September 2023

Bayley–III GSV Technical Supplement

Mark H. Daniel, PhD

Abstract

This report provides a technical description of the Rasch analyses used for each of Bayley-III subtests to develop growth scale values (GSVs). Using joint maximum likelihood estimation, Rasch analysis was done from the results of the standardization sample of 1,700 children between the ages of 1 month and 42 months. Both person ability and item difficulty were identified in the sample. The Rasch logit values that represent ability and difficulty were linearly transformed into GSVs. The change in item success rate as GSV changes is identified and diagnostic information about the Rasch calibration, such as item fit, local independence, and dimensionality, is discussed.

Cite this document as:

Daniel, M. H. (2023). *Bayley–III GSV technical supplement* [GSV Technical Report]. NCS Pearson.

Product Number 66781 v.B

Contents

Test Content and Item Type

The *Bayley Scales of Infant and Toddler Development* (3rd ed.; Bayley–III; Bayley, 2006a) measures cognitive, language, motor, social-emotional, and adaptive behavior characteristics of children ages 1 month to 42 months. The first three of these areas are assessed by five individually administered subtests:

- Cognitive (91 items)
- Language
  - Receptive Communication (49 items)
  - Expressive Communication (48 items)
- Motor
  - Fine Motor (66 items)
  - Gross Motor (72 items)

The items on these subtests are dichotomously scored according to the examiner's judgment of whether the child's performance met the specified criteria. A Rasch scaling of each of these subtests was used to develop growth scale values (GSVs) for each subtest. This report provides a technical description of those Rasch analyses and their results.

Social-emotional skills and adaptive behavior are assessed through a caregiver questionnaire, which is not discussed in this report.

Rasch Analysis and Applications

Method and Sample

Rasch analysis was done with the Winsteps software program (version 5.1.4; Linacre, 2021a), using joint maximum likelihood estimation. The data for the Rasch analysis came from the standardization sample of 1,700 children ages 1 month to 42 months, which was representative of the U.S. population by sex, race/ethnicity, parent education, and geographic region. Approximately ten percent of the children had a diagnosed condition (Down syndrome, cerebral palsy, pervasive developmental disorder, language impairment, or at-risk for developmental delay) or perinatal event (premature birth, low birth weight, small for gestational age, asphyxiation at birth, or prenatal alcohol exposure). Details may be found in the Bayley–III Technical Manual (Bayley, 2006b).

In general, each child started at an age-based starting point expected to be very easy. If they did not set a "basal" by scoring 1 on each of the first three items, they reverted to the next lower starting point, and repeated this process until setting a basal. For all children, testing stopped when the child scored 0 on five consecutive items. Unadministered items were assigned a score of 1 if they preceded the basal or a score of 0 if they followed the discontinue point. Because there were no missing scores in any of the administered items, the Rasch analysis did not include missing data.

Item Difficulty and Person Ability

In the Rasch model, the probability that a person will succeed on a dichotomously scored item depends on the difference between the person's ability and the item's difficulty. Both ability and difficulty are measured on the same scale whose units are called *logits*. Higher logit values correspond to higher abilities and more difficult items. As ability increases relative to difficulty, the probability of success increases. When person ability and item difficulty are equal, the person has an equal (.50) probability of failing or passing the item. When the item difficulty is within two logits of a person's ability, the person's probability of success is between .12 and .88. Items within this range are moderately difficult for the person and provide more information about their ability than items that are very easy or very difficult.

Figure 1 presents a Wright map for each subtest, showing the frequency distributions of item difficulties along with the distribution of abilities of the children in the calibration sample. By comparing the upper and lower distributions, one can see how many moderately difficult items a child with a given ability is likely to encounter on that subtest. Each Bayley-3 subtest contains item difficulties spanning virtually the entire range of abilities of the children in the standardization sample, meaning that nearly all children in this age range will encounter at least a few moderately difficult items.

Figure 1. Wright Map of Person Abilities and Item Difficulties in Logits, by Subtest

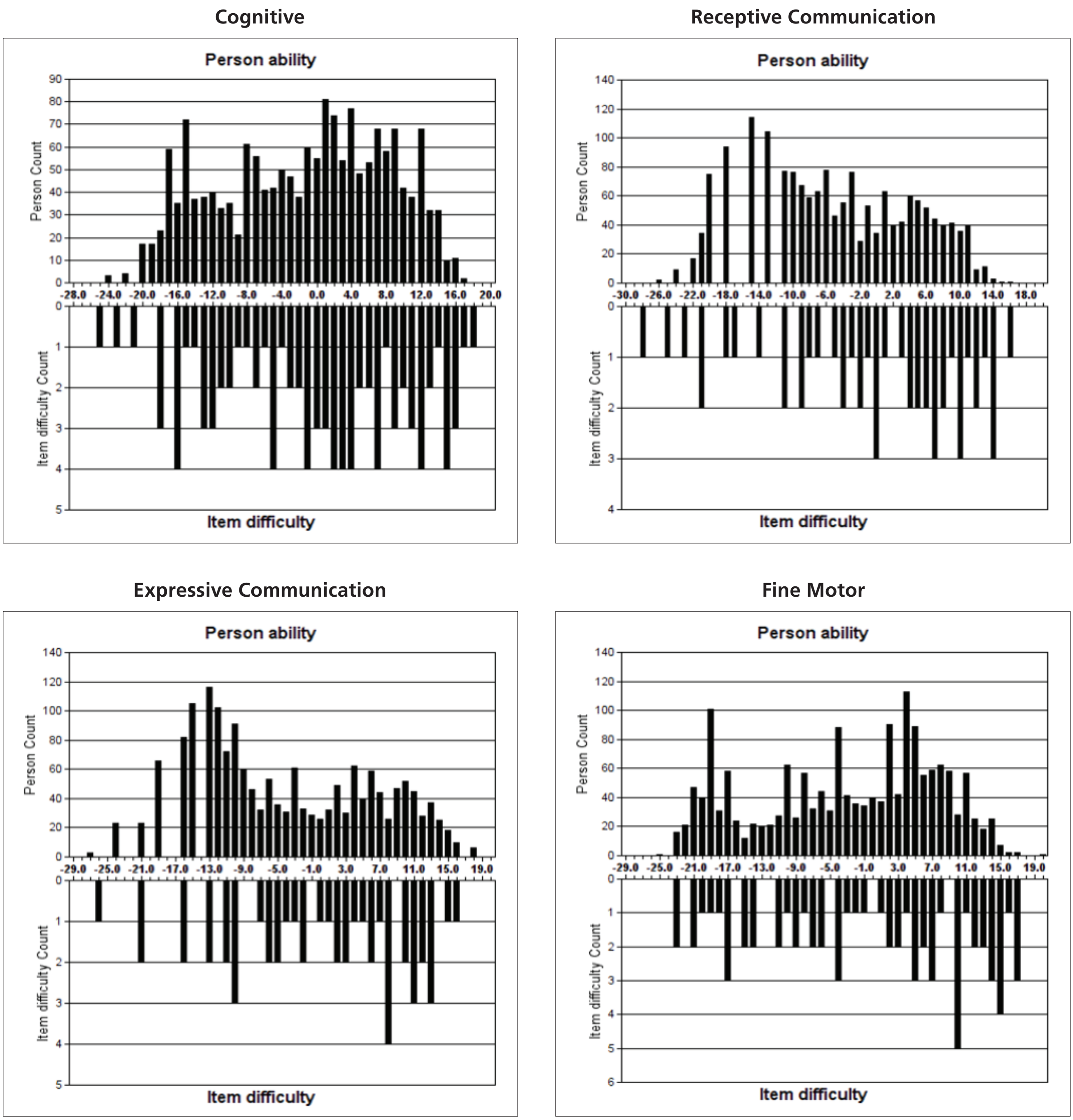
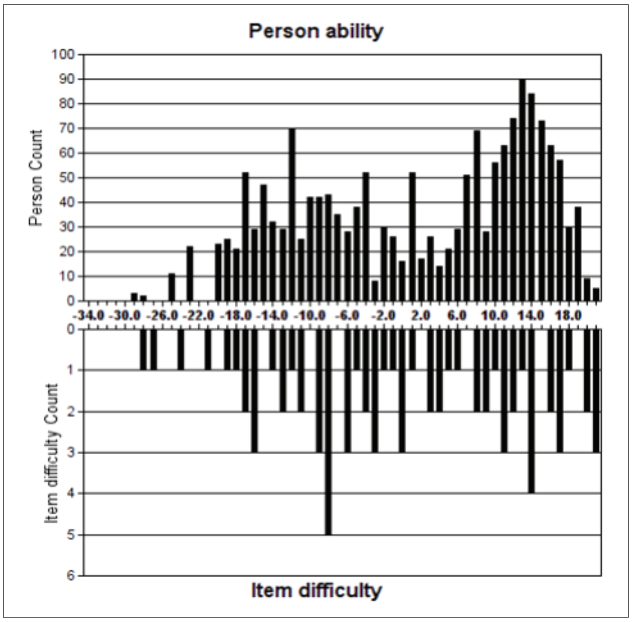


Figure 1. Wright Map of Person Abilities and Item Difficulties in Logits, by Subtest (*continued*)



Transformation of Rasch Ability Scores to GSVs

GSVs are a linear transformation of the logit values of Rasch ability scores: $GSV_i = a \times ability_i + b$. Linear transformation preserves the shape of the ability-score distribution. The coefficient *a* and constant *b* for each Bayley–III subtest, shown in Table 1, were chosen so that GSVs would have a standard deviation of 100 and mean of 500 in the full standardization sample.

Table 1. Parameters for the Linear Transformation of Ability Scores from Logits to GSVs

Subtest	Coefficient (*a*)	Constant (*b*)
Cognitive	10.33	514.1
Receptive Communication	10.65	555.8
Expressive Communication	9.79	540.7
Fine Motor	9.49	529.5
Gross Motor	8.06	488.1

The Appendix presents the GSV corresponding to each raw score on each subtest, and the charts in Figure 2 show the shapes of these relationships. Each chart also includes the frequency distribution of item difficulties in GSV units to illustrate how the density of item difficulties affects the slope of the line: the more items there are in a region of difficulty, the faster the raw score increases as ability increases, because there are more items on which the person can demonstrate their ability.

Figure 2. Raw Score Versus GSV and Frequency Distribution of Item Difficulties, by Subtest

Change in Item Success Probability as GSV Changes

Knowing how a change in GSV affects children's probability of succeeding on a particular test item can be useful for attaching meaning to GSV changes, such as for estimating a minimal clinically important difference (MCID). The relationship between GSV change and change in success probability differs across Bayley–III subtests because they use different coefficients to transform Rasch ability scores to GSVs. Table 2 shows, for each subtest, the probability of item success following various amounts of GSV change, starting from various initial probability values. For example, if a child initially had a probability of .35 of answering a particular Cognitive item correctly, and their GSV increased by 15 points, their new success probability on that item would be .70.

Table 2. Probability of Item Success After Change in GSV, by Subtest and Initial Probability

Cognitive

	Change in GSV												
Initial *p*	**–30**	**–25**	**–20**	**–15**	**–10**	**–5**	**0**	**5**	**10**	**15**	**20**	**25**	**30**
0.99	0.84	0.90	0.93	0.96	0.97	0.98	0.99	0.99					
0.95	0.51	0.63	0.73	0.82	0.88	0.92	0.95	0.97	0.98	0.99	0.99		
0.90	0.33	0.44	0.56	0.68	0.77	0.85	0.90	0.94	0.96	0.97	0.98	0.99	0.99
0.85	0.24	0.34	0.45	0.57	0.68	0.78	0.85	0.90	0.94	0.96	0.98	0.98	0.99
0.80	0.18	0.26	0.37	0.48	0.60	0.71	0.80	0.87	0.91	0.94	0.97	0.98	0.99
0.75	0.14	0.21	0.30	0.41	0.53	0.65	0.75	0.83	0.89	0.93	0.95	0.97	0.98
0.70	0.11	0.17	0.25	0.35	0.47	0.59	0.70	0.79	0.86	0.91	0.94	0.96	0.98
0.65	0.09	0.14	0.21	0.30	0.41	0.53	0.65	0.75	0.83	0.89	0.93	0.95	0.97
0.60	0.08	0.12	0.18	0.26	0.36	0.48	0.60	0.71	0.80	0.87	0.91	0.94	0.96
0.55	0.06	0.10	0.15	0.22	0.32	0.43	0.55	0.66	0.76	0.84	0.89	0.93	0.96
0.50	0.05	0.08	0.13	0.19	0.28	0.38	0.50	0.62	0.72	0.81	0.87	0.92	0.95
0.45	0.04	0.07	0.11	0.16	0.24	0.34	0.45	0.57	0.68	0.78	0.85	0.90	0.94
0.40	0.04	0.06	0.09	0.13	0.20	0.29	0.40	0.52	0.64	0.74	0.82	0.88	0.92
0.35	0.03	0.05	0.07	0.11	0.17	0.25	0.35	0.47	0.59	0.70	0.79	0.86	0.91
0.30	0.02	0.04	0.06	0.09	0.14	0.21	0.30	0.41	0.53	0.65	0.75	0.83	0.89
0.25	0.02	0.03	0.05	0.07	0.11	0.17	0.25	0.35	0.47	0.59	0.70	0.79	0.86
0.20	0.01	0.02	0.03	0.06	0.09	0.13	0.20	0.29	0.40	0.52	0.63	0.74	0.82
0.15	0.01	0.02	0.02	0.04	0.06	0.10	0.15	0.22	0.32	0.43	0.55	0.66	0.76
0.10	0.01	0.01	0.02	0.03	0.04	0.06	0.10	0.15	0.23	0.32	0.44	0.56	0.67
0.05			0.01	0.01	0.02	0.03	0.05	0.08	0.12	0.18	0.27	0.37	0.49
0.01						0.01	0.01	0.02	0.03	0.04	0.07	0.10	0.16

Receptive Communication

	Change in GSV												
Initial *p*	**–30**	**–25**	**–20**	**–15**	**–10**	**–5**	**0**	**5**	**10**	**15**	**20**	**25**	**30**
0.99	0.86	0.90	0.94	0.96	0.97	0.98	0.99	0.99					
0.95	0.53	0.64	0.74	0.82	0.88	0.92	0.95	0.97	0.98	0.99	0.99	0.99	
0.90	0.35	0.46	0.58	0.69	0.78	0.85	0.90	0.94	0.96	0.97	0.98	0.99	0.99
0.85	0.25	0.35	0.46	0.58	0.69	0.78	0.85	0.90	0.94	0.96	0.97	0.98	0.99
0.80	0.19	0.28	0.38	0.49	0.61	0.71	0.80	0.86	0.91	0.94	0.96	0.98	0.99
0.75	0.15	0.22	0.31	0.42	0.54	0.65	0.75	0.83	0.88	0.92	0.95	0.97	0.98
0.70	0.12	0.18	0.26	0.36	0.48	0.59	0.70	0.79	0.86	0.91	0.94	0.96	0.98
0.65	0.10	0.15	0.22	0.31	0.42	0.54	0.65	0.75	0.83	0.88	0.92	0.95	0.97
0.60	0.08	0.13	0.19	0.27	0.37	0.48	0.60	0.71	0.79	0.86	0.91	0.94	0.96
0.55	0.07	0.10	0.16	0.23	0.32	0.43	0.55	0.66	0.76	0.83	0.89	0.93	0.95
0.50	0.06	0.09	0.13	0.20	0.28	0.38	0.50	0.62	0.72	0.80	0.87	0.91	0.94
0.45	0.05	0.07	0.11	0.17	0.24	0.34	0.45	0.57	0.68	0.77	0.84	0.90	0.93
0.40	0.04	0.06	0.09	0.14	0.21	0.29	0.40	0.52	0.63	0.73	0.81	0.87	0.92
0.35	0.03	0.05	0.08	0.12	0.17	0.25	0.35	0.46	0.58	0.69	0.78	0.85	0.90
0.30	0.02	0.04	0.06	0.09	0.14	0.21	0.30	0.41	0.52	0.64	0.74	0.82	0.88
0.25	0.02	0.03	0.05	0.08	0.12	0.17	0.25	0.35	0.46	0.58	0.69	0.78	0.85
0.20	0.01	0.02	0.04	0.06	0.09	0.14	0.20	0.29	0.39	0.51	0.62	0.72	0.81
0.15	0.01	0.02	0.03	0.04	0.06	0.10	0.15	0.22	0.31	0.42	0.54	0.65	0.75
0.10	0.01	0.01	0.02	0.03	0.04	0.06	0.10	0.15	0.22	0.31	0.42	0.54	0.65
0.05		0.01	0.01	0.01	0.02	0.03	0.05	0.08	0.12	0.18	0.26	0.36	0.47
0.01						0.01	0.01	0.02	0.03	0.04	0.06	0.10	0.14

Table 2. Probability of Item Success After Change in GSV, by Subtest and Initial Probability (*continued*)

Expressive Communication

	Change in GSV												
Initial *p*	**–30**	**–25**	**–20**	**–15**	**–10**	**–5**	**0**	**5**	**10**	**15**	**20**	**25**	**30**
0.99	0.82	0.89	0.93	0.96	0.97	0.98	0.99	0.99					
0.95	0.47	0.60	0.71	0.80	0.87	0.92	0.95	0.97	0.98	0.99	0.99		
0.90	0.30	0.41	0.54	0.66	0.76	0.84	0.90	0.94	0.96	0.98	0.99	0.99	0.99
0.85	0.21	0.31	0.42	0.55	0.67	0.77	0.85	0.90	0.94	0.96	0.98	0.99	0.99
0.80	0.16	0.24	0.34	0.46	0.59	0.71	0.80	0.87	0.92	0.95	0.97	0.98	0.99
0.75	0.12	0.19	0.28	0.39	0.52	0.64	0.75	0.83	0.89	0.93	0.96	0.97	0.98
0.70	0.10	0.15	0.23	0.34	0.46	0.58	0.70	0.80	0.87	0.92	0.95	0.97	0.98
0.65	0.08	0.13	0.19	0.29	0.40	0.53	0.65	0.76	0.84	0.90	0.93	0.96	0.98
0.60	0.07	0.10	0.16	0.24	0.35	0.47	0.60	0.71	0.81	0.87	0.92	0.95	0.97
0.55	0.05	0.09	0.14	0.21	0.31	0.42	0.55	0.67	0.77	0.85	0.90	0.94	0.96
0.50	0.04	0.07	0.11	0.18	0.26	0.38	0.50	0.62	0.74	0.82	0.89	0.93	0.96
0.45	0.04	0.06	0.10	0.15	0.23	0.33	0.45	0.58	0.69	0.79	0.86	0.91	0.95
0.40	0.03	0.05	0.08	0.13	0.19	0.29	0.40	0.53	0.65	0.76	0.84	0.90	0.93
0.35	0.02	0.04	0.07	0.10	0.16	0.24	0.35	0.47	0.60	0.71	0.81	0.87	0.92
0.30	0.02	0.03	0.05	0.08	0.13	0.20	0.30	0.42	0.54	0.66	0.77	0.85	0.90
0.25	0.02	0.03	0.04	0.07	0.11	0.17	0.25	0.36	0.48	0.61	0.72	0.81	0.88
0.20	0.01	0.02	0.03	0.05	0.08	0.13	0.20	0.29	0.41	0.54	0.66	0.76	0.84
0.15	0.01	0.01	0.02	0.04	0.06	0.10	0.15	0.23	0.33	0.45	0.58	0.69	0.79
0.10	0.01	0.01	0.01	0.02	0.04	0.06	0.10	0.16	0.24	0.34	0.46	0.59	0.70
0.05			0.01	0.01	0.02	0.03	0.05	0.08	0.13	0.20	0.29	0.40	0.53
0.01						0.01	0.01	0.02	0.03	0.04	0.07	0.11	0.18

Fine Motor

	Change in GSV												
Initial *p*	**–30**	**–25**	**–20**	**–15**	**–10**	**–5**	**0**	**5**	**10**	**15**	**20**	**25**	**30**
0.99	0.81	0.88	0.92	0.95	0.97	0.98	0.99	0.99					
0.95	0.45	0.58	0.70	0.80	0.87	0.92	0.95	0.97	0.98	0.99	0.99		
0.90	0.28	0.39	0.52	0.65	0.76	0.84	0.90	0.94	0.96	0.98	0.99	0.99	
0.85	0.19	0.29	0.41	0.54	0.66	0.77	0.85	0.91	0.94	0.96	0.98	0.99	0.99
0.80	0.14	0.22	0.33	0.45	0.58	0.70	0.80	0.87	0.92	0.95	0.97	0.98	0.99
0.75	0.11	0.18	0.27	0.38	0.51	0.64	0.75	0.84	0.90	0.94	0.96	0.98	0.99
0.70	0.09	0.14	0.22	0.32	0.45	0.58	0.70	0.80	0.87	0.92	0.95	0.97	0.98
0.65	0.07	0.12	0.18	0.28	0.39	0.52	0.65	0.76	0.84	0.90	0.94	0.96	0.98
0.60	0.06	0.10	0.15	0.24	0.34	0.47	0.60	0.72	0.81	0.88	0.93	0.95	0.97
0.55	0.05	0.08	0.13	0.20	0.30	0.42	0.55	0.67	0.78	0.86	0.91	0.94	0.97
0.50	0.04	0.07	0.11	0.17	0.26	0.37	0.50	0.63	0.74	0.83	0.89	0.93	0.96
0.45	0.03	0.06	0.09	0.14	0.22	0.33	0.45	0.58	0.70	0.80	0.87	0.92	0.95
0.40	0.03	0.05	0.07	0.12	0.19	0.28	0.40	0.53	0.66	0.76	0.85	0.90	0.94
0.35	0.02	0.04	0.06	0.10	0.16	0.24	0.35	0.48	0.61	0.72	0.82	0.88	0.93
0.30	0.02	0.03	0.05	0.08	0.13	0.20	0.30	0.42	0.55	0.68	0.78	0.86	0.91
0.25	0.01	0.02	0.04	0.06	0.10	0.16	0.25	0.36	0.49	0.62	0.73	0.82	0.89
0.20	0.01	0.02	0.03	0.05	0.08	0.13	0.20	0.30	0.42	0.55	0.67	0.78	0.86
0.15	0.01	0.01	0.02	0.04	0.06	0.09	0.15	0.23	0.34	0.46	0.59	0.71	0.81
0.10		0.01	0.01	0.02	0.04	0.06	0.10	0.16	0.24	0.35	0.48	0.61	0.72
0.05			0.01	0.01	0.02	0.03	0.05	0.08	0.13	0.20	0.30	0.42	0.55
0.01						0.01	0.01	0.02	0.03	0.05	0.08	0.12	0.19

Table 2. Probability of Item Success After Change in GSV, by Subtest and Initial Probability (*continued*)

Gross Motor

	Change in GSV												
Initial *p*	**–30**	**–25**	**–20**	**–15**	**–10**	**–5**	**0**	**5**	**10**	**15**	**20**	**25**	**30**
0.99	0.71	0.82	0.89	0.94	0.97	0.98	0.99	0.99					
0.95	0.31	0.46	0.61	0.75	0.85	0.91	0.95	0.97	0.99	0.99			
0.90	0.18	0.29	0.43	0.58	0.72	0.83	0.90	0.94	0.97	0.98	0.99		
0.85	0.12	0.20	0.32	0.47	0.62	0.75	0.85	0.91	0.95	0.97	0.99	0.99	
0.80	0.09	0.15	0.25	0.38	0.54	0.68	0.80	0.88	0.93	0.96	0.98	0.99	0.99
0.75	0.07	0.12	0.20	0.32	0.46	0.62	0.75	0.85	0.91	0.95	0.97	0.99	0.99
0.70	0.05	0.09	0.16	0.27	0.40	0.56	0.70	0.81	0.89	0.94	0.97	0.98	0.99
0.65	0.04	0.08	0.13	0.22	0.35	0.50	0.65	0.78	0.87	0.92	0.96	0.98	0.99
0.60	0.04	0.06	0.11	0.19	0.30	0.45	0.60	0.74	0.84	0.91	0.95	0.97	0.98
0.55	0.03	0.05	0.09	0.16	0.26	0.40	0.55	0.69	0.81	0.89	0.94	0.96	0.98
0.50	0.02	0.04	0.08	0.13	0.22	0.35	0.50	0.65	0.78	0.87	0.92	0.96	0.98
0.45	0.02	0.04	0.06	0.11	0.19	0.31	0.45	0.60	0.74	0.84	0.91	0.95	0.97
0.40	0.02	0.03	0.05	0.09	0.16	0.26	0.40	0.55	0.70	0.81	0.89	0.94	0.96
0.35	0.01	0.02	0.04	0.08	0.13	0.22	0.35	0.50	0.65	0.78	0.87	0.92	0.96
0.30	0.01	0.02	0.03	0.06	0.11	0.19	0.30	0.44	0.60	0.73	0.84	0.91	0.95
0.25	0.01	0.01	0.03	0.05	0.09	0.15	0.25	0.38	0.54	0.68	0.80	0.88	0.93
0.20	0.01	0.01	0.02	0.04	0.07	0.12	0.20	0.32	0.46	0.62	0.75	0.85	0.91
0.15		0.01	0.01	0.03	0.05	0.09	0.15	0.25	0.38	0.53	0.68	0.80	0.88
0.10			0.01	0.02	0.03	0.06	0.10	0.17	0.28	0.42	0.57	0.71	0.82
0.05				0.01	0.01	0.03	0.05	0.09	0.15	0.25	0.39	0.54	0.69
0.01						0.01	0.01	0.02	0.03	0.06	0.11	0.18	0.29

Conditional Standard Error of Measurement of GSV

The standard error of measurement (*SEM*) provides an estimate of the amount of error in a child's observed test score. The Bayley-III Technical Manual describes how to evaluate GSV differences. The conditional *SEM* has the same meaning as the traditional *SEM*, but it is specific to each GSV value. Conditional *SEM* is a function of the number of moderately difficult items a child at that ability level will encounter on the subtest. The score from each such encounter provides information about the child's ability, and the more such encounters, the better the estimate of ability and the smaller the conditional *SEM*. This may be seen in Figure 3, showing both the conditional *SEM* and the number of items at each level of ability/difficulty. Conditional *SEM* rises and falls in accordance with the number of items at that GSV level. On every subtest the conditional *SEM* is large for very low or very high GSVs, where the number of items is small. For children with such extreme levels of ability, the test is too difficult or too easy, respectively, to provide highly precise measurement. The Appendix reports the conditional *SEM* of each GSV for each subtest.

Figure 3. Conditional *SEM* Versus GSV, and Frequency Distribution of Item Difficulties, by Subtest

A common application of conditional *SEM*s is to determine whether the difference between two GSVs on the same test (such as a child's GSVs at two points in time) is statistically significant. This is done by dividing the difference by the standard error of the difference:

$$\frac{GSV_1 - GSV_2}{\sqrt{cSEM_1{}^2 + cSEM_2{}^2}}$$

Values of 1.65 and 1.96 indicate significant differences at $p<.10$ and $p<.05$, respectively.

Diagnostic Information About Rasch Calibration

The Rasch model makes several assumptions about how test data behave: item characteristic curves should have similar slope, the test should measure a single dimension of ability, and items should be *locally independent* (i.e., pairs of items should correlate only because they both measure the single ability dimension). This section presents evidence regarding each of these assumptions.

Item Fit

An *item characteristic curve* (ICC) describes how the expected score on an item increases as ability increases. With dichotomous items, the expected score is the probability of success. The Rasch model assumes that every item's ICC is a logistic curve and that all ICCs on a test have the same slope. This assumption can be tested by computing the actual success rate of persons at each ability level and comparing the resulting empirical trend with the expected ICC. To the extent that these differ, the item is said to *misfit*. If the slope of the actual data is steeper than expected (*overfit*), the item does a better job of differentiating between people at higher and lower ability levels—that is, it correlates higher with the overall ability dimension. Conversely, a flatter ICC (*underfit*) means that the item is below-average in its ability to identify different levels of ability. Overfit does not compromise the reliability or validity of scores but can cause underestimation of conditional *SEM*s and affect the accuracy of estimates of success probability (Bond et al., 2021). Underfit, on the other hand, may affect the quality of measurement. It should be noted that underfit also affects raw scores and other scores derived from raw scores, such as standard scores and age equivalents.

Misfit can be measured by several statistics, one of which is *infit mean square*. On any test, the average value of infit mean square is expected to be 1.00; larger values indicate ICCs that are flatter than expected, and smaller values indicate ICCs that are relatively steep. There is no generally accepted standard for the desirable range of infit mean square, but 0.50 to 1.50 (Linacre, 2021b) and 0.75 to 1.30 (Bond et al., 2021) are typical recommendations.

Table 3 reports, for each subtest, the mean and standard deviation of infit mean square and the number of items at different levels of infit mean square. Overall, 98% of items had infit mean square between 0.50 and 1.50, and 82% of values were between 0.76 and 1.30. Most of the values that fell outside the latter range were small (i.e., the items overfit). Three of the four items with infit mean square values smaller than 0.50 belonged to pairs or sets of items that measured different levels of performance on a similar task (crawling or walking). Of the four items with infit mean square values greater than 1.50, two were for incompatible tasks (hands fisted and hands open).

Table 3. Descriptive Statistics for Infit Mean Square, by Subtest

Bayley-III

Subtest	No. items	Infit mean square		Number of items with infit mean square in range:				
		Mean	*SD*	<0.50	0.50–0.75	.76–1.30	1.31–1.50	>1.50
Cognitive	91	0.98	0.18	0	11	78	2	0
Receptive Communication	49	0.95	0.16	0	2	46	1	0
Expressive Communication	48	0.95	0.17	0	3	42	3	0
Fine Motor	66	0.96	0.22	0	11	49	4	2
Gross Motor	72	0.93	0.25	4	12	51	3	2

Local Independence and Item Intercorrelations

The local independence and dimensionality assumptions are both evaluated, in whole or in part, through the intercorrelations of item residuals (the differences between persons' expected and observed item scores). If a person's item performance is solely a function of their level on the underlying ability, then residuals on different items should be uncorrelated. A substantial correlation might reflect the fact that the two items measure a secondary ability in addition to the primary ability; depending on the strength and nature of this secondary ability, the test might be considered multidimensional. Alternatively, there could be a relationship between the content, administration, or scoring of the items. For example, performance on one item might constrain the score on another item or determine whether it is administered; the content of one item might give information useful to solving the other; or both items might require interpreting the same stimulus such as a chart or a reading passage. The term *local dependence* will be used to refer to these latter types of relationships, regardless of whether the item residuals are highly correlated. The effects of local dependence on test usage are generally benign (Bond et al., 2021; Linacre, 2021b). Locally dependent items will tend to overfit, which may cause underestimation of the conditional *SEM* at some score values and may affect the accuracy of estimates of success probability on individual items.

It is desirable to evaluate local dependence first so that any high residual correlations it causes are not misinterpreted as evidence of multidimensionality. For Bayley-III, this was done by examining the content, administration, and scoring procedures of all items, with special attention to pairs of items with highly correlated item residuals. The following locally dependent item sets were found. Except as noted, they consisted of different levels of performance on the same task.

- Cognitive: Seven pairs, three triads, and one four-item set. In one pair, performance on the first item determined whether the second was administered, and in another pair the tasks were partly the same. Correlations between raw residuals on dependent items ranged from .00 to .36 (median = .16).
- Receptive Communication: Three pairs and one triad. Residual correlations: .02 to .35 (median = .13).
- Expressive Communication: Four pairs and two triads. Residual correlations: -.02 to .35 (median = .06).
- Fine Motor: Three pairs, two triads, and four sets of four or five. Two pairs and one triad consisted of highly similar tasks. Residual correlations: -.06 to .46 (median = .02).
- Gross Motor: Five pairs and nine triads. One pair consisted of highly similar tasks. Residual correlations: -.05 to .45 (median = .05).

The effect of these locally dependent items on the subtests' measurement properties was evaluated by consolidating each pair or set of locally dependent items into a single item (by summing their scores), performing a Rasch calibration of the reduced item set, and comparing the resulting ability scores and conditional *SEM*s with the original values for the same raw scores. Table 4 reports these correlations in the Bayley-III standardization sample. The relationships were nearly perfect. The standard deviations of ability scores were 2% to 9% larger when locally dependent items were separate, reflecting the artificially high correlations between them. Average conditional *SEM*s were 1% to 4% larger, meaning that the presence of locally dependent items did not cause underestimation of conditional *SEM*s. Overall, the data indicate that the presence of locally dependent items had a negligible effect on the measurement properties of GSVs.

Table 4. Effects of Summing Locally Dependent Items on Ability Scores and Conditional *SEM*s

Bayley-III

Subtest	Correlation		*SD* of ability scores		Mean conditional *SEM*	
	Ability scores	Conditional *SEM*	Separate	Summed	Separate	Summed
Cognitive	0.9999	0.990	9.61	9.13	0.67	0.65
Receptive Communication	0.9999	0.998	9.37	9.08	1.03	1.02
Expressive Communication	0.9999	0.999	10.21	9.99	0.99	0.98
Fine Motor	0.9997	0.959	10.48	9.67	0.83	0.80
Gross Motor	0.9999	0.989	12.39	11.37	0.81	0.79

In the reduced item sets containing sums of locally dependent items, twelve pairs of items had residual correlations exceeding .40 in absolute value. In five pairs the items had an obvious similarity and might have been treated as dependent (types of play; cutting with scissors; balancing on left or right foot; climbing or descending stairs). For the others, ranging from .40 to .45 in absolute value, the cause of the correlation was not apparent.

Dimensionality

Dimensionality can be evaluated by using a principal components analysis of the item residuals to see if there are any components large enough to constitute secondary ability dimensions. Linacre (2021b) recommends that components with eigenvalues greater than 2.0 are worthy of investigation because they have the "strength" of two or more items. Three features of such components should be examined. One is the percentage of variance accounted for by the component. A second is its impact on the overall score, assessed by scoring each person on three subsets of items: items with high positive loadings on the component, those with high negative loadings, and those in between. If these three subset scores intercorrelate highly, the component has little effect on subtest scores. The third feature to examine is the content of the items with large loadings on the component, to infer what construct the component represents and whether it is outside the conceptual domain of the test. As Smith (2004) notes, "multidimensionality only becomes a problem when data represent two or more dimensions so disparate or distinct that it is no longer clear what dimension the Rasch model is defining (lacks construct validity) or when the different subsets of items would lead to different norm (NR) or criterion-referenced (CR) decisions."

A principal components analysis was performed on the raw item residuals in the reduced item set for each subtest. Four subtests had one component each with an eigenvalue of 2 or greater, but these components were very small, with a maximum eigenvalue of 2.4 and no more than .4 percent of total variance explained. Correlations between item subsets ranged from .89 to 1.00. Components were interpretable and within the construct domain of their subtest: finding a hidden object versus manipulating pieces (Cognitive), utterances versus naming (Expressive Communication), using scissors versus building steps (Fine Motor), and balancing versus using stairs (Gross Motor). These findings indicate that each subtest met the unidimensionality assumption once the effects of procedurally induced local dependence were controlled.

References

Bayley, N. (2006a). *Bayley Scales of Infant and Toddler Development* (3rd ed.). NCS Pearson.

Bayley, N. (2006b). *Bayley Scales of Infant and Toddler Development* (3rd ed.): *Technical manual*. NCS Pearson.

Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences* (4th ed.). Routledge.

Linacre, J. M. (2021a). *Winsteps Rasch measurement computer program*. (Version 5.1.4) [Computer software]. Winsteps.com. Available from https://www.winsteps.com.

Linacre, J. M. (2021b). *Winsteps Rasch measurement computer program user's guide*. (Version 5.1.4). Winsteps.com. Available from https://www.winsteps.com.

Smith, E. V. (2004). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications* (pp. 575–600). JAM Press.

Appendix: Bayley–III GSVs and Conditional *SEM*s Corresponding to Raw Scores

Raw score	Cognitive		Receptive		Expressive		Fine Motor		Gross Motor		Raw score
	GSV	c*SEM*	GSV	c*SEM*	GSV	c*SEM*	GSV	c*SEM*	GSV	c*SEM*	
0	246	21.7	251	25.4	279	21.1	291	18.5	250	16.0	0
1	261	16.7	274	22.1	307	20.9	306	11.8	264	11.1	1
2	285	14.2	296	16.3	335	13.7	318	9.9	282	12.9	2
3	303	12.9	315	12.8	359	17.8	328	9.0	305	14.0	3
4	317	10.6	330	12.5	382	12.4	336	8.6	324	10.6	4
5	326	9.4	347	14.9	396	11.6	344	8.5	335	8.4	5
6	334	8.7	369	15.1	409	11.0	352	8.4	343	7.2	6
7	341	8.3	392	16.2	421	10.5	359	8.2	349	6.6	7
8	348	8.0	416	15.7	431	9.3	366	8.0	354	6.4	8
9	354	7.7	434	12.1	439	8.5	372	8.0	359	6.4	9
10	359	7.6	447	10.8	446	8.4	379	8.2	365	6.8	10
11	365	7.5	457	10.3	454	8.8	386	8.4	371	7.2	11
12	370	7.3	467	10.3	462	9.2	394	8.6	377	7.3	12
13	375	7.1	478	11.1	470	8.8	403	9.2	384	7.1	13
14	380	6.9	490	11.6	478	8.3	412	9.7	390	6.8	14
15	385	6.8	502	10.6	485	8.2	422	9.4	395	6.7	15
16	389	6.8	511	9.7	492	8.5	431	8.8	401	6.5	16
17	394	6.8	520	9.2	500	8.9	439	8.4	406	6.3	17
18	399	6.9	527	8.9	508	8.9	446	7.9	410	6.0	18
19	403	7.1	535	8.6	516	8.9	452	7.7	414	5.6	19
20	409	7.4	542	8.5	525	9.0	458	7.5	418	5.4	20
21	414	7.7	548	8.4	533	9.3	464	7.5	422	5.3	21
22	420	8.1	555	8.5	542	9.3	470	7.5	425	5.3	22
23	427	8.2	562	8.7	551	8.8	476	7.7	429	5.3	23
24	433	8.2	570	9.3	558	8.3	482	7.8	433	5.5	24
25	440	8.0	578	9.6	565	7.8	489	8.0	436	5.6	25
26	446	7.7	587	9.4	571	7.7	496	8.2	440	5.6	26
27	452	7.5	594	8.8	577	7.7	503	8.5	444	5.6	27
28	457	7.3	601	8.3	583	7.8	511	8.9	448	5.6	28
29	462	7.1	608	8.0	590	7.9	520	9.3	452	5.6	29
30	467	7.1	613	7.7	596	7.8	529	9.3	456	5.6	30
31	472	7.1	619	7.5	602	7.7	538	8.7	460	5.6	31
32	477	7.1	624	7.2	608	7.5	545	8.2	464	5.7	32
33	482	7.0	629	7.2	614	7.3	552	7.9	468	5.9	33
34	486	6.9	633	7.2	619	7.2	558	7.8	472	6.0	34
35	491	6.8	638	7.3	625	7.2	564	7.7	477	6.2	35
36	495	6.6	644	7.6	630	7.2	570	7.4	482	6.3	36
37	499	6.4	649	7.7	635	7.1	576	7.0	487	6.4	37
38	503	6.2	655	7.8	640	7.1	581	6.7	492	6.7	38
39	507	6.1	661	7.9	645	7.0	585	6.5	498	7.0	39
40	510	5.9	667	8.0	650	7.0	590	6.5	504	7.2	40
41	513	5.8	673	8.0	655	7.0	594	6.5	510	6.9	41
42	517	5.7	679	8.1	660	7.0	598	6.5	516	6.7	42
43	520	5.7	685	8.2	665	7.2	603	6.5	521	6.7	43
44	523	5.6	691	8.3	671	7.6	607	6.6	527	6.9	44
45	526	5.6	697	8.6	677	8.4	612	6.7	534	7.7	45

Appendix: Bayley–III GSVs and Conditional *SEM*s Corresponding to Raw Scores (*continued*)

Raw score	Cognitive		Receptive		Expressive		Fine Motor		Gross Motor		Raw score
	GSV	c*SEM*	GSV	c*SEM*	GSV	c*SEM*	GSV	c*SEM*	GSV	c*SEM*	
46	529	5.5	703	9.2	686	9.6	617	6.6	542	8.1	46
47	532	5.5	711	10.3	697	11.8	621	6.5	550	7.5	47
48	535	5.5	722	13.1	712	19.0	626	6.5	556	6.7	48
49	537	5.5	741	20.9			630	6.4	561	6.3	49
50	540	5.5					634	6.4	566	6.0	50
51	543	5.5					638	6.4	570	5.8	51
52	546	5.5					643	6.3	574	5.6	52
53	549	5.6					647	6.3	578	5.6	53
54	552	5.6					651	6.2	582	5.6	54
55	555	5.7					654	6.1	586	5.8	55
56	558	5.8					658	6.0	590	5.8	56
57	562	5.9					662	6.0	594	5.8	57
58	565	6.0					665	6.0	598	5.7	58
59	568	6.1					669	6.1	602	5.7	59
60	572	6.1					673	6.2	606	5.7	60
61	576	6.2					677	6.5	611	5.8	61
62	579	6.3					682	6.8	615	6.0	62
63	583	6.4					687	7.4	619	6.0	63
64	587	6.4					694	8.3	624	6.1	64
65	591	6.5					703	10.6	629	6.4	65
66	595	6.5					716	17.9	634	6.6	66
67	599	6.6							639	6.7	67
68	603	6.6							644	6.7	68
69	608	6.5							650	6.8	69
70	612	6.5							656	7.3	70
71	616	6.4							664	9.0	71
72	620	6.4							671	15.2	72
73	624	6.3									73
74	627	6.3									74
75	631	6.3									75
76	635	6.3									76
77	639	6.3									77
78	643	6.4									78
79	647	6.5									79
80	651	6.6									80
81	655	6.6									81
82	659	6.6									82
83	663	6.6									83
84	668	6.7									84
85	672	6.7									85
86	677	6.9									86
87	682	7.2									87
88	687	7.9									88
89	694	8.9									89
90	704	11.6									90
91	713	19.5									91