Developmental Profile 3
DP-3

MANUAL

Gerald D. Alpern, Ph.D.

Illustrations by Joy Allen

wps®
Test with Confidence

1
Introduction

The *Developmental Profile 3* (DP-3) is a completely new, standardized, and updated revision of the DP-II (Alpern, Boll, & Shearer, 1986), a well-established measure of child development. The DP-3 utilizes input from parents or caregivers (as an interview or a checklist) to provide scores in five key areas of development: physical, adaptive behavior, social-emotional, cognitive, and communication. The current version maintains continuity with the DP-II by combining the strong history of the latter instrument with current psychometric standards in test development. The DP-II has been viewed by users as quick, easy, informative, reliable, and valid. It was designed to cover five important developmental areas across a sizable age range and has been widely accepted and used in psychology and education. The DP-3 maintains all of those characteristics while providing a representative normative sample; updated scale names to better describe current content and uses; standard scores; new administration options; updated item content, including revision of items that referred to dated technology and customs; modern statistical scaling techniques; suggested intervention activities; and expanded computer scoring and interpretation.

In the time between the publication of the DP-II and this current version, cultural and technological changes have had a critical impact on both the rate and the nature of the experience of growing up. The technological changes brought on by the computer alone have significantly influenced how the children of today learn, communicate, socialize, function, and play. Differences in culture, such as diverse family patterns and evolving child care institutions, have children living in a very different world from that of just a generation ago. The combined effects of these technological and cultural changes make it important to update the measurement of children's physical, adaptive, social-emotional, cognitive, and communication skills. Changes incorporated into the DP-3 reflect these recent societal shifts.

The DP-3 can be used effectively in a variety of settings and for a variety of purposes, as it provides norm-based scores and information on individual strengths and weaknesses in child development. The DP-3 can be used to assess typically developing children, and also provides a psychometrically sound instrument that can quickly assess for the likelihood of delay. Research suggests that early identification of and intervention for developmental disabilities is essential (Ramey & Ramey, 1998), thus highlighting the need for an efficient and accurate assessment of child development, such as the DP-3. In defining a child with a disability, current federal special education legislation (Individuals with Disabilities Education Improvement Act [IDEA], 2004) delineates the five areas addressed by the DP-3 as domains in which to assess for delay. Furthermore, IDEA and other government programs specifically require that parents be provided with a means to be involved in the assessment of their child by providing developmental information about their child. The DP-3, with its multidimensional evaluation based on a parent interview, meets the criteria outlined by current government regulations.

General Description

The DP-3 provides five scales, each with 34 to 38 items, designed to assess the development and functioning of children from birth through age 12. The DP-3 tests a full range of skills, from serious delay through above-average ability, up through approximately age 7 to 9 years (depending upon the scale). At higher ages, its primary use is in identifying skills that are below average and in providing assurance that the skills of higher functioning children are at least within the normal range. In this sense, effective use of the DP-3 is supported through age 12 years. In addition, the test has application through adulthood for cases where very serious delay is present and there is the need to document adaptive and developmental skills in the range assessed by the DP-3. Five scoring options (standard scores, age equivalents, percentiles, stanines, and descriptive ranges) are available for each of the five scales. Additionally, a General Development score is available (and described in greater detail in chapter 3), which is a composite of the following five areas.

Physical (35 items). This scale measures physical development by determining the child's ability with tasks requiring large- and small-muscle coordination, strength, stamina, flexibility, and sequential motor skills. The items are categorized into those that assess gross-motor skills and those that assess fine-motor skills; see chapter 3 on interpretation for more information.

Adaptive Behavior (37 items). This scale (similar to the Self-Help scale of earlier versions) measures competence, skill, and maturity for coping with the environment. It evaluates the child's ability with tasks such as eating, dressing, functioning independently, and utilizing modern technology.

Social-Emotional (36 items). This scale (similar to the Social scale of earlier versions) measures interpersonal relationship abilities, social and emotional understanding, and functional performance in social situations. Specifically, the scale assesses the manner in which the child relates to friends, relatives, and adults.

Cognitive (38 items). This scale (similar to the Academic scale of earlier versions) measures cognitive abilities in an indirect manner, that is, not by actually measuring intelligence and achievement but by assessing the development of skills necessary for successful academic and intellectual functioning. At younger ages, the scale assesses skills prerequisite to scholastic functioning in academic areas such as reading, writing, arithmetic, and computer use and logic. At the preschool and older levels, actual scholastic abilities are measured.

Communication (34 items). This scale measures expressive and receptive communication skills with both verbal and nonverbal language. The use and understanding of spoken, written, and gestural language are assessed by this scale, as is the ability to use communication devices (e.g., telephone, computer) effectively. The items are categorized into those that measure receptive communication and those that measure expressive communication; see chapter 3 on interpretation for more information.

The DP-3 is designed to be administered as an interview of the child's parent or other caregiver and takes approximately 20 to 40 minutes (Interview Form; WPS Product No. W-462A). Additionally, the DP-3 offers a Parent/Caregiver Checklist version (WPS Product No. W-462B), which consists of the same item content as the Interview Form and can be given to a parent or caregiver to complete alone, and then scored later by the clinician. These administration options are discussed further in chapter 2. Within each of the five content areas, items are arranged in order of difficulty. Scoring is easy, as each item is rated as either *Yes,* indicating that the child possesses the skill, or *No,* indicating that the child has not yet mastered the ability. The responses are tabulated following the interview to obtain raw scores, and the raw scores are then compared to a normative sample of 2,216 individuals to obtain standard scores. These standard scores are consistent with present-day federal, state, and local program requirements. The DP-3 can also be scored using the computer scoring and interpretation report program (WPS Product No. W-462U), which is described in the section titled "PC-Based Scoring and Interpretation for the DP-3" at the end of this manual.

DP-3 Improvements

One major improvement from the DP-II to the DP-3 is the nationally representative normative sample of typically developing children, which very closely approximates the ethnic, geographic, and socioeconomic composition of the U.S. population (U.S. Census Bureau, 2005; see chapter 4 for more information on the standardization sample and study). While the previous version yielded only age equivalents and information about passing rates in the standardization sample, the DP-3 yields standard scores, percentile ranks, and stanines, while retaining age-equivalent scores, thus making it a valuable tool for special education and other evaluations.

In this revision, at the youngest levels ages are broken down into smaller increments in order to capture the rapid developmental growth that occurs during such periods. Additionally, the current version offers norm-referenced scores for children from birth through age 12 years, 11 months. This expanded age range allows for obtaining standard scores for children up through the elementary school level.

Another major improvement in the DP-3 is the updated item content. Items were removed if they no longer reflected current society (e.g., "Can the child strike and light a paper match?"), while items related to technology were added in order to capture recent changes in necessary developmental tasks (e.g., "Does the child purposefully use a mouse, touchpad, or other computerized pointing device to point and click on objects on a computer screen?"). Furthermore, statistical item analysis allowed for deleting redundant items that measured the same developmental skill level as ones that were retained, thus creating an even more efficient and streamlined instrument. However, many older items demonstrated good statistical properties and were retained from the previous version in order to build upon the strengths present in the earlier version. Oftentimes these items were placed in a slightly different location on the scale, based upon the results of the item analyses.

Another improvement includes the expanded interpretation guidelines and clarification of test items and administration in order to provide more accurate guidance for determining what a child needs to demonstrate in order to pass an item. Finally, the current version provides suggested remediation activities tied to each of the DP-3 test items (see Appendix A), enabling clinicians to quickly turn the test results into intervention programming. All of these changes were done while maintaining the *Developmental Profile*'s long history of providing a quick, easy, informative, reliable, and valid developmental assessment.

Principles of Use

The DP-3 Interview Form is easy to administer and score by a person familiar and competent with psychological or educational testing, or by a paraprofessional. Minimal training is necessary for a paraprofessional to produce reliable and valid results. Alternatively, the Parent/Caregiver Checklist may be completed by a caretaker who knows the child well and is able to read and fully understand the items, and later scored and interpreted by a clinician. Interpretation and application of the results require a professional or the supervision of a professional with training and experience in child development, psychology, and/or education. Before using the DP-3, individuals should read and familiarize themselves with the information contained within this manual, including the administration, scoring, development, standardization, and technical properties of this instrument.

Use of the DP-3 in both clinical and research settings should conform to the professional and ethical guidelines developed by the American Psychological Association (2002) and other professional psychological and educational organizations. As with any psychological assessment, this inventory should not be used without the informed consent of the child's parent or legal guardian. Users should also take necessary precautions to safeguard the confidentiality of the test results and to restrict access to results to those with a "need to know." Communication of test results to parents should generally focus on interpretation of the results and their implications rather than on specific scores.

The DP-3 is a valuable tool for any setting in which an efficient measure of any one or all five areas of functional development is needed. It can be used in schools, clinics, hospitals, or any other setting where an evaluation of a child's developmental status, strengths, and weaknesses could be useful. Additionally, the DP-3 can serve as either a screening device or a multidimensional tool used to provide information leading toward the diagnosis of developmental delays or other difficulties.

The DP-3 serves well as a screening tool, such that it determines whether a child requires a more comprehensive diagnostic evaluation. Children who are delayed on any of the five scales might be referred for a more detailed evaluation of that specific area. For the Communication scale, children who fall below a critical cutoff point might be referred for further speech, hearing, visual, or language evaluations. Likewise, psychological or psychiatric evaluations can be selected for children screened as having significant delays on either the Social-Emotional or Adaptive Behavior scale. Referrals for orthopedic, metabolic, or nutritional evaluations could be appropriate for children scoring low on the Physical scale, whereas those scoring low on the Cognitive scale might be referred for comprehensive learning disability testing, or intellectual or achievement evaluations.

The instrument is multidimensional and diagnostic in that it allows the user to efficiently determine precisely how the child compares with his or her peers in five essential areas of development. These norm-based comparisons allow the test user to determine whether a child has a significant delay and would therefore qualify for services. Additionally, the results can help tailor an intervention program to the child's particular intra-individual pattern of development based on scale scores and item analysis. The instrument can also serve as a means of follow-up testing to document the child's progress in an intervention program.

Given its ability to function as either a screening tool or a diagnostic instrument, the DP-3 can be used to accomplish a variety of assessment and educational objectives: (a) to determine eligibility for receiving special education and/or related services; (b) to help plan an individualized educational program (IEP) or individual family service plan (IFSP) consistent with the child's strengths and deficits; and (c) to measure the child's progress by comparing profile scores at the beginning of the school year (pretest) with scores achieved at the end of the school year (posttest). Also, since the DP-3 provides a rapid and accurate measure of development along five dimensions, it can be used as a component in periodic developmental screening programs conducted by health practitioners. Additionally, the DP-3 can effectively be used in research when it is necessary to distinguish between typically developing and delayed children or used as a measure of program evaluation. The measure has a history of successful use in research for classifying participants at different levels of delay (e.g., Glascoe &

Byrne, 1993; Sandler et al., 2000) and in program evaluation (Hebbeler & Gerlach-Downie, 2002; Hur, 1997; Sung, Kim, & Yawkey, 1997).

As with any psychological instrument, the DP-3 should not be used in isolation to diagnose or plan treatment for a child. Instead it should be used in concert with other data, such as information derived from concurrent or former assessments, detailed interviews and history taking, and observations. It can serve as an efficient and economical tool for determining individuals who would benefit from a comprehensive evaluation, but it should not, by itself, be considered a comprehensive evaluation.

The interpretation of DP-3 findings, as with any developmental evaluation, requires consideration of social, cultural, and familial environments in order to evaluate the child's opportunities for acquiring specific developmental skills. This point is further discussed and illustrated in chapter 3. Finally, when interview or self-report formats are used, the adequacy of the informant's responses needs to be considered. Inaccurate perceptions, biases, or deliberate attempts to manipulate findings are all potential sources for inaccurate findings. Strategies and methods for dealing with the validity of interview responses are discussed in both chapter 2 on administration and scoring and in chapter 3 on interpretation.

Contents of This Manual

Chapter 2 of this manual contains instructions for administering and scoring the test, and includes a completed sample of a scale from the hand-scored Interview Form. Guidelines for interpreting DP-3 results are presented in chapter 3. Technical aspects of the test are presented in chapters 4 and 5: chapter 4 reviews the test's initial development and describes how the current version of the test was developed and standardized; chapter 5 discusses the instrument's basic psychometric properties and offers an overview of research that has been conducted with the test. The appendixes present the intervention strategies, normative reference tables, and other tables useful for scoring and interpretation, as well as a section on using DP-II strategies to interpret the DP-3 and another on statistical means of comparing scales to one another.

4
Development and Standardization

Earlier Versions of the *Developmental Profile*

The *Developmental Profile,* along with other measures of child functioning, evolved from the pioneering work of Dr. Alfred Binet and the concept of *mental age* introduced in 1905. His central procedure involved determining age norms for a collection of increasingly difficult academic tasks and then assessing a child's ability to accomplish them. This concept was later applied to social and adaptive functioning by Edgar Doll. Measuring these different areas of functioning allowed for a more comprehensive view of an individual's development and was a precursor to later, more sophisticated multidimensional assessment.

The original *Developmental Profile* combined Binet's age norming of items and Doll's interview techniques into a unique, multidimensional, valid, and reliable assessment of child functioning. Dr. Gerald Alpern, former professor and director of research for Child Psychiatry Services at Indiana University Medical School, and Dr. Thomas Boll created the original *Developmental Profile* five-area approach assessment tool. The practice of assessing the same five developmental areas has subsequently been adopted as a standard requirement for child evaluation by many federal, state, and local government programs.

Development of the items for the original *Developmental Profile* began with a search of the literature and a compilation of behaviors considered to be measurements of age-related developmental competence in each of the five functional areas. The original items were derived from (a) developmental items from actual scales of children's intellectual, physical, social, and language abilities; (b) items compiled from normative data appearing in the child development literature; and (c) original items derived from the multidimensional concepts underlying the test. A number of criteria guided the development process. Items were designed to evaluate observable behaviors, be easily understood by parents and specialists in a variety of disciplines, and be easily administered in a relatively short time period. Pilot work was conducted over the course of 3 years, during which time early versions of the instrument were evaluated by teachers, nurses, psychologists, and psychometrists, and items were dropped or rewritten accordingly.

The original version of the *Developmental Profile* contained 318 items grouped into skill areas and approximate age levels based on the analysis of the literature on child development and the preliminary work with the inventory. In the 1971 standardization study these items were evaluated to confirm that they were placed at the appropriate age levels, possessed a high degree of age discrimination, were accurately responded to by parents, and did not discriminate against children by sex, ethnicity, or socioeconomic status. Item selection and placement in the age categories were accomplished using empirical procedures. Based on the standardization data, items were retained if they were passed by 75% of the students in the appropriate age group, were too difficult for a majority of students in the preceding age group, and were passed by almost all children in the next older age group. The 75% criterion was selected because it represented a clear majority of the children in a given age range while recognizing the variation in individual developmental rates. Items were deleted if there was a discrepant passing rate between males and females. Additionally, items at each age level were deleted if there was a lack of agreement between mothers' reports and observed behavior.

The *Developmental Profile II* (DP-II), published in 1980, represented a refinement of the original inventory (Alpern & Boll, 1972). Items were deleted that assessed functioning levels above 9 years, 6 months or that appeared sexist (e.g., an item asking about gender-stereotypic play). As a result of these changes, the length of the inventory was reduced from 217 items to 186. Other item modifications consisted of wording changes to remove ambiguity and unnecessary use of gender-specific pronouns (i.e., “he” was deleted from all items). Additionally, items and scales were reevaluated to determine their fairness across groups. Standardization data for the DP-II were collected in the early 1970s in a relatively limited geographic region and were not representative of all major ethnic groups in the United States. The normative sample was used to present data on suggested cutoff points for referral, age-equivalent scores, and the percentage of individuals at different ages who passed each item.

The DP-II was updated again in 1986 with the addition of a computer scoring program; however, this current version, the DP-3, represents the first comprehensive revision of the original instrument. Although the DP-II was widely used and appreciated for many reasons, it lacked standard scores and needed updated items and a current representative standardization sample.

DP-3 Revision

The revision of the *Developmental Profile* began with the goal of maintaining the positive aspects of the previous versions. For this reason, many items on the current DP-3 have the same content as in earlier versions. However, many items required updated wording, some items needed to be deleted, and new items needed to be added to reflect the current culture and state of technology.

The initial step in the development of the revised instrument involved conducting a user survey to obtain feedback from a sample of longtime DP-II users. The survey was completed by 147 experienced test users, who responded to a mailed solicitation. School psychologists comprised almost half of the respondents, followed by smaller counts of “other psychologists (clinical, developmental)” and “educational diagnosticians.” Overall, respondents proved to be highly experienced practitioners, with a median of 16 years working in mental health or education. In terms of satisfaction with various components of the DP-II, the following aspects garnered the most negative/neutral (as opposed to positive) ratings: interpretation guidelines in the manual, types of scores provided, language of test items, and age ranges covered. All of these areas were addressed and improved upon during the revision process: interpretation guidelines were expanded and clarified, norm-based standard scores were provided, test items were clarified to give more accurate guidelines as to what exactly a child needs to do to pass an item, and the age ceiling of the test was increased to 12 years, 11 months. While the user survey highlighted important areas needing improvement, it also revealed that people find many strengths in the instrument—particularly related to its clinical utility. Survey results suggested that DP-II users find the test provides valuable data and is quick, easy, and efficient to use. Thus, during the revision and improvement process, the “quick and informative” nature of the test was maintained.

Following the user survey, an archival study was conducted wherein 355 cases were obtained from clinical sites around the country and the items were calibrated by the Rasch one-parameter model (Bond & Fox, 2001; Wright & Stone, 1979) using WINSTEPS (Linacre, 2003). The Rasch model is now used extensively in test development alongside classic statistical procedures. Much of its utility lies in its ability to estimate item difficulty and person ability on the same scale. Rasch measurement is sample-free, meaning that the calibrated item difficulty is the same (accounting for measurement error) regardless of the sample of individuals used to generate the difficulty. The WINSTEPS program yields a logit scale of item difficulty and person ability, with a mean of 0 and a standard deviation of approximately 1. Because this scale is a true interval scale, a 1-logit difference between scores has the same meaning whether the score pair appears near the center or at either extreme of the distribution of scores. The easiest items have negative ability estimates, and the more difficult items have higher positive ability estimates. The relationship between person ability and item difficulty can be described in terms of the probability that a person will succeed on any given item. When the person’s ability is equal to the item difficulty, the person has a 50% chance of succeeding on that item. When the item difficulty is greater than the person’s ability, the chance of success decreases. When the item difficulty is lower than the person’s ability, the chance of success increases.

The most precise measurement occurs when an item has the same measure as the person being assessed, and departures from this ideal in either direction lead to increased measurement error. Thus, a well-constructed developmental scale must include items that span the entire range of ability in the target population (i.e., the scale measures well at the extremes of the person distribution), and within the scale, items must be spread uniformly enough to provide a reasonably precise measurement for all ability levels (i.e., the scale measures well in the center of the person distribution).

Calibrating the items obtained from the archival study using the Rasch model illustrated the ability level at which each item was measuring skills, as well as age gaps in measurement that needed to be filled. For example, it was discovered that additional items were needed on the low end of the Social-Emotional scale because of a gap of many months between two items. This procedure also helped to address a criticism of the DP-II that there was some misplacement of items along the scales (Glascoe & Byrne, 1993).

With the results of the archival study and user survey in hand, the next step was to create new items to address gaps in measurement, replace problematic items, and increase the age ceiling of the test. The initial phase of this process involved a review of the sociological literature to identify recent cultural and technological changes that were not reflected in the items of the original *Developmental Profile.* From this review and an examination of other tests of development, new items were written to reflect many of the societal changes that have impacted U.S. culture over the past quarter century.

The new items were then piloted alongside the existing item set (except for six DP-II items that were deleted due to their outdated content; e.g., "Can the child light a match?"). For the purpose of updating and clarifying some of the older item language, the wording of some of the items retained from the DP-II was slightly altered, while the meaning remained the same. A total of 326 parents participated in the pilot study, and 318 of these parents also completed the new Parent/Caregiver Checklist version of the DP-3. The data gathered from this process were used to generate norms for the Parent/Caregiver Checklist, as described later in this chapter.

In order to test a number of new items (61 to 66 items per scale), while not requiring a parent to answer too many interview questions that were inappropriate for the age of the child, three different forms were utilized in the pilot study. Each form consisted of items considered most appropriate for the age range of the form, as well as additional items that overlapped with the next form. These linked forms enabled the items' difficulty to be estimated over the entire pilot sample. Based on this analysis, items were selected so that each scale spanned the range of expected abilities and provided reasonably "gapless" measurement. Additionally, the information gathered regarding the difficulty level of the items was used in creating the order of test items for the standardization form. Furthermore, pilot study data were analyzed to determine appropriate basal and ceiling rules in order to shorten administration time but not change the obtained score. Results indicated that a basal of five and a ceiling of five met these criteria.

The standardization study was conducted using one research form comprising 36 to 40 items per scale. In order to shorten administration time, suggested starting points and a basal and ceiling procedure were utilized. The data collected in the standardization sample were used to conduct an additional analysis of the items using the Rasch methodology to determine items that were redundant (i.e., added no or little additional measurement) or did not fit well with the Rasch model. The primary criterion for fit in the Rasch model is *infit,* which refers to how well items perform with respect to persons with similar measure (i.e., close to the item difficulty). Items with poor infit were reviewed and some were deleted if they were determined to be nonessential for assessment and intervention. Item difficulty was explored and compared to average ability per age to ensure that items fit with the statistical model, as well as with generally observed developmental stages. All of this was done while keeping in mind the importance of having a screening device that adequately measured development throughout the age range. The item difficulty estimates garnered from the Rasch analysis helped to determine the final order of items on the forms. Based on these analyses, the final forms were created.

Because the individuals in the pilot study responded to items from the DP-II and the DP-3, the two sets of items could be compared. Since six DP-II items were not included in the pilot study, those item responses were estimated based upon the ability calibration from the Rasch analysis of the archival data. Estimates of DP-II raw scores for each scale were obtained and compared to raw scores for each DP-3 scale. It should be noted that due to the overlap of items, these correlations did not consist of independent data. Therefore, the obtained coefficients should be viewed as representing the similarity between versions. Correlation analyses were conducted for the sample as a whole and for each age year with sufficient sample size ($n \geq 30$), which included six analyses from age years 0 through 5. Table 6 displays the number of items per scale on the DP-3, the number of items per scale retained from the DP-II, and the median correlation coefficients between the two item sets, indicating the strong level of continuity between the DP-II and DP-3.

Standardization Study

The standardization sample was obtained by recruiting interviewers from across the United States who had access to typically developing children through schools, neighborhoods, community centers, and so on. These interviewers were screened to ensure they had sufficient training in assessment, and they were provided with a detailed interviewer's manual with complete instructions on the administration of the instrument. There were a total of 59 interviewers from 21 states across the country, representing the four major U.S. Census Bureau regions. The participation of numerous sites helped to ensure that the sample was diversely representative and not influenced by special conditions at one or a few locations. The final standardization sample consisted of 2,216 children. The majority of interviews were conducted with

Table 6
Scales of the *Developmental Profile,* Second and Third Editions

Scale	No. of items on the DP-3	No. of items retained from the DP-II	Median correlations between the DP-II and DP-3[a]
Physical	35	18	.86
Adaptive Behavior	37	20	.89
Social-Emotional	36	22	.87
Cognitive	38	22	.86
Communication	34	23	.89
Total	180	105	

Note. Some items retained from the DP-II have slightly different wording; however, the meaning is comparable.

[a]The displayed correlations are medians obtained from analyses conducted for each age year with $n \geq 30$.

mothers (85%), while the remaining were conducted with fathers (12%) and other relatives (3%).

Table 7 presents the demographic characteristics of the sample, along with corresponding percentages from the U.S. Census (U.S. Census Bureau, 2005) for comparison. In all demographic areas, the sample is distributed similarly to the U.S. population. The only exception is the slight overrepresentation of individuals from the Southern and Midwestern regions of the country; however, examination of the data showed that no systemic differences exist based upon geographic region. Inspection of average scores for groups from the DP-3 standardization sample based on gender, parent education, ethnic background, and region indicated that standard scores for the scales and the General Development score apply acceptably well across a wide spectrum of demographic groups. Additional detail is provided in the upcoming section on moderator variables.

Derivation of Standard Scores

Interview Form

In the evaluation of the means and standard deviations of the raw scores for each of the five DP-3 scales, it became apparent that it was necessary to break down the normative groups into 2-month increments at the youngest ages, due to the rapid changes in development during the first few years of life. For this reason, there are 29 normative reference groups. Table 8 presents the raw score means and standard deviations for each of these 29 groups. Standard scores were created initially by normalizing the raw scores (see Anastasi and Urbina, 1997, p. 62). The original distribution of DP-3 raw scores underwent a nonlinear transformation within each age group so that it would approximately fit a normal curve. The normalized raw scores were converted to standard scores, which have a mean of 100 and a standard deviation of 15. To address minor inconsistencies resulting from sampling artifacts, a smoothing method was employed based upon procedures described in Chambers, Cleveland, Kleiner, and Tukey (1983). In this procedure, for each raw score value, the standard scores for the highest and lowest normative reference groups were left unadjusted. The standard scores for all other age groups were recalculated by averaging each score with the scores for the adjacent younger and older groups. The scores for the adjacent groups were given weights of half that of the target score, allowing the original score to be the most influential. This process addressed the idiosyncrasies present in a few of the normative groups. Once this was done, missing values were assigned based upon interpolation and extrapolation of the existing data points. Techniques similar to the ones utilized here are standard practice in developmental assessments (e.g., *Vineland Adaptive Behavior Scales, 2nd Edition,* Sparrow, Cicchetti, & Balla, 2005; *Developmental Assessment of Young Children,* Voress & Maddox, 1998). These procedures changed the scores very little, as the final standard scores correlated with the normalized scores at .99 for all scales.

The General Development score was derived by adding the standard scores for the five scales. The means and standard deviations of the sums of standard scores were evaluated for each of the 29 normative groups. Although some minor differences in means and standard deviations were evident at different age groups, they were not sufficiently large to make any clinical distinction, and thus the decision was made to calculate the General Development score based on the whole standardization sample.

Parent/Caregiver Checklist

The norms for the Parent/Caregiver Checklist were derived through a process made possible by the use of the Rasch model. Since Rasch measurement is sample-free, and given the comparability of item content, it was possible to use the standardization data to help to

Table 7
Demographic Characteristics of the DP-3 Standardization Sample

	n	Sample %	U.S. Census %[a]
Gender			
Male	1,094	49.4	49.2
Female	1,120	50.5	50.8
Missing	2	0.1	
Race/Ethnic background			
Asian	100	4.5	4.1
Black/African American	299	13.5	12.3
Hispanic/Latino	272	12.3	14.1
Native American	5	0.2	0.8
Native Hawaiian/Pacific Islander	2	0.1	0.1
White	1,474	66.5	67.4
Other	60	2.7	1.5
Missing	4	0.2	
U.S. geographic region			
Northeast	347	15.7	19.0
Midwest	546	24.6	22.9
South	885	39.9	35.6
West	438	19.8	22.5
Parents' educational level			
Less than high school graduate	287	13.0	11.8
High school graduate	620	28.0	30.7
Some college	616	27.8	27.6
Four years of college or more	688	31.1	30.2
Missing	5	0.2	
Age in years and months			
0-0 to 0-3	119		
0-4 to 0-7	110		
0-8 to 0-11	113		
1-0 to 1-3	102		
1-4 to 1-7	78		
1-8 to 1-11	86		
2-0 to 2-3	94		
2-4 to 2-7	88		
2-8 to 2-11	88		
3-0 to 3-5	107		
3-6 to 3-11	95		
4-0 to 4-5	108		
4-6 to 4-11	97		
5-0 to 5-5	111		
5-6 to 5-11	92		
6-0 to 6-5	79		
6-6 to 6-11	85		
7-0 to 7-11	94		
8-0 to 8-11	105		
9-0 to 9-11	99		
10-0 to 10-11	94		
11-0 to 11-11	97		
12-0 to 12-11	75		

Note. N = 2,216.

[a] U.S. Census figures (U.S. Census Bureau, 2005) are based on the U.S. population as a whole, except for the parent education category, which is based on adults aged 25 to 54 (those most likely to be parents of young children).

Table 8
Raw Score Means and Standard Deviations for Each DP-3 Normative Age Group

	Scale									
	Physical		Adaptive Behavior		Social-Emotional		Cognitive		Communication	
Age group	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
0-0 to 0-1	0.5	0.8	1.2	0.9	1.6	0.9	0.5	0.8	0.7	0.7
0-2 to 0-3	1.1	0.8	1.7	1.0	3.0	1.5	1.5	1.1	1.3	0.7
0-4 to 0-5	2.6	1.0	3.6	1.5	4.5	2.1	3.4	1.5	2.3	1.3
0-6 to 0-7	3.7	1.1	5.2	1.4	5.7	1.8	4.3	1.6	3.3	1.5
0-8 to 0-9	5.4	1.6	6.7	1.8	7.9	2.0	6.1	1.6	5.1	1.8
0-10 to 0-11	6.9	1.9	7.9	2.2	8.9	2.3	7.3	2.0	6.1	2.0
1-0 to 1-1	8.6	2.2	9.1	1.6	10.3	2.5	8.7	1.7	7.1	2.3
1-2 to 1-3	9.9	2.3	9.5	1.7	10.8	1.9	9.6	1.8	8.1	1.6
1-4 to 1-5	12.5	2.8	11.5	2.5	12.8	2.0	11.1	1.8	9.4	2.2
1-6 to 1-7	13.0	2.4	12.3	2.5	12.6	2.2	12.2	2.3	10.5	2.0
1-8 to 1-9	13.1	2.4	13.1	2.5	13.9	2.5	12.4	2.2	11.3	2.9
1-10 to 1-11	14.3	2.4	13.8	3.2	14.1	2.8	13.7	2.5	12.4	3.0
2-0 to 2-3	16.6	3.4	15.2	2.4	16.2	2.6	15.7	2.5	15.3	2.8
2-4 to 2-7	17.1	3.4	16.2	2.8	16.4	2.7	16.1	3.1	16.1	3.2
2-8 to 2-11	19.5	4.0	17.5	3.1	17.8	2.5	18.2	3.1	17.4	3.0
3-0 to 3-5	21.1	3.9	19.1	2.7	19.8	2.8	20.2	3.3	19.1	2.9
3-6 to 3-11	23.0	3.5	20.5	3.3	20.8	3.0	21.3	3.4	19.9	3.1
4-0 to 4-5	26.3	3.5	22.9	2.9	23.3	2.9	24.2	2.8	21.8	2.6
4-6 to 4-11	27.1	3.3	24.5	3.1	24.3	3.0	25.2	2.7	23.0	2.7
5-0 to 5-5	28.9	2.9	26.1	3.5	25.1	2.7	27.1	2.9	24.3	2.7
5-6 to 5-11	30.1	2.8	27.2	3.2	26.1	2.7	28.0	2.7	25.4	2.5
6-0 to 6-5	31.7	2.4	29.6	3.1	27.8	2.8	30.6	2.4	28.2	2.3
6-6 to 6-11	32.5	2.6	30.7	2.9	29.0	3.3	31.6	2.5	29.0	2.6
7-0 to 7-11	33.2	1.6	31.8	2.4	29.8	3.2	33.0	1.9	30.1	1.9
8-0 to 8-11	34.2	1.7	33.1	2.3	30.9	2.3	35.1	1.9	30.9	1.8
9-0 to 9-11	34.5	0.9	34.2	2.1	31.9	2.3	36.3	1.7	31.9	1.6
10-0 to 10-11	34.6	1.1	34.8	2.2	32.9	2.6	36.9	1.6	32.4	1.9
11-0 to 11-11	34.8	0.6	35.6	1.5	33.4	2.3	37.4	0.9	32.9	1.5
12-0 to 12-11	34.9	0.5	36.1	1.4	34.1	1.8	37.4	0.9	33.5	1.2

generate the Parent/Caregiver Checklist norms. The first step was to compare the item difficulty between the Interview Form and the Parent/Caregiver Checklist for the 318 individuals who completed both. Differences were found to be small, often lower than the standard error of measurement of the item, and therefore it was determined to be appropriate to use the interview data to contribute to the Parent/Caregiver Checklist norms. The benefit of this was in taking the robustness associated with the larger normative sample and applying it to the alternative form (the Parent/Caregiver Checklist).

The next step involved running the Parent/Caregiver Checklist items answered by a total of 377 individuals through WINSTEPS (Linacre, 2003), along with the Interview Form items for all respondents in the standardization sample (2,216). Thus, the Parent/Caregiver Checklist items were calibrated alongside the Interview Form items. Then the ability level associated with each raw score was generated for each item set (Parent/Caregiver Checklist and Interview Form) for each of the five DP-3 scales. The results indicated that differences between the abilities associated with each raw score value for the two item sets were small. However, some differences did exist, and thus separate norms tables were created. This was accomplished by matching the ability levels of the Interview Form and the Parent/Caregiver Checklist to find the associated raw score and standard score. Users of both tables will notice a great deal of similarity. That is, for many raw score values, the norm-based standard score is the same whether the DP-3 Interview Form or the Parent/Caregiver Checklist was used; however, for some raw score values, the standard score is different.

As with the Interview Form, the General Development score for the Parent/Caregiver Checklist was derived by adding the standard scores for the five scales. This standard score was based on the entire Parent/Caregiver Checklist sample.

Moderator Variables

Relevant moderator variables in the standardization sample were evaluated to determine whether the DP-3 can effectively be used across groups, without bias. The age variable was already accounted for by the creation of multiple normative groups. Variables such as gender, ethnic background, and parent education could affect scores in a way that is unintended. If any of the groups differed greatly on the test, it would cause difficulties with interpretation. Therefore, this section discusses these relevant moderator variables and presents data regarding statistical and meaningful differences. The analyses conducted to determine the impact of moderator variables utilized the standardization study data to compare the standard scores for each of the five DP-3 scales and the General Development score to the average score for the entire standardization sample (100, by definition). Standard scores provide a convenient way to examine data for effect sizes because they have a uniform mean and standard deviation.

It should be noted that in a sample as large as the DP-3 standardization sample, differences between groups are often statistically significant, even if the actual differences are very small. A difference of 2 or 3 standard score points will not make a significant clinical difference, and therefore, effect sizes are evaluated in addition to statistical significance. Effect sizes help to determine whether the statistically significant differences hold any clinical meaning (Cohen, 1992). Effect sizes of 0.1 to 0.3 deviation units (approximately 1 to 4 standard score points) are considered small and not clinically meaningful, effect sizes between 0.3 and 0.5 deviation units (approximately 5 to 8 standard score points) are considered moderate, and effect sizes greater than 0.5 deviation units (or greater than approximately 8 standard score points) are considered large. Effect sizes that are not small suggest that clinically meaningful differences do exist, particularly when a consistent pattern of differences is observed that fits with other knowledge of the group in question. The tables discussed in the following sections display the means for each group of interest on each of the five DP-3 scales. One-sample *t* tests were conducted comparing the group means to the entire standardization sample mean. Any analyses that were statistically significant were flagged, and their effect sizes were calculated.

Gender

The average standard scores for boys and girls in the standardization sample are provided in Table 9. The table reveals that 9 of the 12 comparisons between each group and the expected mean of 100 were significant. Boys tended to score slightly lower than the mean, while girls tended to score slightly higher. However, the effect sizes were all small, ranging from 0.07 to 0.13 (median = 0.09), suggesting that the differences are not clinically important.

Table 9
Average Standard Scores for Boys and Girls in the DP-3 Standardization Sample

Scale	Boys (*n* = 1,094)	Girls (*n* = 1,120)
Physical	100.3	100.3
Adaptive Behavior	98.7* *(es = 0.09)*	101.1* *(es = 0.07)*
Social-Emotional	98.1* *(es = 0.13)*	100.7
Cognitive	98.9* *(es = 0.07)*	101.4* *(es = 0.09)*
Communication	98.2* *(es = 0.12)*	101.4* *(es = 0.09)*
General Development score	98.5* *(es = 0.10)*	101.5* *(es = 0.10)*

*p < .01 for a one-sample t test comparing the obtained value with the expected mean of 100. Numbers in parentheses are effect sizes.

Table 10
Average Standard Scores by Parent Education Level in the DP-3 Standardization Sample

Scale	Not high school graduate (n = 287)	High school graduate (n = 620)	Some college (n = 616)	College graduate or more (n = 688)
Physical	99.5	99.0	100.3	101.8* *(es = 0.12)*
Adaptive Behavior	100.7	99.2	99.7	100.4
Social-Emotional	100.2	97.3* *(es = 0.18)*	98.5* *(es = 0.10)*	101.7* *(es = 0.11)*
Cognitive	96.4* *(es = 0.24)*	97.7* *(es = 0.15)*	100.3	103.9* *(es = 0.26)*
Communication	97.7* *(es = 0.15)*	98.0* *(es = 0.13)*	100.3	102.1* *(es = 0.14)*
General Development score	98.6	97.6* *(es = 0.16)*	99.9	102.9* *(es = 0.19)*

*p < .01 for a one-sample t test comparing the obtained value with the expected mean of 100. Numbers in parentheses are effect sizes.

Parent Education Level

Average standard scores based on parent education level are presented in Table 10. As with gender, many of the comparisons were found to be statistically significant; however, none of the comparisons exceeded a small effect size and thus were not considered to be clinically significant. Effect sizes for the 12 significant comparisons ranged from 0.10 to 0.26 (median = 0.15). The highest parent education group (bachelor's degree or higher) obtained the highest means; however, those differences were not large enough to draw meaningful conclusions with regard to interpretation of DP-3 results in clinical practice. Nonetheless, a relationship between socioeconomic status and child development has been established (e.g., Reading, 2004), and therefore the observed differences in means support the validity of the test.

Ethnicity

Table 11 presents the average standard scores for five of the seven ethnic groups. There were only two individuals in the standardization sample who self-identified as Native Hawaiian and five who self-identified as Native American. These numbers were too small for an examination of means to lead to any reliable conclusions. For this reason, these two ethnic groups were excluded from the table. Of the 30 comparisons, only 4 were clinically significant, suggesting that, overall, a single set of DP-3 norms is similarly valid across ethnic groups. One of these clinically significant differences was found to exceed the threshold of small effect sizes (*es* = 0.41). On the Social-Emotional scale, Asians had a significantly lower average score. However, Asians represent one of the smaller ethnic groups, and therefore it is likely that this difference is due to sampling anomalies rather than to

Table 11
Average Standard Scores for Children in Various Ethnic Groups in the DP-3 Standardization Sample

Scale	Asian (n = 100)	Black/ African American (n = 299)	Hispanic/ Latino (n = 272)	White (n = 1,474)	Other (n = 60)
Physical	98.0	100.9	97.9* *(es = 0.14)*	100.7	101.5
Adaptive Behavior	98.7	100.7	99.5	99.9	98.8
Social-Emotional	93.9* *(es = 0.41)*	100.0	97.8* *(es = 0.15)*	99.8	102.0
Cognitive	99.5	98.5	98.0	100.9	102.4
Communication	98.6	98.4	98.5	100.6	99.5
General Development score	96.9	99.7	97.8* *(es = 0.15)*	100.6	101.3

Note. Due to the small cell sizes for the Native American and Native Hawaiian/Pacific Islander samples, the results for those groups are excluded from this table as they are not reliable enough to interpret.

*p < .01 for a one-sample t test comparing the obtained value with the expected mean of 100. Numbers in parentheses are effect sizes.

actual differences. Especially given that a pattern of differences was not observed, it is reasonable to interpret the Social-Emotional scores for Asians in the same way as for other ethnic groups.

Overall, the standard scores based on the entire standardization sample, stratified by age, apply well to a number of different demographic groups. Differences that were found between groups were generally small and lacked any pattern that suggested they should be interpreted in a different way or were worthy of separate norms. This suggests that the DP-3 can be validly used by all groups included in the standardization sample. Although these variables were not examined for the Parent/Caregiver Checklist, the high level of similarity between the two forms allows for the assumption that the Parent/Caregiver Checklist can also be effectively used with different demographic groups. It should be noted that although the Parent/Caregiver Checklist sample was not designed to be nationally representative, it comprised individuals who varied in gender, ethnicity, parent education, region, and age.

5
Technical Properties

This chapter describes the psychometric properties of the DP-3 and the earlier versions of the *Developmental Profile*. The first section reviews the reliability of the current version. The following section delineates the validity studies on the DP-3 and then reviews the research that has been conducted on the original *Developmental Profile* and the DP-II since the 1970s.

Reliability

The reliability of a test refers to the extent to which the results are dependable and relatively free from error. That is, an individual should obtain a similar score on repeated testing occasions under varying circumstances of administration. Adequate reliability is necessary for a test user to feel confident in using the scores to describe a child's developmental functioning. Two important types of reliability are described here: internal consistency and test-retest.

Internal Consistency Reliability

Internal consistency reliability refers to the extent to which the items on each scale represent a common underlying construct, in this case, one of five areas of child development. For the purposes of this analysis, as well as for the calculation of the standard error of measurement, the standardization and clinical samples were combined. Utilizing this combined sample increased the variance of the test (see the discussion of ceiling effects for the standardization sample in chapter 3) and better represents the population for which the test will be used.

Internal consistency can be estimated in multiple ways, but the one most appropriate for a test with a developmental gradient is the split-half method. In this procedure, items on each scale are separated into two halves by alternating consecutive items. The resulting Pearson correlation from these two halves is adjusted using the Spearman-Brown formula to estimate the reliability based on the full length of the scale, rather than half of it (Anastasi & Urbina, 1997). Internal consistency estimates for the five DP-3 scales and the General Development score at each age year are presented in Table 12. (Separate internal consistency estimates for the standardization sample and the clinical sample can be found in Tables B31 and B32, respectively.) It can be seen that all of the correlations are above .80, indicating that they range from good to excellent. Two thirds of the correlations are .90 or above. Thus, these internal consistency estimates support the strong reliability of the DP-3.

Test-Retest Reliability

Test-retest reliability represents the stability of DP-3 scores over time and involves administering the test to the same parents on two occasions and then correlating the scores from each administration. Sixty-six individuals from the standardization sample were administered the DP-3 Interview Form a second time, with an average interval of 2 weeks (range = 13 to 18 days) between administrations. Two examiners participated in this study, one in the Southern region of the United States and one in the Midwestern region. Individuals in the test-retest study varied by age, gender, ethnicity, and education level. Correlation coefficients are presented in Table 13 and indicate that the test-retest correlations range from .81 to .92 for the five scales and the General Development score, representing good reliability over time across different ages and demographic groups.

Standard Error of Measurement *(SEM)*

The standard error of measurement *(SEM)* utilizes a measure of reliability in order to approximate the measurement of error in a score and thus the amount that an observed score differs from the "true score," assuming the test contained no error. *SEM* values for the combined standardization and clinical sample are presented in Table 14 by age and scale as standard score points. (Separate

Table 12
Internal Consistency Estimates of the DP-3 Scales by Age Year for the Combined Sample

		Scale					
Age year	***n***	**Physical**	**Adaptive Behavior**	**Social-Emotional**	**Cognitive**	**Communication**	**General Development score**
0	361	.93	.92	.89	.89	.90	.97
1	296	.89	.91	.88	.87	.88	.97
2	342	.89	.91	.93	.91	.93	.97
3	259	.86	.88	.92	.90	.91	.97
4	269	.90	.88	.94	.93	.90	.97
5	238	.87	.86	.86	.89	.86	.95
6	179	.93	.91	.88	.93	.92	.98
7	102	.94	.91	.93	.92	.90	.98
8	124	.96	.91	.84	.93	.91	.98
9	123	.92	.90	.82	.92	.88	.97
10	109	.96	.90	.86	.88	.87	.96
11	122	.99	.96	.94	.97	.96	.99
12	90	.97	.96	.89	.86	.94	.98
Median		.93	.91	.89	.91	.90	.97

Note. *N* = 2,614.

SEM values for the standardization sample and the clinical sample can be found in Tables B33 and B34, respectively.) The *SEM* is calculated using the following equation:

$SEM = SD\ \sqrt{1 - r},$ where *SD* = standard deviation
and *r* = reliability

For the DP-3, the value of 15 is used for the standard deviation, and the internal consistency reliability for each scale at each age year is used. The *SEM* values range from 1.51 to 6.30, with most of the *SEM*s ranging from approximately 3 to 5 standard score points.

The *SEM* values can be most practically used to calculate confidence intervals, which represent the range within which an individual's true score lies with a certain level of probability. Different levels of confidence can be utilized, but the 95% level is recommended here, as it is the basic standard in psychology. Convenient rules of thumb are offered to the user as an approximation of the 95% confidence bands calculated for the combined standardization and clinical sample. These rules of thumb tend to be relatively conservative estimates. For all five DP-3 scales, a confidence range of ±10 standard score points is recommended for use with children aged 0-0 to 5-11, and a confidence range of ±9 standard score points is recommended for use with children aged 6-0 to 12-11. For the General Development score, a confidence range of ±5 standard score points is appropriate for all ages. For example, a child aged 4 years, 5 months who obtains a score of 90 on the Social-Emotional scale will have a confidence band of approximately ±10, and therefore there is a 95% probability that the child's true raw score lies in the range of 80 to 100.

Table 13
Test-Retest Reliability for DP-3 Scores

Scale	Two-week interval
Physical	.86
Adaptive Behavior	.82
Social-Emotional	.81
Cognitive	.88
Communication	.82
General Development score	.92

Note. *N* = 66.

For test users who wish to know the exact confidence intervals, Tables B35, B36, and B37 display the 95% and 90% confidence intervals for the combined standardization and clinical sample, the standardization sample alone, and the clinical sample alone, respectively. The confidence intervals are displayed as plus and minus a specific number. To use the tables, the user must read down to find the child's age year and across to the relevant DP-3 scale, then add and subtract the confidence interval number from the obtained standard score. For example, if a child aged 2 years, 5 months is tested and

Table 14
Standard Errors of Measurement for the DP-3 Scales by Age Year for the Combined Sample

		Scale					
Age year	*n*	Physical	Adaptive Behavior	Social-Emotional	Cognitive	Communication	General Development score
0	361	4.10	4.22	5.08	4.97	4.67	2.40
1	296	5.08	4.42	5.17	5.38	5.22	2.40
2	342	4.94	4.57	3.99	4.39	3.92	2.40
3	259	5.53	5.23	4.22	4.86	4.44	2.40
4	269	4.72	5.15	3.58	3.91	4.67	2.40
5	238	5.49	5.66	5.64	4.96	5.70	3.26
6	179	3.94	4.41	5.11	4.05	4.26	2.14
7	102	3.62	4.59	3.94	4.30	4.73	2.14
8	124	2.86	4.39	6.00	4.07	4.45	1.85
9	123	4.13	4.80	6.30	4.15	5.30	2.40
10	109	3.16	4.72	5.54	5.27	5.33	2.86
11	122	1.51	2.84	3.81	2.45	2.98	1.51
12	90	2.73	3.14	5.00	5.59	3.70	2.14
Median		4.10	4.57	5.08	4.39	4.67	2.40

Note. N = 2,614. Standard errors of measurement are reported in standard score units.

found to have a standard score of 85 on the Social-Emotional scale, Table B35 is referenced to find that the confidence interval of ±8 is used to determine that 95 times out of 100 the child's standard score will fall between 77 and 93. Age year 2 is used in this example, as it includes children aged 2-0 to 2-11.

Reliability of the Parent/Caregiver Checklist

Internal consistency estimates were calculated for the five DP-3 scales and the General Development score on the Parent/Caregiver Checklist in the same manner as for the Interview Form. However, some age years were combined due to the smaller sample size. Table 15 presents the correlations based upon the combined group of typically developing and clinical children. As with the Interview Form, the correlations all range from good to excellent (.79 to .99). Thus the strong reliability found for the Interview Form holds for the Parent/Caregiver Checklist as well.

*SEM*s and confidence intervals were also calculated for the Parent/Caregiver Checklist; these can be found in Tables C31 and C32, respectively. Alternatively, for the confidence intervals the rules of thumb outlined for the Interview Form can be applied. For all five DP-3 scales, a confidence range of ±10 is recommended for use with children aged 0-0 to 5-11, and a confidence range of ±9 is recommended for use with children aged 6-0 to 12-11. For the General Development score, a confidence range of ±5 is appropriate for all ages.

Validity

The validity of a test refers to its ability to accurately measure what it is designed to measure. The examination of validity is an ongoing process, and this chapter presents information describing the validity studies conducted during the development and standardization process of the DP-3 and reviews the research literature for validity evidence from previous versions of the test. Various types of validity have been examined and are described in the following sections.

Content Validity

Content validity refers to the utilization of appropriate item content to measure the area of interest. Therefore, an attempt to build content validity into the *Developmental Profile* was made from the outset. During the initial development stages of the original instrument, the literature and existing measures were surveyed to identify and define the broad spectrum of developmental skills. These were categorized into five skill areas reflecting a multidimensional view of child development. The selection and development of the items were conducted to ensure that items were age appropriate and

Table 15
Internal Consistency Estimates of the DP-3 Parent/Caregiver Checklist Scales by Age Year

		Scale					General
Age year	*n*	Physical	Adaptive Behavior	Social-Emotional	Cognitive	Communication	General Development score
0	67	.88	.91	.86	.91	.91	.97
1	53	.87	.83	.84	.91	.90	.97
2	56	.96	.95	.94	.92	.94	.98
3	42	.83	.81	.84	.79	.82	.96
4	32	.90	.80	.90	.90	.93	.96
5	33	.94	.95	.86	.96	.94	.97
6	33	.95	.87	.95	.96	.96	.99
7–8	35	.96	.93	.91	.96	.92	.98
9–10	39	.99	.95	.97	.97	.97	.99
11–12	42	.92	.91	.92	.95	.95	.97
Median		.94	.91	.90	.92	.93	.97

Note. *N* = 432. Some age years were combined due to small cell size.

representative of their respective skill area. Item development included an extended period of field testing; teachers serving handicapped children used the instrument to assess individual children and to plan and implement skill-based curricula. Their responses to the instrument's clarity and usefulness for designing and evaluating instructional interventions provided a check on the content validity of the inventory. Additionally, the fact that raw scores consistently increase as the child's age increases provides evidence that the DP-3 accurately measures relevant developmental content, as development is expected to increase with age.

Construct Validity

Construct validity is measured by examining the structural characteristics of the scales through the use of interscale correlations, factor analysis, and item response theory analysis. It is also assessed through the relationship between the DP-3 and other psychological tests. Construct validity is supported when high correlations are found between the DP-3 and measures designed to assess similar constructs and when lower correlations are observed between the DP-3 and measures of different psychological characteristics.

Structural characteristics. Interscale correlations for the five DP-3 scales and the General Development score were calculated for each age year. Results indicated that although most were similar across age groups, at ages 8 and above the Physical scale was less strongly correlated to most of the other scales. This is likely related to the fact that the Physical scale has a lower ceiling compared to the other scales of the test. Table 16 displays interscale correlation results based upon the entire standardization sample. These were calculated

Table 16
Interscale Correlations in the DP-3 Standardization Sample

Scale	Physical	Adaptive Behavior	Social-Emotional	Cognitive	Communication	General Development score
Physical	–					
Adaptive Behavior	.49	–				
Social-Emotional	.42	.54	–			
Cognitive	.44	.47	.51	–		
Communication	.39	.44	.48	.59	–	
General Development score	.71	.77	.78	.79	.75	–

Note. *N* = 2,216.

using the standard scores obtained on each of the five DP-3 scales and the General Development score. The scales all exhibit correlations in the moderate range, which is not unanticipated. Given the fact that each scale represents one aspect of child development, it is expected that the scales would be related to one another. However, each scale has a higher correlation with the General Development score than with any of the other scales, and the correlations between the five scales are lower than the reliability estimates for each scale. This provides support for the separate scoring and interpretation of the five scales.

Items were also analyzed to determine the correlations with their assigned scales, each of the other four DP-3 scales, and the General Development score. As expected, due to the interrelated nature of aspects of child development, the analysis revealed that items tended to correlate well with all scales and the General Development score. As mentioned, each of the five scales can be viewed as representing one aspect of general child development, with the breakdown by scale useful for interpretation and remediation planning.

To further examine the structure of the DP-3, an exploratory common factor analysis with oblimin rotation was conducted with all 180 items using the standardization sample. Results indicated that items tended to load primarily onto one dominant factor. Although other factors emerged, the first factor appears to represent a general development factor. Interestingly, item loadings onto this first factor were similar across scales for items at the same difficulty level. This type of result wherein the factor analysis yielded variation based on difficulty level (item endorsement) is a frequent result when using dichotomous variables (Floyd & Widaman, 1995) and thus is not specific to the DP-3. This finding also likely results from the fact that the items have a steep difficulty gradient.

A second exploratory common factor analysis was done with the standard scores for each of the five scales for all of the individuals in the standardization sample, and the results indicated that all scales loaded onto a single factor, with loadings ranging from .61 to .74.

Structural characteristics of the items and scales were also examined using Rasch methodology. As described in chapter 4, the Rasch model estimates item difficulty and person ability on the same metric. The Rasch measures (as expressed in logits) were examined for the five scales of the DP-3. Table 17 presents the range (in logits) of item difficulty and person ability for each of the five scales. It can be seen that for all of the scales, the range of person ability extends slightly below and slightly above the range of item difficulty. This is not unexpected, as the skills tested by the first few items on each scale (representing the earliest measurable developmental tasks) are generally not performed by newborns. Additionally, major tasks of child development are generally achieved during the elementary school years, and thus each scale has a ceiling that was hit by some of the individuals in the standardization sample. The fact that the ranges of person abilities and item difficulties for each scale are similar to one another reveals that the items do a good job of measurement within the desired skill range.

In addition to examining the range of item difficulties, the progression of difficulty throughout each scale was explored. Findings indicated that gaps between items were generally less than one logit on all scales (a few items scattered throughout were between one to two logits), indicating good distribution of item difficulties. Some larger gaps (greater than two logits) were noted at the low end of the scales, which is expected given steep developmental increases in the early months. One logit indicates a different age spread depending on the child's age in a way that is consistent with the progression of development. In the early months, one logit tends to equal approximately 1 month, during the preschool years one logit represents approximately 3 to 5 months, and in the elementary school years one logit represents about 1 year. These findings lend support to the ability of all five scales to measure development through the age range of the test in a dependable fashion.

Table 17
Rasch Model Information From the DP-3 Standardization Sample

Scale	Range of item difficulties	Range of person abilities
Physical	–25.67 to 10.72	–26.70 to 12.67
Adaptive Behavior	–23.24 to 11.74	–24.09 to 13.50
Social-Emotional	–15.31 to 9.21	–16.30 to 11.38
Cognitive	–16.92 to 15.38	–18.27 to 16.51
Communication	–21.65 to 13.30	–22.52 to 14.90

Note. Ranges are presented in logits, which are on an equal interval scale and have a mean of 0 and a standard deviation of 1.

Finally, the functioning of the items on each was examined by looking at the items' fit with the measurement model. Across all five scales (180 items), only 6 items had an infit statistic exceeding 1.30, and thus the vast majority of items evidenced very good levels of fit.

Construct validity through the relationship between the DP-3 and other tests. To further validate the DP-3 scales, scores were compared to those obtained from related tests. The samples for these concurrent validity studies were drawn from subsamples of the DP-3 clinical sample. To be part of the clinical sample, individuals needed to have a behavioral, emotional, developmental, or other problem severe enough to warrant referral for services. A total of 398 children from 16 sites comprised the clinical sample, with diagnoses including developmental delay, mental retardation, maternal in utero drug use, autism, traumatic brain injury, physical/medical disabilities, Down syndrome (and other chromosomal disorders), cerebral palsy, visual impairment, hearing impairment, Attention-Deficit/Hyperactivity Disorder (ADHD), Oppositional Defiant Disorder, Conduct Disorder, adjustment disorders, mood disorders, speech delays, and learning disabilities. The sample ranged in age from 4 months to 12 years, 11 months and varied adequately by ethnicity and parent education level. The sample had approximately twice the number of boys as girls, which is consistent with general research findings of higher rates of developmental and other disabilities among males (e.g., Cuffe, Moore, & McKeown, 2005; Fombonne, 1999).

Tests of development. The *Vineland Adaptive Behavior Scales, Second Edition* (Vineland II; Sparrow et al., 2005) was administered to the parents of 89 individuals from the clinical sample. The Vineland II data were gathered from six different sites, and the children ranged in age from 1-10 to 12-11. The Vineland II is a comprehensive measure of adaptive behavior. Its format is similar to that of the DP-3, as it utilizes a parent interview method as a means of obtaining information. Additionally, it measures skills in four of the five DP-3 domains. Table 18 displays the correlations between the four Vineland II domains and the Adaptive Behavior Composite and the five DP-3 scales and the General Development score. The correlations for the DP-3 scales and Vineland II domains most similar in content are bolded. Correlations are moderate to high for all comparisons, ranging from .42 to .85. Correlations are similar across scales and domains, pointing to the construct of general child development that appears to underlie the different areas of functioning. However, three of the four (bolded) correlations between scales/domains of similar content are higher than the correlations between the same Vineland II domain and the other four DP-3 scales. This suggests that despite the high correlations across all areas, the DP-3 content scales do measure certain specific aspects of development in a similar way to the Vineland II.

The *Developmental Assessment of Young Children* (DAYC; Voress & Maddox, 1998) was administered to 139 parents from seven of the clinical sites. The DAYC is a collection of five subtests that measure the same areas of functioning as the DP-3 with a greater number of items. The DAYC is designed for use with children aged birth through 5 years, 11 months. The correlations between the DAYC subtests and the General Development Quotient and the five DP-3 scales and the General Development score are displayed in Table 19. Correlations were found to be

Table 18
Correlations Between the DP-3 and the *Vineland Adaptive Behavior Scales, Second Edition* (Vineland II)

	DP-3					
Vineland II	Physical	Adaptive Behavior	Social-Emotional	Cognitive	Communication	General Development score
Communication	.59	.55	.79	.73	**.82**	.72
Daily Living Skills	.71	**.68**	.78	.68	.66	.75
Socialization	.53	.53	**.76**	.67	.72	.70
Motor Skills[a]	**.85**	.69	.53	.42	.56	.59
Adaptive Behavior Composite	.69	.67	.85	.78	.84	.81

Note. N = 89 for all Vineland II scales except Motor Skills. Bold type indicates expected correlation based on similar content.
[a]Motor Skills domain is administered only to children up through age 6; *n* = 28.

Table 19
Correlations Between the DP-3 and the *Developmental Assessment of Young Children* (DAYC)

	DP-3					
DAYC	Physical	Adaptive Behavior	Social-Emotional	Cognitive	Communication	General Development score
Cognitive	.60	.68	.61	**.68**	.61	.68
Communication	.46	.63	.66	.68	**.71**	.68
Social-Emotional	.52	.64	**.68**	.64	.64	.64
Physical Development	**.69**	.56	.49	.53	.44	.53
Adaptive Behavior	.50	**.64**	.60	.63	.56	.63
General Development Quotient	.65	.72	.68	.72	.66	**.72**

Note. N = 139. Bold type indicates expected correlation based on similar content.

moderate, ranging from .44 to .72 across all scales. For all five content areas measured by both tests, the correlations were highest between the scales assessing similar content. This again points to the utility of the DP-3 scales in providing information about each of the five areas of development.

Tests of specific domain areas. In addition to comparing the DP-3 with other tests of general development, data were also gathered from tests that examine areas specific to only one of the DP-3 scales. For the following two tests, only certain scores are presented, as additional data were not available from the sites.

The *Preschool Language Scales, Fourth Edition* (PLS-4; Zimmerman, Steiner, & Pond, 2002) is designed to measure the receptive and expressive language of young children. Scores were obtained for Auditory Comprehension and Expressive Communication for 37 children and were correlated with the five DP-3 scales and the General Development score. Table 20 presents these correlations. The expected correlations were between the Communication scale of the DP-3 and both PLS-4 scores, and these were found to be moderate, at .48 and .53. Correlations with the Social-Emotional and Cognitive scales and the General Development score yielded some correlations at similar levels, which is not entirely unexpected, as language has a social and cognitive component and is related to child development as a whole.

The *Peabody Developmental Motor Scales, Second Edition* (PDMS-2; Folio & Fewell, 2002) assesses fine- and gross-motor skills in young children. Scores from the two fine-motor subtests, Grasping and Visual-Motor Integration, were available for 23 of the children in the clinical sample. Table 21 illustrates the correlations for the five DP-3 scales and the General Development score with these two subtests. As would be expected, moderate correlations were found between the two PDMS-2 subtests

Table 20
Correlations Between the DP-3 and the *Preschool Language Scales, Fourth Edition* (PLS-4)

	DP-3					
PLS-4	Physical	Adaptive Behavior	Social-Emotional	Cognitive	Communication	General Development score
Auditory Comprehension	.28	.40	.53	.36	**.48**	.45
Expressive Communication	.20	.34	.50	.49	**.53**	.48

Note. N = 37. Bold type indicates expected correlation based on similar content.

Table 21
Correlations Between the DP-3 and the *Peabody Developmental Motor Scales, Second Edition* (PDMS-2)

	DP-3					
PDMS-2	Physical	Adaptive Behavior	Social-Emotional	Cognitive	Communication	General Development score
Grasping	**.56**	.69	.84	.70	.60	.72
Visual-Motor Integration	**.71**	.76	.65	.65	.51	.66

Note. *N* = 23. Bold type indicates expected correlation based on similar content.

and the Physical scale of the DP-3 (.56 and .71). In fact, moderate-to-high correlations were found with all of the DP-3 scores. These findings should be viewed with caution due to the small sample size. However, the results were included to further illustrate the ability of the DP-3 to measure constructs similarly to other widely used tests of development.

Discriminant Validity

The clinical sample was used to illustrate that the DP-3 can effectively discriminate between typically developing children and children with a clinical problem. Initially, the clinical sample of 398 individuals was examined as a whole and compared to the expected mean of the standardization sample (100, by definition). Table 22 illustrates that for all five scales and the General Development score, the clinical sample was both statistically and meaningfully different from the standardization sample. The effect sizes ranged from 1.5 to 2.2 (median = 1.8), indicating differences far above the cutoff for a large effect size. The clinical sample was then divided into three groups based upon diagnosis. The first group consisted of children who had developmental delays and disorders, maternal drug use, mental retardation, and physical/medical/chromosomal disabilities; the second group comprised individuals with emotional, behavioral, and adjustment disorders; and the third group was made up of children with speech and learning disabilities. Analyses determined that the standard score means were not significantly different for the second and third groups, and therefore they were combined into a single group.

Table 23 displays *t* test results comparing the first group and the combined group on standard scores for the five DP-3 scales and the General Development score. The results illustrate that the group consisting of children with developmental delays and related difficulties had group means that were significantly lower than those for the group consisting of children with behavioral,

Table 22
Average Standard Scores in the DP-3 Clinical Sample

Scale	Mean standard score
Physical	77.1* *(es = 1.5)*
Adaptive Behavior	73.5* *(es = 1.8)*
Social-Emotional	70.2* *(es = 2.0)*
Cognitive	74.2* *(es = 1.7)*
Communication	73.4* *(es = 1.8)*
General Development score	67.4* *(es = 2.2)*

Note. *N* = 398.

*p < .01 for a one-sample *t*-test comparing the obtained value with the standardization mean of 100. Numbers in parentheses are effect sizes.

Table 23
Average Standard Scores for Two Groups in the DP-3 Clinical Sample

Scale	Developmental delays[a] (*n* = 242)	Other problems[b] (*n* = 153)	*t*
Physical	68.5	90.4	10.66*
Adaptive Behavior	67.5	83.1	7.98*
Social-Emotional	61.3	84.0	10.96*
Cognitive	65.2	88.2	10.35*
Communication	67.1	83.6	8.48*
General Development score	59.0	80.5	11.00*

Note. Three individuals in the clinical sample were missing specific diagnoses and thus were not included in the above groups.

[a] This group includes children with developmental delays and disorders, maternal drug use, mental retardation, and physical/medical/chromosomal disabilities.

[b] This group includes children with emotional, behavioral, and adjustment disorders, and speech and learning disabilities.

*p < .001 for an independent-sample *t* test comparing the two groups.

emotional, speech, or learning problems. The means for the first group all fell into the Delayed range, while the means for the other group ranged from Below Average to Average. The group with developmentally delayed children would be expected to demonstrate more pervasive and deficient scores across all scales, while the other group would be expected to have lower scores in specific skill areas. For example, the mean score on the Physical scale for the second group was in the Average range, suggesting that their disabilities are more specific to their diagnoses (emotional/behavioral problems, speech and learning disabilities) and do not greatly impact other areas of development. These results support the validity of the DP-3 in its ability to distinguish between two types of clinical difficulties. This is especially important given that the measure is designed to detect developmental delays, and these data illustrate that it does just that.

Validity of the Parent/Caregiver Checklist

Given the high level of similarity of content and measurement between the Interview Form and the Parent/Caregiver Checklist, it is reasonable to apply the validity evidence from the Interview Form to the Parent/Caregiver Checklist version. However, some additional validity studies were conducted to specifically examine the relationship between the DP-3 Parent/Caregiver Checklist and other parent report measures on related constructs.

Construct validity. The Vineland II (Sparrow et al., 2005) Parent/Caregiver rating form was completed by the 99 parents who also completed the DP-3 Parent/Caregiver Checklist. This sample consisted of both typically developing children (52) and clinically diagnosed children (47). Table 24 displays the correlations for the four Vineland II domains and the Adaptive Behavior Composite with each of the five DP-3 scales and the General Development score. The correlations between the DP-3 scales and the Vineland II domains most similar in content are bolded; in some cases these were the highest correlations, in others they were not. Correlations are moderate to high for all comparisons, ranging from .40 to .78. Similar to the study evaluating the correlations between the Interview Forms of the Vineland II and the DP-3, correlations are similar across scales and domains, pointing to the construct of general child development that appears to underlie the different areas of functioning. The relationship between these two tests lends support to the utility of the Parent/Caregiver Checklist version of the DP-3.

The *Adaptive Behavior Assessment System, Second Edition* (ABAS-II; Harrison & Oakland, 2003) is a comprehensive measure of an individual's adaptive skills and has a form that can be completed by a parent or primary caregiver. Raw scores were available for a sample of 150 typically developing children ranging in age from 3 to 12 whose parent also completed the DP-3 Parent/Caregiver Checklist. Raw scores for the five DP-3 Parent/Caregiver Checklist scales were correlated with nine skill areas of the ABAS-II, and the results are displayed in Table 25. The correlations were found to be moderate to high across scales, ranging from .45 to .87 (median = .66). This is not surprising, because although adaptive behavior is one scale of the DP-3, it is a broad construct closely related to general development.

Discriminant validity. As with the Interview Form of the DP-3, the parents of a clinical sample of children were given the Parent/Caregiver Checklist version of the DP-3 to complete. This sample consisted of 56 individuals

Table 24
Correlations Between the DP-3 Parent/Caregiver Checklist and the Vineland II

	DP-3					
Vineland II	Physical	Adaptive Behavior	Social-Emotional	Cognitive	Communication	General Development score
Communication	.44	.48	.61	.74	**.67**	.69
Daily Living Skills	.40	**.61**	.65	.69	.67	.71
Socialization	.44	.40	**.65**	.69	.57	.67
Motor Skills[a]	**.74**	.53	.53	.58	.41	.66
Adaptive Behavior Composite	.52	.56	.69	.78	.68	.78

Note. *N* = 99 for all Vineland II scales except Motor Skills. Bold type indicates expected correlation based on similar content.
[a] Motor Skills domain is administered only to children up through age 6; *n* = 60.

Table 25
Correlations Between the DP-3 Parent/Caregiver Checklist and the *Adaptive Behavior Assessment System, Second Edition* (ABAS-II)

	DP-3				
ABAS-II	Physical	Adaptive Behavior	Social-Emotional	Cognitive	Communication
Communication	.67	.67	.65	.73	.70
Community Use	.53	.60	.61	.68	.58
Functional Academics	.71	.79	.77	.87	.81
Home Living	.57	.71	.69	.69	.62
Health and Safety	.63	.68	.61	.72	.62
Leisure	.56	.57	.60	.61	.53
Self-Care	.72	.77	.70	.75	.68
Self-Direction	.61	.66	.68	.70	.61
Social	.45	.46	.49	.50	.48

Note. N = 150.

with developmental delays, mental retardation, Down syndrome, behavior/emotional disorders, and ADHD. As with the clinical sample for the Interview Form, the initial step was to compare the clinical sample of 56 individuals to the expected mean of the standardization sample (100, by definition). Table 26 illustrates that for all five scales and the General Development score, the clinical sample showed clinically meaningful differences in the expected direction. The effect sizes ranged from 1.0 to 1.6 (median = 1.3), indicating differences far above the cutoff for a large effect size. Therefore, the Parent/Caregiver Checklist version of the DP-3 is also capable of effectively distinguishing between typically developing children and those with a clinical diagnosis.

When the sample was split into different diagnostic categories of developmentally delayed (Group 1) and emotional/behavioral problems and ADHD (Group 2), some meaningful differences were observed. For example, the mean standard scores on the Physical scale (65.7 and 99.3, respectively) illustrate a statistical and clinically meaningful difference that would be expected, as children with pervasive developmental problems would likely have physical development deficits, while those with emotional and behavioral problems would not necessarily evidence such difficulties. The difference in Cognitive scale standard scores was also significant and clinically meaningful (70.4 and 88.6, respectively). All other scales indicated differences in standard scores in the expected direction; however, they were not statistically significant, in part because the analyses were limited by the small sample size (n = 31 for Group 1 and n = 25 for Group 2). Despite the sample size limitations, these results indicate that the Parent/Caregiver Checklist performs similarly to the Interview Form when used to differentiate levels of clinical impairment.

Table 26
Average Standard Scores in the DP-3 Parent/Caregiver Checklist Clinical Sample

Scale	Mean standard score
Physical	80.7* *(es = 1.3)*
Adaptive Behavior	84.7* *(es = 1.0)*
Social-Emotional	79.8* *(es = 1.3)*
Cognitive	78.5* *(es = 1.4)*
Communication	83.7* *(es = 1.1)*
General Development score	76.5* *(es = 1.6)*

Note. N = 56.

*p < .01 for a one-sample t-test comparing the obtained value with the standardization mean of 100. Numbers in parentheses are effect sizes.

Validity Evidence From Earlier Versions of the *Developmental Profile*

Analyses described thus far were conducted using the standardization and clinical samples for the current version of the *Developmental Profile,* the DP-3. However, the research base lends additional validity support to the instrument through studies using the original *Developmental Profile* and the DP-II.

Construct validity. Construct validity can refer to evidence derived from analysis of the structure of a measure, as well as from studies that determine a positive association between the measure of interest and another measure intended to assess a similar construct, and a lesser association with tests designed to measure different characteristics.

Structural characteristics. Quay and Steele (1998) explored the factor structure of the DP-II, using teachers as informants. In the process of validating another instrument, the authors ran two factor analyses with the

DP-II scales. The first used a sample of 127 prekindergarten children, and the second had a sample of 180 kindergarten children. They did a principal component extraction and found that in both analyses all scales loaded on a single factor, with factor loadings ranging from .79 to .88. A similar result was found with the DP-3, wherein all scales loaded on one factor.

Relationship between parents' estimates and other methods. A number of studies have specifically evaluated the accuracy of parent reports on the *Developmental Profile* when compared to direct assessment of the child and when compared to teacher ratings. The accuracy of parent reports of children's performance has been a focus of a number of studies using previous versions of the *Developmental Profile*. Interest in the accuracy of information provided by parents is multifaceted. Parent reports can provide an efficient, cost-effective method for collecting student information, so long as it is accurate. Additionally, information on child behaviors provided by a parent may reflect valid differences in the child's performance in different settings and may provide useful information about the parent's perception of his or her child that can aid intervention methods.

A study was conducted to examine an educational program in England designed to increase the independence skills of young children (aged 3 to 4) with cerebral palsy (Hur, 1997). Both the DP-II and the Vineland (Sparrow, Balla, & Cicchetti, 1984) teacher report were used. Although statistical analyses were not done to compare these two instruments, results of analyses revealed that teacher responses on the Vineland were similar to parent responses on the DP-II in terms of indicating improvement over time in multiple domains of development.

A study of parents' accuracy as informants was conducted with 57 children aged 36 to 72 months (M = 57.05 months) drawn from two preschools in suburban Cincinnati (Wenker, 1977). Mothers in small groups completed the *Developmental Profile,* along with six global statements about their children's developmental functioning. Teachers also provided global estimates of the children's functioning level. The mothers' estimates on the inventory were highly correlated with the teachers' global estimates; coefficients ranged from .57 to .77 on estimates of the same ability (e.g., physical) and all were significant (p < .001).

In a large-scale study of the impact of Head Start programs on handicapped children, a methodological issue of concern was the validity of parents' reports of their children's abilities (Applied Management Sciences, 1978). A pilot study was conducted to evaluate the validity of parent reports. Parents' estimates of their children's abilities were obtained in interviews using the *Developmental Profile.* A direct assessment of the children's functioning level was obtained using the *Learning Accomplishment Profile*–Diagnostic Form (Sanford, 1974). Data were collected on both measures for 126 children and the scores were correlated. The correlations, ranging from .46 to .71, demonstrated a moderate relationship between the scales on the two measures. The magnitudes of the correlations were considered adequate to justify the use of parent reports as the outcome measure in the evaluation study.

Two studies found significant differences between parents' estimates of their children's performance and other objective measures. Walker, Sieg, Quick, and Boll (1975) used parents as informants to complete the *Developmental Profile* and then assessed the children directly. Children in the study were diagnosed as having diabetes, cystic fibrosis, asthma, or hearing impairments. When parents' predictions were compared with actual child performance, only 2 parents accurately (i.e., within 6 months) estimated their children's performance on all five scales. Four parents were accurate on four scales; 12 were accurate on three scales; 9 were accurate on two scales; 7 were accurate on only one scale; and 1 parent missed on all five scales. Inaccuracies ranged from a 36-month underestimate of Physical age to a 58-month overestimate on the Communication scale. Parents were accurate 49% of the time. When their estimates were inaccurate, parents were more likely to overestimate (31%) than underestimate (20%) their children's abilities.

A second study investigating the estimates of parents of physically handicapped children obtained similar results (Tavormina, Boll, Dunn, Luscomb, & Taylor, 1977). Personality and family functioning measures were administered to 143 mothers and 107 fathers of diabetic, asthmatic, cystic fibrotic, and hearing impaired children. Estimates of the children's performance was made by both parents individually and compared to actual performance of the children on the test items. Parents were generally unrealistic in their estimates, tending to overestimate performance on the Self-Help, Communication, and Academic scales of the *Developmental Profile,* while underestimating ability on the Social scale.

A final study involved 80 preschool children involved in a home-based program and a center-based program for the handicapped (Montgomery, 1980). The purpose of the study was to examine the relationship between the mother's locus of control and child progress in a prescriptive intervention program. In addition to the main focus of the study, mothers' and teachers' assessments of the children using the *Developmental Profile* and the Marshalltown *Behavioral Developmental Profile* (Donahue et al., 1973) revealed that pretest ratings were significantly different, but posttest ratings showed no differences.

Taken as a whole, these studies suggest that while oftentimes parents provide accurate information about their children's performance, they may also at times provide over- or underestimations. Usually when the parent report differs from the teacher report or direct assessment of the child, the parent tended to overestimate his or her child's skills.

Relationship to other measures. Other studies have looked specifically at the relationship between child development, as measured by the *Developmental Profile,* and other, related areas of functioning. Wachs and De Remer (1978) investigated the relationship between cognitive functioning and adaptive behavior in young developmentally delayed children. Twenty-five infant and preschool children, ranging in age from 11 to 30 months and attending an intervention program, were assessed with a Piagetian measure of intellectual abilities, the *Infant Psychological Development Scale* (Uzgiris & Hunt, 1975), and the *Developmental Profile*. The instruments were administered in the child's home on two separate occasions within the same week by two different examiners. Canonical correlations were computed with age partialled out; results suggested that object permanence and foresight are significantly related to self-help and social skills.

In Wenker's study of 57 preschool children, described previously, the relationships between the five scales on the *Developmental Profile* and three other instruments were investigated. The *Slosson Intelligence Test* (Slosson, 1963) correlated highly (range = .68 to .77) with all five scales. The highest correlation was obtained with the Academic scale (*r* = .77) and the next highest correlation was with the Communication scale (*r* = .72), as one might expect since the Slosson is a verbal intelligence test. The *Developmental Test of Visual-Motor Integration* (VMI; Beery & Buktenica, 1967) had moderate to high correlations with the five scales (range = .52 to .68). Surprisingly, the highest correlations were with the Academic and Self-Help scales (*r* = .68 and .65, respectively), and the lowest correlation was with the Physical scale (*r* = .52). This may be due to the mix of gross- and fine-motor items on the Physical scale and to the presence of drawing tasks on the Academic scale. Finally, the *Goodenough-Harris Drawing Test* (Harris, 1963) was highly correlated with the Self-Help and Academic scales (*r* = .65 and .70, respectively). It was moderately correlated with the remaining three scales (range = .52 to .63).

Higgins, Kiefert, and Lewis (1979) compared scores on the *Developmental Profile* with two measures of intelligence. A total of 113 developmentally delayed children ranging in age from 8 months to 6 years, 9 months were administered the *Stanford-Binet Intelligence Scale* (Terman & Merrill, 1973) or the *Cattell Infant Intelligence Scale* (Cattell, 1960), depending on age. Using parents as informants, the *Developmental Profile* was also administered for each child. The mental age obtained on the intelligence scale was correlated with the average functioning age on the *Developmental Profile,* resulting in a coefficient of .85. A coefficient of .86 was obtained between the mental age on the intelligence measures and the developmental age on the Academic scale of the *Developmental Profile*. Both coefficients were significant at the .001 level, indicating a strong relationship.

Although other studies found strong relationships between the *Developmental Profile* and measures of intelligence, a study by Bloom and Zelko (1994) determined that there are differences between the constructs of development and intelligence. They engaged in a research study with 117 children who were referred due to suspected developmental delay. The study was designed to evaluate and compare measures of adaptive behavior and development with those of cognitive functioning. All five scales of the DP-II were administered to the parents, although the Self-Help and Social scales were the focus of the study. Results indicated that the participants' mean age scores on the DP-II were delayed compared to the chronological age of the children. However, there was a great deal of variability among the children, as many who were found to have mild or moderate cognitive delay by an IQ measure fell into the normal ranges on the Self-Help and Social scales of the DP-II. The authors concluded that the results support the importance of evaluating adaptive skills in addition to assessing cognitive abilities in individuals suspected of delay, as there are differences between the two constructs. Additionally, they noted that the DP-II is a measure that is useful for the "differential assessment of multiple facets of adaptive functioning in mental retardation" (p. 264).

The studies comparing the *Developmental Profile* with other measures reveal that in most cases the *Developmental Profile* scales were moderately related to similar measures. There was some inconsistency in the relationship between the *Developmental Profile* and measures of intelligence, suggesting this to be a useful area for future research.

Intervention studies. An initial purpose motivating the development of the inventory was to provide educational and mental health professionals with information for instruction. Thus, it is not surprising that one of the major areas of interest in the *Developmental Profile* has been its ability to measure change following some treatment. The following studies provide additional validity for the *Developmental Profile,* as the construct validity is supported if the measure can demonstrate the ability to assess relevant changes in development.

Hebbeler and Gerlach-Downie (2002) used the DP-II as part of a study evaluating, over the course of 3 years, the effectiveness of a home visitation program for children who were at risk for later difficulties. The sample consisted of 21 case study families, and the DP-II was used as a measure of child development. Other measures included the *Bayley Scales of Infant Development* (Bayley, 1993) and the *Peabody Picture Vocabulary Test, Revised* (Dunn & Dunn, 1981). Although the home visitors reported that the children were doing well, the assessment data revealed that many children had low scores on all three assessments. Thus, the researchers determined that their home visitation program did not have the desired effect. However, the DP-II was utilized in an additional way; depending upon the child's DP-II scores, parents were encouraged to ask for additional help. Thus it appears that the DP-II can be useful for providing parents with information regarding the development of their children, so they can then seek appropriate services.

Another study utilizing the DP-II was conducted by Sung, Kim, and Yawkey (1997), in which they used a Spanish translation of the DP-II to evaluate an intervention to promote parent involvement in children's education. The intervention was a home visitation program designed to provide culturally and linguistically diverse parents with skills to help their young children learn. The participants were 29 Puerto Rican parents with children in kindergarten, and the DP-II was administered at pre- and posttest (6 months apart) to assess the parents' understanding of their children's development. During the program, parents were taught and encouraged to communicate with their children with regard to school. At posttest, the children in both the experimental and control groups had significantly higher DP-II scores; however, the scores for the children in the experimental group were significantly higher. The authors concluded that this result represents greater understanding on the part of the parents of their children's development and learning.

Sandler et al. (2000) conducted a study of the use of a drug to address a hypothesized cause of regressive-onset autism in which it was necessary to classify the participants' developmental levels. The DP-II was used to provide developmental ages in the areas of communication, socialization, and self-help in order to compare the children's levels of functioning to their chronological ages. The results indicated that the drug provided short-term improvement, but these gains were not evident at follow-up.

The *Developmental Profile* was also used as part of an evaluation of a developmental child care program in Korea (Lee, 1993). Participants were the mothers of 32 children, aged 2 years, 4 months to 3 years. The *Developmental Profile* was given as a pre- and posttest, separated by the 9-month program. Results indicated that all of the children involved in the program improved in all five areas of development over the 9-month period, while some of the children in the comparison group appeared to have experienced some positive and negative effects. The *Developmental Profile* results helped the researchers conclude that children can be assisted in developing normally when they are cared for by professional caregivers.

Cooke, Ruskus, Apolloni, and Peck (1981) evaluated the progress of handicapped and nonhandicapped children in integrated and segregated preschool programs over a 3-year period. Developmental gains were assessed using the *Peabody Picture Vocabulary Test* (Dunn, 1965), the *Vineland Social Maturity Scale, Revised* (Doll, 1935), and the *Developmental Profile*. Data from the first-year evaluation, involving 60 children, indicated that handicapped children showed significant change only on the Academic scale of the *Developmental Profile*. Gains for this group were not related to setting; however, nonhandicapped children in segregated settings made significant gains on more measures than children in integrated settings. The second-year evaluation replicated this study with 97 children and found that handicapped children made significant gains on all dependent measures, and nonhandicapped children in the integrated setting made significant gains on the Vineland and the Physical scale. Nonhandicapped children in segregated settings had significant gains on the Social scale. The third-year evaluation included 117 preschool children and found that handicapped children made significant gains on the Vineland and the Self-Help, Social, and Academic scales of the *Developmental Profile,* regardless of setting. Nonhandicapped children in integrated settings made significant gains on all scales of the *Developmental Profile,* whereas nonhandicapped children in segregated settings made significant gains only on the Self-Help and Social scales and on the Vineland. Results of the analysis of covariance indicated that significant gains were made by the nonhandicapped group only in integrated settings. Handicapped children in the integrated setting made significant gains only on the Communication scale. The authors concluded that integration of handicapped children in preschool settings alone is inadequate; internal educational strategies designed to encourage interaction, systematic collection of student data, and structured interventions are needed as well.

Developmental Associates (1977) evaluated the effects of the Child and Family Resource Program (CFRP), a federally funded project designed to increase parents' coping skills and knowledge of parenting skills and child growth, and to foster child growth and development through site-sponsored direct service programs. Assessment data were collected from 1,058 families from 10 CFRP sites in the fall of 1976 and the spring of 1977. Three groups were compared: (a) families enrolled in CFRP prior to 1976 and with one child in Head Start in 1976; (b) families not enrolled in CFRP, but with one child in Head Start in 1976; and (c) families with a child less than 1-year-old as of October 1976. Child data included information from the *Developmental Profile* and the *McCarthy Scales of Children's Abilities* (McCarthy, 1972). Results indicated that some significant effects were obtained. Self-Help scores were significantly higher for non-CFRP children at a single site. Differences on the Social scale were reported for two sites; the non-CFRP group had the higher mean at one site and the lower mean at the other site. Overall, the results were inconclusive, suggesting no significant differences between the group receiving the CFRP program and the comparison group. This failure to find program differences may have been a result of the variety of direct services offered at the different sites.

Schortinghaus and Frohman (1974) compared the effectiveness of professionals and paraprofessionals in a home training program for 37 handicapped preschool children in a rural area. The program consisted of a weekly 1 1/2-hour demonstration lesson in the child's home and a weekly list of curriculum activities for the parent. Children were pretested in the fall and posttested 8 months later on the Academic and Communication scales of the *Developmental Profile*. Although children were assigned to instructors on a geographical basis, age and IQ data seemed to indicate that the groups were comparable. Analysis of variance was used to compare gains in months for the 21 children instructed by paraprofessionals and the 16 children instructed by professionals. Differences were not significant for Communication gains; however, children instructed by paraprofessionals made significantly greater gains on the Academic scale than did children instructed by professionals. Study results suggest that paraprofessionals can function effectively as home trainers after receiving adequate training and with weekly access to a home training specialist.

The *Developmental Profile* was intended to provide multidimensional information on a child's functioning level. The intervention studies described herein provide support for the utility of the instrument in planning and evaluating instructional programs in a variety of contexts for a variety of purposes.

Discriminant validity. Discriminant validity is evidenced when a measure can effectively categorize individuals expected to perform differently on a test. One study in this area was conducted by Factor, Perry, and Freeman (1990) to investigate family stress and child functioning, and their relationship to the utilization of respite care services. Participants were 36 families with an autistic child or a child with a diagnosis of Pervasive Developmental Disorder, Not Otherwise Specified. The DP-II, completed by the staff and the parents conjointly, was used to assess child functioning. The DP-II results indicated that the functional level of the child in the families who utilized the respite care was lower than in the families who did not use the care. Respite users' children had significantly lower scores on the Social, Communication, and Academic scales, suggesting that parents whose children have a higher level of need tend to utilize respite care to a greater extent. In this case, the DP-II was able to discriminate between individuals with disparate levels of impairment.

Another study was conducted by Pulsifer et al. (2004), who used the DP-II as one component of a large battery of tests in a study evaluating the success of hemispherectomy in 71 children who had severe seizures. Results indicated that the DP-II scores were significantly different depending upon the etiology of the seizures. The etiologies were associated with differing IQ scores, and thus the DP-II was effective at differentiating among mentally delayed individuals at different levels of impairment (e.g., mild versus severe retardation). Additionally, children with Rasmussen syndrome, who had a right-hemispherectomy, tended to score higher on both IQ and the DP-II Communication and Academic scales than the left-hemispherectomy patients.

Greenberg and Marvin (1979) examined patterns of attachment in profoundly deaf preschool children, aged 3 to 5, with hearing parents using the *Developmental Profile*. Estimates of each child's communication skills were obtained from interviews in which the mother responded to items from the Communication scale, which had been modified to include signing as verbal behavior. The interaction of mother and child was observed during instructional tasks and rated on a 7-point scale based on a subscale of the *Index of Communicative Competence* (Schlesinger & Meadow, 1972). The children were split into high- and low-competence groups using the median rating. High-communication children received significantly higher scores ($p < .02$) on the Communication scale.

The studies described here lend validity support to the ability of the *Developmental Profile* to effectively distinguish between relevant groups of children.

Criterion validity. One means of demonstrating criterion validity is to provide evidence that a test can predict a related measure at a later point in time. For example, Harris & Fagley (1987) conducted a study exploring the predictive utility of the *Developmental Profile* with a group of autistic preschoolers. Parents of 29 autistic children were administered the *Developmental Profile* interview at intake; 4 to 7 years later they were asked to complete a questionnaire regarding their child's current level of functioning. Pearson product moment correlations were conducted between the child's

Developmental Profile scale score and the mother's report of functioning in the same domain 4 to 7 years later. Results indicated that for all five developmental areas, correlations were significant and moderate, ranging from .43 to .61 (median = .54). The authors concluded that the *Developmental Profile* was useful for assessing the development of autistic preschoolers due to its ease of administration and high level of predictive ability. This study suggests that the *Developmental Profile* has some ability to predict later functioning.

Summary and Directions for Future Research

The research presented in this chapter supports the strong reliability and validity of the DP-3 for measuring children's development in five essential functional areas. The reliability evidence, as measured by test-retest and internal consistency studies, supports the use of the Interview Form; the internal consistency analyses for the Parent/Caregiver Checklist reveal that it too can function well as an instrument of development. The validity evidence illustrates that the DP-3 is an overall measure of general child development that is broken down into important skill areas for assessment, interpretation, and treatment planning. Further evaluation of the structure of the measure would be beneficial in future research studies. The validity research comparing both the Interview Form and the Parent/Caregiver Checklist to other measures of development and adaptive behavior reveal that the DP-3, while shorter in length and administration time, measures constructs similarly to other established psychological tests. Additionally, it can discriminate effectively not only between developmentally delayed and typically developing children, but also between children with different levels of clinical impairment. Furthermore, the base of research utilizing earlier versions of the *Developmental Profile* lends support to the strength of this instrument over time and in a variety of contexts.

Although a great deal of research already exists, establishing validity is an ongoing process, and therefore further research using the DP-3 with different populations, in different contexts, and for different purposes will contribute well to the research base. Additionally, the Parent/Caregiver Checklist is a new component of the *Developmental Profile,* and while initial research supports its use, additional studies will further illustrate the ways in which it can be useful. In particular, research exploring the use of the Parent/Caregiver Checklist with different demographic groups would be informative.

Another area for suggested future research involves the evaluation of the Interview Form as a means of gathering clinical information about the child, the parent/caregiver, and family interactions. Such clinical information can often be gained from using an interview format (and is one reason why this is the primary method of administration for the DP-3); however, it would be useful for research studies to formally address this issue.