Chapter 5
An Empirical Example of Change Analysis by Linking Longitudinal Item Response Data From Multiple Tests

John J. McArdle and Kevin J. Grimm

Linking, *equating*, and *calibrating* refer to a series of statistical methods for comparing scores from tests (scales, measures, etc.) that do not contain the same exact set of measurements but presume to measure the same underlying construct. Lord (1955a,1955b) provided one of the first examples of this kind where one test (x) was administered to 1,200 people, while two other tests (y_1 & y_2) were each only administered to a different half of the group. The resulting data and analysis were reprinted in Cudeck (2000), who showed that the assumption of a single factor model for all three tests (x, y_1, y_2) made it possible to identify a maximum likelihood estimator of the correlation among the two variables that were never measured on the same persons (y_1 & y_2). In contemporary terms the common score (x) served as an anchor for the correlation of the other two scores, and this simple design is one version of what is termed a *nonequivalent anchor test* (von Davier, Holland, & Thayer, 2004b).

There has been a great deal of work on similar incomplete data problems at the level of items. The introduction of item response methods led to improved linking techniques (e.g., common-item equating, common-person equating) as item response models have built-in linking mechanisms for incomplete data (Embretson, 1996). Most of the recent work on this topic has been summarized in Dorans, Pommerich, and Holland (2007), and Dorans (2007) provided a good readable overview of linking scores. Dorans examined the general assumptions of different data collection designs and gave explicit definitions of equating, calibrating, and linking. Dorans also provided a compelling example of the importance of adequate linking using multiple health outcome instruments, and how an individual's health

J.J. McArdle (✉)
Dept. of Psychology, University of Southern California, SGM 501 3620 South McClintock Ave, Los Angeles, CA 90089, USA
e-mail: jmcardle@usc.edu

K.J. Grimm
University of California, 1 Shields Ave, Davis, CA 95616, USA
e-mail: kjgrimm@ucdavis.edu

A.A. von Davier (ed.), *Statistical Models for Test Equating, Scaling, and Linking*,
Statistics for Social and Behavioral Sciences,
DOI 10.1007/978-0-387-98138-3_5,

may be misunderstood if alternative tests presumed to measure the same construct fail to do so.

The research we present here has far fewer consequences because it is not intended for high-stakes decision making. Instead, we attempt to use the new approaches in item linking to deal with a perplexing problem in lifespan research—we ask, "How can we get a reasonable measure of the same construct when the tests themselves are changing over age and time?" The approach we present here is intended to be useful for research into the dynamics of aging but is not intended as a placement device or as an improved marker of health.

5.1 Challenges in Lifespan Developmental Research

Examining change over extended periods of time or during critical developmental periods where the expression of the construct changes is a complex undertaking. Often the measurement of the construct must change to adequately capture the construct. In these situations changes in measurement and changes in the construct are difficult to separate. One empirical example comes from the Intergenerational Studies (IGS) of Human Development, a collection of three studies initiated at the University of California-Berkeley in 1928. A main interest of the IGS was to examine the growth and change of cognitive abilities during infancy, childhood, adolescence, and adulthood. In the IGS, cognitive abilities have been measured with a variety of tests across the 70 years of the study, including the California First-Year Mental Scale (Bayley, 1933), California Preschool Mental Scale (Jaffa, 1934), Stanford-Binet (Terman, 1916), Stanford-Binet Form L and Form M (Terman & Merrill, 1937), Wechsler-Bellevue Intelligence Scale (Wechsler, 1946), Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1955), and the Wechsler Adult Intelligence Scale-Revised (WAIS-R; Wechsler, 1981). These measures were chosen because they were the best available age-appropriate tests.

Changes in cognitive abilities could be measured for each developmental period. For example, Figure 5.1 is a series of longitudinal plots for specific developmental periods where the same cognitive test was administered. Plots A and B in Figure 5.1 are of the California First-Year Mental Scale and California Preschool Mental Scale for participants in the Berkeley Growth Study and Berkeley Guidance-Control Study, respectively. These two plots cover a similar developmental period (i.e., birth through age 5) using different cognitive tests but generally show a pattern of rapid increase. Plot C in Figure 5.1 shows mental age from the series of Stanford-Binet tests (i.e., 1916, Form L, Form M), mostly collected from ages 6–17. Plot D in Figure 5.1 shows Block Design scores from the Wechsler-Bellevue Intelligence Scale measured from ages 16–27 years and shows a period of slight growth and stability. Plot E in Figure 5.1 shows Block Design scores from the WAIS, which was administered once; therefore individual change patterns cannot be captured. Finally, Plot F in Figure 5.1 shows Block Design scores from the WAIS-R and shows stability in the change pattern and large between-person differences therein. It is important to note that the Block Design scores from the Wechsler tests are

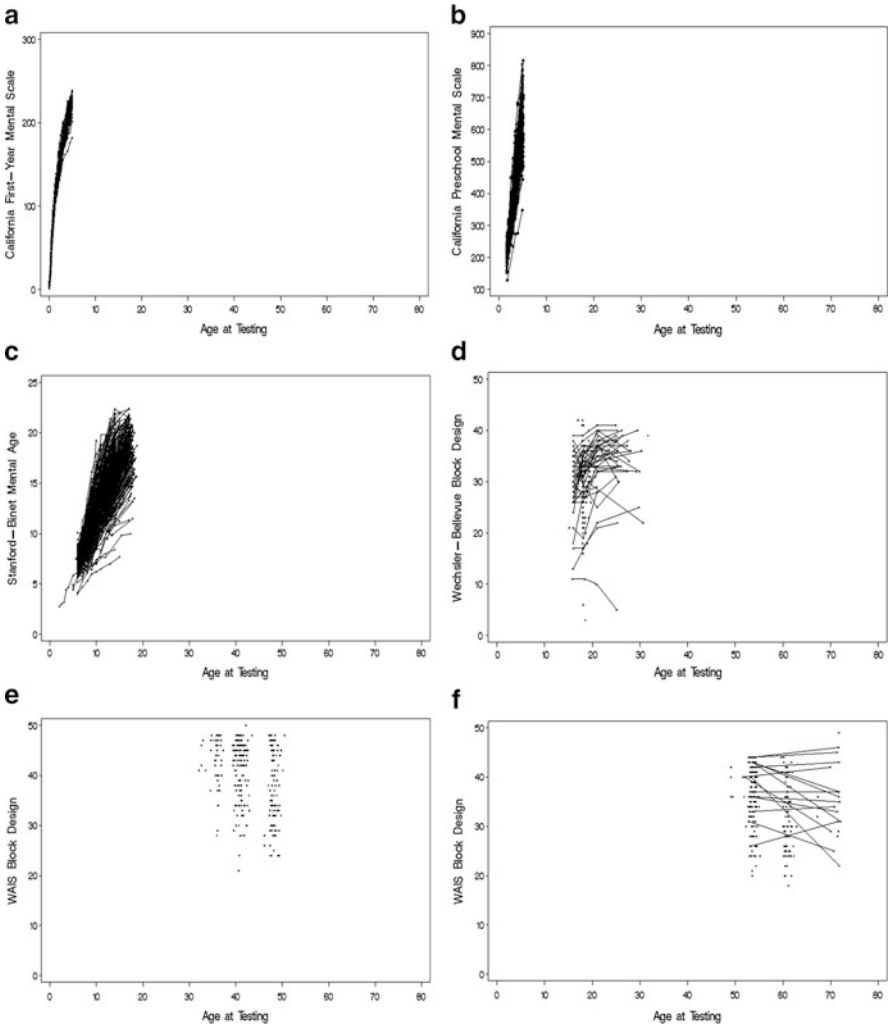

Fig. 5.1 Longitudinal plot of (**a**) First-Year Mental Scale total score, (**b**) Preschool Mental Scale total score, (**c**) Stanford-Binet mental age, (**d**) Wechsler-Bellevue Block Design, (**e**) Wechsler Adult Intelligence Scale (WAIS) Block Design, and (**f**) WAIS-Revised (WAIS-R) Block Design

not on a common scale; however, some items are identical across edition of the Wechsler. Additionally, the tests described above may be measuring different constructs at the total score level, even though they are all intelligence tests.

An alternative view regarding the developmental process is that it is continuously evolving. Thus, by selecting items from different tests that measure a common construct and scaling them with an appropriate model, a *lifespan* trajectory of specific cognitive abilities may be represented. Scaling items, in essence, would equate items from different tests, recognizing their differences in level of

difficulty and relationship with the underlying construct. In order to link the tests, there must be sufficient item overlap within and between test forms. Item overlap occurs because different tests were administered at the same occasion and because different tests contain the same items.

In this chapter we describe an example of using item response linking procedures with scarce longitudinal item-level data collected over a 70-year period to help understand and evaluate theories of cognitive growth and decline. Data for this project come from the IGS and the Bradway-McArdle Longitudinal Study (BMLS). We realize that IGS and BMLS data are weak in terms of equating but recognize their longitudinal strength. Building on the data's strength, we link items measuring nonverbal intelligence and model within-person changes in the lifespan development of nonverbal intelligence and the between-person differences therein.

5.2 Longitudinal Item-Level Data

Longitudinal studies provide an added dimension (e.g., time/age) to consider when linking item-level data. Measurement invariance is tremendously important in longitudinal studies as researchers are most interested in studying change; however, measurement invariance is often overlooked or assumed because the same test is often administered in longitudinal studies. In many instances in longitudinal studies it is not reasonable to administer the same test (see Edwards & Wirth, 2009; McArdle, Grimm, Hamagami, Bowles, & Meredith, 2009). Potential reasons include age appropriateness, improved and revised tests become available, and prior poor experiences. For example, the Child-Behavior Checklist (Achenbach & Rescorla, 2001), an informant-based behavior rating scale, is an often-used measure of behavior problems in children and has two age-appropriate forms. One form is appropriate for children between the ages of 1½ and 5 years, whereas the second form is appropriate for children between 6 and 18 years old. The two forms share the same dimensions of behavior problems (e.g., internalizing and externalizing) and share several items, but they do have items specific to each age-appropriate form. Also, there are items that both forms share but are expected to relate to the underlying dimensions in different ways (e.g., a question about crying is part of the internalizing dimension in the 6–18 form but not part of the internalizing or externalizing dimensions in the 1½–5 form).

5.3 Method

We start with descriptions of the samples and measures, an overview of the item response and longitudinal models, and results from fitting these models to the longitudinal data. Table 5.1 contains information regarding the cognitive tests that were administered at each age for each sample.

Table 5.1 Summary of Measurement Occasions for Each Sample

Age	Berkeley Growth	Guidance-Control	Oakland Growth	Bradway-McArdle Longitudinal
2–5 ½	–	–	–	SB-L, SB-M (139)
6	1916 SB (60)	1916 SB (205)	–	–
7	1916 SB (47), SB-L (8)	1916 SB (204)	–	–
8	SB-L (51)	SB-L (187)	–	–
9	SB-L (53)	SB-L (94), SB-M (98)	–	–
10	SB-M (53)	SB-L (102), SB-M (88)	–	–
11	SB-L (48)	SB-L (77)	–	–
12	SB-M (50)	SB-L (90), SB-M (43)	1916 SB (192)	–
13–14	SB-L (42)	SB-L (82), SB-M (97)	–	SB-L (111)
15	–	SB-M (51)	–	–
16	WB-I (48)	–	–	–
17	SB-M (44)	–	SB-M (147)	–
18	WB-I (41)	WB-I (157)	–	–
21	WB-I (37)	–	–	–
25	WB-I (25)	–	–	–
29	–	–	–	WAIS, SB-L (110)
36	WAIS (54)	–	–	–
40	–	WAIS (156)	–	WAIS, SB-LM (48)
50	–	–	WAIS (103)	–
53	WAIS-R (41)	WAIS-R (118)	–	WAIS (53)
60	–	–	WAIS-R (78)	–
63	–	–	–	WAIS (48)
72	WAIS-R (31)	–	–	–

Note: Sample sizes are contained within parentheses next to the test name; more than one sample size within a single testing age denotes different participants. 1916 SB = 1916 edition of the Stanford-Binet, SB-L = Revised Stanford-Binet Form L, SB-M = Revised Stanford-Binet Form M, SB-LM = Revised Stanford-Binet Form LM, WB = Wechsler-Bellevue Intelligence Scale, WAIS = Wechsler Adult Intelligence Scale, WAIS-R = Wechsler Adult Intelligence Scale-Revised

5.3.1 Berkeley Growth Study

The Berkeley Growth Study was initiated by Nancy Bayley in 1928. Sixty-one infants were enrolled between September 15, 1928, and May 15, 1929, to trace early mental, motor, and physical development during the first years of life. An additional 13 infants were enrolled in the Berkeley Growth Study within 3 years of the start of the study, bringing the final sample size to 74.

Data collection in the Berkeley Growth Study began within 4 days of an infant's birth as anthropometric, neurological, and physiological measurements were made in the hospital by pediatricians. Participating infants were assessed at the Institute of Human Development at the University of California-Berkeley every month from 1–15 months of age, every 3 months from 18–36 months of age, and then annually from 4–18 years of age. In adulthood the participants were measured at 21, 25, 30, 36, 52, and 72 years of age. The Berkeley Growth Study was the most measurement-intensive IGS study.

5.3.2 Berkeley Guidance-Control Study

The Berkeley Guidance-Control Study began in early 1928 under the leadership of Jean Macfarlane. The 248 original participants in the Berkeley Guidance-Control Study were drawn from a survey of every third birth in Berkeley from January 1, 1928, through June 30, 1929. The initial nature of the Berkeley Guidance-Control Study was a 6-year project with goals of (a) documenting the frequency and occurrence of behavior and personality problems in a cross-sectional sample of young children during the preschool years, (b) identifying the biological and environmental factors associated with the presence or absence of such behavioral problems, and (c) estimating the effects of guidance activities with the parents of these children.

Monthly home visits began when infants were 3 months old and continued through 18 months of age. When the infants were 21 months of age, half of the sample ($n = 124$) was assigned to the guidance condition, and the remaining half of the sample ($n = 124$) was assigned to the control condition. Parents of the infants in the guidance condition engaged in intensive discussions with public health nurses and other project staff. An initial, intensive assessment of the infants and their parents was conducted at 21 months. Thereafter, infants and parents were interviewed and tested every 6 months from the child's age of 2–4 years and then annually from 5–18 years of age. In adulthood, the Berkeley Guidance-Control Study participants were assessed at ages 30, 40, and 52.

5.3.3 Oakland Growth Study

The Oakland Growth Study began in 1931 under the guidance of Harold Jones, Mary Jones, and Herbert Stolz. A total of 212 students attending five elementary schools in Oakland, California, were enrolled into the study. The goal of the Oakland Growth Study was to study normal adolescence, particularly physical and physiological maturation and peer relationships. Initial measurements were taken in 1932 when the participants ranged in age from 10–12 years. Participants in the Oakland Growth Study were assessed semiannually during the six years of junior and senior high school. In adulthood, the participants were assessed at ages 38, 48, and 60.

5.3.4 Bradway-McArdle Longitudinal Study

The Bradway-McArdle Longitudinal Study began in 1931 when 139 children aged 2½ to 5 years were tested as part of the standardization of the Revised Stanford-Binet (Terman & Merrill, 1937). The sample was tested by Katherine Bradway

(Bradway, 1944, 1945a. 1945b) with the Revised Stanford-Binet in 1941. The sample was tested in 1957, 1969, 1984, and 1992. McArdle, Hamagami, Meredith, and Bradway (2000) coordinated the last three waves of data collection. It is important to note that this sample took Forms L and M of the Stanford-Binet in the same occasion in 1931; Form L of the Stanford-Binet and the WAIS in 1957; Form LM of the Stanford-Binet and the WAIS in 1969, and the WAIS and additional WAIS-R items in 1992.

5.3.5 Measures of Nonverbal Intelligence

The cognitive measures administered in these studies and examined here include the 1916 Stanford-Binet (Terman, 1916), Revised Stanford-Binet (Form L, Form M, & Form LM; Terman & Merrill, 1937, 1960), Wechsler-Bellevue Intelligence Scale (Wechsler, 1946), WAIS (Wechsler, 1955), and the WAIS-R (Wechsler, 1981). From these scales, nonverbal intelligence items were selected. For the Stanford-Binet tests, item selection was based on a categorization conducted by Bradway (1945). A list of nonverbal items selected from the Stanford-Binet tests is presented in Table 5.2. The first column of Table 5.2 contains a running total of the number of items that measure nonverbal intelligence from the Stanford-Binet tests. As seen, 65 items from the Stanford-Binet measure nonverbal intelligence. The second column contains the name of the item, and the third through sixth columns contain the Stanford-Binet item number if the item appeared on the edition of the test. These columns were left blank if the item did not appear on the edition. Items on the Stanford-Binet tests are grouped by age appropriateness (as opposed to construct), and item numbers reflect this. For example, II-1 means this item is the first item from the age 2 items. Table 5.2 shows the level (or lack thereof) of item overlap across editions of the Stanford-Binet. Each edition of the Stanford-Binet contains items that are unique and shared with other editions.

For the Wechsler tests, items from Picture Completion, Picture Arrangement, Block Design, and Object Assembly were chosen to represent nonverbal intelligence. A list of the nonverbal items from the Wechsler tests is presented in Table 5.3. As in Table 5.2, the first column is a running total of the number of items selected from the Wechsler series of tests. As seen, 68 distinct items were selected. Column 2 contains the subscale from which the item comes from, and columns 3–5 indicate the item number from each edition of the Wechsler intelligence scales. Columns were left blank if the edition did not contain the item allowing for the examination of item overlap across Wechsler editions. It is important to note that, in several cases, the scoring of items had to be modified to be considered equivalent because of different time bonuses. In the Wechsler series of intelligence tests, the scoring system from the WAIS was adopted for the Wechsler-Bellevue and WAIS-R where appropriate. Several items were similar on the surface and in name but were slightly different in presentation or scoring.

Table 5.2 Nonverbal Items From the Stanford-Binet Intelligence Scales and Their Overlap Across Test Forms

Item number	Item	1916 Stanford-Binet	Stanford-Binet Form L	Stanford-Binet Form M	Stanford-Binet Form LM
01	Delayed Response	–	–	II-1	II-2
02	Form Board (1)	–	II-1	II-4	II-1
03	Block Tower	–	II-4	–	II-4
04	Motor Coordination (1)	–	–	IIH-2	–
05	Form Board - Rotated (1)	–	IIH-6	–	–
06	Stringing Beads (2)	–	–	IIH-A	–
07	Vertical Line	–	–	III-4	III-6
08	Stringing Beads (4)	–	III-1	–	III-1
09	Block Bridge	–	III-3	III-1	III-3
10	Circle (1)	–	III-5	–	III-5
11	Form Board - Rotated (2)	–	III-A	III-A	IIH-A
12	Patience: Pict (1)	–	–	IIIH-2	IIIH-2
13	Animal Pict. (4)	–	–	IIIH-3	IIIH-3
14	Sorting Buttons	–	–	IIIH-5	IIIH-5
15	Matching Obj. (3)	–	–	IIIH-A	–
16	Cross	–	IIIH-A	–	–
17	Stringing Beads (7)	–	–	IV-2	–
18	Compar. Lines	IV-1	–	–	–
19	Discrimination of Forms (3)	IV-2	–	–	–
20	Pict. Comp. (Man)	–	IV-3	–	–
21	Discrimination of Forms (8)	–	IV-5	–	IV-5
22	Animal Pict. (6)	–	–	IV-A	–
23	Animal Pict. (7)	–	–	IVH-1	–
24	Pict. Compl. (Bird)	–	–	IVH-4	–
25	Pict. Compar. (3)	–	IVH-3	–	–
26	Patience: Pict. (2)	V-5	–	IVH-A	–
27	Pictorial Sim. & Dif II (9)	–	–	V-3	V-5
28	Patience Rec. (2)	–	–	V-4	V-6
29	Pict. Compl. (Man)	–	V-1	–	V-1
30	Folding Triangle	–	V-2	–	V-2
31	Square (1)	IV-4	V-4	–	V-4
32	Mut. Pict. (3)	VI-2	–	V-6	–
33	Mut. Pict. (4)	–	–	–	VI-3
34	Knot	VII-4	V-A	V-A	V-A
35	Bead Chain I	–	VI-2	VI-2	–
36	Mut. Pict. (4)	–	VI-3	–	–
37	Pict. Compar. (5)	–	VI-5	–	–
38	Pict. Absurd. I (3)	VII-2	VII-1	–	–
39	Pict. Absurd. I (4)	–	–	–	VII-1
40	Diamond (2)	–	VII-3	–	–
41	Diamond (1)	VII-6	–	–	VII-3
42	Pict. Absurd. I (2)	–	–	VII-3	–
43	Ball & field	VIII-1	–	–	–
44	Paper Cutting I (1)	–	IX-1	–	IX-1
45	Pict. Absurd. II	–	X-2	–	–
46	Absurditites (4)	X-2	-	-	-
47	Pict. Absurd. II	–	–	XII-5	XII-3
48	Plan of Search	–	XIII-1	XIII-1	XIII-1

(*continued*)

Table 5.2 (continued)

Item number	Item	1916 Stanford-Binet	Stanford-Binet Form L	Stanford-Binet Form M	Stanford-Binet Form LM
49	Paper Cutting I (2)	–	XIII-3	–	XIII-A
50	Reasoning	–	–	XIV-1	XIV-3
51	Induction	XIV-2	XIV-2	–	XIV-2
52	Pict. Absurd. III	–	XIV-3	XIV-2	–
53	Ingenuity (1)	–	XIV-4	XIV-5	XIV-4
54	Codes (1.5)	–	AA-2	AA-4	–
55	Ingenuity (2)	–	AA-6	AA-2	AA-2
56	Directions I (4)	–	–	AA-6	AA-6
57	Paper Cutting	–	–	AA-8	AA-A
58	Boxes (3)	AA-4	SAI-2	–	–
59	Enclosed Box (4)	–	–	–	SAI-2
60	Ingenuity (3)	–	–	SAII-2	SAII-4
61	Codes II (1)	–	–	SAII-5	SAII-A
62	Code	AA-6	–	–	–
63	Paper Cutting II	SA-2	SAIII-4	–	–
64	Reasoning	–	SAIII-5	–	SAIII-5
65	Ingenuity	SA-6	–	–	–

Note: Item numbers represent the age level and item number; for example, II-1 is the first item at the age 2 level. IIH = age 2½ items; AA = Average Adult level; SAI = Superior Adult I; SAII = Superior Adult II; SAIII = Superior Adult III; -A = represents alternative item

These items were treated as distinct instead of requiring assumptions regarding their equivalence. This data collation leads to 3,566 people-occasions measured on 130 items.

5.4 Models

5.4.1 Measurement Models

We focus on a strong measurement model to account for the within-time relationships among the nonverbal intelligence items. We begin with a longitudinal one-parameter logistic (1PL) or Rasch model (Rasch, 1960). A longitudinal 1PL model can be written as

$$\ln\left(\frac{P(X_i[t]=1)_n}{1-P(X_i[t]=1)_n}\right)=\theta[t]_n-\beta_i \tag{5.1}$$

where $\theta[t]_n$ is person n's ability at time t, β_i is item i's difficulty parameter, and $P(X_i[t]=1)_n$ is the probability that person n answered item i correctly at time t given the person's ability and item's difficulty. The longitudinal 1PL model was

Table 5.3 Nonverbal Items From the Wechsler Intelligence Scales and Their Overlap Across Test Forms

Subscale & item number	Item	Wechsler-Bellevue	Wechsler Adult Intelligence Scale	Wechsler Adult Intelligence Scale-Revised
01	Picture Completion	–	1	1
02	Picture Completion	8	–	–
03	Picture Completion	–	–	2
04	Picture Completion	–	–	3
05	Picture Completion	4	5	4
06	Picture Completion	–	4	–
07	Picture Completion	–	–	5
08	Picture Completion	10	6	6
09	Picture Completion	–	7	7
10	Picture Completion	–	–	8
11	Picture Completion	–	9	–
12	Picture Completion	–	–	9
13	Picture Completion	–	12	–
14	Picture Completion	–	–	10
15	Picture Completion	11	16	–
16	Picture Completion	–	–	11
17	Picture Completion	5	15	–
18	Picture Completion	–	–	12
19	Picture Completion	–	8	13
20	Picture Completion	15	18	14
21	Picture Completion	–	–	15
22	Picture Completion	–	–	16
23	Picture Completion	–	–	17
24	Picture Completion	–	19	18
25	Picture Completion	14	21	19
26	Picture Completion	–	20	20
27	Picture Completion	6	2	–
28	Picture Completion	1	3	–
29	Picture Completion	13	10	–
30	Picture Completion	–	11	–
31	Picture Completion	–	13	–
32	Picture Completion	7	14	–
33	Picture Completion	–	17	–
34	Picture Completion	2	–	–
35	Picture Completion	3	–	–
36	Picture Completion	9	–	–
37	Picture Completion	12	–	–
38	Block Design	–	1 (Time = 60)	1 (Time = 60)
39	Block Design	–	2 (Time = 60)	2 (Time = 60)
40	Block Design	1 (Time = 75)	3 (Time = 60)	
41	Block Design	2 (Time = 75)	4 (Time = 60)	3 (Time = 60)
42	Block Design	3 (Time = 75)	5 (Time = 60)	4 (Time = 60)
43	Block Design	4 (Time = 75)	6 (Time = 60)	5 (Time = 60)
44	Block Design	5 (Time = 150)	7 (Time = 120)	6 (Time = 120)
45	Block Design	6 (Time = 150)	8 (Time = 120)	7 (Time = 120)
46	Block Design	–	9 (Time = 120)	8 (Time = 120)
47	Block Design	–	10 (Time = 120)	9 (Time = 120)
48	Block Design	7 (Time = 196)	–	–
49	Picture Arrangement	–	1 (Time = 60)	–
50	Picture Arrangement	1 (Time = 60)	2 (Time = 60)	1 (Time = 60)
51	Picture Arrangement	2 (Time = 60)	3 (Time = 60)	–
52	Picture Arrangement	–	4 (Time = 60)	4 (Time = 60)
53	Picture Arrangement	–	5 (Time = 60)	–

(*continued*)

Table 5.3 (continued)

Subscale & item number	Item	Wechsler-Bellevue	Wechsler Adult Intelligence Scale	Wechsler Adult Intelligence Scale-Revised
54	Picture Arrangement	4 (Time = 120)	6 (Time = 60)	2 (Time = 60)
55	Picture Arrangement	6 (Time = 120)	7 (Time = 120)	–
56	Picture Arrangement	5 (Time = 120)	8 (Time = 120)	10 (Time = 120)
57	Picture Arrangement	-	–	3 (Time = 60)
58	Picture Arrangement	–	–	6 (Time = 90)
59	Picture Arrangement	–	–	7 (Time = 90)
60	Picture Arrangement	–	–	9 (Time = 120)
61	Picture Arrangement	3 (Time = 60)	–	–
62	Picture Arrangement	–	–	5 (Time = 90)
63	Picture Arrangement	–	–	8 (Time = 90)
64	Object Assembly	1 (Time = 120)	1 (Time = 120)	1 (Time = 120)
65	Object Assembly	–	2 (Time = 120)	2 (Time = 120)
66	Object Assembly	2 (Time = 180)	–	–
67	Object Assembly	3 (Time = 180)	3 (Time = 180)	3 (Time = 180)
68	Object Assembly	–	4 (Time = 180)	4 (Time = 180)

then extended to accommodate multicategory (polytomous) response formats because several items from the Wechsler tests have partial credit scoring. This model can be written as

$$\ln\left(\frac{P(X_i[t]=x)_n}{1-P(X_i[t]=x)_n}\right)=\theta[t]_n-\delta_{ij} \tag{5.2}$$

where $P(X_i[t]=x)_n$ is the probability the response of person n to item i is in category x, given the response is either in category x or $x-1$, and δ_{ij} is the step-difficulty for step j of item i. This measurement model is a straightforward longitudinal extension of Masters's partial-credit model (Masters, 1982). In both equations, we note that person ability is time dependent and item difficulty (or step difficulty) does not depend on time. It is also important to note that we are going to estimate $\theta[t]_n$ for each person at each measurement occasion.

5.4.2 Longitudinal Models

To model lifespan changes in nonverbal ability, we use growth curves with an interest in exponential change patterns, as exponential patterns have been found to adequately fit lifespan changes in a variety of cognitive abilities (see McArdle, Ferrer-Caja, Hamagami, & Woodcock, 2002; McArdle et al., 2009). The basic growth curve for the ability estimates can be written as

$$\begin{aligned}\theta[t]_n &= g_{0n}+g_{1n}(A[t])+e[t]_n\\ g_{0n} &= \mu_0+d_{0n}\\ g_{1n} &= \mu_1+d_{1n}\end{aligned} \tag{5.3}$$

where g_{0n} is the intercept for person n, g_{1n} is the slope for person n, A is a vector of basis functions, and $e[t]_n$ is the time-dependent (Level 1) residual term. The Level 1 intercept and slope are decomposed into a sample-level (Level 2) means (μ_0 and μ_1) and individual deviations (d_{0n} and d_{1n}) from the sample-level means. Individual deviations are assumed to be multivariate normally distributed with a mean of 0, variances (σ_0^2 and σ_1^2), and a covariance (σ_{01}). The time-dependent residuals are assumed to be normally distributed with a mean of 0 and a single variance (σ_e^2). We fit the growth curves with the following basis functions: level only ($A[t]=0$), linear ($A[t]=t$), exponential ($A[t] = (1 - e^{-\pi_g \cdot t})$), and dual exponential ($A[t] = (e^{-\pi_d \cdot t} - e^{-\pi_g \cdot t})$). The dual exponential was of specific interest because the model captures growth and decline as π_d is the decline rate and π_g is the growth rate. With this model, we can test whether the decline rate is significantly different from zero ($\pi_d \neq 0$). The decline rate is an important parameter in the lifespan development of cognitive abilities because a significant decline rate indicates ability declines as adults age.

5.4.3 Analysis Plan

There are several alternative ways to analyze these kinds of data using these types of models (see McArdle et al., 2009). For clarity here, we only present a two-phase approach for analyzing the longitudinal item-level data. In the first phase the measurement model (Equation 5.2) is fit to the longitudinal item-level data, without capitalizing on the repeated measures nature of the data. The benefit of this first analysis is that it provides a linking equation for conversion of the observed score pattern of any test (items administered) at any occasion. In any case where a person has responded to a series of items, theoretical ability scores ($\theta[t]_n$) from the overall measurement model are estimated and treated as observed longitudinal data in the second phase. The explicit assumption of invariance of the construct over time is similar to those made for scales using metric factorial invariance (see McArdle, 1994, 2007). However, we recognize that the lack of common overlapping items within occasions makes it difficult to reject this strict invariance hypothesis, so our original substantive choice of item content is critical.

In the second phase the linked scores (ability estimates) from Step 1 are treated as observed data for each person at each time and the within-person changes in the ability estimates are modeled using growth curves (Equation 5.3). On the other hand, the combined model (Equations 5.2 and 5.3) can be estimated, and this approach is often seen as optimal, as it produces easy-to-use parameter estimates and allows the modeling of nonverbal ability as a latent entity instead of an observed (estimated) entity.

Although we do not want to treat this two-phase approach as optimal, it certainly is practical. This two-phase approach is not optimal from a statistical point of view—the nonverbal ability scores have to be estimated using the prior model assumptions, and these assumptions are likely to have some faults. However, as we

demonstrate here, this two-phase approach is simple, is computationally efficient, and allows exploration of longitudinal patterns in the ability estimates from the first step (as in McArdle et al., 2009). It is possible to create a joint estimation of the scores within the longitudinal growth models (as in McArdle et al., 2009), and these and more complex programming scripts can be downloaded from http://psychology.ucdavis.edu/labs/Grimm/personal/downloads.html.

5.5 Results

5.5.1 Step 1: Nonverbal Ability Estimates

A total of 65 items from the Stanford-Binet were deemed to measure nonverbal intelligence, 68 items (37 Picture Completion, 11 Block Design, 15 Picture Arrangement, and 5 Object Assembly) from the Wechsler intelligence tests, and a total of 3,566 person-occasions. The partial-credit model was fit to these data; ability estimates were calculated for 3,184 (nonextreme) person-occasions, and item difficulties were calculated for 123 (nonextreme) items. By fitting the partial-credit model to the item-level data in this way, we assumed the item parameters did not vary across time. That is, item difficulty and discrimination were the same for a given item regardless of age, year, and occasion.

Fit of the partial credit model was evaluated in terms of the item fit. Commonly used item fit indices are termed INFIT, OUTFIT, and point biserial correlation. INFIT and OUTFIT consider the amount of noise when the ability level of the participant is close to and far from the item difficulty, respectively. There is an expected amount of noise for each item and person, based on the probabilistic nature of the model. When the amount of noise is as expected, INFIT and OUTFIT statistics will be 1.0. If a person or item is acting too predictably, the person/item is considered muted and the INFIT and OUTFIT will show this by being considerably less than one, while noisy people/items will have values greater than one. Generally acceptable limits on INFIT and OUTFIT statistics are 0.8 – 1.2 with 0.6 – 1.4 being liberal boundaries. It's important to note that the OUTFIT statistic may be unreliable for extreme cases (easiest and hardest). The point biserial correlation is another indication of item fit as it is the correlation between the participant's probability of correctly answering the item and participant's score on the test, basically determining whether people with higher overall scores tend to correctly answer the item. The fit of the nonverbal items to the partial credit model was generally good as only five items showed misfit based on liberal boundaries of INFIT. Based on OUTFIT, 29 items showed misfit with most items showing less noise than expected, which may be a function of repeated testing. Point biserial correlations were positive for 114 items; negative point biserial correlations were found for nine items. Items with negative biserial correlations tended to have few responses.

Person reliability was generally high (.92); however, it might have been overestimated because the repeated measures nature of the data was not accounted for in

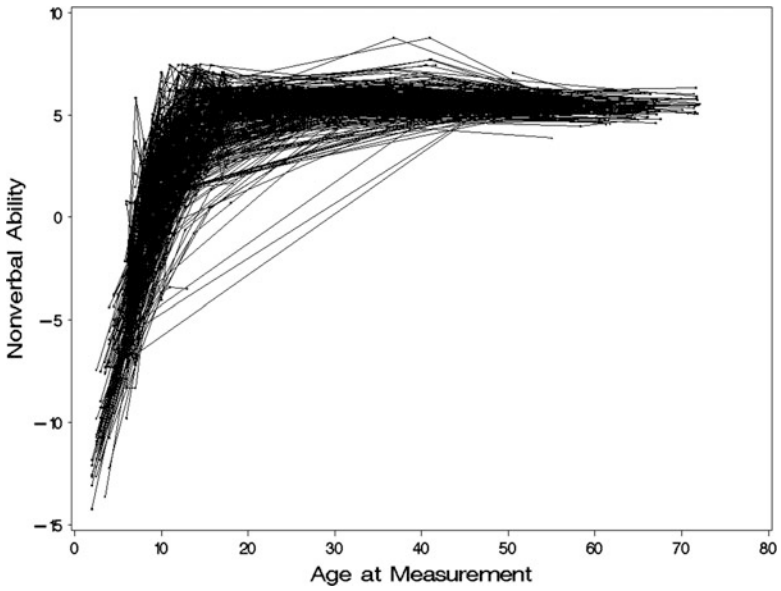

Fig. 5.2 Longitudinal plot of nonverbal ability estimates

this model. Estimates of nonverbal ability were calculated for each person at each occasion using the partial-credit model. The ability estimates were a simple function of the items administered at each occasion and the participant's responses to those items, ignoring age at measurement. The scaling of the ability estimates was such that the average item difficulty was 0 (i.e., $\Sigma\beta\iota = 0$), and between- and within-person differences were scaled in a logit metric reflecting linear probability changes.

After calculating nonverbal ability estimates for each person at each occasion, they were plotted against the persons' age at testing in the lifespan trajectory plot displayed in Figure 5.2. The lifespan trajectories of nonverbal ability are represented for each person, and the ability estimates have, under the partial-credit model, the same interpretation at ages ranging from 2–72. From Figure 5.2, it is easy to see that the trajectories of nonverbal ability rose rapidly through childhood, decelerated during adolescence, flattened out during adulthood, and potentially show a slow but terminal decline into older adulthood. Most importantly, there appear to be sizable individual differences in both the level of nonverbal ability and between-person differences in those changes across the lifespan.

5.5.2 Step 2: Growth Modeling of Nonverbal Ability

Several growth models (i.e., level only, linear, single exponential, dual exponential) were fit to the ability estimates from the partial-credit model. The level-only model provided baseline fit statistics for comparison purposes to determine whether there

Table 5.4 Parameter Estimates From the Dual Exponential Growth Model Fit to the Nonverbal Ability Estimates

Parameter	Parameter estimate	Standard error
Fixed effects		
Intercept (η_0)	5.46	.045
Slope (η_1)	0.71	.098
Growth rate (π_g)	0.20	.006
Decline rate (π_d)	0.09	.011
Random effects		
Intercept (σ_0^2)	0.57	.063
Slope (σ_1^2)	0.01	.002
Intercept-slope covariance (ρ_{01})	−0.02	.007
Residual (σ_e^2)	1.09	.033

were systematic changes in nonverbal ability span. The linear and quadratic models did not converge, indicating they were not appropriate for these data. The single and dual exponential models fit better than the level-only model, and the dual exponential model fit significantly better than the single exponential model (Δ-*2LL* = 81, Δ*parms* = 1), indicating that nonverbal ability declined during older adulthood.

Parameter estimates from the dual exponential model are contained in Table 5.4. The average rate of change was positive (η_1 = .71), and there was significant variation in the average rate of change (${\sigma_1}^2$ = .10). The mean intercept, centered at 20 years, was 5.46, and there was significant variation in nonverbal ability at this age (${\sigma_0}^2$ = .57). There was a negative covariance (σ_{01} = -.02; ρ_{01} = -.31) between the intercept and rate of change such that participants with more nonverbal ability at age 20 tended to have slower rates of change. The expected mean (and between-person deviations) of the age-based latent curve of nonverbal ability is displayed in Figure 5.3. Figure 5.3 shows the sharp increases during childhood before changes in nonverbal ability decelerated, peaked around 30 years of age, and slowly declined through older adulthood. It also becomes apparent that even though the decline rate was significantly different from zero, there was only a small amount of decline in nonverbal ability for participants in these studies.

5.6 Discussion

Analyzing item-level data in longitudinal studies could become the norm in applied longitudinal research because of its many practical benefits, including the possibility of using longitudinal data where the scales have changed (see McArdle et al., 2009). Of course, there are also many limitations to analyzing item-level data, some of which researchers and publishers may need to overcome. That is, it would be possible to create "translation tables" from larger scale cross-sectional studies using common items and apply these to our inevitably smaller longitudinal studies.

The benefits of using item-level data in longitudinal studies include the potential reduction practice effects by administering different tests at different occasions, checks and tests of item drift and differential item functioning across time

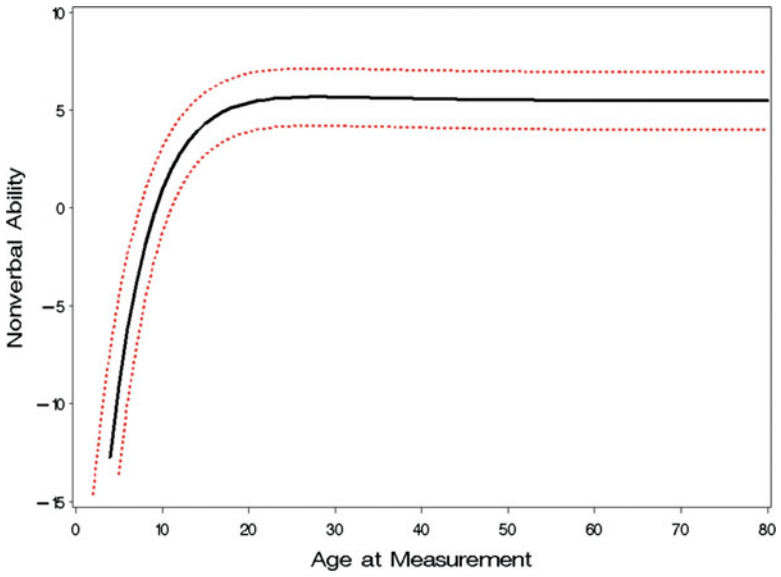

Fig. 5.3 Predicted trajectory for nonverbal ability based on the dual exponential growth model

and age, more precise estimates of ability compared to simple sum scores, and information regarding the relative magnitudes of measurement error within and across time. One way to reduce practice effects is to administer a new test (and therefore items) at each successive occasion. This would reduce item-specific retest effects, which may be contributing to the lack of significant decline in nonverbal ability reported here. Next, tests of measurement equivalence or the invariance of item discrimination and difficulty (and thresholds) parameters can be examined across time or age to make sure the measurement properties are stable. Items may become more difficult or may not relate to the underlying construct in the same way across time.

There are many ways to consider the utility of factorial invariance in longitudinal studies (McArdle, 2007). In one approach to this problem, Horn and McArdle (1992) and McArdle and Cattell (1994) treated factorial invariance as a desirable property—so desirable that the search was extended (over groups) to allow resulting invariant functions, which are highly complex representations of the original data. In contrast, in the approach suggested by Horn, McArdle, and Mason (1983) and Edwards and Wirth (2009), the lack of measurement invariance is not necessarily a bad result, and it largely becomes problematic if measurement invariance is assumed and not tested. Unfortunately, this is exactly the problem of the typical longitudinal study with completely changing measurements, where it becomes difficult to provide formal tests of hypotheses. What is needed in new longitudinal studies is more appreciation of the utility of overlapping items from one time to the

next, because as of now there is no formal way to go back in time to add such useful items (F. M. Lord, personal communication, June 1977).

In this context, longitudinal analysis of item-level data helps our precision in two ways. First, our estimates of ability are more precise since we are only focusing on the items that each participant answered and their response pattern, as opposed to making assumptions regarding the correctness and incorrectness of items that were not administered to the participant based on starting and stopping rules common to intelligence tests. Second, using item-level data allows for estimating the standard error of measurement for each individual response pattern at each occasion. This information can be used to weight data in the estimation of statistical models to provide more precise estimates of important model parameters.

Drawbacks of using item-level data in longitudinal research stem from sample size restrictions and availability of user-friendly software for combining item response models with higher order statistical models to examine applied research questions. Item response models often have many parameters to estimate, which are poorly estimated with small and nonrepresentative samples—the types of samples that are often found in psychological and longitudinal research. One way to overcome this problem is for researchers and test makers to publish item parameters. In this situation, item parameters can be fixed to known values and the benefits of item response models are carried forward to the examination of the applied research question, without having to estimate item parameters with a small and potentially biased sample.

This research is intended to raise new questions about the optimal use of item responses in longitudinal data. For example, it is clear that dropping some items at later occasions is a reasonable technique, especially if the item does not have high discriminatory power. It is also clear that dropping items is reasonable when there is a large mismatch between item difficulty and person ability, and we see this in the administration of commonly used cognitive assessments with their built-in starting and stopping rules. However, it is not yet clear how much statistical power will be lost in the longitudinal assessments if items are dropped, even though they had been presented at earlier occasions. Careful study of the existing item-level longitudinal data can be useful in the determination of what is most reasonable for future studies. But we hope it is also obvious that the goal of this study is to improve the scales used in lifespan dynamics research. This research does not deal with the more difficult situation faced by high-stakes testing, and this longitudinal item-linking approach will certainly need to be improved before these critical issues can be considered.

5.7 Concluding Remarks

Longitudinal data are not ideal for equating tests because of item-level practice effects, item drift, and changes in ability level. Ideally, equating multiple tests would be conducted with large and appropriately aged samples measured at appropriate time periods. However, given the nature of the IGS and BMLS data, this was

not possible. As mentioned, the data described here were weak for linking multiple tests forms. Sample sizes were small at any given occasion, especially when multiple test forms were administered (e.g., $n = 110$ for the BMLS when the WAIS and Stanford-Binet Form L were administered). Additionally, only age-appropriate items were administered at any measurement occasion. Thus, equating at a given occasion would have led to highly unstable item parameters. Instead, we utilized the longitudinal strength of the data and fit an item response model to all of available data leading to more stable estimates of item parameters that are on the same scale. Finally, we leaned heavily on a very simple item response model with strong assumptions (e.g., longitudinal measurement invariance, equal discrimination) that were untestable given our limited data. To the extent that model assumptions are not met by our data, our results are misleading.