Reliability and Score Differences

8 Chapter

Introduction

Consistency of measurement across items and testing occasions is an important attribute of good instruments. Because important decisions about the treatment or placement of children may rely on information from the M-P-R, several different types of reliability estimates were calculated. Both classical test-theory and item-response theory (IRT) approaches to reliability estimation were employed. Since some users of the M-P-R may be less familiar with IRT approaches to reliability, classical reliability coefficients were seen as an essential adjunct to IRT information.

When testing professionals employ conventional norm-referenced scales, information such as standard scores and classical reliability coefficients are used in setting confidence intervals around scores reported to teachers, parents and others. When Rasch-based growth scales are employed, the user is referred to Appendix E where the standard errors are provided for each possible score. (As the confidence interval is truly different for each and every score, the Rasch confidence intervals are preferable.) The section on Test-Information Curves provides a visual representation of the measurement error and the accuracy at the different growth scores which is valuable for those using the Rasch-based growth scores, though it is also important that these curves are examined by others who may make critical decisions about children. Whichever approach is used, this chapter presents estimates of the internal consistency, test-retest reliability, standard error of measurement and test-information curves for the various scales and measures in the M-P-R.

Internal Consistency Reliability Coefficients

Figure 8.1 presents estimates of the internal-consistency reliability of the Cognitive Battery scales. Importantly, because Cronbach's (1957) alpha coefficients were employed, the estimates in **Figure 8.1** are estimates of the lower bound of internal reliability (Lord, 1980). This is a measure that examines whether the items test the same content, based upon consistency of response. Note that each of the scales spans a large range of ages, but that each estimate was calculated on the children within the designated age group of the normative sample. Thus, the reliability estimates are not inflated by developmental growth, as they would if calculated on the entire normative sample. The reliability coefficients in **Figure 8.1** are the basis for the standard error of measurement, significance of score differences (standard error of the difference), and other important elements of classical test interpretation. For these reasons, lower-bound estimates of internal consistency were deemed most beneficial to prevent over interpretation of small differences between scores. Because most of the coefficients exceed .90, the level of reliability for M-P-R scales is considered excellent, and useful for making informed decisions about individual children (Gregory, 1996, p. 100). Fundamentally, high internal consistency means that the items within the M-P-R scales are highly intercorrelated and form a homogeneous collection of indicators of the targeted developmental ability.

The Memory and Speed supplementary scales are the shortest of the Cognitive Battery (40 and 19 items, respectively, compared to 59-238 for the other scales) and shorter scales are typically lower in reliability than longer scales. However, even reliabilities in the .77 range are found in prominent cognitive batteries (e.g., some of the subtests of the Wechsler scales, Wechsler, 1991). As compared to the longer scales, the Memory and Speed scales should be used for generating clinical hypotheses, and not for decisions about individual children's placement or diagnosis. Due to the difficulties in accurately measuring memory and speed in infants, the scales begin around 18 months of age with norms at 23 months of age. Also, as shown in **Figure 8.1**, Receptive Language begins at 13 months, due to the lesser degree

Figure 8.1 Internal Consistency (Alpha) Reliability Coefficients for the Cognitive Battery Scales of the M-P-R

Age Group in Months	Number of Subjects	Cognitive Battery Scales						
		Developmental Index	Cognitive	Receptive Language	Fine Motor	Memory	Speed	Visual Motor
0 – 12	195	.98	.96	—	.94	—	—	.95
13 - 24	161	.97	.93	.93	.90	.70	.70	.92
25 - 48	340	.98	.95	.96	.92	.84	.84	.90
49+	372	.97	.92	.91	.87	.74	.74	.82
Average Reliability		.98	.94	.94	.91	.77	.77	.91

Figure 8.2 Internal Consistency (Alpha) Reliability Coefficients for the Gross Motor Battery and Social-Emotional, Self-Help/Adaptive and Language Scales of the M-P-R

Age Group in Months	Number of Subjects	Remaining Scales				
		Gross Motor Total	Social-Emotional Development	Self-Help/ Adaptive	Language Total	Expressive Language
0 – 12	195	.96	.94	.94	—	—
13 - 24	161	.94	.90	.93	.98	.97
25 - 48	340	.92	.94	.94	.98	.97
49+	372	.90	.92	.93	.97	.96
Average reliability		.93	.93	.94	.98	.97

of vocabulary development in infants, and the difficulties in measuring "communication", without vocabulary development.

Figure 8.2 shows the internal consistency estimates for the Gross Motor Battery total, the Social-Emotional, Self-Help/Adaptive and Language scales. Again, the reliabilities are very strong, and allow the examiner confidence in using the scales in making decisions about individual children (Gregory, 1996).

Conventional Standard Errors of Measurement

Standard errors of measurement (SEM) for M-P-R standard scores, using classical test theory assumptions, are presented in **Figure 8.3** (Cognitive Battery Scales) and **Figure 8.4** (Gross Motor Battery, Social-Emotional, Self-Help, and Language Scales, *see following page*). The conventional SEM is calculated from the internal consistency reliability of the scale (standard deviation of the standard scores—15—times the square root of 1.0 minus the reliability coefficient). Note that for Rasch-based growth scores, each score value has a standard error associated with it (see section on Growth Scores). The conventional standard errors of measurement for the M-P-R scales are useful for placing conventional confidence intervals around obtained scores and, generally, for examiners and parents/ guardians to appreciate the degree of error in any test score. One traditional method of forming confidence intervals is to add or subtract a numerical amount (based on the SEM) from a given obtained score. Intervals constructed from one SEM (added or subtracted from the obtained score) create a range of scores wherein the "true score" (error free theoretical score of the child) is expected to occur 68 per cent of the time. This is the classic percentage of the normal curve located within one SD of the mean. If an individual were retested numerous times (hypothetically), his or her obtained scores would theoretically form a normal curve with SEM as the standard deviation of the obtained scores. For greater confidence, a 95 per cent confidence interval can be formed from just under two SEM's (1.96 times SEM, added or subtracted from the obtained score).

An example of the construction of SEM confidence intervals is as follows.

> An 11-month-old girl was administered the Cognitive scale. She obtained a standard score of 110. As shown in **Figure 8.3**, the SEM for the Cognitive scale, at this age, is 3.00 (15 times the square root of [1.0 minus the reliability of .96]). A 68 per cent confidence interval around the obtained score would be 107 to 113, and a 95 per cent confidence interval would be approximately 104 to 116. The parents could be told, with 95 per cent certainty, the child's true score is in the 104 to 116 range. This will help them to understand the error factor in all such test scores.

Figure 8.3 Conventional Standard Errors of Measurement for the Cognitive Battery Scales of the M-P-R

Age Group in Months	Number of Subjects	Cognitive Battery Scales						
		Developmental Index	Cognitive	Receptive Language	Fine Motor	Memory	Speed	Visual Motor
0 – 12	195	2.12	3.00	—	3.67	—	—	3.35
13 - 24	161	2.60	3.97	3.97	4.74	8.26	8.26	4.24
25 - 48	340	2.12	3.35	3.00	4.24	6.00	6.00	4.74
49+	372	2.60	4.24	4.50	5.71	7.65	7.65	6.36

Figure 8.4 Conventional Standard Errors of Measurement for the Gross Motor Battery and Social-Emotional, Self-Help/Adaptive and Language Scales of the M-P-R

Age Group in Months	Number of Subjects	Remaining Scales				
		Gross Motor Total	Social-Emotional Development	Self-Help/ Adaptive	Language Total	Expressive Language
0 – 12	195	3.00	3.67	3.67	—	—
13 - 24	161	3.67	4.74	3.97	2.12	2.60
25 - 48	340	4.24	3.67	3.67	2.12	2.60
49+	372	4.74	4.24	3.97	2.60	3.00

Table 8.5 Scaled Score Means, Standard Deviations and Test-Retest Correlations for the M-P-R Developmental Scales, Minus Expressive Language

	Cognitive Battery Scales									
	Developmental Index	Cognitive	Receptive Language	Fine Motor	Visual Motor	Speed	Memory	Gross Motor	Social-Emotional	Self-Help
Correlation	.89	.87	.90	.90	.90	.84	.89	.88	.89	.84
Time 1 Mean	94.3	97.8	98.4	99.7	98.2	94.6	104.0	103.9	102.9	105.3
Time 1 SD	11.9	18.4	14.1	11.2	10.5	28.9	8.3	8.1	19.1	14.2
Time 2 Mean	94.0	99.4	100.6	100.8	101.4	92.3	105.8	102.5	101.5	104.7
Time 2 SD	32.7	21.2	14.6	10.8	12.9	32.2	10.6	10.8	15.8	11.0

Test-Retest Reliability Coefficients

A sample of 41 children, ages 3 to 70 (median 44) were administered the Cognitive Battery on two occasions, an average of just over 3 weeks apart. Means, standard deviations and correlations are shown in **Figure 8.5** for all of the developmental scales. The number of subjects in each age group were 0-12 months - 29 per cent, 13 to 24 months - 14 per cent, 25-47 months - 11 per cent and 48 to 78 months - 46 per cent, respectively. Some of the cognitive scales show very small practice effects (shifts in mean scores from test to retest), as is frequently found on developmental tests (e.g., Bayley, 1993). In this case, there was very little change from the first (Time 1) to the second (Time 2) testing.

Reliability of Temperament Measures, Parent and Examiner

The measures addressed above are developmental in nature. A stable, gradual increase in score, with age, would be expected for these scales. Reliability, for these measures, means the scales measure very similar content (alpha) and that the behaviors would not change across time (test-retest).

Alphas for the measures are shown in **Figure 8.6** and **Figure 8.7**. As one would expect, these reliabilities are not terribly high for the very young children, but much higher for children 18 months of age and older.

Figure 8.8 presents the retest correlations for the Examiner Rating Scales, for children 18 months of age and older. As part of the standardization testing for the M-P-R, examiners completed the Examiner Rating Scale following the completion of the Cognitive Battery. The diagonal elements are the scale reliabilities. One should note that "angry" is substantially lower than "organized". This does not necessarily mean this scale exhibits measurement problems, but that children who are angry in the testing setting

Figure 8.6 Internal Consistencies (alpha) for Examiner Rating Scale

	Examiner Rating Scale	
Scales	Under 18 months	18 months Up
Attention	.49	
Fearful	.86	
Emotionality	.73	
Angry		.78
Active		.90
Organized		.95

Figure 8.7 Internal Consistencies (alpha) for Temperament Rating Scales

	Parent Temperament Rating	
Scales	Under 18 mo.	18 mo. Up
Easy	.77	.94
Difficult	.67	.81
Fearful		.68

Figure 8.8: Examiner Rating Scale Test-Retest Reliability, 3 Weeks (Diagonal elements are reliability, upper right are first test left column by retest top row.)

	Retest		
Test	**Angry**	**Active**	**Organized**
Angry	.53	-.33	-.09
Active		.66	.19
Organized			.90

may not have an angry temperament, and are angry much of the time. They may have been upset as a "state" rather than as a "trait". Something may have happened at that particular time, such as they were not allowed to go out and play. However, children who are "organized" in the testing setting, are also likely to be "organized" at the retest setting, and are likely to have an "organized" temperament.

Infant Language and Memory Scales

The Infant Language and Memory scales are just small scales that assess children's emerging social communicative responses (infant language) and infant memory (including object permanence). These are not central scales of the M-P-R, but rather are ways to examine these emerging abilities in research studies, and to better understand a child's performance. Behaviors like this, assessed in very young children, on very short measures, tend not to be as stable as measures of cognitive abilities, especially at older ages. Still, they provide valuable information for certain purposes. The internal consistency (alpha) for the brief Infant Language scale is .62. (This scale is not added into the 12 - item Parent Report of Expressive Language, but is meant to supplement that data for interpretative purposes.) The internal consistency (alpha) for the Infant Memory scale is .74. This is a scale that isn't added into the Memory scale in the Cognitive Battery, due to an absence of memory items in the first half of the cognitive measure, at age level one.

Growth Scale (Rasch) Test Information Curves

Figure 8.9 shows the test information curves for the Developmental Index and several other Growth Scales,

Figure 8.9 Information Curves for Developmental Scales

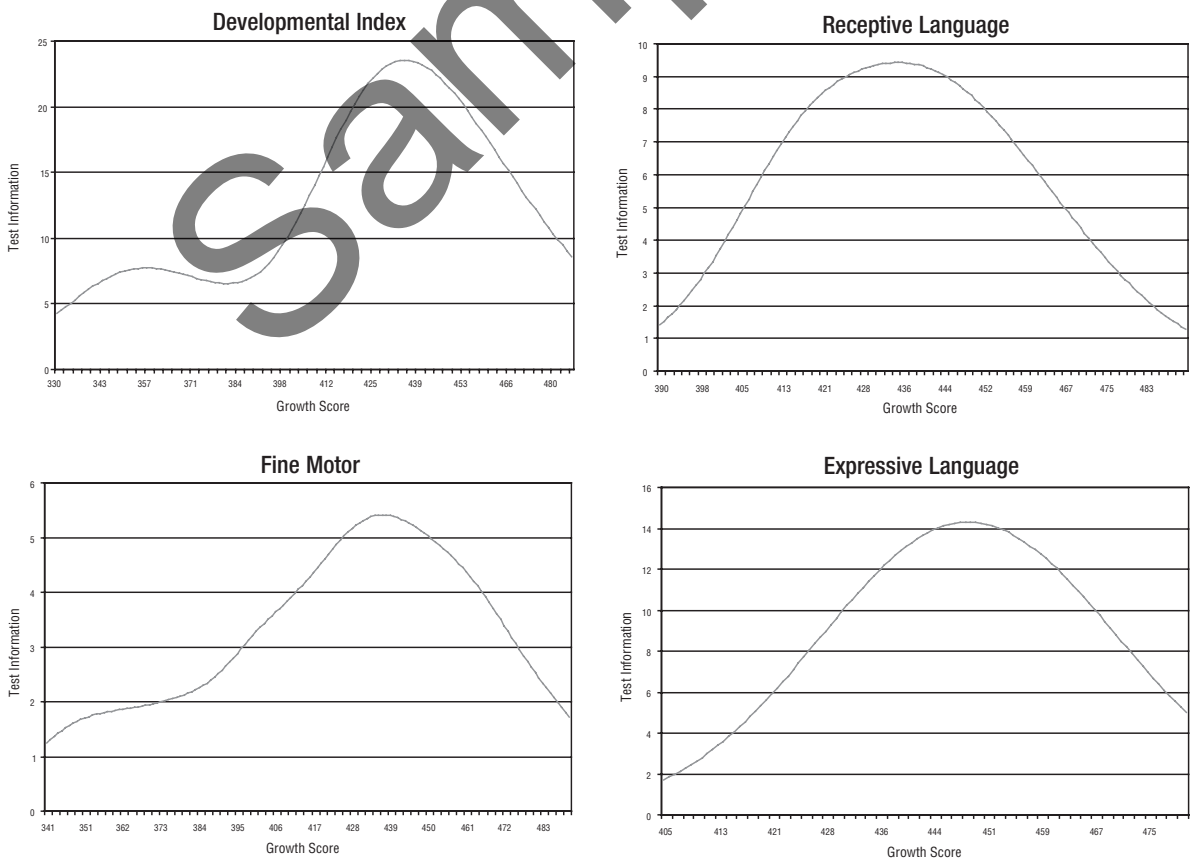
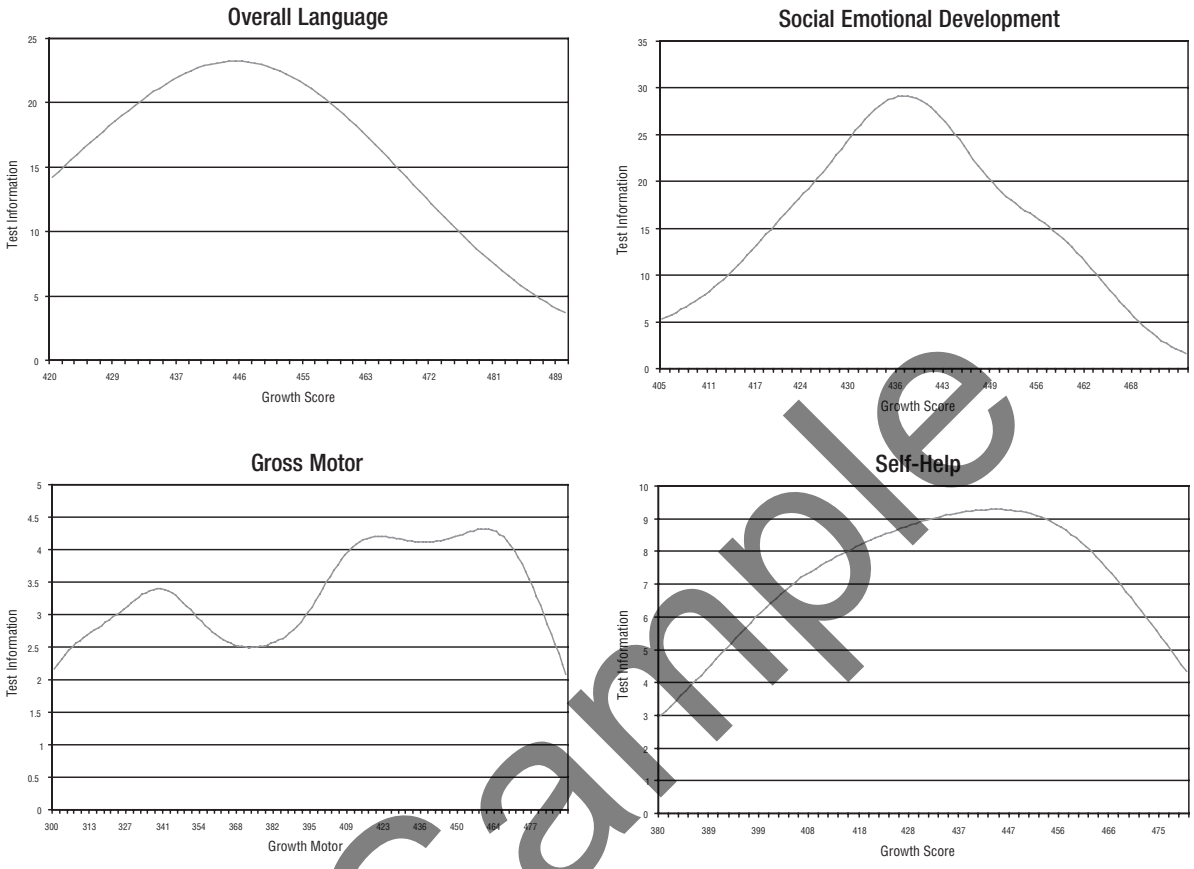


Figure 8.9 *Continued* Information Curves for Developmental Scales



Note: One of the reasons for the depression of the information curve for young children in the Developmental Index is that Receptive Language does not begin until part of the way through Age 1.

consisting of all Rasch-scaled items. These figures clearly show the ranges where there is the greatest accuracy in assessing the individuals score. Unlike the traditional standard error of the measure which is identical across all test score values, here it can clearly be seen that there is greater accuracy (lower standard error of the measure) in the middle of the range and lower accuracy in the extremities. While not a graph or standard errors, similar data is available in the standard error table for the Raw to Rasch conversions (See Appendix). Knowing where the greatest accuracy levels exist can be important when more permanent, irreversible decisions are being made with information from a test such as the M-P-R scale scores. As you can see in these graphs, most of the curves are not excessively steep and have a fairly flat peak. This means they have good discrimination across a range of ability (age). It also shows that test information is lower for very young children and therefore, for that group, test data should be held as being less reliable. One should use the Rasch standard error of the measure to assure a sufficiently broad range of possible true scores. It also shows that, unless someone needs to look at changes in growth scores across time or if the child's abilities are several years delayed, children should be moved on to other measures at 6-years of age.

Differences Between Standard Scores

A description of the use of score differences in the interpretation of the M-P-R was provided in Chapter 6. Tables for identifying score differences — both statistical significance and frequency of differences in the normative sample— are provided in **Figure 6.2** and the Appendix. This section of the manual describes first the formula employed in the development of those tables. Some

Figure 8.10a Score Differences for Statistical Significance (Cognitive Battery Scales)

Age in Months	Standard Score Differences Required for Statistical Significance at the .05 Level by Age Group for the Cognitive Battery M-P-R Scales					
	Cognitive vs. Receptive Language	Cognitive vs. Fine Motor	Fine Motor vs. Receptive Language	Visual Motor vs. Speed	Visual Motor vs. Memory	Memory vs Speed
0 - 12	—-	9.29	—-	—-	—-	—-
13 - 24	11.00	12.12	12.12	18.20	18.20	22.90
25 - 48	8.81	10.59	10.18	14.99	14.99	16.63
49+	12.12	13.94	14.25	19.50	19.50	21.20

Note: See elsewhere in Chapter 8 for details on the calculation and use of these values.

tables that arise from these formula can be useful in the interpretation of test data.

Figures 8.10a-e employed the same age categories, and the same internal consistency reliability coefficients as those included in the earlier sections of this chapter. The age groupings were 0 - 12, 13 - 24, 25 - 48, and 49 months or more. For ease of use, the differences between scores, is listed at the .68 level. Usually, one will want to approximately double this value (multiply it by either 1.96 or 2) because of its widely recognized status. Also, avoidance of the clinical-hypothesis levels of .10 or .15 helps to reduce over-interpretation of differences. Differences needed for significance were calculated by employing a normal curve value, Z (set at 1.96 for the .05 level) and the standard errors of the difference (SEM) of each score, as shown in Formula 8.1 below. This is discussed early in Chapter 6.

Formula 8.1:
SEDiff. = 15* [SQRT(SEM1*SEM1 + SEM2*SEM2)]

The formula for the standard error of measurement is the familiar one shown in Formula 8.2, below.

Formula 8.2:
SEM = 15*SQRT (1.0 - reliability)

Where reliability (**r**) is the internal consistency estimate for a given age group and the 15 is listed as the standard deviation for all the developmental scores of M-P-R including the Developmental Index, and indeed it is always very close to 15.

For frequency of differences, a partial table is provided here, in **Figure 8.1**. The frequency distributions of absolute values of differences were tabulated for each set of difference scores and the cumulative distributions expressed in percentiles for each pair. Clinically meaningful percentiles were identified and listed in the appendix table—the 1, 2, 3, 5,10, 15 and 25 percentile points. These percentile points are the most relevant for clinical interpretation, because it is the rarity of the difference (low percentile points) that is of greatest importance clinically. As typically found in other cognitive batteries, a rule of thumb is to accept differences as clinically important when they occur in 15 per cent or less of the normative sample. For example, for the difference between the Cognitive and Fine Motor scales, a difference of at least 13 points is needed to form clinical hypotheses. Note that the value 13 is not quite significant at the .05 level as shown in the table. To consider both significance and frequency, a magnitude of at least 14 points of difference is therefore, needed. For decisions about children's placements or sensitive diagnoses based on differences, you should

Figure 8.10b Score Differences for Statistical Significance (Cognitive Battery with Additional Scales)

Age in Months	Standard Score Differences Required for Statistical Significance at the .05 Level by Age Group for the Differences Between Scores on the Cognitive Battery and Other M-P-R Measures				
	Cognitive vs. Gross Motor	Cognitive vs. Total Language	Expressive Language vs Receptive Language	Social-Emotional vs. Self-Help/Adaptive	Fine Motor vs. Gross Motor
0 - 12	—-	8.32	—-	10.17	9.29
13 - 24	10.60	8.82	9.30	12.12	11.75
25 - 48	10.59	7.77	7.78	10.17	11.75
49+	12.46	9.75	10.60	11.38	14.55

Note: See elsewhere in Chapter 8 for details of the calculation and use of these values.

Figure 8.10c Score Differences at Several Significance Levels (Cognitive Battery Scales)

	Values of Standard Score Differences at Various Percentiles of the Normative Sample of the M-P-R Language Scales used for All Ages of Children			
Percentile	Cognitive vs. Gross Motor	Cognitive vs. Fine Motor	Social-Emotional vs. Self-Help/Adaptive	Fine Motor vs. Gross Motor
25%	19	11	15	20
15%	**24**	**13**	**19**	**25**
10%	28	15	22	29
5%	**32**	**19**	**28**	**35**
3%	36	22	33	38
2%	40	26	38	40
1%	44	30	42	43

Note. For ages 1 to 78 months. See Chapter 8 for details of The calculation and use of these values. Clinical tradition suggests using difference values that are relatively rare (15% for hypotheses or 5% or less for sensitive decisions).

Figure 8.10d Score Differences at Several Significance Levels (Cognitive Battery Scales with Language Scales)

	Values of Standard Score Differences at Various Percentiles of the Normative Sample of the M-P-R Language Scales used for Ages 12 months and above			
Percentile	Cognitive vs. Receptive Language	Cognitive vs. Total Language	Fine Motor vs. Receptive Language	Expressive Language vs. Receptive Language
25%	9	15	11	16
15%	12	19	14	20
10%	15	22	17	23
5%	18	27	21	28
3%	20	32	24	32
2%	21	34	28	34
1%	28	38	31	38

Note. Use for ages 12 months and above. See Chapter 8 for details of the calculation and use of these values. Clinical tradition suggests using difference values that are relatively rare (15% for hypotheses or 5% or less for sensitive decisions).

Figure 8.10e Score Differences at Several Significance Levels (Supplemental Scales)

	Values of Standard Score Differences at Various Percentiles of the Normative Sample for the Supplementary M-P-R Scales in the Cognitive Battery for use with Ages 12 Months and Above		
Percentile	Visual Motor vs. Speed	Visual Motor vs. Memory	Memory vs. Speed
25%	14	13	14
15%	18	16	17
10%	22	17	19
5%	26	23	22
3%	28	26	25
2%	32	28	28
1%	34	32	34

Note. Use with ages 12 months and above only. See Chapter 8 for Details of calculation and use of these values. Clinical tradition suggests using difference values that are relatively rare (15% for hypotheses or 5% and less for sensitive decisions).

use only rare differences (5 per cent or less in frequency), such as 19 points in this case (which is clearly statistically significant).

Factor Analyses of the Examiner Rating of Test Session Behavior

A lengthy series of exploratory factor analyses were conducted in the final selection of items for the Examiner Rating of Test Session Behavior. Analyses at the item level were conducted first, using principal component and maximum-likelihood factor analyses with varimax and direct-oblimin rotations. These were followed by principal-axis factor analyses of the subscales to verify the second-order factors that constitute the composite scores. All analyses were conducted on typical children from the standardization sample. Each series of analysis will be described, with emphasis on the subscale analyses. It was felt that the (extensive) details of the item factor analyses went beyond the scope of this test

manual, and that the composite scores should be emphasized because they have wider applicability, and higher reliability, than the subscales.

Because the Examiner Rating scale had the full standardization sample of the M-P-R, matching the size (N = 1,068) of the standardization sample, extensive analyses at each of several age levels were possible. The first stage of data analysis was a comparison of principal-component analyses (equamax rotations) the first 15 items by themselves which were aimed at younger children and infants. Fewer responses (N = 195) were available there versus the larger sample for older age items. Items were those from the original 62-item version of the Examiner Rating Scale employed in the standardization study for the M-P-R (see Chapter 7). Exploratory analyses using maximum-likelihood extraction with direct-oblimin rotation were also inspected. A set of fairly reliable subscales were identified by the size of factor loadings from these analyses, as shown in **Figure 8.11**. These scales, varying in reliability from .78 to .94, were identified clusters of items that were "consensus" across the various types of factor analyses calculated. In each analysis, 3 factors emerged for the younger children and 5 factors emerged for the older children. Factors were identified based on scree plot, eigenvalues greater than 1.0, and more than 3 items per factor with factor loadings above .40. Subsequent internal consistency reliability analyses (alpha along with item-total correlations) for the identified subscales (see **Figure 8.12**).

Because the subscales in **Figure 8.10** were shown to correlate among themselves, and because Rasch scaling of the items (Wright & Lineacre, 1999) suggested two primary dimensions, higher-order factor analyses (Gorsuch, 1983) were conducted on the subscales. As shown also in **Figure 8.11**, the "positively" labeled scales (organized, compliant, active, and attentive) were clustered together on a factor labeled "Easy Child." The other scales (angry, emotional, and fearful) were found to cluster on a second major factor labeled "Difficult Child." Some evidence was present in the factor analyses and in the Rasch scaling that the "fearfulness" items could be separated from the other Difficult Child subscales. The reliability of these new longer, higher order scales were then calculated and showed excellent levels of reliability (see **Figure 8.2** and **Figure 8.12**).

Figure 8.11 Factor Loadings on Higher-Level Factor Analysis of Clear Internally Consistent Content Subscales, Identified in Item Factor Analyses of the Examiner Rating of Test Session Behavior

Content Domains	Reliability	Easy Child Factor	Difficult Child Factor
Organized/ Cooperative	.94	.83	
Compliant	.79	.77	
Active & Eager to Work	.89	.76	(-.46)
Attentive	.86	.72	(-.51)
Angry & Oppositional	.78		.71
Emotional	.84		.83
Fearful	.83		.72

Figure 8.12 Internal Consistency Reliability (Alpha) of the Easy Child, Difficult Child, and Fearfulness Scales of the Examiner Rating of Test Session Behavior

Age Levels (Months)	Easy Child	Difficult Child	Fearfulness
18 – 24	.90	.78	.82
25 – 48	.94	.85	.86
49 – 78	.94	.64	.68
Average	.92	.83	.81

Validity and Research Studies

Chapter 9

Evidence of test validity traditionally includes content-related, criterion-related and construct-related evidence, collected over time, from a series of research studies. Such evidence should be examined for each of the *interpretations* of test scores (e.g., multiple uses of the same test score), rather than for the test or test scores *per se* (American Psychological Association, 1999). Messick (1980) emphasized the overall importance of construct-related evidence, but also introduced some new categories, such as "consequential validity" which is the study of the effects of using test results on the people and organizations that use tests. Thus, there is not a single index of validity, such as a correlation with another test, that can represent the validity of all possible test interpretations. This chapter presents the validity research conducted on the M-P-R, to date.

Content-Related Evidence of Validity

Gregory (1996) stated that "content validity is determined by the degree to which the questions, tasks, or items on a test are representative of the universe of behavior the test was designed to sample (p. 108)." Several types of systematic sampling of item content, examination of the "fit" of each item to the unidimensional construct underlying a given scale, and other types of evidence will be described for the M-P-R.

The need for accurate assessment of a wide range of children, including those who are developing typically, developing in exceptional ways, and those from different cultures, point to the need for *toy-based and "hands on"* assessment methods. This type of approach is needed to maximize motivation and avoid culturally biased responding. The M-P-R emphasizes the use of toys and manipulatives to assess various dimensions of cognitive, language, motor, and behavioral development. Developmental abilities are defined as those behavioral and cognitive skills that show a "growth" trend, in typical children, across the first several years of life. Many of these skills can be tested with toys, blocks, pictures, and "game-like" tasks, for purposes of "child friendly" assessment. The original Merrill Palmer Scales (Stutsman, 1931) provided a "hands on" assessment of developmental abilities for children ages 18 to 60 months. The original scales were widely used in special education, especially for children with hearing difficulties, speech and language delays and autism. Hence, the M-P-R was designed to have some of the same content and usefulness as the original scale, but with extensions into infancy, and expanded coverage of all areas required by IDEA (1997) legislation.

The project to revise, expand, and standardize the M-P-R, has been a multi-year effort, with some Federal grant support and extensive field testing. Each task was carefully scrutinized by the 57 examiners in the tryout phase and the 100 examiners in the standardization phase, to assure that the materials and subtests in each of the scales would be effective, and easily administered to children. After completing their assigned cases, examiners completed a series of questionnaires to obtain content and procedural ratings of each subtest. Only subtests with uniformly high ratings were retained without revision (see Chapter 7). Some of the subtests of the Tryout Edition, such as Action Words and Secret Code were deleted because they required verbal responses (expressive language).

The items of the original Merrill-Palmer Scales were analyzed and categorized according to Carroll's (1993) catalog of cognitive factors. It became apparent the original items included a wide array of factors, collected from existing tests by Stutsman (1931), developed by various researchers. The major facets represented in the Merrill-Palmer Scales included fluid reasoning; various aspects of visualization ability; short-term memory; quantitative reasoning; visual matching; fine motor development; speed of processing (e.g., quickness in puzzle assembly), and expressive and receptive language development. Because these were precisely the types of developmental subdomains needed for a comprehensive cognitive assessment, additional items were developed for nearly all these existing areas of content.

After certain items were eliminated for modernization and child-safety purposes (e.g., the 1930's style puzzle pictures, the button task and the "pink tower), major classifications of items were expanded into full subtests. The area of expressive language (e.g., in the original "action agents" verbal fluency questions) was separated from the Cognitive Battery, after the Tryout Edition, a separate Expressive Language scale was developed. Separating Expressive Language from the Cognitive Battery allowed the M-P-R to continue the traditional role of the original Merrill-Palmer Scales in the assessment of children with communication difficulties. Also, to assure consistent content and construct validity, Teaching Items, at the starting points of each subtest, were designed to verify that children would clearly understand the assessment task, prior to completing each type of item. The total collection of subtests, particularly in the Cognitive Battery, was designed to tap as many cognitive, fine motor and receptive language abilities as possible, within

testing-time constraints. The M-P-R was, thus, designed on the basis of a unifying model of abilities (Carroll, 1993) with the Cognitive Battery emphasizing limited English language expression. In addition, Spanish language examiner prompts were added, to increase the usage of the test among children with limited English proficiency.

Pilot versions of subtests, in the age range of 1-18 months, were examined with particular focus on identifying effective toys for infant assessment. Pilot testing was followed by a comprehensive Tryout Edition administration to more than 500 children, ages 1 month to 7 years of age. Finally, 1068 (additional data was collected to allow for the analysis of Native-American performance but are not included in the 1068 figure) children were tested on an improved Standardization Edition, using a national sample representative of SES and ethnic backgrounds. Studies of internal reliability, factor analyses, Rasch item and test scaling, and test bias were conducted on both the Tryout and Standardization Editions, and were used to establish subtests reflective of major cognitive, motor, language, and behavior scales with high internal consistency. Cross-battery correlations (e.g., with the Bayley Scales of Infant Development, Stanford-Binet 5th Edition, Leiter-R, etc.) were employed to verify the existence of expected construct and content dimensions. Content experts, representing various ethnic minorities, and psychologists (including the 57 examiners in the national field tryout and 100 examiners in the national standardization testing), familiar with culturally-diverse examinees, inspected all items. Items with poor internal consistency (item-total correlations), high indexes of item bias, or poor ratings by examiners and experts, were eliminated from the final published version of the test. For a complete discussion of these activities, see earlier chapters, including Chapter 7. Chapter 7 also reviewed some of the developmental steps related to content validation. Consult Chapter 7 for a discussion of the research basis for the Cognitive Battery, the Gross Motor Scale, and the various Rating Scales, which underwent careful development and review by experts and examiners, in both Tryout and Standardization Editions.

Kamphaus (1993) has emphasized the importance of including empirical item analysis as a category of evidence for content validity. Chapter 7 discussed several aspects of the conventional item analyses conducted on both the Tryout and Standardization Editions of the M-P-R. Additionally, extensive analyses of items, using item-response theory, were completed at both phases of test development, and in the final adjustment of tasks for the published edition. As Reckase (1996), Lord (1980), and Wright and Stone (1979) have emphasized, item selection with IRT methodology can provide a more uniform measurement (in the sense of measurement error or "test information") up and down the scale continuum. This was achieved in the M-P-R by employing the Rasch (1980) program WINSTEPS (Lineacre, 2002) to examine comprehensively the characteristics of items. Two major types of evidence for item acceptability are particularly related to content validity: a) the fit of the items to unidimensional scales within each content domain (and within each ability domain such as cognitive, receptive language, fine motor), and b) absence of differential item functioning (DIF, Holland & Wainer, 1993), which provides evidence of item and test fairness across gender and ethnic groups.

Exceptional fit to the Rasch 1-parameter logistic model was obtained following the screening of poorly fitting items. For example, the entire set of Developmental Index (Cognitive Battery) items were analyzed in one large analysis of 327 items with the normative sample (N=1,068). Measures of model-data-fit for the Rasch model included mean-square-fit indexes (information-weighted "in fit" and outlier weighted "out fit") and their standardized equivalents. Each of these values, along with item discrimination (item-to-total correlations) was inspected for all items; with misfitting items removed in most cases. In the Cognitive Battery, for example, only 11 items (3%) out of 327 had standardized outfit values exceeding 3.0, and only 23 items (11%) out of 327 exceeded 2.0. Standardized in-fit values showed a similar trend. Mean-square-fit indexes were originally described by Wright and Stone (1979) and detailed by Linacre and Wright (1990).

Through studies of differential item functioning (DIF), the refined scales of this published edition of M-P-R showed exceptional similarity in the correlations of difficulty parameters for each subgroup. In other words, groups were quite similar in the ordering of items on the underlying Rasch scales, denoting fairness of measurement across groups. As explained in Hambleton and Swaminathan (1985), item bias can be examined by comparing item parameters across subgroups. In the case of the M-P-R, separate Rasch-difficulty calibrations were derived for pairs of subgroups including male and female, Caucasian and African-American and Caucasian and Hispanic groups. Separate calibrations for each item were then plotted on item scatterplots, and the pattern of

plotted points examined in comparison to a 45-degree line and a best-fitting line. Confidence intervals, based on standardized values, were then used to identify any outliers from the linear relationships that signal consistency across groups. Only a very small number of items were detected as falling outside the confidence zones for these scatterplots, and most of these items were removed from the M-P-R.

Both the model-data-fit analyses and the item-bias studies are related to content validity in the following way: Except for rare instances, the M-P-R items consistently measure the underlying construct dimensions (developmental abilities) of the intended task, and are relatively free from the influence of extraneous variables such as ethnic differences. In other words, the items appear to measure developmental abilities directly, without the interference of "nuisance" variables. Thus, content-related evidence of validity was established by a combination of careful item-response theory analysis; item selection or item development based on review of the literature; scaling verification (to establish consistency with development theory); expert review, and empirical studies of internal consistency.

Criterion-Related Evidence of Validity

The word "criterion" is typically defined as an important societal indicator or outcome such as "success" (e.g., reaching a sales quota, or acceptable performance as measured by grades in graduate school) or "special status" (e.g., developmental delay, or "qualified for special-education services"). Criterion-related evidence of validity examines the statistical relationships between test score interpretations (e.g., cutting scores on the DI scale) and important societal criteria. Evidence of classification accuracy of M-P-R scores, often considered criterion-related evidence, is presented in the section on Consequential Validity in this chapter. The correlations between M-P-R scores and various criterion tests will be discussed first, followed by statistical comparisons of various atypical groups. Finally, a section on consequential validity evidence follows.

Correlation with the Bayley Scales of Infant Development Second Edition

A sample of 34 children, ages 1 to 39 months, was administered both the Bayley Scales (BSID-2) and the M-P-R Developmental Scales. The sample included 44% females and 56% males with a mixture of geographic and ethnic backgrounds. Caucasians composed 44% of the sample, 6% were African-American and 50% were Hispanic. The sample was composed of all typical (normative) cases. Parental education level was 28% less than high school (HS), 22% HS or GED, and 50% post-HS.

Figure 9.1 presents the means, standard deviations and correlations between the main Bayley score (the Mental Scale) and M-P-R scores. The correlation between the Mental Scale and the overall cognitive composite score (Developmental Index) of the M-P-R is high, .92, as are the correlations with the Fine Motor, Receptive Language, Expressive Language and Memory scores, despite the small sample. For the Bayley Motor Scale, only the correlation with gross motor is high, which is what would be predicted. Because the Bayley is the most widely used developmental instrument in North America, the correlations provide strong criterion-related evidence of validity.

Figure 9.1 Means, Standard Deviations and Standard Score Correlations between the Bayley Scales of Infant Development (2nd Edition) Mental Scale and M-P-R Cognitive Battery Scores

			Correlations, M-P-R Rasch to Bayley Raw	
Score	Mean	SD	Mental	Motor
M-P-R Scores				
Developmental Index	102.32	12.27	.92	.57
Cognitive	101.70	10.52	.76	.32
Fine Motor	103.95	14.74	.86	.49
Receptive Language	98.74	10.58	.92	.61
Expressive Language	103.00	15.45	.98	.36
Memory	95.35	12.58	.85	.76
Speed of Processing	104.23	12.69	.48	.07
Visual Motor	100.62	11.12	.65	.23
Gross Motor	103.92	15.32	.51	.81
Social-Emotional Dev.	92.53	22.51	.79	.54
Self-Help/Adaptive	100.49	17.53	.56	.31
Bayley Mental Scale	99.60	19.59		
Bayley Motor	99.70	12.70		

Note: Number of subjects = 24 Mental Scale (correlations with Expressive Language and Parent Report Scales are lower)

When M-P-R and Bayley Mental Scale Standard Scores were correlated, overall the four cognitive domain correlations were strong. The highest correlation (other than the Developmental Index with its high item number) was the receptive language scale. This was somewhat surprising as the Bayley is not extremely verbally based. An additional study was conducted for 17 children from the Bayley/M-P-R sample who were given the Motor Scale of the Bayley. The Bayley Motor Scale correlated .81

with the M-P-R Gross Motor total score. Means of the two scores showed a slightly higher mean for the M-P-R score (99.7 for the Bayley Motor Scale and 103.92 for the M-P-R Gross Motor total) with slightly higher standard deviations for the M-P-R (Bayley 12.7 and M-P-R 15.32). Generally the correlations were as would be expected with the highest correlation with the gross motor scale of the M-P-R (.81) and moderately high correlation for fine motor (.49). One can speculate that the high receptive language correlation (.61) was due to the verbal instructions for motor performance, though the high correlation with memory is unclear.

Correlation with the Leiter International Performance Scale—Revised

A sample of 40 children, ages 37 to 78 months (median 58 months), were given both the Leiter-R Brief IQ (4 subtests of the Visualization and Reasoning Battery) and the M-P-R Cognitive Battery. The sample included 56% females, 44% males, and a mixture of geographic and ethnic backgrounds. Caucasians composed 81% of the sample and the rest (19%) were African-American. The sample was all typical (normative) cases and had parental education levels of 19% less than high school (HS), 19% HS or GED, and 62% post-HS.

All were significant beyond .01 level except Social-Emotional Development and Self Help/Adaptive standard score correlations.

Figure 9.2 Means, Standard Deviations and Correlations Between the Leiter-R Brief IQ and M-P-R Cognitive Battery Scores

			Correlation with	
			Leiter-R	Standard
Score	Mean	SD	Rasch	Score
M-P-R Scores				
Developmental Index	97.60	15.40	.97	.94
Cognitive	98.83	16.42	.94	.89
Fine Motor	98.61	15.43	.96	.86
Receptive Language	98.06	14.98	.95	.85
Memory	98.93	12.31	.91	.75
Speed of Processing	101.75	13.41	.93	.83
Visual Motor	96.04	15.62	.96	.84
Language Total	103.06	16.81	.84	.84
Expressive Language	98.81	16.12	.71	.50
Gross Motor Total	107.08	12.69	.70	.53
Social-Emotional Developmental Total	99.08	15.02	.76	.48
Self-Help/Adaptive Total	107.17	12.18	.75	.34
Leiter-R Brief IQ	103.25	11.70		

Figure 9.2 presents the means, standard deviations and correlations between Leiter-R Brief IQ score and M-P-R scores. The correlation between the Leiter-R Brief IQ and the Developmental Index (DI) of M-P-R is quite high (.97 and .94). The Leiter-R score is about 5 points higher than the M-P-R DI score, which may be due to the exclusive non-verbal (and cognitive) content of the Leiter-R. As an indicator of construct validity, note that the correlation of Leiter-R and M-P-R Receptive Language (which requires only nonverbal item responses) is higher than its correlation with the M-P-R Expressive Language scale, or the Language Total score.

Other Correlative Studies with Cognitive Batteries

A sample of 45 children was tested on both the M-P-R and selected tasks from the Stanford-Binet Intelligence Scale—Fifth Edition (SB5, Roid, 2003). The Abbreviated Battery of SB5 was administered by giving the first two subtests, Object-Series/Matrices and Vocabulary in conjunction with the M-P-R. The SB5 results were converted to "change sensitive scores" (CSS), the Rasch-based scores similar to the M-P-R growth scores.

All of the correlations between the scales were significant beyond the .01 level. Correlations with the SB5 for two age groups are in the range (.76 - .90) with Gross Motor and Parent Reports slightly lower. often observed in correlational studies across cognitive batteries (Roid & Miller, 1997). Another typical pattern is the higher correlation of receptive language scales as compared to expressive language scales (Roid, 2003). This data showed that there was similarity in the general pool of cognitive ability tapped by both tests.

Demographics and Mean Scores of Special Groups

There were a total of 6 special criterion groups examined with the M-P-R and the demographics of each group are described below. For each child in each group, examiners obtained documentation from parents, guardians, teachers, or officials in local agencies to verify the diagnosis or group membership. In nearly all cases, scores from nationally standardized instruments (e.g., published IQ, motor skills, and language tests) were available along with agency placements or DSM-IV criteria for each child. Informed consent was obtained for all parents prior to testing. Following the demographic description of each subsample, tables of mean scores for the Cognitive and Gross Motor Batteries are presented showing the mean profiles of each group. A further study of premature

Figure 9.3 Correlations for 2 and 3 Year Olds (Row 1), 4 to 6 Year Olds (Row 2) and M-P-R Standard Score Means and Standard Deviations for total Sample Between M-P-R Growth Scores and SB-5 Abbreviated Battery Change-Sensitive Score

	Scales											
Value Type	Dev. Index	Cog.	Fine Motor	Recept. Lang.	Visual Motor	Speed	Memory	Expr. Lang.	Total Lang.	Gross Motor	Social Emot.	Self-Help
r-2 & 3	.80	.81	.83	.75	.74	.87	.83	.60	.85	.57	.63	.74
r-4 & up	.86	.90	.78	.66	.74	.87	.83	.78	.72	.50	.82	.81
Mean	104.4	104.9	106.8	104.2	105.5	103.4	103.9	103.9	102.1	96.3	102.7	104.0
SD	13.8	12.0	12.1	13.5	11.4	14.1	14.1	16.5	16.4	14.5	17.4	13.6

infants is presented in the section "Construct-Related Evidence of Validity."

1. **Significant Cognitive Delay (Mild, Moderate or Severe Mental Retardation)**
   The M-P-R was administered to 32 children who were diagnosed with severe cognitive delay or mental retardation. Their age ranged from 18 to 70 months (median = 56). The sample was made up of 34% female and 66% male children, with 16% of the parents having less than a high school education, 59% had a high school diploma or a GED, and 25% of the parents received an education beyond high school. These children were divided among all four regions of the U.S., with most (78%) living in the West. The ethnic diversity among the participants consisted of 53% Caucasian, 9% African-American, 6% Asian, and 31% Hispanic descent.

2. **Premature Infants**
   The M-P-R was administered to 39 children who were born 37 weeks or less gestational age, as reported by the parents who were informed of the prematurity by their physicians. Their age ranged from 2 to 35 months (median = 11). The sample consisted of 36% female and 64% male children, with 23% of parents having a high school diploma or a GED, and 77% having an education beyond high school. These children were divided among all four regions of the U.S., with the highest percentage (49%) from the South. The ethnic diversity among the participants consisted of 56% Caucasian, 10% African-American, 3% Asian, and 31% Hispanic descent.

3. **Severe Speech/Language Delay**
   The M-P-R was administered to 43 children with documented delays in either speech or language development. Their age ranged from 24 to 73 months (median = 47 months). The sample was made up of 35% female and 65% male children, with 9% of their parents receiving less than a high school education, 33% having a high school diploma or a GED, and 58% receiving an education beyond high school. These children were divided regionally, with 2% living in the Northeast, 33% living in the Midwest, 49% residing in the South, and 16% living in the West. The sample included 70% Caucasian and 12% African-American racial origin as well as 18% Hispanic ethnicity.

4. **Deafness or Severe Hard-of-Hearing Conditions**
   There were 18 children with deafness or severe hearing difficulties. The sample included children ages 10 to 78 months (median 49 months) with 44% females and 56% males. Parents of these children were diverse in educational background, with 6% of the parents having less than a high school education, 50% having high school diplomas or GED, and 44% of having education beyond high school. Geographically, there were 6% from the Midwest, 67% from the South, and 27% from the West. Of the 18 children, 72% were of Caucasian origin, 11% African-American origin, and 17% Hispanic ethnicity.

5. **Severe Motor Delay or Deviation**
   The M-P-R was administered to a small sample of 15 children who had severe motor delays or deviations (Cerebral Palsy, etc.). Their age ranged from 10 to 77 months (median = 50 months). The sample consisted of 60% female and 40% male children, with 20% of parents having a less than high school education, 33% with high school diploma or a GED, and 47% having an education beyond high school. These children were divided among three regions of the U.S. (excluding the Northeast), with the highest percentage (60%) from the West. The ethnic diversity among the participants consisted of 73% Caucasian non-Hispanic and 27% Hispanic origin.
6. **Autistic Spectrum Disorders**
   There were 14 children with well-documented diagnoses of autism or autistic spectrum disorders. The sample included children ages 36 to 75 months

(median 53 months) with 29% females and 71% males. Parents of these children were diverse in educational background, with 21% of the parents having less than a high school education, 29% having high school diplomas or GED, and 50% of having education beyond high school. Geographically, there were 29% from the Midwest, 21% from the South, and 50% from the West. Of the 14 children, 36% were of Caucasian origin, 14% African-American, 14% Asian, and 36% Hispanic ethnicity.

Figure 9.4 presents the mean standard scores for each of the scales completed by the special-group samples described above. Interesting patterns of means are found, but the reader is cautioned to be reserved in interpreting patterns, due to the small N size of some of the groups, and due to the mixed nature of age distributions between groups. The means of the cognitive delay group are uniformly low, as expected, and those of the deafness/hard-of-hearing group and speech/language delay group are low on Receptive Language, as expected. Below average scores in the speech delay. motor-delay, and autism groups may be due to histories of learning difficulties and/or the presence of neurological and multiple-handicapping conditions in some instances. For most scores, the standard deviation (SD) in these special groups exceeds the normative SD of 15 because of the presence of some severe delays or disabilities. Variations in SD across small samples are often due to chance sampling fluctuations and should be interpreted with caution.

Evidence of Consequential Validity

The concept of consequential validity was introduced by Messick (1980) to emphasize the social consequences of using tests to make decisions about people and institutional usage of specific test interpretations. Social consequences include the effects on children who are diagnosed with conditions that may have pejorative labels (e.g., mental retardation) and the effects of using particular score levels or score differences to place groups of children in special education programs. Readers should note that these consequences are due to certain uses, interpretations, or regulations concerning test scores, and not necessarily due to the characteristics of the test scores themselves. In other words, evidence of consequential validity must ultimately be collected by examiners and local-agencies that use the M-P-R, and is often beyond the control of the test developer, authors or publisher. Specifically, the developers can make recommendations on test use and give interpretive cautions, but actual uses of the test in the field may vary greatly from those recommendations or cautions. Users of M-P-R have an ethical responsibility to study the effects of using particular score interpretations at the local level; thus contributing to evidence of consequential validity.

The Merrill-Palmer has traditionally been used for important decisions related to the assessment of children in special education and children who show developmental delays in cognitive, motor, or language abilities. Following ethical guidelines (e.g, American

Figure 9.4 M-P-R Standard Score Means and Standard Deviations (SD) for Special Groups

Developmental Group	Means (SD) for Each Scale							
	DI	C	FM	RL	M	S	VM	GM*
1. Cognitive Delay	42.0	49.6	47.0	——	53.6	60.3	48.9	54.3
	(25.7)	(26.4)	(26.5)		(26.0)	(24.1)	(22.7)	(26.3)
2. Premature Infants	93.5	94.2	90.8	——	——	——	92.1	90.9
	(22.5)	(22.3)	(19.6)				(18.0)	(24.0)
3. Speech Delay	80.8	85.8	87.2	76.4	87.3	86.5	88.1	90.6
	(20.3)	(22.7)	(21.5)	(15.2)	(21.1)	(19.0)	(20.1)	(22.0)
4. Deaf/Hard-of-Hearing	84.7	92.5	94.4	72.8	95.4	93.4	95.5	83.9
	(20.5)	(21.9)	(21.0)	(19.5)	(20.1)	(18.9)	(21.6)	(25.4)
5. Motor Delay	70.6	76.5	72.3	73.4	73.9	80.6	74,3	70.9
	(32.6)	(35.8)	(33.9)	(25.4)	(31.0)	(38.8)	(33.3)	(31.3)
6. Autistic Spectrum	52.5	58.7	58.9	54.2	68.3	69.4	57.1	75.5
	(24.9)	(30.7)	(26.3)	(15.6)	(22.9)	(23.2)	(23.2)	(15.0)

*DI = Developmental Index, C = Cognitive, FM = Fine Motor, RL = Receptive Language, M = Memory, S = Speed, VM = Visual Memory, GM = Gross Motor

Note. For descriptions of sample see the section "Demographics and Mean Scores of Special Groups."

Psychological Association, 2002), test scores such as those from M-P-R should never be used in isolation to make important decisions about children. Multiple sources of data, including multiple tests, observations by parents and teachers, and meetings between examiners and parents should all be employed in the pursuit of accurate decisions about children. However, because test scores are often perceived as important elements of the decision process, studies of their accuracy remain important. Thus, this section presents studies of the accuracy of certain M-P-R score usage on the classification of children into special groups, and the consequences of measurement error on that classification.

Classification Accuracy of M-P-R Scores for Various Decisions

Classification accuracy is defined as the number of correct classifications identified by a specific "cutpoint" on a test scale (e.g., an IQ less than 70 for mental retardation) or other measurement scale. Correct classifications ("hits") include both "positive" classifications (as in the medical meaning of identifying a person "positive" for an illness) and "negative" classifications (as in the absence of illness). Classification research relies on the presence of data on two kinds of previously identified subjects: "typicals" who have an absence of any delay or disability condition, and "atypicals" who have a documented delays or disabilities that have been verified by tests or criteria independent of the M-P-R (e.g., cognitive delay established by multiple criteria including low adaptive-behavior level and Developmental Index less than 70). As is done in medical research, the following categories of classification statistics were calculated:

- *Total Hit Rate* - The total percentage of children correctly classified, whether correctly classified as atypical or typical children.
- *Specificity* - The percentage of children previously identified as typical who are correctly classified.
- *Sensitivity* - The percentage of children previously identified as atypical who are correctly classified.
- *False Negatives* - The total percentage of children who are truly atypical, but incorrectly classified as typical.
- *False Positives* - The total percentage of children who are truly typical, but incorrectly classified as atypical.

The seriousness and social consequences of false identification often varies with different decisions in different settings. If a social stigma is connected with the decision (e.g., mental retardation), the false positive error can have lasting ramifications and possible negative social consequences. If the decision is either positive or reversible, such as placement in a temporary home treatment program, false positive errors may be less problematic and, in fact, the false negative (not identifying the child who needs treatment) may be more serious.

As part of the standardization study for M-P-R, documented cases of cognitive delay, premature birth, and speech or language delay were identified and tested. When these atypical cases were contrasted with matching typical cases from the M-P-R normative sample, classification studies were possible and are reported in the sections that follow. First, classification statistics will be presented for the identification of cognitive delay (mental retardation), delays due to prematurity, and severe speech/language delay.

Identification of Cognitive Delay

Although M-P-R scores should NOT be used as the sole criteria for identifying cognitive delay, studying their accuracy for such classifications remains important, as discussed previously. Classification statistics were derived by forming two groups of children: 32 atypical cases previously identified with cognitive delay (see section "Significant Cognitive Delay" in the previous "Special Groups" section), and a selected sample of 448 typical children from the M-P-R normative sample. The typical sample was restricted in age range (37 to 70 months) to match the age range of the atypical sample, but still showed the characteristics of a nationally representative sample of children as described in Chapter 7. The section on "Demographics and Mean Scores of Special Groups" in the previous part of this chapter presented the demographics of the atypical sample for these analyses. The M-P-R score cutpoint was set between scores of 69 and 70 on the Cognitive, Fine Motor, and Receptive Language scales, using the cutpoint usually employed by school districts for

Figure 9.5 Diagram of Classification Accuracy Categories

Identification by M-P-R-R Score			
	Typical	Atypical	
Real Typical	**A** True Negative	**B** False Positive	**A** + **B**
Real Atypical	**C** False Negative	**D** True Positive	**C** + **D**
			Total **A** + **B** + **C** + **D**

(**A** + **B**) / Total = Total Hit Rate
A / (**A** + **B**) = Specificity
D / (**C** + **D**) = Sensitivity

Figure 9.6 Percentages of Classification for M-P-R Cognitive Battery Scores Using Cutpoints to Identify Cognitive Delay (Mental Retardation), Delays Due to Prematurity, and Speech/Language Delays with the M-P-R Cognitive Battery

Score	Cutpoint	Total Hits	Specificity	Sensitivity	False Positive	False Negative
Cognitive Delay						
Cognitive Scale	(< 70)	98.5	98.8	91.7	1.1	0.4
Fine Motor Scale	(< 70)	99.3	99.5	95.8	0.4	0.2
Receptive Language	(< 70)	99.6	99.8	95.7	0.2	0.2
Delay Due to Prematurity						
Fine Motor Scale	(< 86)	78.3	81.5	39.1	17.1	4.6
Speech/Language Delay						
Receptive Language	(< 85)	83.9	84.8	69.0	14.3	1.8
Receptive Language	(< 80)	89.6	91.5	57.1	8.0	2.4

the identification of mental retardation (two standard deviations below the mean of 100). **Figure 9.6** shows the number of subjects, and the classification statistics for the identification of cognitive delay.

Figure 9.6 shows an excellent level of classification accuracy for the identification of cognitive delay (mild or greater mental retardation) by using a cutpoint of 70 on any of the three scales listed. Very few subjects were incorrectly classified (only 7, 3, and 2 children for the three scales, respectively, out of the total of 480 cases in the analysis). In comparison to the other classifications shown in Figure 9.6,all percentages are quite excellent including the false positive (typical cases incorrectly called atypical) and false negative (atypical cases incorrectly called typical) percentages.

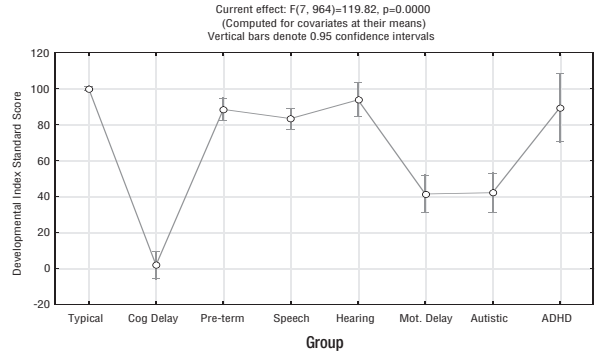
Identification of Delay in Premature (Pre-term) Infants

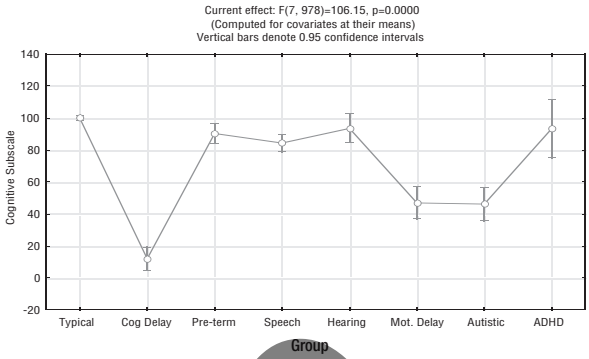
Because premature infants are often found to have delays in cognitive, fine motor, and language abilities (until some of them "catch up" later in childhood), the accuracy of using M-P-R with this special population was studied. Unlike cognitive delay, one would not expect that a test would have a high sensitivity in identifying premature children. The reason is that the majority of premature children do catch up with age peers and we are just identifying those that do not catch up. Full data and background information was available for 39 atypical (premature) cases and 281 typical cases in the same age range. See the section on "Demographics and Mean Scores of Special Groups" for a description of the atypical sample. The typical cases were drawn from the M-P-R normative sample, ages 2 to 35 months of age and mother's education high-school or greater, to match the characteristics of the atypical sample. **Figure 9.6** shows a lower degree of classification accuracy for the identification of delays in premature infants as compared to the identification of cognitive delay. The Fine Motor scale is used in this classification because it had the lowest mean score in the special group of documented prematurity (see **Figure 9.4**). After experimenting with various cutpoints, **Figure 9.6** shows the recommended cutpoint between 85 and 86 on the Fine Motor scale. The sensitivity (identification accuracy within the premature infant group) was particularly low (39%) and the percentage of false positive errors were quite high (17%). Thus, M-P-R should be used with caution if the sole purpose is to use standard scores with a cut-point methodology to screen premature infants. Better uses of the M-P-R would be for annual assessments of premature infants to track progress on the M-P-R growth scales.

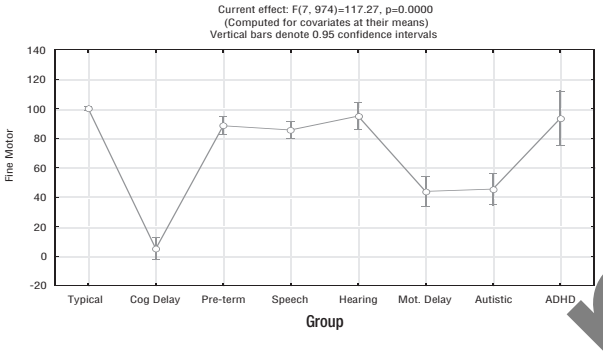
Identification of Speech/Language Delay

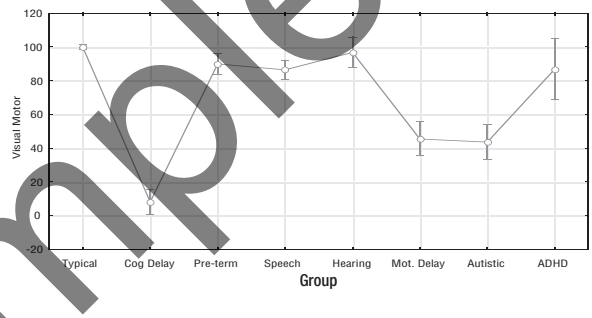
Given the nature of speech and language delays, an M-P-R language scale was chosen for the analyses. The largest number of subjects was available for the Receptive Language (RL) scale (698 typical and 43 atypical cases), as compared to the Expressive or Total Language scales; hence RL scores were used in the study reported in Figure 9.6. See the section on "Demographics and Mean Scores of Special Groups" for a description of the 43 cases with speech/language delays. The 698 typical cases were drawn from the M-P-R normative sample, ages 24 to 73 months of age, to match the age range of the atypical sample. Several cutpoints were investigated and those at 80 and 85 produced the best results. The cutpoint at 85 was the more accurate than the cutpoint of 80 within the documented cases of speech/language delay (69.0), although it resulted in a sizeable (14%) rate of false positive errors. The cutpoint at 80 was less accurate in the atypical sample (57.1%) but showed reduced false positive errors (8%).

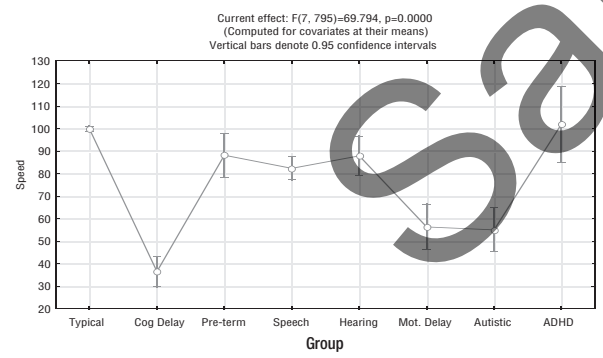
Patterns of Performance Across Different Scales for Special Populations

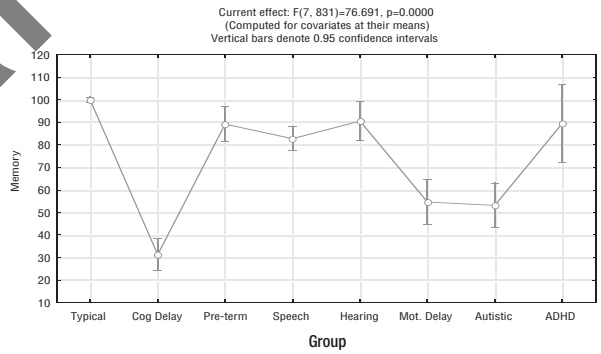


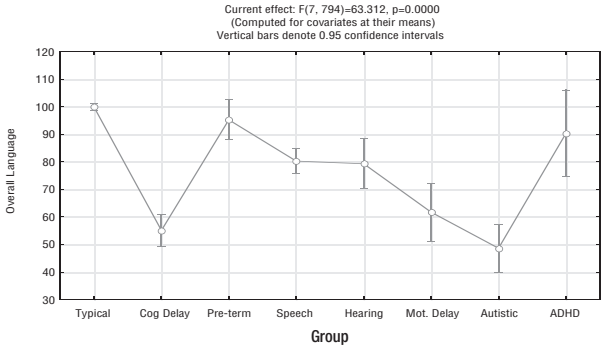


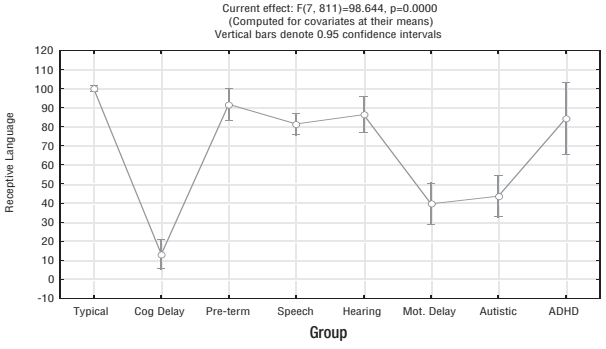


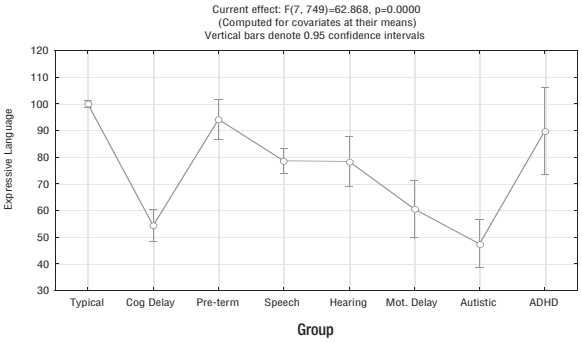


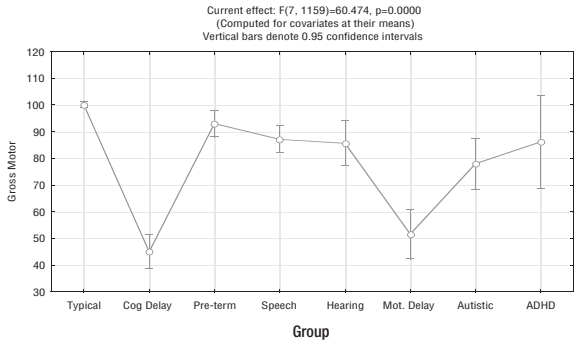


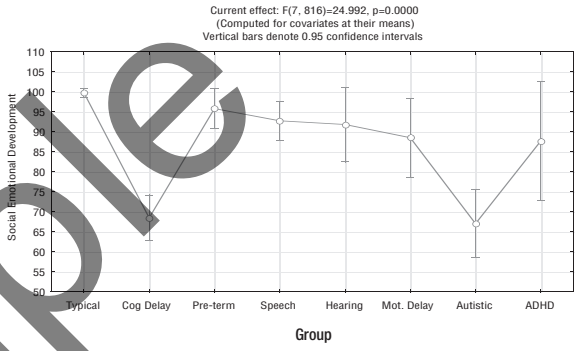












Profiles of Special Populations

One way to examine the classification of populations is through an examination of the patterns of the strengths and deficits of the special populations. Tests should provide results that are consistent with the research data and clinical examination. For example: one would expect that low scores on cognitive tests and depression in other scales would be found in individuals with developmental disabilities/retardation and those with speech delays should have their greatest deficits in the language areas of the assessments. The series of graphs shows the patterns of performance across the different scales for a number of different special populations.

Summary of Classification Results and Recommended Procedures

The user should be professionally cautious when attempting to use the M-P-R scores in isolation in the identification of cognitive delay, delays in premature infants, or speech/language delays. The applications of M-P-R with the clearest, positive consequential validity include use of the cognitive score in identification of cognitive delay. In addition, the M-P-R scores show power to suggest clinical hypotheses such as delay due to prematurity, speech/language delays, and other developmental delays. However, users should carefully study the degree of error shown in Figure 9.6 and take this information into account when attempting to make decisions about children with M-P-R scores. Of all the comparisons shown in Figure 9.6, the use of M-P-R to identify cognitive delay is clearly the most superior and is recommended when employed within the context of multiple sources of data on each child.

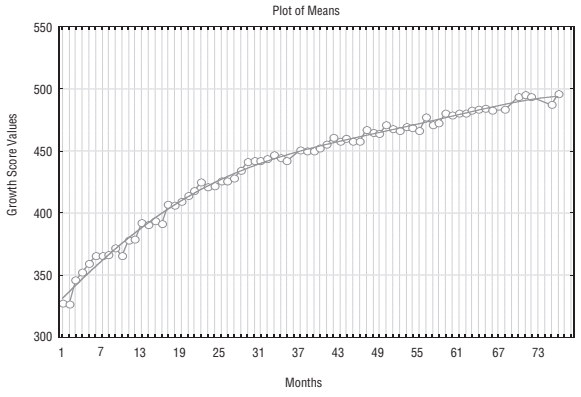
Construct-Related Evidence of Validity

Construct-related evidence of validity is gathered from a variety of correlational and experimental studies aimed at demonstrating that the instrument truly measures the construct intended. Cronbach and Meehl (1955) defined construct validity as a complex of evidence, gained from a lengthy series of research studies, showing that the instrument measures the construct intended. The research must show empirical trends and effects predicted by the construct in relationship to other constructs in a "nomological net" of hypothesized relationships. Types of evidence include age trends, differences induced by experimental intervention, convergent and discriminant correlations with other measures of the same and different construct-measures (Campbell & Fiske, 1959), verification of scale dimensionality, and other studies. Many of these studies were completed on the Tryout and Standardization Editions of the M-P-R and more will be likely to appear in the research literature following the initial publication of the test.

Age Trends

Because cognitive processes are known to mature in conjunction with the development of the brain and neurological functions, measures of cognitive skills typically follow "growth curves." The M-P-R was developed with growth curves in mind and all items in the Pilot and Tryout phases were inspected using percentage-correct data on each item for each age group. Items not showing sensitivity to age were inspected closely and most were discarded unless they served a

Figure 9.7 Line Graph Showing Mean Scores on M-P-R of Developmental Index Growth Scores for Age Groups (1 to 78 Months) with Fitted Quadratic and Cubic Trend Curves



function of introducing or concluding a task. Details of item development and the cycles of revision in the M-P-R project were provided in Chapter 7. A quick examination of the normative tables in Appendices A will show that the median raw scores on each of the tasks increased systematically across age groups in the M-P-R normative sample. These consistent age trends reflect the growth curves embedded in the raw data of the M-P-R standardization.

As shown in Chapter 6, the M-P-R growth scales provide the clearest example of the developmental trend in the data — the scale location of each age group. To document the age-trend in the standardization sample, mean scores for various age groups were calculated for the overall Developmental Index growth score. **Figure 9.7** shows the graph of the means across age groups. The age groups in Figure 9.7 were defined as the children at each month of age, where age-group samples ranged from only 3 cases to as many as 20 (median 11 cases). Various smooth curves were fitted to the age-group data in **Figure 9.7**, with quadratic and cubic trends showing excellent fit (multiple squared correlation between means and age in months of .99).

Because the Rasch-based growth scale has some excellent equal-interval properties, it is very useful for showing growth trends across age groups. The trend in the progression of Growth-Scale means was fitted to quadratic and cubic regression lines, as shown in **Figure 9.7**, producing a huge F-value to test the significance of the regression, with a multiple squared correlations of .99. The growth curve shows an increasing upward trend through the age range, with some slowing of growth after 36 months of age.

Cross-Battery Age Equivalence on the Growth Scale

Evidence of construct validity for the M-P-R Developmental Index (Growth score conversion) is provided by a comparison of age equivalent values across test batteries. Rasch-based scores similar to the M-P-R Growth Scale have been employed on a number of other published tests. Beginning with the development of the W-scale (Woodcock & Dahl, 1973), all of the editions of the Woodcock-Johnson Psychoeducational Battery (Woodcock & Johnson, 1977; 1989; Woodcock, McGrew & Mather, 2001) have used the Rasch-based W-scale centered at 500 for age 10 with an expansion factor of 9.1024, as used on the M-P-R. The Leiter-R (Roid & Miller, 1997) employed the same scaling for its Rasch-based scores, called "Growth Scales" as in the M-P-R. The Stanford-Binet Intelligence Scales, Fifth Edition (SB5, Roid, 2003a) also employed Rasch-based scores in the same metric, calling them "change sensitive scores" (CSS). In each case, the Rasch-based scores were used to obtain age equivalent values. **Figure 9.8** shows the age equivalent values for various editions of the published tests described in this section. The nonverbal tests such as Leiter-R relate well to the nonverbal portions of other tests (e.g., the SB5 Nonverbal CSS) and are, as expected, slightly lower in developmental age equivalent as compared to the more verbal tests or scales. The M-P-R age equivalents are particularly consistent with the Woodcock-Johnson Revised edition; thus demonstrating construct validity for the M-P-R Developmental Index scores.

Figure 9.8 Rasch-based Age Equivalence Values for M-P-R and Several Other Tests

	Growth score values				
Age in Months	M-P-R DI	Leiter-R	SB5-Full	SB5-NV	WJ-R
6	358	—-	—-	—-	—-
12	386	—-	—-	—-	—-
24	425	425	430	425	425
36	449	440	447	441	447
48	464	453	460	453	463
60	478	464	470	465	474

Note: M-P-R DI is the M-P-R Developmental Index overall growth score conversion to age equivalents. Leiter-R is the Brief-IQ Growth score values. SB5-Full is the Full Scale IQ equivalent. SB5-NV is the Nonverbal IQ equivalent. WJ-R values are age equivalents (e.g., 3 years 0 months for 36 months) on the 1989 edition for which printed age equivalent tables were provided.

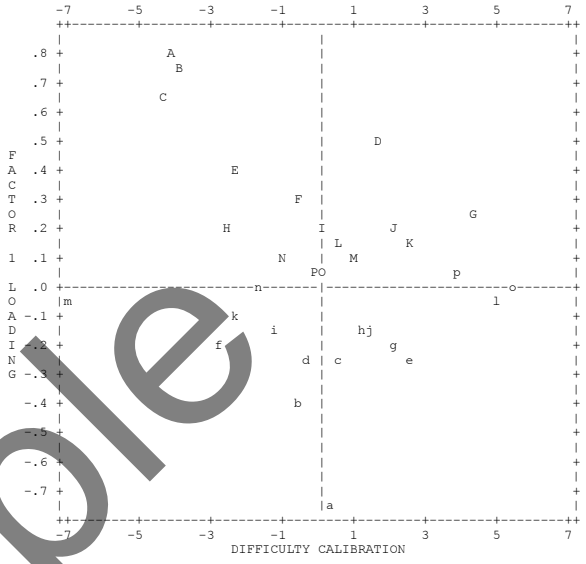
Factor Analyses of the Tryout Edition

As presented previously in Chapter 7, a series of exploratory factor analyses were conducted on the typical children in the tryout sample, using the SPSS factor analysis routines. Multiple analyses were conducted on data from Levels 3, 4, and 5 where measures of a wide array of cognitive abilities were available from the data. Total scores from tasks representing cognitive, fine motor, and language skills were standardized using means and standard deviations of 6-month age groups (e.g., 36-41, 42-47, 48-53 months, etc.) and converting the resulting scores to a scaled-score metric (mean 10, SD 3). A range of principal components analyses, principal axis and maximum likelihood factor analyses were conducted using both varimax (orthogonal) and promax (oblique) factor rotations. Subsamples of approximately 80 subjects at each of three age levels (3, 4, and 5) were examined for factor structure among the tasks at their age-appropriate level. Criteria for factor extraction were eigenvalues greater than 1.0 and rotation to simple structure using squared multiple correlations in the diagonal for the principal axis and maximum likelihood methods. Factor identification and naming was based on significant factor loadings (greater than .45 and highest loading in each row of the factor matrix). At Level 3, three factors were identified accounting for about 70% of the variance in the scores: 1) a mixture of fluid reasoning and language-based crystallized knowledge, 2) a visual-spatial factor composed of puzzle and block-stacking tasks, and 3) a timed/motor performance factor. At Level 4, four factors were found: 1) a mixed crystallized (language) and short-term memory factor, 2) a visualization factor (hidden figures), 3) speed of processing (time to assemble puzzles), and 4) a fine-motor factor on puzzle assembly accuracy. At Level 5, fluid reasoning, crystallized (picture vocabulary), and fine motor (design copying) factors emerged. This series of factor analyses established that the new M-P-R was multi-factorial in composition, showed differentiation among the cognitive abilities measured, and promised to provide important subscores for memory, visualization, fine motor, and language.

Principal Components Residual Analyses to Establish Unidimensionality in the Rasch Scales

As part of the item-response theory analyses of each of the major scales in the M-P-R, principal components analyses of the residual variance among the items were conducted for all the M-P-R scales. Using the WINSTEPS program (Lineacre, 2002), items with residual variance loadings were identified from Rasch analyses of

Figure 9.9 Principal Components (Standardized Residual) Factor Plot of M-P-R Fine Motor Items from the Standardization Edition Showing Locations of Items on the Factor by Difficulty Dimensions



the M-P-R Standardization Edition. These items showed evidence of measuring dimensions outside the unidimensional factor underlying the Rasch scaling of the majority of items. For example, approximately 10 items measuring dimensions other than cognition were identified in scaling analyses of the 170 Cognitive scale standardization items by conducting principal components analyses of the residual variance remaining after extraction of the major cognitive dimension underlying the 170 items. **Figure 9.9** shows an example of the standardized residual factor plots printed for each M-P-R scale. In the top left corner of **Figure 9.9** are items labeled A, B, and C that showed evidence of loading on a residual factor separate from the main measurement dimension (fine motor ability) underlying the rest of the items. Items such as A, B, and C were removed from the M-P-R for the final, published edition to create greater unidimensionality to the M-P-R scale. (See Chapter 7 for more details of the analyses conducted on the M-P-R Standardization Edition leading to the Final Edition.)

Factor Analyses of the Final Edition: Conventional Exploratory Analyses of Scores

When the hypothesized construct dimensions underlying a composite test score (e.g., the Developmental Index of M-P-R) are visible from exploratory factor analysis, construct-related evidence of validity for that

test score is demonstrated. A series of principal-axis factor analyses were performed on the major composite scores of the M-P-R. Employing the conventional method used for published cognitive and developmental tests, analyses were conducted on random samples of typical subjects in the normative sample at various age levels. As recommended in the factor-analysis literature (e.g., Gorsuch, 1983), samples of 100 or more were formed by combining data into three age levels: 13 to 36 months (N = 100), 37 to 54 months (N = 141), and 55 to 78 months (N = 134). Children 12 months of age or less were excluded because the Receptive and Expressive Language scores were not available in this age group (due to the lack of sufficient development of language at these young ages). This strategy allowed the analyses to include all the scores in the M-P-R Developmental Index-Cognitive, Receptive Language, and Fine Motor scores.

The factor analyses employed squared multiple correlations as communality estimates in the diagonal of the correlation matrices and relied on *scree* tests (Cattell, 1966) to assist in the selection of the number of factors. Equamax rotations to simple structure were employed with the principal-axis factor extraction so that the three main factors emerging in the analyses would be treated equally in terms of factor variance. As shown in **Figure 9.10**, results clearly showed the presence of the developmental index factor along with factors labeled Motor/Self-Help and Social/Adaptive. Justification for the combination of Cognitive, Fine Motor, and Receptive Language scores into the Developmental Index are clearly shown by the high factor loadings for these scores on the first factor in **Figure 9.10**. Patterns of factor loadings varied somewhat across age groups, partially due to the different collections of tasks employed at each age and partially due to developmental trends. For example, the low loading of the Gross Motor score on the Motor/Self-Help factor at ages 13 to 36 months may be due to the greater interrelationship among all abilities at those ages (as shown from the loadings on Expressive Language and Cognitive scores at that age level). Also, the Self-Help/Adaptive score is known to measure a diverse collection of skills, shifting from adaptive behaviors in the younger ages to more self-care behaviors in the upper age levels. The moderate and varied relationship of Expressive Language (EL) to the developmental index factor makes intuitive sense, given the relationship.

Note. Small factor loadings of .20 and below are not shown. Loadings in bold print are the hypothesized variables defining each factor. Loadings in parenthesis are less significant, being below .30, and may have occurred due to sampling or measurement error. The number of subjects in each age group was 100, 141, and 134, respectively, with a total sample of 375 children.

Dimensionality of Non-Developmental Scales

As reported in the previous chapter, extensive factor analysis of the factor structures of these rating scales demonstrated their dimensionality. These studies demonstrate that the dimensionality of the sub-scales was reliable and relationships between scales is provided.

Figure 9.10 Factor Loadings of Major M-P-R Scores on the Developmental Index, Motor/Self-Help, and Social/Adaptive Factors across Three Age Levels

	Developmental Index				Motor/Self-Help				Social/Adaptive			
Score	13-36	37-54	55+	All	13-36	37-54	55+	All	13-36	37-54	55+	All
Cognitive	**.86**	**.87**	**.86**	**.89**	.38	.31	(.22)				(.22)	
Fine Motor	**.85**	**.65**	**.81**	**.78**	(.27)	.38	(.26)	.31			(.21)	
Receptive Language	**.89**	**.75**	**.84**	**.81**	(.29)					(.23)	(.25)	
Expressive Language	(.28)	.37	.62	.48	.60	.41	.27				.41	.32
Gross Motor						**.44**	**.68**	**.43**	.37		(.22)	
Social-Emotional			(.21)		.34		(.26)		**.72**	**.84**	**.80**	**.75**
Self-Help/ Adaptive						.68	.45	.62	**.75**	**.64**	**.57**	**.58**

Note: Small factor loadings of .20 and below are not shown. Loadings in bold print are the hypothesized variables defining each factor. Loadings in parenthesis are less significant, being below .30, and may have occurred due to sampling or measurement error. The number of subjects in each age group was 100, 141, and 134, respectively, with a total sample of 375 children.

Unidimensionality of the Other M-P-R Scales.
Extensive Rasch scaling analyses were conducted on the Social-Emotional and Self-Help/Adaptive scales of the M-P-R. To establish unidimensionality of these scales, principal component analyses of the residual variance among the items were completed, similar to the analysis reported in **Figure 9.9** earlier in this chapter. Any items showing deviation from the central variance among the items was flagged and considered for deletion from the scales, as demonstrated for the outlier items in **Figure 9.9**. These analyses provided evidence of construct validity for these additional M-P-R scales.

Summary

Based on all the analyses presented in this chapter, the M-P-R shows consistent evidence of validity from content-analysis studies with extensive item analysis data. Also, extensive criterion-related studies showed excellent results for concurrent correlations and the classification accuracy in identifying cognitive delay. Finally, several analyses of age trends, consistency of age equivalents across published test batteries, dimensionality of scores, and other data showed strong evidence of construct-related validity. Researchers are encouraged to continue studies of construct-related issues, particularly with future cross-battery factor analyses.