5

Technical Properties

This chapter presents research supporting the reliability and validity of the DP-4. The analyses discussed in this chapter are based on the standardization and clinical samples described in Chapter 4.

Reliability

The reliability of a test score refers to the extent to which the score is consistent and relatively free from error. That is, a child 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.

This section describes five approaches in which the reliability of the DP-4 was estimated: internal consistency, test–retest reliability, interrater reliability (two respondents rating the same child using the same form), cross-form consistency (two respondents rating the same child using two different forms), and alternate-form consistency (the same respondent rating the same child on two different forms).

Internal Consistency

Internal consistency refers to the extent to which all items in a test or scale consistently measure the same ability or trait. Internal consistency can be estimated in several ways, but the method most frequently used for a test with a developmental gradient is the *split-half method* (Cronbach, 1970). In this procedure, items on each scale are separated into halves by alternating consecutive items. The resulting correlation from these two halves is adjusted using the Spearman-Brown formula (Anastasi & Urbina, 1997).

Internal consistency reliabilities were calculated using raw scores for each of the five DP-4 scales. Alternatively, the internal consistency reliability for the General Development Score (which is a combination of standard scores of the five scales) was estimated using the formula for reliability of linear combinations (Nunnally & Bernstein, 1994). Internal consistency estimates for the five DP-4 scales and the General Development Score at each age year (or age ranges for older years) are presented in Tables 5.1, 5.2, 5.3, and 5.4 for the four forms (Parent/Caregiver Interview, Parent/Caregiver Checklist, Teacher Checklist, and Clinician Rating). Note that different age groupings for reliability analyses were used on different forms to ensure a sufficient number of individuals in each group.

Table 5.1. Internal Consistency Coefficients and Standard Errors of Measurement: Parent/Caregiver Interview Form

		DP-4 Scale											
		Physical		Adaptive Behavior		Social–Emotional		Cognitive		Communication		General Development Score	
Age (in years)	*n*	*r*	*SEM*	*r*	*SEM*	*r*	*SEM*	*r*	*SEM*	*r*	*SEM*	*r*	*SEM*
0	211	.93	3.97	.86	5.69	.87	5.41	.89	5.07	.90	4.72	.89	5.01
1	215	.88	5.20	.87	5.38	.80	6.70	.91	4.60	.87	5.51	.86	5.52
2	194	.87	5.41	.88	5.09	.87	5.46	.87	5.37	.90	4.66	.88	5.17
3	177	.90	4.74	.89	4.91	.92	4.11	.90	4.75	.92	4.35	.91	4.47
4	182	.92	4.24	.91	4.48	.92	4.32	.91	4.48	.93	3.92	.92	4.20
5	200	.91	4.50	.88	5.26	.92	4.22	.92	4.15	.91	4.48	.91	4.50
6	194	.91	4.50	.89	5.06	.94	3.79	.95	3.39	.90	4.65	.92	4.23
7	122	.90	4.74	.89	4.99	.93	3.85	.95	3.39	.88	5.19	.91	4.44
8	121	.92	4.24	.94	3.81	.94	3.65	.95	3.19	.94	3.80	.94	3.70
9	109	.93	3.97	.92	4.21	.89	5.06	.95	3.33	.89	4.88	.92	4.31
10	104	.98	2.12	.95	3.28	.92	4.20	.97	2.59	.91	4.61	.95	3.47
11 to 12	196	.89	4.97	.93	3.85	.89	5.01	.95	3.48	.90	4.70	.91	4.39
13 to 16	126	.96	3.00	.88	5.25	.91	4.56	.86	5.56	.90	4.86	.90	4.69
17 to 21	108	.95	3.35	.87	5.36	.91	4.48	.90	4.76	.94	3.80	.92	4.32

Note. Split-half reliability correlation coefficients (*r*) were adjusted using the Spearman-Brown formula for all scales.
Reliability coefficients of the General Development Score were calculated using the reliability of linear combinations (Nunnally & Bernstein, 1994).
SEM = *SD* $\sqrt{(1 - r)}$, where *SD* is the standard deviation of the standard score unit (15) and *r* is the reliability coefficient.

Table 5.2. Internal Consistency Coefficients and Standard Errors of Measurement: Parent/Caregiver Checklist Form

		DP-4 Scale											
		Physical		Adaptive Behavior		Social–Emotional		Cognitive		Communication		General Development Score	
Age (in years)	*n*	*r*	*SEM*	*r*	*SEM*	*r*	*SEM*	*r*	*SEM*	*r*	*SEM*	*r*	*SEM*
0	50	.94	3.75	.85	5.78	.85	5.84	.82	6.34	.88	5.10	.87	5.37
1	39	.79	6.87	.85	5.85	.76	7.27	.85	5.74	.86	5.62	.82	6.30
2	54	.92	4.28	.90	4.76	.87	5.50	.90	4.74	.93	3.83	.91	4.62
3	50	.91	4.49	.94	3.56	.95	3.50	.94	3.80	.96	3.16	.94	3.56
4	47	.95	3.43	.93	3.92	.96	2.87	.95	3.39	.95	3.20	.96	2.91
5	44	.89	4.90	.91	4.57	.94	3.77	.91	4.53	.86	5.70	.90	4.65
6 to 7	70	.92	4.26	.91	4.49	.93	3.91	.94	3.76	.84	6.03	.91	4.46
8 to 12	113	.93	3.83	.90	4.62	.87	5.40	.94	3.72	.90	4.73	.91	4.44
13 to 21	79	.90	4.72	.80	6.75	.91	4.43	.89	4.92	.92	4.36	.89	4.97

Note. Split-half reliability correlation coefficients (*r*) were adjusted using the Spearman-Brown formula for all scales.
Reliability coefficients of the General Development Score were calculated using the reliability of linear combinations (Nunnally & Bernstein, 1994).
SEM = *SD* $\sqrt{(1 - r)}$, where *SD* is the standard deviation of the standard score unit (15) and *r* is the reliability coefficient.

Table 5.3. Internal Consistency Coefficients and Standard Errors of Measurement:
Teacher Checklist Form

		DP-4 Scale											
		Physical		Adaptive Behavior		Social–Emotional		Cognitive		Communication		General Development Score	
Age (in years)	*n*	*r*	*SEM*	*r*	*SEM*	*r*	*SEM*	*r*	*SEM*	*r*	*SEM*	*r*	*SEM*
2	109	.92	4.22	.90	4.84	.90	4.78	.90	4.81	.93	3.98	.91	4.47
3	136	.92	4.30	.89	5.03	.95	3.36	.92	4.25	.92	4.30	.92	4.19
4	146	.94	3.77	.91	4.55	.93	3.95	.91	4.56	.92	4.18	.93	3.94
5	153	.89	4.96	.87	5.48	.90	4.85	.91	4.56	.88	5.11	.89	4.94
6	156	.92	4.14	.91	4.62	.93	3.83	.92	4.27	.91	4.61	.92	4.23
7	109	.89	4.89	.93	3.84	.87	5.32	.94	3.60	.86	5.56	.90	4.64
8	102	.93	4.09	.92	4.31	.89	4.96	.96	3.13	.88	5.26	.92	4.35
9	87	.79	6.91	.88	5.28	.93	4.03	.89	5.05	.86	5.53	.87	5.40
10	83	.92	4.26	.92	4.30	.90	4.74	.94	3.79	.89	5.08	.92	4.35
11 to 12	165	.96	3.03	.96	3.17	.94	3.74	.94	3.74	.95	3.39	.95	3.34
13 to 21	191	.70	8.17	.76	7.37	.90	4.77	.79	6.79	.84	6.08	.79	6.74

Note. Split-half reliability correlation coefficients (*r*) were adjusted using the Spearman-Brown formula for all scales.
Reliability coefficients of the General Development Score were calculated using the reliability of linear combinations (Nunnally & Bernstein, 1994).
$SEM = SD\sqrt{(1-r)}$, where *SD* is the standard deviation of the standard score unit (15) and *r* is the reliability coefficient.

Table 5.4. Internal Consistency Coefficients:
Clinician Rating Form

		DP-4 Scale				
Age (in years)	*n*	Physical	Adaptive Behavior	Social–Emotional	Cognitive	Communication
2 to 3	41	.96	.80	.89	.91	.95
4 to 5	75	.95	.90	.92	.96	.93
6 to 7	47	.94	.93	.94	.94	.97
8 to 12	50	.96	.97	.96	.97	.95
13 to 21	63	.93	.94	.94	.97	.90

Note. Split-half reliability correlation coefficients (*r*) were adjusted using the Spearman-Brown formula for all scales.

Internal consistency reliability estimates for the Parent/Caregiver Interview, Parent/Caregiver Checklist, and Clinician Rating forms were nearly all ≥.80, with two exceptions, .76 and .79, on the Parent/Caregiver Checklist. This threshold indicates good reliability. Many coefficients are above .90, which is considered excellent. The internal consistency reliability estimates on the Teacher Checklist form also ranged from good to excellent, with the exception of a few that were ≥.70 at the older ages.

Standard Error of Measurement

The standard error of measurement (*SEM*) statistic translates a reliability estimate into more practical terms by providing an index of how close an individual's observed score is likely to be to the "true" score that would be obtained if there were no measurement error. The *SEM* is calculated using the following equation: $SEM = SD\sqrt{1-r}$, where *SD* is the standard deviation of the scale and *r* is the reliability of the scale.

The *SEM* values can be converted into confidence intervals that give a range of scores that likely contain the true score. For example, the 95% confidence interval represents the range of scores around the observed score that has a 95% probability of containing the true score. The tables that present the internal consistency estimates (Tables 5.1, 5.2, and 5.3) also present the age-stratified *SEM* values for the DP-4 scales on the Parent/Caregiver Interview, Parent/Caregiver Checklist, and Teacher Checklist forms. *SEM*s are not presented for the Clinician Rating form because it does not yield standard scores, which are required for calculation of the *SEM*.

A practical application of the *SEM* is to derive the confidence values that are provided in raw-score-to-standard-score-conversion tables in the appendix (Tables A.1, A.2, B.1, B.2, C.1, and C.2). The confidence values are expressed in standard score units and rounded to the nearest whole number. The procedure for using confidence values to determine confidence intervals is presented in Chapter 2, and interpretation of confidence intervals is presented in Chapter 3.

Test–Retest Reliability

Test–retest reliability represents the stability of DP-4 scores for the same child over time and involves administering the measure to the same respondent(s) on two occasions. Correlations between the mean standard scores at Time 1 and Time 2, as well as the effect size of the difference between these scores, were calculated for the Parent/Caregiver Interview, Parent/Caregiver Checklist, and Teacher Checklist forms.

The DP-4 retest studies included a total of 74 Parent/Caregiver Interviews administered for typically developing children from the standardization sample, with subsets of 33 Parent/Caregiver Checklist forms and 57 Teacher Checklist forms. In the overall sample of 74 cases, children ranged in age from 0 to 21 years (*M* = 7.9 years, *SD* = 5.7). The sample was 47% male and 53% female, with 41% Hispanic, 58% White, and 1% other ethnicities. In terms of head-of-household education level, 37% had a high school diploma or lower, and 25% had a bachelor's degree or higher. For the Parent/Caregiver Checklist subset of 33 cases, 45% were male and 55% female, ranging in age from 0 years to 21 years (*M* = 10.4, *SD* = 7.3). The ethnic composition of the sample was 55% Hispanic and 45% White. Forty-nine percent of parents had a high school diploma or less, and 36% had a bachelor's degree or higher. The Teacher Checklist subset of 57 cases ranged in age from 2 years to 14 years (*M* = 5.7, *SD* = 2.5), with 47% males and 53% females. The ethnic composition was 21% Hispanic, 77% White, and 2% other ethnicities. Twenty-three percent of cases had a parent with a high school diploma or less, and 30% had a bachelor's degree or higher.

After the initial administration (Time 1), forms were administered a second time (Time 2) to the same respondent, with an average interval of 2 weeks between administrations. Over intervals of such brief durations, test scores are not expected to change appreciably. However, scores may change as a result of random variations in ratings of behavior.

Paired sample t-tests were performed for each pair of standard scores (e.g., Physical Scale score at Time 1 and Time 2). Effect sizes of the difference between each pair of scores were then computed using the formula for Cohen's *d*. Effect sizes help gauge whether group differences are large enough to be considered clinically meaningful. An effect size of 0.2 is considered small, 0.5 medium, and 0.8 large (Cohen, 1992). By convention, a clinically meaningful effect size is at least medium (0.5) in magnitude.

Effect sizes for the differences between scores at Time 1 and Time 2 are presented in Tables 5.5, 5.6, and 5.7. For the Parent/Caregiver Interview form they ranged from 0.03 to 0.07. For the Parent/Caregiver Checklist form, the range was 0.04 to 0.22. Finally, the range of effect sizes for the Teacher Checklist form was 0.04 to 0.25. For each set of effect sizes, no clinically meaningful differences between administrations were found. Overall, effect sizes of the difference between mean scores were small, supporting the good test–retest reliability of the DP-4.

Correlation coefficients are also presented in Tables 5.5, 5.6, and 5.7. Results indicate that, for the five scales and the General Development Score, the test–retest correlations range from .65 to .84 for the Parent/Caregiver Interview form, .55 to .84 for the Parent/Caregiver Checklist form, and .70 to .86 for the Teacher Checklist form. These results show that the test–retest reliability of DP-4 scores is moderate to high, making it acceptable for clinical use and consistent with that of other behavior rating scales.

Table 5.5. Test–Retest Reliability:
Parent/Caregiver Interview Form

DP-4 scale/General Development Score	Time 1		Time 2		Effect size[a]	*r*
	Mean	*SD*	Mean	*SD*		
Physical	104.70	11.99	105.50	10.82	0.07	.65
Adaptive Behavior	105.34	12.30	105.91	11.66	0.05	.77
Social–Emotional	105.70	16.31	106.57	13.65	0.06	.77
Cognitive	103.70	12.56	104.09	13.34	0.03	.70
Communication	106.49	10.95	105.99	11.99	0.04	.66
General Development Score	102.42	9.88	102.78	9.25	0.04	.84

Note. n = 74. Means and *SD*s are expressed in standard score units (*M* = 100, *SD* = 15).
[a] Effect size (Cohen's *d*) = Absolute value of difference between Time 1 and Time 2, divided by pooled *SD*.

Table 5.6. Test–Retest Reliability:
Parent/Caregiver Checklist Form

DP-4 scale/General Development Score	Time 1		Time 2		Effect size[a]	*r*
	Mean	*SD*	Mean	*SD*		
Physical	96.67	13.27	101.58	9.36	0.22	.55
Adaptive Behavior	99.64	12.88	100.70	11.78	0.04	.84
Social–Emotional	97.76	11.76	99.61	10.98	0.08	.66
Cognitive	101.70	13.05	99.36	10.71	0.10	.83
Communication	102.39	13.16	105.36	12.75	0.11	.72
General Development Score	96.64	10.84	98.33	9.33	0.08	.80

Note. n = 33. Means and *SD*s are expressed in standard score units (*M* = 100, *SD* = 15).
[a] Effect size (Cohen's *d*) = Absolute value of the difference between Time 1 and Time 2, divided by pooled *SD*.

Table 5.7. Test–Retest Reliability:
Teacher Checklist Form

DP-4 scale/General Development Score	Time 1		Time 2		Effect size[a]	*r*
	Mean	*SD*	Mean	*SD*		
Physical	105.70	11.19	107.09	9.67	0.13	.74
Adaptive Behavior	108.19	10.99	108.68	11.81	0.04	.70
Social–Emotional	107.53	10.88	108.94	12.43	0.12	.70
Cognitive	101.62	13.09	104.13	16.32	0.17	.74
Communication	105.13	13.39	108.57	14.18	0.25	.79
General Development Score	103.34	10.64	105.53	12.49	0.19	.86

Note. n = 57. Means and *SD*s are expressed in standard score units (*M* = 100, *SD* = 15).
[a] Effect size (Cohen's *d*) = Absolute value of the difference between Time 1 and Time 2, divided by pooled *SD*.

Interrater Reliability

Another method of evaluating reliability is to examine the relationship between scores obtained from different respondents who completed the same form (e.g., two parents completing the Parent/Caregiver Interview). Two interrater reliability studies were conducted on the DP-4: one with the Parent/Caregiver Interview (*n* = 57) and the other with the Parent/Caregiver Checklist (*n* = 38). The Parent/Caregiver Interview sample was 50% male and 50% female, ranging in age from 0 to 19 years (*M* = 7.9, *SD* = 6.4). The ethnic composition of the sample was 28% Hispanic, 47% Black, 18% White, and 7% other ethnicities. Forty-seven percent of parents had a high school diploma or less, while 10% had a bachelor's degree or higher. The Parent/Caregiver Checklist sample was 65% female and 35% male, ranging in age from 6 months to 12 years (*M* = 4.6, *SD* = 3.4). The ethnic composition of the sample was 26% Hispanic, 8% Black, 52% White, and 14% other ethnicities. Forty-five percent of parents had an educational level of a high school diploma or less, while 26% had a bachelor's degree or higher.

Interrater reliability was estimated using the intraclass correlation coefficient. Table 5.8 presents the intraclass correlation coefficients for both sets of data, ranging from .60 to .92 for the Parent/Caregiver Interview, and .73 to .86 for the Parent/Caregiver Checklist. These results indicate a high level of agreement overall between different respondents who completed the same form for identical cases.

Cross-Form Consistency

Cross-form consistency refers to studies in which two respondents rate an individual on two different forms (e.g., Parent/Caregiver Interview and Teacher Checklist forms). Two cross-form consistency studies were conducted on the DP-4: one comparing the Parent/Caregiver Interview with the Teacher Checklist (*n* = 1,408) and another comparing the Parent/Caregiver Checklist with the Teacher Checklist (*n* = 387).

The expectation is that there will be a moderate association between the scores provided by different respondents who complete different forms. The cross-form ratings may vary because the two respondents are providing responses based on observations of the child in different environments and under different conditions.

The cases containing both a Parent/Caregiver Interview and Teacher Checklist ranged in age from 2 to 21 years (*M* = 7.7, *SD* = 4.3). These cases were 51% male and 49% female, with an ethnic composition of 28% Hispanic, 20% Black, 42% White, and 10% other ethnicities. Forty percent had a parent with a high school diploma or less, and 31% with a bachelor's degree or higher. The cases where both the Parent/Caregiver Checklist and Teacher Checklist were completed on the same child (*n* = 387) ranged in age from 2 to 21 years (*M* = 7.6, *SD* = 4.7). Males comprised 51% and females 49% of the sample, and the ethnic composition was 20% Hispanic, 27% Black, 38% White, and 15% other ethnicities. Thirty-two percent of parents had a high school diploma or less, and 41% had a bachelor's degree or higher.

Tables 5.9 and 5.10 present the results of these analyses. The first analysis looked at the effect size of the difference scores between the two forms' scale standard scores and General Development Score. The second examined the correlation between the scores. Results from the comparisons of 1,408 cases on the Parent/Caregiver Interview and Teacher Checklist revealed small effect sizes, ranging from 0.01 to 0.10, indicating small differences between the two sets of forms. Scale correlations were moderate, ranging from .57 to .68, and the General Development Scores were correlated at .99. Results from the comparisons of 387 cases on the Parent/Caregiver Checklist and Teacher Checklist yielded similarly small effect sizes, ranging from 0.01 to 0.12. Scale correlations were moderate, ranging from .62 to .70, and the General Development Score correlation was also moderate at .73.

These results indicate that different raters' responses may yield similar scores, demonstrating the reliability of the DP-4. In addition, the differences between ratings by multiple respondents will provide more breadth of information about the child.

Table 5.8. Interrater Reliability:
Parent/Caregiver Interview and Parent/Caregiver Checklist Forms

DP-4 scale/General Development Score	Parent/Caregiver Interview	Parent/Caregiver Checklist
	r	*r*
Physical	.84	.73
Adaptive Behavior	.60	.83
Social–Emotional	.85	.79
Cognitive	.91	.85
Communication	.89	.73
General Development Score	.92	.86

Note. Parent/Caregiver Interview *n* = 57. Parent/Caregiver Checklist *n* = 38.

Table 5.9. Cross-Form Consistency:
Parent/Caregiver Interview and Teacher Checklist Forms

DP-4 scale/General Development Score	Parent/Caregiver Interview		Teacher Checklist		Effect size[a]	*r*
	Mean	*SD*	Mean	*SD*		
Physical	102.90	18.33	102.63	13.72	0.02	.61
Adaptive Behavior	101.44	17.82	103.14	15.22	0.10	.62
Social–Emotional	101.29	18.82	101.68	14.62	0.02	.57
Cognitive	102.22	18.52	102.43	15.90	0.01	.68
Communication	102.47	16.32	103.97	15.10	0.10	.59
General Development Score	99.20	15.66	99.00	17.25	0.01	.99

Note. n = 1,408. Means and *SD*s are expressed in standard score units (*M* = 100, *SD* = 15).

[a] Effect size (Cohen's *d*) = Absolute value of the difference between Parent/Caregiver Interview mean and Teacher Checklist mean, divided by pooled *SD*.

Table 5.10. Cross-Form Consistency:
Parent/Caregiver Checklist and Teacher Checklist Forms

DP-4 scale/General Development Score	Parent/Caregiver Checklist		Teacher Checklist		Effect size[a]	*r*
	Mean	*SD*	Mean	*SD*		
Physical	102.52	21.21	102.31	15.25	0.01	.65
Adaptive Behavior	102.31	20.12	102.13	15.65	0.01	.62
Social–Emotional	103.22	23.50	100.91	15.86	0.12	.64
Cognitive	103.05	21.93	100.67	16.65	0.12	.70
Communication	101.39	19.28	103.35	15.70	0.11	.65
General Development Score	100.03	16.17	98.98	16.09	0.07	.73

Note. n = 387. Means and *SD*s are expressed in standard score units (*M* = 100, *SD* = 15).

[a] Effect size (Cohen's *d*) = Absolute value of the difference between Parent/Caregiver Checklist mean and Teacher Checklist mean, divided by pooled *SD*.

Alternate-Forms Reliability

Alternate-forms reliability is established when the same person completes two different forms on the same child. In 532 cases of the DP-4 standardization study, the same parent completed both the Parent/Caregiver Interview and Parent/Caregiver Checklist on the same child. This sample was 51% male and 49% female, ranging in age from 0 to 21 years (*M* = 6.6, *SD* = 5.3). The ethnic composition was 22% Hispanic, 23% Black, 39% White, and 16% other ethnicities. Thirty percent of parents had a high school diploma or less and 44% had a bachelor's degree or higher.

Two analyses were conducted, and the results are presented in Table 5.11. The first analysis looked at the effect size of the difference between the two forms' standard scores. The second analysis examined the correlation between the scores. The resulting effect sizes were all small, ranging from 0.05 to 0.18, suggesting little meaning in the differences between the scores. Correlation coefficients were all high; scale scores ranged from .80 to .83, and the General Development Score correlated at .86. These results suggest that similar scores will result from the use of either form.

Table 5.11. Alternate-Forms Reliability:
Parent/Caregiver Interview and Parent/Caregiver Checklist Forms

DP-4 scale/General Development Score	Parent/Caregiver Interview		Parent/Caregiver Checklist		Effect size[a]	*r*
	Mean	*SD*	Mean	*SD*		
Physical	101.61	19.52	102.75	19.91	0.06	.83
Adaptive Behavior	99.67	18.34	101.80	18.65	0.12	.80
Social–Emotional	99.52	19.46	103.10	21.83	0.17	.82
Cognitive	99.67	18.56	103.12	20.17	0.18	.80
Communication	100.78	17.26	101.64	18.42	0.05	.81
General Development Score	97.33	16.52	100.02	15.01	0.17	.86

Note. n = 532. Means and *SD*s are expressed in standard score units (*M* = 100, *SD* = 15).

[a] Effect size (Cohen's *d*) = Absolute value of the difference between Parent/Caregiver Interview mean and Parent/Caregiver Checklist mean, divided by pooled *SD*.

Validity

The validity of a test refers to its ability to accurately measure what it is designed to measure. This chapter presents information describing the validity studies conducted for the DP-4. The types of validity that will be discussed are content, construct, convergent, and validity evidence based on clinical groups. The capacity of the DP-4 to detect developmental deficits at various cutoff scores will also be discussed.

Given the high level of similarity of content and measurement between the Parent/Caregiver Interview form and the other three forms (Parent/Caregiver Checklist, Teacher Checklist, and Clinician Rating), it is reasonable to apply the validity evidence from the Parent/Caregiver Interview form to the other three forms. However, some additional validity studies were conducted to specifically examine the relationship between the DP-4 Parent/Caregiver Checklist and Teacher Checklist and other parent- and teacher-reported measures on related constructs.

Content Validity

Content validity refers to the utilization of appropriate item content to measure the construct 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 representative of their respective skill area, a practice that has been continued with each revision.

Feedback from clinicians during the user survey described in Chapter 4 contributes to the content validity of the DP-4. Additionally, the fact that raw scores consistently increase as the child's age increases provides evidence that the DP-4 accurately measures relevant developmental content, as development is expected to increase with age. Figure 5.1 illustrates the pattern of raw scores for each scale across each age year in the DP-4 standardization sample. These raw scores were eventually converted to standard scores through the process of norming, which required that scores were "flattened out" across certain age groups (see Chapter 4). Figure 5.2 depicts the final DP-4 age groups that were chosen for this process. This overall pattern demonstrates the content validity of the DP-4 as a measure of development that increases over time.

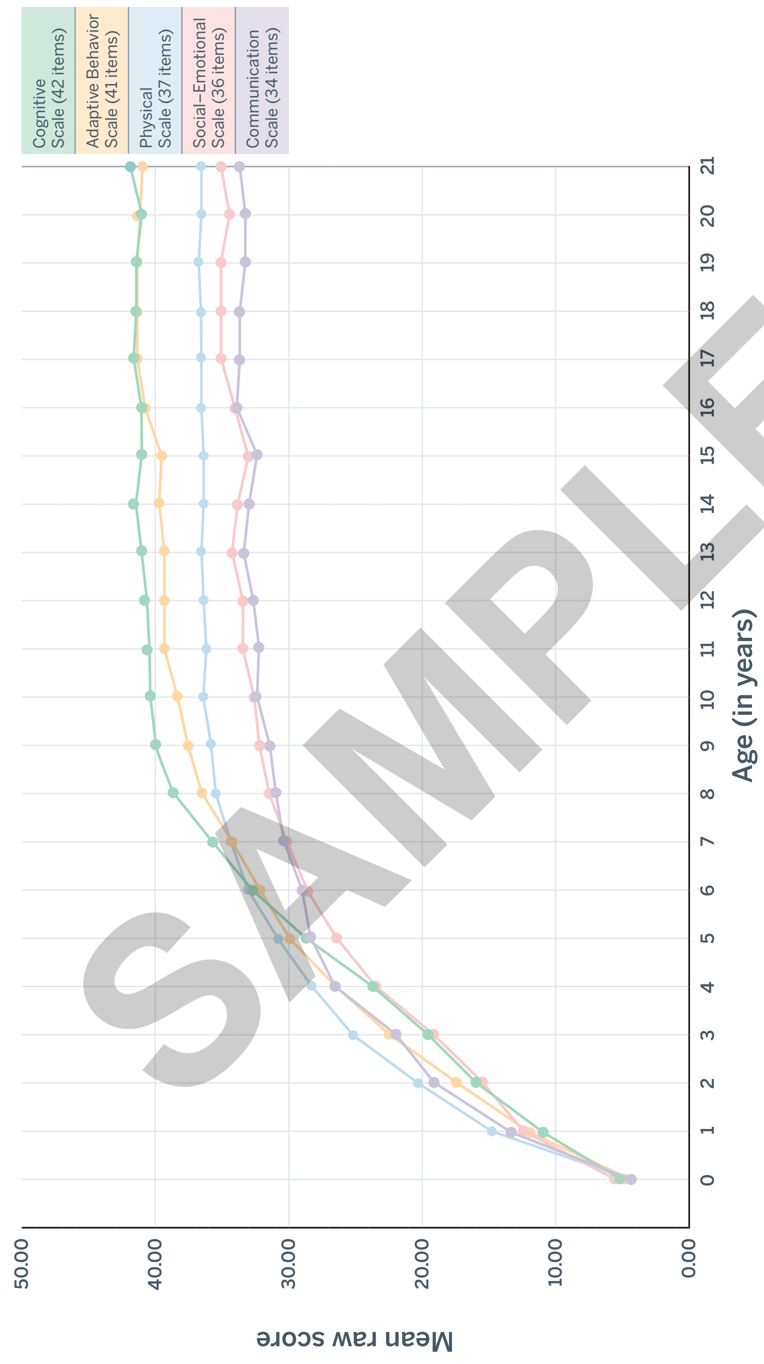
Construct Validity

Construct validity for the DP-4 was measured by examining the structural characteristics of the scales through the use of factor analysis, interscale correlations, item-to-scale (i.e., item-total) correlations, and Rasch analysis.

Factor analysis To examine the structure of the DP-4, an exploratory common-factor analysis with oblimin rotation was conducted with all 190 items using the standardization sample data. The common factor approach was selected because it allows for sources of variance (e.g., measurement error) other than the extracted factors. Oblimin rotation was chosen because it assumes correlated factors, which is a theoretically and empirically reasonable assumption for the DP-4. 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, which is to be expected of a measure with scales that are correlated with one another. Correlations among scales are discussed in the next section.

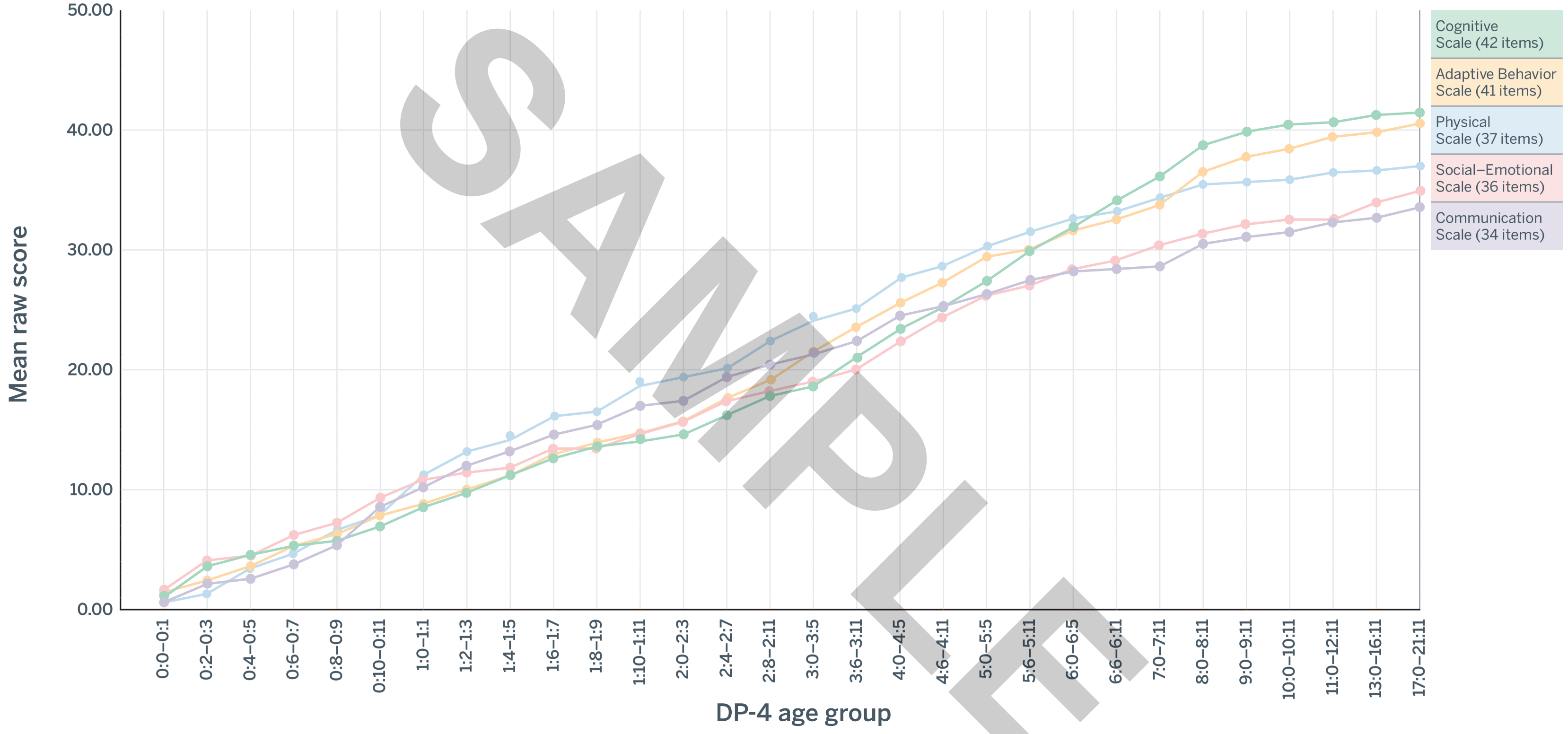
Interscale correlation analysis Determining the interrelatedness of the scale scores through interscale correlations helps establish whether the scales can be viewed as measurements of separate aspects of overall development. The theoretical structure of the DP-4 suggests that the scales, considered together, have both shared and unique variance, and thus should intercorrelate at moderate levels. The scales should in turn show stronger associations with the General Development Score than with each other.

Interscale correlations for the five DP-4 scales and the General Development Score were calculated for standard scores in each sample (Parent/Caregiver Interview, Parent/Caregiver Checklist, and Teacher Checklist). As expected, strong correlations were found between the scales and the General



Note. Because each scale has a different number of items, the mean raw score for each scale will have different maximum thresholds.

Figure 5.1. *Developmental Curve by Age (in Years)*



Note. Because each scale has a different number of items, the mean raw score for each scale will have different maximum thresholds.

Figure 5.2. *Developmental Curve by DP-4 Age Group*

Development Score for each sample. Additionally, scales exhibit mostly moderate correlations with each other. Tables 5.12, 5.13, and 5.14 display interscale correlation results for the Parent/Caregiver Interview, Parent/Caregiver Checklist, and Teacher Checklist samples. As can be expected, all scales exhibit correlations in the moderate to high ranges. Given 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 internal consistency coefficients for each scale (reported in Tables 5.1, 5.2, and 5.3). These results provide support for the separate scoring and interpretation of the five scales.

Item-total correlation analysis All 190 items were also analyzed in an item-total correlation analysis to determine their correlations with their assigned scales, each of the other four DP-4 scales, and the General Development Score. In a set of empirically distinct subscales, it is expected that an item will correlate more strongly with the subscale that contains it than with any other subscale. This outcome occurred on 70% of the Physical, 80% of Adaptive Behavior, 89% of the Social–Emotional, 69% of the Cognitive, and 79% of Communication scale items. Thus, most of the items correlated most strongly with their own scales, but there was also a high level of intercorrelation with other scales. This finding supports the utility of separate scales for interpretation, despite the fact that all areas of development are closely related.

Taken together, the interscale correlation comparisons, along with the item-total correlation analysis, suggest that the DP-4 scales are only marginally separable, on an empirical basis. That is, these scales tap a general construct of development, with the breakdown by scale and item providing guidance for interpretation and remediation planning.

Rasch analysis The final analysis conducted to establish construct validity of the DP-4 utilized the Rasch methodology to examine the item coverage over the range of abilities intended to be measured by the DP-4. This type of analysis is possible because Rasch item and person measures are expressed in the same metric. Comparison of the ranges of item difficulties and person abilities for each of the five scales revealed that the range of person ability extends just below and just 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 mastered during the elementary school years, and the DP-4 standardization sample includes individuals up to the age of 21 years, 11 months. These ranges are also similar to one another across scales, indicating that the items do a good job of measurement within the desired skill range.

Convergent Validity

The convergent validation method examines a test's relationship to other measures of similar constructs. Strong correlations with convergent measure are considered to be supportive of the construct validity of the test under study.

To validate the DP-4 scales against those of other similar measures, scores were compared to those obtained from four related tests: (a) the Developmental Profile 3 [DP-3], (b) the Vineland Adaptive Behavior Scales, Third Edition [Vineland-3], (c) the Developmental Assessment of Young Children, Second Edition [DAYC-2], and (d) the Adaptive Behavior Assessment System, Third Edition [ABAS®-3].

Developmental Profile 3 (DP-3) Thirty-seven individuals were administered both the DP-3 and the DP-4 interview forms. Sixty-three percent of the children in this study were male and 37% were female, all ranging in age from 2 years to 12 years (M = 6.6, SD = 3.2). The ethnic composition of the sample was 26% Hispanic, 24% Black, 45% White, and 5% other ethnicities. Eighteen percent of parents had a high school diploma or less, and 47% had a bachelor's degree or higher.

Table 5.15 displays the correlations between the DP-4 scale standard scores and those from the DP-3. Correlations between the same scales on each test are displayed in bold and ranged from .80 to .89 (all correlations were significant at $p < .01$). As a general measure of development, the General Development Scores of the DP-3 and DP-4 were significantly correlated ($r = .93$). These scale and General Development Score correlations indicate strong relationships between scores, and thus underscore the similarity of item content between the DP-3 and DP-4 revision.

Table 5.12. Interscale Correlations in the DP-4 Standardization Sample: Parent/Caregiver Interview Form

DP-4 scale/General Development Score	PHY	ADP	SOC	COG	COM	GDS
Physical	—	—	—	—	—	—
Adaptive Behavior	.66	—	—	—	—	—
Social–Emotional	.56	.68	—	—	—	—
Cognitive	.55	.63	.61	—	—	—
Communication	.53	.65	.67	.68	—	—
General Development Score	.79	.87	.85	.83	.84	—

Note. *n* = 2,259. All correlations are significant at *p* < .01.
PHY = Physical; **ADP** = Adaptive Behavior; **SOC** = Social–Emotional; **COG** = Cognitive; **COM** = Communication; **GDS** = General Development Score

Table 5.13. Interscale Correlations in the DP-4 Standardization Sample: Parent/Caregiver Checklist Form

DP-4 scale/General Development Score	PHY	ADP	SOC	COG	COM	GDS
Physical	—	—	—	—	—	—
Adaptive Behavior	.71	—	—	—	—	—
Social–Emotional	.64	.75	—	—	—	—
Cognitive	.66	.71	.66	—	—	—
Communication	.61	.70	.71	.70	—	—
General Development Score	.83	.88	.87	.85	.86	—

Note. *n* = 542. All correlations are significant at *p* < .01.
PHY = Physical; **ADP** = Adaptive Behavior; **SOC** = Social–Emotional; **COG** = Cognitive; **COM** = Communication; **GDS** = General Development Score

Table 5.14. Interscale Correlations in the DP-4 Standardization Sample: Teacher Checklist Form

DP-4 scale/General Development Score	PHY	ADP	SOC	COG	COM	GDS
Physical	—	—	—	—	—	—
Adaptive Behavior	.67	—	—	—	—	—
Social–Emotional	.59	.75	—	—	—	—
Cognitive	.58	.70	.63	—	—	—
Communication	.59	.70	.74	.73	—	—
General Development Score	.79	.89	.87	.86	.88	—

Note. *n* = 1,437. All correlations are significant at *p* < .01.
PHY = Physical; **ADP** = Adaptive Behavior; **SOC** = Social–Emotional; **COG** = Cognitive; **COM** = Communication; **GDS** = General Development Score

Table 5.15. Correlations Between DP-3 and DP-4 Standard Scores

	DP-3 scale/General Development Score					
DP-4 scale/General Development Score	Physical	Adaptive Behavior	Social–Emotional	Cognitive	Communication	General Development Score
Physical	**.89**	.71	.68	.60	.65	.83
Adaptive Behavior	.82	**.80**	.81	.63	.72	.86
Social–Emotional	.64	.68	**.85**	.62	.68	.80
Cognitive	.57	.66	.70	**.83**	.86	.81
Communication	.70	.75	.82	.77	**.89**	.88
General Development Score	.81	.81	.87	.79	.86	**.93**

Note. n = 37. All correlations are significant at $p < .01$.

Vineland-3 The Vineland-3 is a comprehensive measure of adaptive behavior. Its format is similar to that of the DP-4, as it utilizes a parent interview method as a means of obtaining information, as well as parent and teacher ratings. Additionally, it measures skills in four of the five DP-4 areas (no Cognitive Scale is included in the Vineland-3). Tables 5.16, 5.17, and 5.18 display correlations between the DP-4 and Vineland-3 scores.

Interview forms The Vineland-3 Interview and DP-4 Parent/Caregiver Interview forms were administered to the parents of 105 children. This sample included 63% males and 37% females, ranging in age from 2 years to 17 years (M = 6.4, SD = 3.7). The ethnic composition of the sample was 22% Hispanic, 15% Black, 54% White, and 9% other ethnicities. Twenty-one percent of parents had a high school diploma or less, and 42% had a bachelor's degree or higher.

Table 5.16 displays the correlations between the Vineland-3 and DP-4 scores from this sample. The correlations for the domains and scales that are most similar in content are displayed in bold type. Overall, these correlation coefficients are moderate to high in magnitude (ranging from .79 to .86) across scales and domains, indicating a relationship that would be expected between different measures with similar item content. Across the two measures, the bolded correlations between scores with similar item content are higher than the correlations between scores with nonsimilar item content.

Parent forms The Vineland-3 Parent/Caregiver rating form was completed by 54 parents who also completed the DP-4 Parent/Caregiver Checklist. The sample was composed of 79% males and 21% females ranging in age from 3 years to 20 years (M = 7.5, SD = 5.7). The ethnic composition of the sample was 15% Hispanic, 21% Black, 57% White, and 7% other ethnicities. Twenty-eight percent of parents had a high school diploma or less and 36% had a bachelor's degree or higher.

Table 5.17 displays the correlations between the DP-4 scale scores and the Vineland-3 domain scores for the parent rating forms. The correlations of the domains and scales that are most similar in content are displayed in bold type. In three of four instances, the correlations in bold were the highest. Correlations are moderate to high for all comparisons, ranging from .65 to .86. Similar to the study evaluating the correlations between the interview forms of the Vineland-3 and the DP-4, the correlations document a relationship that would be expected between measures that have similar item content.

Teacher forms The teacher version of the DP-4 and Vineland-3 were administered to teachers of 128 children. Seventy-four percent of these cases were male and 26% were female, ranging in age from 3 years to 18 years (M = 6.3, SD = 3.9). The ethnic composition of this sample was 21% Hispanic, 23% Black, 48% White, and 8% other ethnicities. Thirty-three percent of parents had a high school diploma or less, and 30% had a bachelor's degree or higher.

Table 5.18 shows the correlations between the Vineland-3 and the DP-4 scores on the teacher forms. The findings were similar to those for the Parent/Caregiver Interview and Parent/Caregiver Checklist forms. The correlations in bold were higher than correlations between scores from scales with dissimilar content. These correlations were moderate to high, ranging from .68 to .79.

Table 5.16. Correlations Between the DP-4 and the Vineland Adaptive Behavior Scales, Third Edition (Vineland-3): Interview Forms

	DP-4 Parent/Caregiver Interview form					
Vineland-3: Interview form	**Physical**	**Adaptive Behavior**	**Social–Emotional**	**Cognitive**	**Communication**	**General Development Score**
Communication	.75	.81	.75	.79	**.84**	.86
Daily Living Skills	.82	**.85**	.77	.73	.80	.86
Socialization	.71	.79	**.79**	.71	.74	.81
Motor Skills	**.86**	.82	.63	.68	.75	.83
Adaptive Behavior Composite	.79	.84	.80	.77	.83	.87

Note. $n = 105$. Bold type indicates expected correlation based on similar content.
All correlations are significant at $p < .01$.

Table 5.17. Correlations Between the DP-4 and the Vineland Adaptive Behavior Scales, Third Edition (Vineland-3): Parent Forms

	DP-4 Parent/Caregiver Checklist form					
Vineland-3: Parent form	**Physical**	**Adaptive Behavior**	**Social–Emotional**	**Cognitive**	**Communication**	**General Development Score**
Communication	.67	.75	.75	.79	**.83**	.82
Daily Living Skills	.75	**.80**	.74	.68	.74	.84
Socialization	.65	.74	**.79**	.70	.76	.81
Motor Skills	**.86**	.86	.69	.66	.72	.81
Adaptive Behavior Composite	.70	.78	.76	.73	.77	.84

Note. $n = 54$. Bold type indicates expected correlation based on similar content.
All correlations are significant at $p < .01$.

Table 5.18. Correlations Between the DP-4 and the Vineland Adaptive Behavior Scales, Third Edition (Vineland-3): Teacher Forms

	DP-4 Teacher Checklist form					
Vineland-3: Teacher form	**Physical**	**Adaptive Behavior**	**Social–Emotional**	**Cognitive**	**Communication**	**General Development Score**
Communication	.65	.71	.65	.68	**.76**	.76
Daily Living Skills	.74	**.75**	.67	.66	.71	.78
Socialization	.63	.66	**.68**	.55	.67	.70
Motor Skills	**.79**	.72	.52	.55	.58	.70
Adaptive Behavior Composite	.70	.76	.73	.67	.77	.80

Note. $n = 128$. Bold type indicates expected correlation based on similar content.
All correlations are significant at $p < .01$.

Overall, the convergent validity of the DP-4 is supported by the similar pattern of correlational findings between the DP-4 and Vineland-3 across three methods of administration (interview, parent rating, and teacher rating).

Developmental Assessment of Young Children, Second Edition (DAYC-2) The DAYC-2 is a collection of five subtests that measure the same areas of functioning as the DP-4 (Physical, Adaptive Behavior, Social–Emotional, Cognitive, and Communication). The DAYC-2 is designed for use with children aged birth through 5 years, 11 months. The DAYC-2 and the DP-4 Parent/Caregiver Interview forms were administered to 37 parents. The sample included 57% males and 43% females, ranging in age from 2 years to 6 years (*M* = 3.9, *SD* = 1.1). The ethnic composition of the cases was 46% Hispanic, 24% Black, 24% White, and 6% other ethnicities. Forty-three percent of parents had a high school diploma or less, and 32% had a bachelor's degree or higher.

The correlations between the DAYC-2 subtest scores and the DP-4 scale scores are shown in Table 5.19. Correlations between the five areas shared by both measures (displayed in bold type) were found to be moderate, ranging from .49 to .67. As with the Vineland-3 study, the correlations between DP-4 and DAYC-2 scores document a relationship that would be expected between different measures with similar item content. Overall, the results from this analysis support the validity of the DP-4 as a measure of child development in the same manner as another measure of child development.

Adaptive Behavior Assessment System, Third Edition (ABAS-3) The ABAS-3 is a comprehensive measure of an individual's adaptive skills, with rating forms that can be completed by a parent/caregiver or teacher.

Parent forms The ABAS-3 has two forms that may be completed by parents; one for children aged birth to 5 years (Parent/Primary Caregiver form) and the other for children aged 5 to 21 years (Parent form). Both forms include nine adaptive skill areas, of which six (Community Use, Home Living, Health and Safety, Leisure, Self-Care, and Self-Direction) are related to the Adaptive Behavior Scale of the DP-4. Of the remaining ABAS-3 skill areas, Functional Academics relates to the Cognitive Scale of the DP-4, the Communication area relates to the DP-4 scale of the same name, and the Social area relates to the Social–Emotional Scale of the DP-4.

One of the two ABAS-3 parent rating forms, depending on the child's age, was administered to 95 parents who also completed the DP-4 Parent/Caregiver Checklist form. This sample included cases of children who were 77% male and 23% female, with ages ranging from 2 years to 20 years (*M* = 6.5, *SD* = 4.6). The ethnic composition of the group was 13% Hispanic, 17% Black, 64% White, and 6% other ethnicities. Eighteen percent of parents had a high school diploma or less and 49% had a bachelor's degree or higher.

Table 5.19. Correlations Between the DP-4 and the Developmental Assessment of Young Children, Second Edition (DAYC-2)

	DP-4 Parent/Caregiver Interview form					
DAYC-2 subtest	Physical	Adaptive Behavior	Social–Emotional	Cognitive	Communication	General Development Score
Cognitive	.29	.41*	.52**	**.57****	.41*	.51**
Communication	.28	.49**	.69**	.71**	**.58****	.63**
Social–Emotional	.23	.47**	**.67****	.39*	.30	.46**
Physical Development	**.49****	.49**	.42*	.52**	.37*	.54**
Adaptive Behavior	.50**	**.62****	.56**	.62**	.55**	.67**
General Developmental Index (GDI)	—	—	—	—	—	**.64****

Note. n = 37. Bold type indicates expected correlation based on similar content.
**Correlation is significant at *p* < .01.
*Correlation is significant at *p* < .05.

Table 5.20 shows the correlations between the nine adaptive skill areas of the ABAS-3 parent forms and the five scales of the DP-4; related areas are displayed in bold type. Correlations of the related scales were found to be moderate, ranging from .45 to .68. This pattern is not unexpected, since adaptive behavior is a broad construct closely related to the five main areas of development measured by the DP-4.

Teacher forms The ABAS-3 includes the same nine skill areas in each of its two teacher rating forms: the Teacher/Daycare Provider form for younger children (ages 2–5) and Teacher form for older children (ages 5–21). Either of the two forms was completed by 56 teachers. This sample included cases of children who were 75% male and 25% female, ranging in age from 3 years to 20 years (*M* = 7.5, *SD* = 5.5). The ethnic composition of the sample was 14% Hispanic, 16% Black, 63% White, and 7% other ethnicities. Twenty-three percent of parents had a high school diploma or less, and 37% had a bachelor's degree or higher.

Table 5.21 shows the correlations between the ABAS-3 teacher rating forms and DP-4 Teacher Checklist form. Findings were similar to those found with the parent rating forms. Correlations of related scales were moderate and ranged from .45 to .77, once again supporting the notion that the DP-4 is related to the construct of adaptive behavior, as well as to related domains of development as measured by the ABAS-3.

Table 5.20. Correlations Between the DP-4 and the ABAS-3: Parent Forms

	DP-4 Parent/Caregiver Checklist form				
ABAS-3: Parent forms	Physical	Adaptive Behavior	Social–Emotional	Cognitive	Communication
Communication	0.43	0.45	0.53	0.49	**0.67**
Community Use (older children only)	0.51	**0.45**	0.53	0.51	0.51
Functional Academics	0.40	0.46	0.37	**0.60**	0.53
Home Living	0.64	**0.61**	0.60	0.37	0.51
Health and Safety	0.66	**0.61**	0.62	0.45	0.58
Leisure	0.49	**0.47**	0.64	0.41	0.56
Self-Care	0.64	**0.68**	0.50	0.44	0.54
Self-Direction	0.52	**0.49**	0.51	0.40	0.47
Social	0.45	0.48	**0.56**	0.34	0.49

Note. n = 95. Bold type indicates expected correlation based on similar content.
All correlations are significant at *p* < .01.

Table 5.21. Correlations Between the DP-4 and the ABAS-3: Teacher Forms

	DP-4 Teacher Checklist form				
ABAS-3: Teacher forms	Physical	Adaptive Behavior	Social–Emotional	Cognitive	Communication
Communication	.43**	.58**	.56**	.52**	**.77****
Community Use (older children only)	.39*	**.45***	.37*	.32	.36
Functional Academics	.51**	.52**	.25	**.54****	.42**
Home Living	.56**	**.69****	.60**	.46**	.58**
Health and Safety	.51**	**.64****	.58**	.46**	.64**
Leisure	.46**	**.58****	.54**	.36**	.50**
Self-Care	.55**	**.76****	.51**	.48**	.56**
Self-Direction	.42**	**.55****	.59**	.45**	.43**
Social	.26	.45**	**.50****	.30*	.63**

Note. n = 56. Bold type indicates expected correlation based on similar content.
**Correlation is significant at *p* < .01.
*Correlation is significant at *p* < .05.

Validity Evidence Based on Clinical Group Comparisons

The DP-4 clinical sample described in Chapter 4 was used to illustrate that the DP-4 can effectively discriminate between typically developing children and children with a clinical disorder. Standard scores from children in the clinical sample were compared to those of a matched group from the standardization sample. Cases were matched based on age, gender, ethnicity, and region.

Table 5.22 illustrates that, for all five scales and the General Development Score, the mean scores for the clinical sample were both statistically and meaningfully lower than those for the typically developing group sample. The effect sizes ranged from 1.26 to 1.79. By convention, these effect sizes represent clinically meaningful differences between the two groups and support the expectation that the DP-4 scores discriminate between typically developing children and those with a clinical diagnosis.

The clinical sample was then divided into groups based on diagnosis. The diagnoses most likely to occur in cases where the DP-4 is administered, namely intellectual disability, developmental delay, and autism spectrum disorder, were compared separately with their own matched control groups. Table 5.23 displays paired t-test results and resulting effect sizes comparing the diagnostic groups on standard scores for the five DP-4 scales and the General Development Score. The results illustrate that the groups showed significant differences, both statistically and clinically, in scores across all five scales, further supporting the validity of the DP-4 in its ability to distinguish between different types of developmental difficulties.

Detection of Skill Deficits

Conditional probability analyses (also known as receiver operating characteristic [ROC] curves) were run to determine the capacity of the DP-4 to detect deficits in child development and functioning at various cutoff scores. This analysis included the following clinical groups: autism spectrum disorder, developmental delay, intellectual disability, visual impairment, physical disability, and other disability. The clinical groups were combined and compared to typically developing children. Results indicated that the DP-4 General Development Score (area under ROC curve = .950, p < .001) provided statistically significant improvement over chance in detecting the disorders present in the clinical group. The sample for this analysis included 2,051 typically developing children and 348 children with clinical diagnoses.

Table 5.22. Descriptive Statistics and Effect Sizes for Overall Clinical Group and Matched Typically Developing Group

DP-4 scale/General Development Score	Clinical group		Typically developing group		Effect size[a]
	Mean	*SD*	Mean	*SD*	
Physical	79.42	25.60	105.40	15.74	1.26
Adaptive Behavior	75.33	21.45	104.89	15.15	1.62
Social–Emotional	75.07	20.03	104.97	17.06	1.61
Cognitive	76.25	23.06	106.05	17.37	1.47
Communication	77.15	21.48	106.45	13.41	1.68
General Development Score	73.16	19.90	102.80	13.25	1.79

Note. n = 348. Means and *SD*s are expressed in standard score units (M = 100, SD = 15). All pairs of means differ significantly, p <.001.

[a] Effect size (Cohen's d) = Absolute value of the difference between typically developing group mean and clinical group mean, divided by pooled *SD*.

Table 5.23. Descriptive Statistics and Effect Sizes for Clinical Groups by Diagnosis and Matched Typically Developing Groups

DP-4 scale/General Development Score by diagnostic group	Clinical group		Typically developing group		
	Mean	*SD*	Mean	*SD*	Effect size[a]
Intellectual disability					
Physical	76.08	24.18	103.19	15.53	1.37
Adaptive Behavior	70.75	19.51	104.39	14.42	1.98
Social–Emotional	73.78	19.62	102.11	17.77	1.52
Cognitive	64.94	18.65	99.81	13.27	2.18
Communication	68.19	16.04	104.14	13.00	2.48
General Development Score	66.89	17.08	99.89	13.19	2.18
Developmental delay					
Physical	79.13	23.61	108.83	19.99	1.36
Adaptive Behavior	76.98	20.22	107.58	17.23	1.63
Social–Emotional	75.09	16.43	107.80	18.65	1.86
Cognitive	72.33	19.59	106.48	22.17	1.64
Communication	78.28	17.09	110.73	15.30	2.00
General Development Score	72.77	16.83	105.58	15.55	2.03
Autism spectrum disorder					
Physical	79.07	21.79	106.55	15.01	1.49
Adaptive Behavior	73.53	18.53	104.73	15.13	1.85
Social–Emotional	63.69	16.17	104.88	16.88	2.49
Cognitive	79.65	24.62	108.70	16.92	1.40
Communication	74.12	21.57	105.30	11.98	1.86
General Development Score	70.51	17.76	103.28	12.72	2.15

Note. Intellectual disability *n* = 64; developmental delay *n* = 64; autism spectrum disorder *n* = 74. Means and *SD*s are expressed in standard score units (*M* = 100, *SD* = 15). All pairs of means differ significantly per t-test results, *p* < .001.

[a] Effect size (Cohen's *d*) = Absolute value of the difference between typically developing group mean and clinical group mean, divided by pooled *SD*.

Table 5.24 displays the sensitivity and specificity associated with various standard score values of the DP-4 General Development Score. Sensitivity refers to a test's capacity to detect true positive cases of the deficit in question. Specificity refers to a test's capacity to exclude true negative cases (persons who do not have the deficit in question). Betz et al. (2013) recommend providing sensitivity and specificity results for multiple values so that clinicians can choose a cutoff score that is best suited to their clinical setting. The values presented in Table 5.24 are representative of the range of sensitivity and specificity found in current assessments of child development.

Table 5.24. Conditional Probability Analysis for Detection of Clinical Cases

Standard score cutoff value	Sensitivity	Specificity
70	.56	.99
75	.68	.98
80	.77	.96
85	.84	.93
90	.89	.87

Note. Sample analyzed included 348 clinically diagnosed children and 2,051 typically developing children.

To illustrate, Table 5.24 shows that at a cutoff of 85 (one standard deviation below the mean), the DP-4 General Development Score has sensitivity of .84 and specificity of .93. In practical terms, this means that 84% of the children with clinical diagnoses associated with developmental delays had standard scores less than 85, whereas 93% of the typically developing children had standard scores of 86 or greater. It is worth noting that only 1% or fewer of typically developing children had standard scores of 70 or less (≥2 *SD* below the mean) as shown by the specificity of .99 at the cutoff value of 70. However, due to the variability inherent in clinical data, only the most severely impaired children will be identified as having a developmental delay when using a strict cutoff value of 70 (sensitivity of .56). This finding demonstrates that when a child's General Development Score is ≤70, there is a strong probability that the child has significant development problems that require intervention, though using a higher cutoff value will help to identify more children with developmental delays.

These results serve as a reminder that, at any level of test score deficit, there is a risk of under- or over-identifying children who are in need of intervention. Although the DP-4 provides a measurement of child development and functioning, results should not be used in isolation for diagnosis or treatment planning. Instead, these results should be used in concert with other data (e.g., other assessment results, parent and teacher interview, review of available records, direct observation).

Summary and Directions for Future Research

This chapter has described the psychometric studies conducted to support the publication of the DP-4. Reliability was examined from several perspectives, and the DP-4 scores performed well on indexes of internal consistency, test–retest reliability, interrater reliability, cross-form consistency and alternate-form reliability. An exploratory factor analysis showed acceptable fit with models of child development, supporting a general factor of development with related scales. Similarly, the DP-4 scales correlate in expected ways with one another, as well as with other tests of development, thereby yielding evidence of convergent validity. Finally, the DP-4 standard scores distinguish typically developing individuals from those in the clinical population.

As with all measures, treatment outcome research is needed to expand the range of validity evidence for the DP-4. Such research should include studies that assess individuals with developmental disorders and other related disabilities, before and after intervention. These studies will further help to validate the DP-4 as a critical tool in the assessment of development and developmental delays, as well as in evaluating the effectiveness of intervention efforts.