# 4

# Development and Standardization

This chapter describes the development of the DP-4, beginning with a discussion of its theoretical background and earlier versions. The chapter then details the research studies that were conducted and samples that were collected to standardize and validate the DP-4 forms. It ends with a discussion of the methods used to derive the DP-4 standard scores, age equivalencies, growth scores, Start and Stop rules, and values for determining significant differences in scores.



## Earlier Versions of the Developmental Profile

The first psychometric measures of child functioning evolved from the pioneering work of Alfred Binet, who introduced the concept of *mental age*. Binet's central procedure involved determining age norms for a collection of increasingly difficult academic tasks and then assessing the ability of children to accomplish these tasks. 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 (Alpern & Boll, 1972) incorporated Binet's age norming of items and Doll's interview techniques into a multidimensional assessment of children's functioning. The assessment of five separate areas of development (physical, adaptive behavior, social–emotional, cognitive, and communication) became a standard of practice, as well as a requirement for child evaluation for many federal, state, and local government agencies. Development of the items was based on what was known at the time in terms of the literature, other measures reflecting the same five areas, and clinical observations of age-related developmental competence. Items were designed to reflect observable behaviors, to be understood by parents as well as specialists in a variety of disciplines, and to be administered in a relatively short time period. The original version contained 318 items grouped into skill areas and approximate age levels. Items were designed so they reflected an appropriate age progression; possessed a high degree of age discrimination; were accurately responded to by parents; and did not discriminate against children by gender, ethnicity, or socioeconomic status.

The Developmental Profile II (DP-II; Alpern et al., 1980), represented a refinement of the original 1972 inventory and was a widely used and well-received instrument. Outdated items were deleted or modified (e.g., items were rewritten if they appeared to embody outdated or ambiguous references to gender or ethnicity). In addition, items were dropped if they referred to developmental milestones normally achieved after 9 years, 6 months of age. As a result of these changes, the length of the inventory was reduced to 186 items. 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 standardization sample was used to derive cutoff points for referral, age-equivalent scores, and percentages of children at different ages who passed each item. The Developmental Profile II (DP-II; Alpern et al., 1986) was again updated with the addition of a computer scoring program.

The Developmental Profile 3 (DP-3; Alpern, 2007) represented the first comprehensive revision of the original instrument. The DP-3 retained the strengths of the DP-II while adding a representative normative sample, updated item content, updated scale names, updated scoring options, new items, modern statistical scaling techniques, suggestions for interventions, and expanded computer scoring and interpretation. Additionally, the DP-3 included smaller increments of age stratification to capture the rapid developmental growth that occurs at younger ages. Norm-referenced standard scores were provided over an expanded age range of birth through 12 years, 11 months. Item content was updated by deleting outdated items (e.g., "Can the child strike and light a paper match?") and adding items related to technology use (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?"). The DP-3 also included remediation activities for each of the items.

# Development of the DP-4

The DP-4 is intended to improve on previous revisions by adding a new normative sample, an expanded age range (up to 21 years, 11 months), two new forms (the Teacher Checklist and Clinician Rating forms), updated item content, a new scoring option (growth scores), and an additional option for interpretation (rater comparisons). It was developed through the research studies and sample collection processes described in this section.

## User Survey

An initial step in the development of the DP-4 was to gauge the experience of users of the DP-3. A survey was sent to users, which included questions about settings and applications in which the DP-3 is used, ages and clinical problems of clients being assessed, items needing revision, and potential new content areas (e.g., technology use). Almost a third of respondents were school psychologists, followed by smaller numbers of occupational therapists, speech and language pathologists, and those in other clinical, medical, and educational fields (e.g., licensed psychologists, pediatricians, educational diagnosticians). More than half of the respondents reported 10 or more years of experience working in clinical or educational settings.

Overall, results from the user study showed that the DP-3 successfully met most needs of clinicians. The feedback from the survey led to the improvements mentioned above: the creation of the Teacher Checklist and Clinician Rating forms, the expansion of the age range, and the generation of new items to address specific suggestions from the survey respondents. The new items included items representing very easy as well as more difficult developmental tasks, which improve precision of measurement at the lower and upper bounds of the DP-4 age range. Some retained items from the DP-3 were reworded to be culturally sensitive, gender neutral, and inclusive of children who are deaf and hard of hearing.

## Pilot Study

A pilot study was conducted to evaluate the psychometric characteristics of the new items alongside those retained from the DP-3. Among the 307 items tested in the pilot study, 80 were

brand new and 35 others were significantly reworded DP-3 items. Data were collected via an online data collection platform, and a total of 348 Parent/Caregiver Checklist forms, with a subset of 176 Teacher Checklist forms, were completed.

An additional feature of the pilot study was the inclusion of items that asked respondents to indicate the degree of their confidence in their answers. After responding to each item, teachers and parents selected one of the following choices: "I'm definite about my answer," "I'm unsure about my answer," and "I completely guessed on my answer." This was done to inform the development of the new Teacher Checklist form in order to identify items on which teachers might express low confidence that they could observe the skill or behavior under question. Such items were eventually excluded from the Teacher Checklist form.

The pilot sample included 348 typically developing individuals, across the age range of 6 years, 6 months to 21 years, 11 months. The sample was 22% Black, 16% Hispanic, 55% White, 1% Asian, and 6% Other. Females comprised 54% of the sample, while 46% were male, and 25% had parents who had not attended any college.

Data from the pilot study were analyzed to determine the item set that would be standardized and to order those items by difficulty on the standardization forms. Analytic techniques included examination of item score and preliminary scale score distributions, and item response theory (IRT) analysis to evaluate item difficulty,[1] bias, and fit to the measurement model.

The IRT approach most suited to the data was the Rasch one-parameter model (Bond & Fox, 2001; Wright & Stone, 1979). Data were analyzed using the software program jMetrik (Meyer, 2014). The Rasch model's 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 difficulties are on the same scale (accounting for measurement error) regardless of the sample of individuals used to generate the difficulties. The jMetrik program yields a logit scale of item difficulty and person ability, with a mean of 0 and a standard deviation of 1. Because this scale is a true interval scale, a 1-logit difference between scores has identical meaning regardless of whether the score pair occurs near the center or on 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 chances of success decrease, and when the item difficulty is lower than the person's ability, the chances of success increase.

Items that fail to demonstrate the predicted relationship between difficulty and person ability, in terms of probability of success, are said to have poor fit to the Rasch measurement model. Specifically, the parameter of *infit* refers to item fit evaluated with respect to persons with a similar Rasch measure to the item under study (i.e., the difference between the person ability measure and the item difficulty measure is relatively small). Items with poor infit were dropped from the DP-4.

The Rasch item difficulty measures were used to order items by difficulty on the standardization forms, and to ensure sufficient, nonredundant item coverage over the DP-4's intended ranges of age and development. Under the Rasch model, the most precise measurement occurs when an item has the same (or numerically similar) measure of difficulty as the person's ability measure, 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 abilities in the target population (i.e., it measures well at the extremes of the person distribution). Within the scale, items must spread uniformly enough to provide reasonably precise measurement for all ability levels (i.e., the scales measure well in the center of the person distribution).

Finally, potential bias across key demographic variables was investigated using the differential item functioning (DIF) methodology. DIF can determine whether individuals who have similar person ability

[1] For the DP-4, the concept of *item difficulty* refers to a child's ability to perform a skill measured by a DP-4 item. The child's ability (or lack thereof) causes the respondent to choose *Yes* or *No* as a response to the item. In this way, item difficulty on the DP-4 is an index of how *difficult* it is for a respondent to choose *Yes* and thus confirm that the child possesses the skill in question.

but belong to different demographic groups (e.g., those defined by gender, ethnicity, and socioeconomic status) differ unexpectedly in their performance on certain items. Such a result suggests that those items are biased by demographic status. Any item that showed evidence of significant bias via DIF was dropped from the DP-4.

These pilot study analyses yielded a set of 249 items to be tested in the standardization study. These items were implemented in the Parent/Caregiver Interview, Parent/Caregiver Checklist, Teacher Checklist, and Clinician Rating forms. The wording on the Parent/Caregiver Interview, the Parent/Caregiver Checklist, and Clinician Rating forms was identical (with the exception of replacing "the child" with "your child" on the Parent/Caregiver Checklist form). For the Teacher Checklist form, some items that did not apply to the classroom setting were removed, while others were reworded to better fit this setting.

## Standardization and Validation Data Collection

The standardization study, which included research on the clinical validity of the DP-4, comprised data from more than 3,000 administrations of the four forms. The purpose of this study was threefold:

1. Collect a nationally representative sample of typically developing children to use for developing the DP-4 norms.
2. Collect a clinical sample of children who have clinical diagnoses and who were receiving special services. This sample was used in the DP-4 validity studies described in Chapter 5. Sixty percent of these cases were also used in the standardization sample.
3. Apply the analytic procedures described previously in this chapter to reduce the 249-item standardization set to a final, optimized item set suitable for publication.

Data were collected by 65 data collectors in 28 states across all four U.S. Census regions. Online forms administered via the online platform were the primary means of data collection, though paper forms were used when necessary. For the Parent/Caregiver Interview form, the data collector conducted the interview in person and then entered the responses on the form.

The standardization and clinical samples were obtained by recruiting data collectors throughout the United States who had access to families of individuals aged birth through 21 years, 11 months. For the purposes of this study, typically developing children were defined as those who did not have a diagnosed moderate to severe disability. Children with mild disabilities who spent most of their day in a general education classroom were also included in the typically developing sample. Clinical cases were defined as children who had a moderate to severe diagnosis and spent most (i.e., more than 50%) of their time in a special education setting. To include children for whom the DP-4 will most likely be administered, clinical cases were included in the normative sample based on their age and primary diagnosis.

### *Standardization Sample*

The Parent/Caregiver Interview standardization sample consisted of 2,259 cases, of whom 2,051 were considered typically developing and 208 had clinical diagnoses.

Most interviews were administered in English; however, 55 of the typically developing Parent/Caregiver Interview cases were from primarily Spanish-speaking families. The Spanish version of the form was used for these cases (see discussion below about the development of the Spanish DP-4 forms). All 55 cases were matched with English-speakers in the sample based on age, gender, region, and ethnicity. Paired-sample t-tests showed no significant differences between the two groups on their scale standard scores and the General Development Score, and no effect size was larger than 0.21, indicating small differences in magnitude between the groups.

The majority (74%) of interviews were conducted with the subjects' mothers, while the remaining interviews were conducted with fathers (8%) and other relatives (18%). Table 4.1 presents the demographic characteristics of the standardization sample, along with corresponding percentages from the U.S. Census (U.S. Bureau of the Census, 2012). Table 4.2 provides the age groups that were used for stratification in the norming process. As shown in the table, the sample's distribution is similar to that of the U.S. population across all demographic areas. Proportions of all demographic categories represented were within approximately 5% of the U.S. population at the time data were collected, consistent with a guideline suggested by Andersson (2005).

**Table 4.1.** Demographic Characteristics of the DP-4 Standardization Sample: Parent/Caregiver Interview Form

| Characteristic | *n* | % of sample | U.S. Census %[a] |
|---|---|---|---|
| **Gender** | | | |
| Male | 1,165 | 51.6 | 49.2 |
| Female | 1,094 | 48.4 | 50.8 |
| **Race/Ethnicity** | | | |
| Asian | 62 | 2.7 | 4.6 |
| Black/African American | 347 | 15.4 | 13.9 |
| Hispanic Origin | 624 | 27.6 | 23.5 |
| Native Hawaiian/Pacific Islander | 10 | 0.4 | — |
| American Indian/Alaska Native | 7 | 0.3 | — |
| White | 1,057 | 46.8 | 52.7 |
| Other | 152 | 6.7 | 5.3 |
| **Parents' educational level** | | | |
| No high school diploma | 199 | 8.8 | 11.9 |
| High school graduate | 573 | 25.4 | 26.4 |
| Some college | 682 | 30.2 | 30.5 |
| Bachelor's degree or higher | 805 | 35.6 | 31.2 |
| **U.S. geographic region** | | | |
| Northeast | 378 | 16.7 | 17.7 |
| South | 876 | 38.8 | 37.5 |
| Midwest | 537 | 23.8 | 21.3 |
| West | 468 | 20.7 | 23.5 |

*Note.* $N$ = 2,259. Due to missing data, not all totals will sum to $N$. Due to rounding, total percentages may not equal 100.0%.
[a]U.S. Census Bureau (2012). Race/Ethnicity based on ages 0–21; parents' educational level based on ages 25–64 (those most likely to have children within the DP-4 age range); gender and region based on the general population.

**Table 4.2.** Age Ranges of the DP-4 Standardization Sample: Parent/Caregiver Interview Form

| Characteristic | *n* | % of sample |
|---|---|---|
| **Age range** | | |
| 0:0–0:1 | 26 | 1.2 |
| 0:2–0:3 | 26 | 1.2 |
| 0:4–0:5 | 42 | 1.9 |
| 0:6–0:7 | 40 | 1.8 |
| 0:8–0:9 | 35 | 1.5 |
| 0:10–0:11 | 42 | 1.9 |
| 1:0–1:1 | 37 | 1.6 |
| 1:2–1:3 | 43 | 1.9 |
| 1:4–1:5 | 33 | 1.5 |
| 1:6–1:7 | 34 | 1.5 |
| 1:8–1:9 | 38 | 1.7 |
| 1:10–1:11 | 30 | 1.3 |
| 2:0–2:3 | 71 | 3.1 |
| 2:4–2:7 | 64 | 2.8 |
| 2:8–2:11 | 59 | 2.6 |
| 3:0–3:5 | 84 | 3.7 |
| 3:6–3:11 | 93 | 4.1 |
| 4:0–4:5 | 92 | 4.1 |
| 4:6–4:11 | 90 | 4.0 |
| 5:0–5:5 | 107 | 4.7 |
| 5:6–5:11 | 93 | 4.1 |
| 6:0–6:5 | 105 | 4.6 |
| 6:6–6:11 | 89 | 3.9 |
| 7:0–7:11 | 122 | 5.4 |
| 8:0–8:11 | 121 | 5.4 |
| 9:0–9:11 | 109 | 4.8 |
| 10:0–10:11 | 104 | 4.6 |
| 11:0–12:11 | 196 | 8.7 |
| 13:0–16:11 | 126 | 5.6 |
| 17:0–21:11 | 108 | 4.8 |

*Note.* $N = 2{,}259$. Due to missing data, not all totals will sum to $N$. Due to rounding, total percentages may not equal 100.0%.

Data for the Parent/Caregiver Checklist and Teacher Checklist forms were collected on subsets of this sample. The Parent/Caregiver Checklist sample included 543 cases (see Table 4.3), of which 97 were clinical. This sample was used to link with the Parent/Caregiver Interview sample to create norms (see Derivation of the DP-4 Scores section later in this chapter for details on this process). The Teacher Checklist sample, summarized in Table 4.4, included 1,437 cases, 173 of which were clinical. The proportions of gender, ethnicity, and parent education demographic variables for each of these samples resembled those found in the larger Parent/Caregiver Interview sample. Cases varied in terms of their geographic region, though all four regions were adequately accounted for. Age groups were adequately represented beyond the age of 2 years (the starting age group for the Teacher Checklist form). Again, since the Parent/Caregiver Checklist sample was rather small, the linking process to the larger Parent/Caregiver Interview sample compensated for the small counts of cases in the younger ages.

**Table 4.3.** Demographic Characteristics of the DP-4 Standardization Sample: Parent/Caregiver Checklist Form

| Characteristic | *n* | % of sample |
|---|---|---|
| **Gender** | | |
| Male | 279 | 51.4 |
| Female | 264 | 48.6 |
| **Race/Ethnicity** | | |
| Asian | 19 | 3.5 |
| Black/African American | 124 | 22.8 |
| Hispanic Origin | 118 | 21.7 |
| Native Hawaiian/Pacific Islander | 1 | 0.2 |
| American Indian/Alaska Native | 2 | 0.4 |
| White | 213 | 39.2 |
| Other | 66 | 12.2 |
| **Parents' educational level** | | |
| No high school diploma | 27 | 5.0 |
| High school graduate | 107 | 19.7 |
| Some college | 117 | 21.5 |
| Bachelor's degree or higher | 195 | 35.9 |
| **U.S. geographic region** | | |
| Northeast | 59 | 10.9 |
| South | 241 | 44.4 |
| Midwest | 211 | 38.9 |
| West | 32 | 5.9 |

*Note.* $N = 543$. Due to missing data, not all totals will sum to $N$. Due to rounding, total percentages may not equal 100.0%.

**Table 4.4.** Demographic Characteristics of the DP-4 Standardization Sample: Teacher Checklist Form

| Characteristic | *n* | % of sample |
|---|---|---|
| **Gender** | | |
| Male | 734 | 51.1 |
| Female | 703 | 48.9 |
| **Race/Ethnicity** | | |
| Asian | 33 | 2.3 |
| Black/African American | 289 | 20.1 |
| Hispanic Origin | 368 | 25.6 |
| Native Hawaiian/Pacific Islander | 3 | 0.2 |
| American Indian/Alaska Native | 3 | 0.2 |
| White | 648 | 45.1 |
| Other | 93 | 6.5 |
| **Respondents' educational level** | | |
| No high school diploma | 160 | 11.1 |
| High school graduate | 412 | 28.7 |
| Some college | 402 | 28.0 |
| Bachelor's degree or higher | 434 | 30.2 |
| **U.S. geographic region** | | |
| Northeast | 254 | 17.7 |
| South | 545 | 37.9 |
| Midwest | 408 | 28.4 |
| West | 230 | 16.0 |

*Note.* $N$ = 1,437. Due to missing data, not all totals will sum to $N$. Due to rounding, total percentages may not equal 100.0%.

### *Clinical Sample*

The clinical sample consisted of 348 children diagnosed with a behavioral, emotional, developmental, or other disorder severe enough to warrant referral for services. Diagnoses included developmental disorder, intellectual disability, autism, attention-deficit/hyperactivity disorder (ADHD), hearing impairment, learning disability, mood disorder, speech/language impairment, visual impairment, physical disability, and a general "other" category. A total of 348 Parent/Caregiver Interview, 293 Teacher Checklist, 179 Parent/Caregiver Checklist, and 276 Clinician Rating forms were collected for the clinical sample. The majority (85%) of interviews were conducted with the subjects' mothers, and the rest with fathers and other relatives. The sample ranged in age from 1 year, 4 months to 21 years, 11 months and was diverse in terms of 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 (Boyle et al., 2011). Table 4.5 presents the demographic characteristics of the clinical sample, including percentages of primary diagnoses. Although the clinical sample was not expected to replicate the U.S. Census demographic distribution due to the inclusion criteria, the sample does offer some diversity in terms of the demographic variables described here.

The clinical sample was used for several purposes. Selected cases were included in the normative sample (as described above). All cases were used for the validity studies described in Chapter 5, and those cases with Clinician Rating forms were used to determine the growth scores for that form. Unlike the other three forms (Parent/Caregiver Interview, Parent/Caregiver Checklist, and Teacher Checklist), it was not feasible to collect Clinician Rating data on typically developing individuals, and therefore it was not possible to create normative referenced scores for the Clinician Rating form. This is because the clinician must be sufficiently familiar with the child in order to complete the form, and clinicians would not know typically developing children well enough to do so.

**Table 4.5.** Demographic Characteristics of the DP-4 Clinical Sample

| Characteristic | *n* | % of sample |
|---|---|---|
| **Gender** | | |
| Male | 232 | 66.7 |
| Female | 116 | 33.3 |
| **Race/Ethnicity** | | |
| Asian | 9 | 2.6 |
| Black/African American | 78 | 22.4 |
| Hispanic Origin | 69 | 19.8 |
| Native Hawaiian/Pacific Islander | 4 | 1.1 |
| American Indian/Alaska Native | 2 | 0.6 |
| White | 165 | 47.4 |
| Other | 21 | 6.0 |
| **Parents' educational level** | | |
| No high school diploma | 29 | 8.3 |
| High school graduate | 88 | 25.3 |
| Some college | 90 | 25.9 |
| Bachelor's degree or higher | 141 | 40.5 |

*Note.* $N = 348$. Due to missing data, not all totals will sum to $N$. Due to rounding, total percentages may not equal 100.0%.

*Table 4.5 continued on next page*

**Table 4.5.** Demographic Characteristics of the DP-4 Clinical Sample *(continued)*

| Characteristic | *n* | % of sample |
|---|---|---|
| **U.S. geographic region** | | |
| Northeast | 69 | 19.8 |
| South | 141 | 40.5 |
| Midwest | 91 | 26.1 |
| West | 47 | 13.5 |
| **Age range** | | |
| 2:0–2:3 | 5 | 1.4 |
| 2:4–2:7 | 6 | 1.7 |
| 2:8–2:11 | 3 | 0.9 |
| 3:0–3:5 | 18 | 5.2 |
| 3:6–3:11 | 23 | 6.6 |
| 4:0–4:5 | 24 | 6.9 |
| 4:6–4:11 | 22 | 6.3 |
| 5:0–5:5 | 18 | 5.2 |
| 5:6–5:11 | 22 | 6.3 |
| 6:0–6:5 | 17 | 4.9 |
| 6:6–6:11 | 17 | 4.9 |
| 7:0–7:11 | 30 | 8.6 |
| 8:0–8:11 | 17 | 4.9 |
| 9:0–9:11 | 15 | 4.3 |
| 10:0–10:11 | 18 | 5.2 |
| 11:0–11:11 | 11 | 3.2 |
| 12:0–12:11 | 9 | 2.6 |
| 13:0–14:11 | 37 | 10.6 |
| 15:0–16:11 | 10 | 2.9 |
| 17:0–18:11 | 21 | 6.0 |
| 19:0–21:11 | 5 | 1.4 |
| **Primary diagnosis**[a] | | |
| Attention-deficit/hyperactivity disorder | 19 | 5.5 |
| Autism | 80 | 23.0 |
| Developmental disorder | 64 | 18.4 |
| Hearing impairment | 42 | 12.1 |
| Intellectual disability | 40 | 11.5 |
| Learning disability | 7 | 2.0 |
| Mood disorder | 23 | 6.6 |
| Other | 3 | 0.9 |
| Physical disability | 46 | 13.2 |
| Speech/Language impairment | 8 | 2.3 |
| Visual impairment | 16 | 4.6 |

*Note. N* = 348. Due to missing data, not all totals will sum to *N*. Due to rounding, total percentages may not equal 100.0%.
[a]Many cases included comorbid diagnoses.

## Final Item Selection

A total of 190 items were retained for the Parent/Caregiver Interview, Parent/Caregiver Checklist, and Clinician Rating forms for publication. Only 180 items were retained for the Teacher Checklist, due to the deletion of 10 items that reflected skills usually not observable in the classroom. Using Rasch methods, the final item set for the Parent/Caregiver Interview form was ordered by difficulty, and the same item ordering was then used on the other three forms.

It was determined that some items could appear, with equal empirical justification, on either the Cognitive Scale or the Communication Scale. Rational content analysis was used to determine the final scale assignments for these items. Statistical evaluation of these assignments was also conducted to ensure these items fit their assigned scales. Items that referred to the activities of reading, writing, or technology use, and which did not also involve direct communication with another person, were assigned to the Cognitive Scale. An example would be items in which performance would not be affected by the physical presence of another person (e.g., typing on a computer, recognizing a printed name). Conversely, items that do involve directly expressing or receiving a message from another person were assigned to the Communication Scale.

In total, the resulting 190 items of the DP-4 comprise an item pool that is 57% original DP-3 items (many with minor wording changes), 10% revised items, and 33% new items. Correlations between the DP-3 and DP-4 ranged from .80 to .93 across all five scales and the General Development Score, supporting a strong relationship between the two versions (see Table 5.15 in Chapter 5).

## Spanish-Language Versions

Once the DP-4 item set and form instructions were finalized for publication, Spanish-language versions of each form were developed. The items and instructions were translated into Spanish by a clinical psychologist with extensive translation experience and expertise in test development. These versions were then independently back-translated into English, and the back-translations were reviewed by the author and publishing staff. The author provided feedback to the translator, who adjusted the translations accordingly to produce the finalized Spanish forms. This iterative process was designed to result in Spanish translations that would be understood by the widest range of Spanish speakers.

# Derivation of the DP-4 Scores

## Derivation of Standard Scores

Raw-to-standard-score-conversion tables were created for three forms: the Parent/Caregiver Interview, Parent/Caregiver Checklist, and Teacher Checklist.

**Parent/Caregiver Interview and Teacher Checklist** Examination of the raw score distributions for the five domain scale scores on the Parent/Caregiver Interview form revealed the need for finely grained age stratification, especially at the youngest ages. For this reason, there are 30 normative age groups (see Table 1.1 in Chapter 1) on the DP-4 Parent/Caregiver Interview form (as well as the Parent/Caregiver Checklist form). For the Teacher Checklist form, the age starts at 2 years; thus there are 18 normative groups.

To construct the normative age groups, raw score means and standard deviations were examined to determine whether similarities existed with the age groups from the DP-3, and to determine an optimal age-stratification scheme. As expected based on the typical progression seen in development, raw scores on each scale increased most rapidly at the youngest ages and then continued to increase through the elementary school years, though less steeply. The original distribution of DP-4 raw scores underwent a nonlinear transformation within each age group so that it would approximately fit a normal curve. The estimated smoothing curves of the DP-4 standard scores conformed to simple growth curve expectations; that is, third-order polynomials. The normalized raw scores were converted to standard scores, which have a mean of 100 and a standard deviation of 15. Interpolation was used to establish consistent data points along the developmental curve. Some manual hand-smoothing was required at the extremes of the standard score distributions to ensure the expected progression of scores when a child transitions from one age group to the next.

The General Development Score is a standard score 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 normative groups to determine if age-stratification was necessary for the General Development Score. Although some minor variability in means and standard deviations was evident across age groups, it was not large enough to justify age-stratification of the General Development Score norms.

**Parent/Caregiver Checklist** Since the Parent/Caregiver Checklist contains the same item content as the Parent/Caregiver Interview, a smaller standardization subsample was collected for the former and the larger Parent/Caregiver Interview sample was used to generate separate norms. This was achieved by Rasch techniques, which were used to draw on the item and person calibrations from the larger sample (standardized on 2,259 cases) in order to generate the separate age-stratified norms for the Parent/Caregiver Checklist form.

The first step was to evaluate equivalence between the item parameters of the Parent/Caregiver Interview and Parent/Caregiver Checklist forms for the 536 individuals who were administered both forms. This process was necessary to determine if the item parameters on the two forms were similar enough for the larger Parent/Caregiver Interview sample to be used to generate norms for the smaller Parent/Caregiver Checklist sample. Utilizing the *robust z method* provided in jMetrik, differences were found to be nonsignificant and often lower than the standard error of measurement of each item. This process verified that the calibrations from the larger sample were appropriate baseline measures for generating age-stratified norms in the smaller sample.

The next step was to determine if the norms from the larger sample could be applied to the Parent/Caregiver Checklist as is, or if new norms were needed. This step involved linking the Parent/Caregiver Interview form data to the Parent/Caregiver Checklist form data. Items in each scale on both forms were run through jMetrik concurrently to yield new item difficulty and person ability parameters. Small differences were found between the new parameters, thus requiring the statistical process of true score equating to derive separate Parent/Caregiver

Checklist norms. It is worth noting that the resulting norms for two of the Parent/Caregiver Checklist scales (Physical and Communication), are identical to those on the Parent/Caregiver Interview. However, they are presented in separate tables alongside the other scales for consistency. The General Development Score for the Parent/Caregiver Checklist was derived by the same method as the General Development Score for the Parent/Caregiver Interview.

## Derivation of Age-Equivalent Scores

Age-equivalent scores are provided for the DP-4 Parent/Caregiver Interview (Appendix Table A.3), Parent/Caregiver Checklist (Appendix Table B.3), and Teacher Checklist (Appendix Table C.3) forms.

The age equivalent represents the age at which a particular raw score is the average score. Age equivalencies for the DP-4 were derived by determining the raw score that corresponded to a standard score of 100 (or closest to 100) for each age group. This process was repeated for each of the three forms mentioned above.

## Derivation of Growth Scores

The process of deriving growth scores applies to all four DP-4 forms: Parent/Caregiver Interview, Parent/Caregiver Checklist, Teacher Checklist, and Clinician Rating. Thus, each form has its own set of growth scores.

As part of the Rasch analysis described previously in this chapter, each case received a Rasch ability score for each form. This ability score represents a child's performance on all items on the given form, and thus is an index of overall development as measured by each DP-4 scale. The Rasch person ability score was transformed from its native logit scale to the DP-4 growth score, which has a mean of 500 and a standard deviation of 25.

Unlike the standard score, the growth score is not a norm-referenced score. That is, the growth score does not represent a direct comparison of the child to the performance of typically developing, same-age peers. Rather, it is a score that reflects the child's own ability (in this sense, it is analogous to a raw score).

The growth score is an absolute measure of ability, in the same sense that a yardstick is an absolute measure of length in the physical realm. A child's growth score remains constant regardless of whether they are being compared to typically developing children or those with special needs. This property of "sample-free" measurement is a hallmark of the Rasch methodology. The growth score is better suited than the standard score for evaluating change over time, because it contains no variance introduced by the statistical reference to a same-age peer group.

Chapter 2 references the tables in the appendix (Tables A.4, B.4, C.4, and D.1) that can be used to convert a child's total raw score on each scale into a growth score for each of the four forms. These tables are derived from the Rasch difficulties of the items using the Newton–Raphson procedure (Wright & Stone, 1979). The appendix tables provide a growth score for the "perfect" total raw scores of 0 (failure on all items) and a maximum score for each scale (success on all items). By definition, "perfect" performance embodies no variance, and thus corresponding person-abilities cannot be calculated directly using the Rasch model. The tabled values for the "perfect" scores are thus extrapolated estimates. The differences between the growth scores for the nearest pair of adjacent total raw scores were used to extrapolate tabled values for the "perfect" total raw scores of 0 and the maximum. For example, the growth score for a raw score of 0 is calculated by subtracting the difference between growth scores for 1 and 2 from the growth score for 1. The growth scores for each maximum score were derived in an analogous manner.

Growth scores are available for all four DP-4 forms, but they are the only score provided by the Clinician Rating form. Growth scores on the DP-4 are intended for tracking progress over time, comparing the Clinician Rating form to the other three forms (i.e., rater comparisons), and determining differences between scales on the Clinician Rating form (see Chapter 2).

# Establishing the Start and Stop Rules

The Parent/Caregiver Interview form is the only DP-4 form that makes use of *Start and Stop rules*. For the other three forms, all items are completed by all respondents. The only exception is with the Clinician Rating form, to which the clinician may choose to apply Start and Stop rules described here, at their own discretion.

On measures such as the DP-4, in which items are ordered in terms of difficulty, Start and Stop rules enable an administration procedure that is efficient and yields precise measurement. The Start rule is applied by beginning the administration at the item (the *start item*) that represents typical development for children of the same age as the child being assessed. This allows the clinician to skip easier items on the DP-4, thus shortening administration time. This method operates on the assumption that these easier items would have been scored *Yes*, had they been administered.

To test this assumption, the clinician applies the Start rule, which requires that a certain number of consecutive *Yes* responses are achieved, starting with the start item and working upward. When this rule is satisfied, the clinician can continue administering more difficult items and assign *Yes* scores to all items below the start item. If the examiner encounters a *No* response before the Start rule is satisfied, they then administer items downward from the start item until achieving a streak of consecutive *Yes* responses that satisfy the rule.

Analogously, the Stop rule operates on the assumption that testing can be discontinued after a certain number of consecutive *No* responses, because all items more difficult than the last *No* item in that streak would have been scored *No*, had they been administered. Applying the Stop rule permits the examiner to skip more difficult items, further reducing administration time.

A goal in developing the DP-4 was to determine if the same Start and Stop rules from the DP-3 could be carried forward to the DP-4, to preserve continuity of administration procedures between the two versions. The DP-3 start item is determined by the child's assignment to one of four age ranges: 0:0–1:11, 2:0–3:11, 4:0–5:11, and 6:0 and older. The DP-3 Start rule is five consecutive items scored *Yes*; the Stop rule is five consecutive items scored *No*.

To determine the viability of these rules for the DP-4, the Parent/Caregiver Interview standardization sample was grouped into the four DP-3 Start rule age ranges. The distributions of streaks of five consecutive *Yes* responses were examined across these four age groups. For each case, the "highest" such streak (the one comprising the most difficult items) was located, and all items less difficult than this streak were rescored to *Yes*. The distributions of streaks of five consecutive *No* responses were also treated in a similar way. For each case, the "lowest" such streak (the one comprising the least difficult items) was located and all items more difficult than this streak were rescored to *No*.

The rescored item responses were summed (*Yes* = 1, *No* = 0) to yield new raw scores, which were compared to the original raw scores (prior to rescoring based on Start/Stop rules). These procedures were repeated for the five DP-4 scale scores. Across all age groups and DP-4 scales, the rescored and original raw scores correlated at $r \geq .97$. In addition, within each scale and age group, average mean differences between rescored and original scores were ≤1.15 raw score points.

These analyses suggest a level of similarity between the rescored and original raw scores. For this reason, it was determined that the DP-3 age-based start items and Start/Stop rules could be carried forward to the DP-4 without compromising precision of measurement.

# Determining the Significance of Score Differences

Critical values serve as thresholds for interpreting whether score differences are statistically significant. For each set of comparisons between DP-4 scales, administrations, and raters, critical values were determined by the following method described by Anastasi and Urbina (1997). The formula calculates a standard error of differences that sets a criterion for determining the significance of a difference score:

$$SE_{diff} = \sqrt{(SEM_1)^2 + (SEM_2)^2}$$

where $SEM_1$ and $SEM_2$ are the standard errors of measurement for score 1 and score 2, respectively. Once calculated, the $SE_{diff}$ is then multiplied by the desired level of significance (1.96 for $p < .05$; 2.58 for $p < .01$) to obtain the cutoff number, or *critical value*, associated with a significant difference. Statistical significance was set at .05, which indicates that the probability of a score difference (or any differences greater than those listed in the tables) occurring by chance is less than 5%. Critical values are provided for different age groups.

In addition to critical values for determining statistical significance between scores, base rates help to determine whether these differences are clinically meaningful. Base rates are available for most scale comparisons and rater comparisons, and were derived by determining the frequencies of how often these differences occurred in the standardization sample. There are instances where the score differences are lower, not higher, than the critical value. These lower differences resulted from low variability in the sample, and thus do not meet the threshold for statistically significant differences. Frequencies are provided for score differences that occurred 25%, 20%, 10%, 5%, and 1% of the time across all age groups.