and 95% confidence intervals are much more psychometrically sound.

Finally, there is the issue of broad-band versus narrow-band screening tests and questionnaires. Broad-band instruments such as the Survey of Well Being of Young Children (Sheldrick et al., 2019) that measures milestones, social-emotional function, ASD, and family risk factors, or the Developmental Indicators for the Assessment of Learning (4th ed.; Mardell & Goldenberg, 2011) are useful when the referral reason is more general because they can pinpoint

previously unidentified areas of concern. Good practice would be to start with use of a broad-band instrument. Some tests such as the Bayley-4 (Bayley & Aylward, 2019a) contain a blend of components: broad developmental composite scores, specific measurement of adaptive and social–emotional functioning, a sensory processing checklist, and an ASD checklist. Narrow-band instruments applicable to developmental domains (e.g., cognitive, motor, language) or specific disorders (e.g., ASD) could then refine diagnoses. Examples of more focused, narrow-band instruments include the Peabody Picture Vocabulary Test-5, the CELF Preschool-2 (both addressing language issues), as well as most of the ADHD rating scales used with older children.

### Consideration of the Child's History

Awareness of possible sequelae frequently associated with a particular disease, medical condition, or perinatal issue is important. Bias should be avoided, but knowing the potential problems that could occur (based on the child's history) would likely ensure that the deficit would not go undetected.

There are three types of risk categorizations that should be considered: *established risk* (e.g., Down syndrome [DS], Rett syndrome, Fragile X), *medical/biologic risk* (prematurity, birth asphyxia), and *environmental risk* (low-SES household, poor stimulation). For example, with an established risk such as DS, intellectual disability is likely, and children with DS often score higher on developmental tests early in infancy than later on. This is because their deficits in expressive language and abstract usage become more apparent as they age. In Jenny's case, medical/biologic risk is a concern.

Examiners should be cognizant of this type of information with regard to DS as well as testing characteristics of other disorders to avoid giving parents false expectations and to also make them aware of things to monitor. In general, the likelihood of moderate to severe developmental problems is highest in the established risk groups, varies in infants at biologic risk (depending on the *type* and *degree* of biologic risk), and often is reflected in language delays in children who are at environmental risk. Also note that delays or disorders can be identified in as many as 10% of children who do not fall into any of these risk categories.

The medical, developmental, and intervention history of the child and family is important when assessing all domains of development. Consideration of environmental influences is also significant particularly when one is concerned with "experiential bias" (Aylward, 2020; Bayley & Aylward, 2019a, 2019b). This term, used within developmental assessment, means that experience with or exposure to certain tasks and experiences or, conversely, lack of such exposure, can significantly affect a toddler's test performance. This effect, positive or negative, can cause an incorrect reading of the child's capabilities. This is especially true with language development.

Variation in the rate of language development in the pediatric population is substantial, again underscoring the difficulty in the distinction between delays and disorders. Approximately 16% of young children demonstrate initial language delays, and difficulties (disorders) will persist in approximately half (8%) of them. When considering language, practitioners should document family history of language problems, the child's hearing ability and history, language stimulation in the home, and any loss or plateau in language skills.

Receptive language is not necessarily an area of deficit in children like Jenny who are born preterm, or in their full-term counterparts, although language processing, expressive communication, and verbal working memory can be (Aylward, 2005, 2020). A child Jenny's age should be able to put together two words, have a 50-word vocabulary, and be intelligible at least half the time to adults who are unfamiliar with the child. Both the receptive and expressive communication subtests of the Bayley-4 should be administered and compared. If either of these is positive, more specific language-based tests should follow.

Cognitive abilities are also of concern, and cognitive disorders often first present as language delays. For children born extremely prematurely, mean group IQ/DQ's generally decrease by 1.5 to 2.5 points per week below 32 weeks. There is a risk for low-average to borderline cognitive abilities due to IVH, disruption of normal brain development due to extreme prematurity, and continued low-grade hypoxia caused by BPD (the need for supplemental oxygen beyond 36 weeks gestational age). A toddler who is born extremely preterm is also at risk later on for "high prevalence/low-severity dysfunctions" (ADHD, learning disabilities, neuropsychological dysfunction [including executive dysfunction]; Aylward, 2002, 2005). The Bayley-4 Cognitive scale should also be administered, with the examiner looking for both areas of weakness as well as those of more optimal performance (Aylward, 2020).

Motor problems should be assessed, particularly in very premature (28–31 weeks gestational age) and extremely premature infants (< 28 weeks). More than 75% of extremely premature children experience deficits in visual–motor integration and writing. There is also a high percentage of cerebral palsy and developmental coordination disorders, which are first reported as "clumsiness." This is particularly the case with Grade IV IVH. Grades I and II reflect minimal bleeding in the germinal matrix of the lateral ventricles and are considered mild. Grade III is more severe and indicates more blood in the ventricle and distention. Grade IV is typically asymmetrical and bleeding extends into the brain. Children with Grade III have a 35%–55% rate of disability, while those with Grade IV have more than a 90% chance of cognitive and motor deficits (Aylward, 2005). This risk warrants administration of the fine and gross motor subscales of the Bayley-4. Again, if these are positive, referral for more specific motor

evaluation by occupational and/or physical therapists and explicit interventions are recommended.

When considering an infant or young child's medical history, a popular misconception involves the meaning of Apgar scores. The Apgar score is based on five physiologic items scored 0–2: skin color, heart rate, reflex irritability, tone, and respiration. Scores of  $\leq 3$  are considered worrisome, while scores of 4–6 are borderline. Although they are often used to predict outcomes, these scores were not designed for that purpose and thus perform poorly in that regard. This is especially the case with children born preterm. Nonetheless, many clinicians put much weight on these scores, not realizing that only when they are extremely low for a sustained period in conjunction with other indicators of hypoxia/ischemia that they might be meaningful (Rudiger & Rozycki, 2020).

### Delay and Disorder in Context

Serial assessment can add some clarification to the parents' concerns about whether their child manifests delay or a disorder (deficit). Baseline assessment is needed, but multiple touch points are much more informative than a one-time assessment. If problems continue across several assessments, the likelihood of a disorder or deficit increases.

To discern the overall nature and severity of a developmental problem, clinicians should consider changes *within* developmental domains as well as changes in the relationships *between* these domains. In other words, clinicians should evaluate how the child is progressing or not progressing within a developmental domain such as language, but also compare changes in the level of function of this domain to performance in subdomains or other domains such as cognitive or motor functions as well. This is called a “dissociation” when a significant difference occurs in the evolving rates of two developmental domains such as a major discrepancy between receptive and expressive language abilities, or crawling but not being able to sit independently. A dissociation by itself is not abnormal, but it suggests an increased probability of developing a disorder later. What can be assessed also changes with age, evolving from the neurologic → motor → sensorimotor → cognitive function (Aylward, 2009, 2020). The breadth of assessment expands corresponding to an increase in age. By 2½ years of age, cognitive function and language are two major growth areas.

### Lessons Learned Regarding Developmental Assessment

- It is difficult to distinguish between a developmental delay and a disorder. Serial assessment and detection of patterns of developmental problems can help clarify this distinction.

- The clinician must carefully select tests based on the purpose of the assessment. The strengths and weaknesses of each test should be identified, knowing that no test is perfect.
- Developmental assessment should lead to intervention. Simply identifying a delay is not adequate.
- Developmental assessment has four components: (a) administration of structured items, (b) direct observation of behaviors and milestones, (c) use of caregiver report and participation, and (d) integration of the child and family's history. All four components are necessary for a thorough assessment. Proper use of these components distinguishes the clinician from the technician.

### References

- Aylward, G. P. (2002). Cognitive and neuropsychological outcomes: More than IQ scores. *Mental Retardation and Developmental Disabilities Research Reviews*, 8, 234–240. <https://doi.org/10.1002/mrdd.10043>
- Aylward, G. P. (2005). Neurodevelopmental outcomes of infants born prematurely. *Journal of Developmental and Behavioral Pediatrics*, 26, 427–440.
- Aylward, G. P. (2009). Developmental screening and assessment: What are we thinking? *Journal of Developmental and Behavioral Pediatrics*, 30, 169–173. <https://doi.org/10.1097/DBP.0b013e31819f1c3e>
- Aylward, G. P. (2020). *Bayley 4: Clinical use and interpretation*. Cambridge, MA: Academic Press.
- Aylward, G. P., & Zhu, J. J. (2019). *The Bayley Scales: Clarification for clinicians and researchers*. (pp. 1–12). Bloomington, MN: NCS Pearson.
- Bayley, N., & Aylward, G. P. (2019a). *Bayley Scales of Infant and Toddler Development: Administration manual* (4th ed.). Bloomington, MN: NCS Pearson.
- Bayley, N., & Aylward, G. P. (2019b). *Bayley Scales of Infant and Toddler Development: Technical manual* (4th ed.). Bloomington, MN: NCS Pearson.
- Berry, M. J., Saito-Benz, M., Gray, C., Dyson, R. M., Dellabarca, P., Ebmeier, S., Foley, D., Elder, D. E., & Richardson, V. F. (2017). Outcomes of 23- and 24-weeks gestation infants in Wellington, New Zealand: A single centre experience. *Scientific Reports*, 7, 1–8. <https://doi.org/10.1038/s41598-017-12911-5>
- Bricker, D., & Squires, J. (2009). *The Ages and Stages Questionnaire* (3rd ed). Baltimore, MD: Brookes Publishing.
- Feldman, H. M. (2020). How young children learn language and speech. *Pediatrics in Review*, 40, 398–410.
- Glascoc, F. P. (1998). *Collaborating with parents: Using parents' evaluation of developmental status to detect and address developmental and behavioral problems*. Nashville, TN: Ellsworth & Vandemere Press.
- Hoekstra, R. E., Ferrara T. B., Couser, R. J., Payne, N., & Connett J. E. (2004). Survival and long-term neurodevelopmental outcome of extremely premature infants born at 23–26 weeks' gestational age at a tertiary center. *Pediatrics*, 113, e1–e6. <https://doi.org/10.1542/peds.113.1.e1>
- Lipkin, P. H., Macias, M. M., Chen, B. B, Coury, D., Gottschlich, E., Hyman, S. L., Sisk, B., Wolfe, A., & Levy, S. E. (2020). Trends in

- pediatricians' developmental screening: 2002–2016. *Pediatrics*, 145(4), e20190851. <https://doi.org/10.1542/peds.2019-0851>
- Mardell, C., & Goldenberg, D. S. (2011). *Developmental Indicators for the Assessment of Learning* (4th ed.). Bloomington, MN: Pearson.
- Robins, D. L., Fein, D., & Barton, M. (2009). *Modified Checklist for Autism in Toddlers, revised, with follow-up (M-CHAT-R/F)*<sup>TM</sup>. <https://mchatscreen.com>
- Rudiger, M., & Rozycki, H. J. (2020). It's time to reevaluate the Apgar score. *JAMA Pediatrics*, 174, 321–322. <https://doi.org/10.1001/jamapediatrics.2019.6016>
- Semel, E., Wiig, E. H., & Secord, W. A. (2004). *The Clinical Evaluation of Language Fundamentals® Preschool-2 (CELF Preschool-2)*. Bloomington, MN: Pearson.
- Sheldrick, R. C., Schlichting, L. E., Berger, B., Clyne A., Penscheng, N., Perrin, E. C., & Vivier, P. M. (2019). Establishing new norms for developmental milestones. *Pediatrics*, 144(6), e20190374. <https://doi.org/10.1542/peds.2019-0374>
- Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2011). *The Preschool Language Scales-5*. Bloomington, MN: Pearson.
- Glen P. Aylward**, PhD, ABPP, is Professor Emeritus of Pediatrics & Psychiatry, Southern Illinois University School of Medicine and director of NICU Developmental Follow-Up. He is a fellow of six American Psychological Association (APA) divisions, past president of APA Division 54 and the Society for Developmental and Behavioral Pediatrics, and associate editor of the *JDBP*. He is the author of the *Bayley Scales of Infant and Toddler Development* (4th ed., 2019) and *Bayley-4: Guide for Clinical Use and Interpretation* (2020).