Modeling Life-Span Growth Curves of Cognition Using Longitudinal Data With Multiple Samples and Changing Scales of Measurement

John J. McArdle University of Southern California Kevin J. Grimm University of California, Davis

Fumiaki Hamagami University of Virginia Ryan P. Bowles
Michigan State University

William Meredith University of California, Berkeley

The authors use multiple-sample longitudinal data from different test batteries to examine propositions about changes in constructs over the life span. The data come from 3 classic studies on intellectual abilities in which, in combination, 441 persons were repeatedly measured as many as 16 times over 70 years. They measured cognitive constructs of vocabulary and memory using 8 age-appropriate intelligence test batteries and explore possible linkage of these scales using item response theory (IRT). They simultaneously estimated the parameters of both IRT and latent curve models based on a joint model likelihood approach (i.e., NLMIXED and WINBUGS). They included group differences in the model to examine potential interindividual differences in levels and change. The resulting longitudinal invariant Rasch test analyses lead to a few new methodological suggestions for dealing with repeated constructs based on changing measurements in developmental studies.

Keywords: latent curve analysis, item response modeling, changing longitudinal measurements, life-span cognition, multilevel item analysis

Classical research on cognitive abilities has provided information about the growth and decline of intellectual abilities over the life span (Cattell, 1941, 1998; Horn, 1988, 1998). Many recent analyses of this topic use some form of longitudinal mixed-effects, multilevel latent curve models

(McArdle, 1986, 1988; McArdle, Ferrer-Caja, Hamagami, & Woodcock, 2002; McArdle & Nesselroade, 2003; Meredith & Tisak, 1990). One of the basic measurement assumptions of all latent curve models is *longitudinal measurement equivalence*; that is, the same unidimensional at-

John J. McArdle, Department of Psychology, University of Southern California; Kevin J. Grimm, Department of Psychology, University of California, Davis; Fumiaki Hamagami, Center for Advanced Study of Teaching and Learning, University of Virginia; Ryan P. Bowles, Department of Psychology, Michigan State University; William Meredith, Department of Psychology, University of California, Berkeley.

John J. McArdle was supported by National Institute on Aging (NIA) Grants AG02695, AG04704, and AG07137. Kevin J. Grimm was supported in part by NIA Training Grant T32 AG20500-01 and by an Institute of Education Sciences training grant from the University of Virginia when he attended the university. Ryan P. Bowles was supported by NIA Training Grant T32 AG20500-01 when he attended the University of Virginia.

This work is dedicated to our late coauthor, William Meredith, who had the original idea for the Longitudinal Item Response Theory type analysis. Of course, these analyses would not have been possible

without the cooperation of the participants of the Bradway-Longitudinal Study and the Inter-Generational Studies. The extensive recoding of the item level was carried out by the Kevin J. Grimm in collaboration with Anna Hammarskjold. We also thank our colleagues, John L. Horn, John R. Nesselroade, Karen M. Schmidt, and Carol A. Prescott, for their assistance in this research and for their comments on drafts of this article. Early versions of this article have been presented at the Society of Multivariate Experimental Psychology, University of Virginia (October 2002); at the Working Conference on the Future of Longitudinal Studies, Institute of Human Development, University of California, Berkeley (March 2003); and at the annual meeting of the Gerontological Society of America, San Diego, CA (November 2003).

Correspondence concerning this article should be addressed to John J. McArdle, Department of Psychology, University of Southern California, 3620 South McClintock Avenue, SGM 501, Los Angeles, CA 90089-1061. E-mail: jmcardle@usc.edu

tribute is measured on the same persons with the same scale of measurement at every occasion (see Smith, 1996; Smith & Suh, 2003; van der Linden, 1994; Wilson, 2005; Wright & Linacre, 1994). Tests of these assumptions start by measuring the same variables at each occasion and considering tests of factorial invariance (e.g., McArdle, 2007). However, the classical requirements of exactly equivalent scales of measurement is often impractical and not often achieved. These measurement issues have been raised in classic treatments of the analysis of change (e.g., Cattell, 1966; Harris, 1963; Wohlwill, 1973) but have not fully been resolved (e.g., Burr & Nesselroade, 1990; Collins & Sayer, 2001).

One creative solution to this problem of changing scales was illustrated in the work of Bayley (1956) in her analysis of data from the seminal Berkeley Growth Study. Individual growth curves of mental abilities from birth to age 26 years were plotted for a selected set of male (see Figure 1A) and female participants (see Figure 1B). In the early stages of this data collection (circa 1929), Bayley (among many others) assumed that any measurement occasion within each study should incorporate an age-appropriate intelligence test, that is, a version of the Stanford-Binet Intelligence Scale (Becker, 2003; McNemar, 1942) at ages 6-17, then the Wechsler-Bellevue Intelligence Scale (D. R. Duncan & Barrett, 1961; Wechsler, 1941) at ages 16-26. Although these tests measure specific intellectual abilities, they are not administered or scored in the same way and may measure different intellectual abilities at the same or different ages. However, Bayley was interested in using the statistical techniques applied to physical growth curves, so she created the individual growth curves represented in Figure 1 by

adjusting the means and standard deviations of different mental ability tests at different ages into a common metric (based on z scores formed at age 16). As Bayley suggested,

They are not in "absolute" units, but they do give a general picture of growth relative to the status of this group at 16 years. These curves, too, are less regular than the height curves, but perhaps no less regular than the weight curves. One gets the impression both of differences in rates of maturing and of differences in inherent capacity. (p. 66)

She also noticed the striking gender differences in dispersion of the resulting curves.

This classic study can be considered an early application of what is now termed linked or mapped measurement scaling of growth data. The practical scaling method used by Bayley permitted the analysis of fundamental features regarding growth curves of cognition and appeared to put mental growth on the same scientific footing as physical growth. Nonetheless, not all researchers were convinced by the merits of this approach. In one important critique Wohl-will (1973) suggested,

Yet the pooling of data as conceptually diverse as Wechsler–Bellevue raw scores and Stanford–Binet mental age scores is surely suspect. For the reasons previously indicated, growth functions based on the latter are altogether artifactual, so that pooling intelligence test scores from this scale together with other intelligence test scores can hardly be expected to yield useful information concerning the growth function. (pp. 75–76)

These critiques highlight important technical concerns about the possible and most appropriate ways to examine these issues. In this article, we use concepts from item

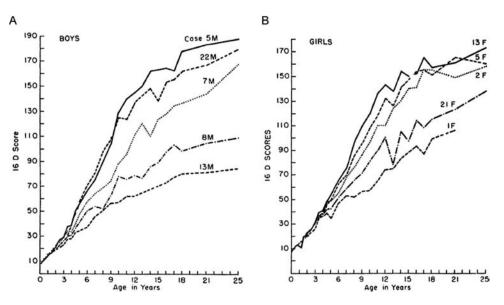


Figure 1. Growth curves of intellectual abilities from the Berkeley Growth Study of Bayley (1956; age 16 D scores). From "Individual Patterns of Development," by N. Bayley, 1956, *Child Development*, 27, p. 67. Copyright 1956 by Wiley-Blackwell.

response theory (IRT) to create measurement linkages for tests even though the same measurement device was not used on all occasions. We merge IRT with concepts from latent curve modeling (LCM) for examining growth and change over age using data pooled from multiple longitudinal samples.

Longitudinal Growth and Change Modeling

A great deal of prior work on structural equation modeling (SEM) has been devoted to the problems of longitudinal analysis. This includes important work on the autoregressive simplex models, as well as new ways to deal with common factors and measurement error in panel data (Horn & McArdle, 1980; Jöreskog & Sörbom, 1979; Sörbom, 1975; Wiley & Wiley, 1970). Meredith and Tisak (1990) proved how the Tuckerized curve models (Tucker, 1958; cf. Rao, 1958) could be represented and fitted with SEM based on restricted common factors. These SEM representations of growth curve models offered the possibility of representing a wide range of alternative models and quickly led to other methodological and substantive studies (McArdle, 1986, 1988, 1989). The LCM approach to modeling change has since been expanded upon and used by many researchers (e.g., S. C. Duncan & Duncan, 1995; Fan, 2003; Metha & West, 2000; Muthén & Curran, 1997; Tisak & Tisak, 1996; Willett & Sayer, 1994). LCMs can now be estimated from observed raw score longitudinal data that are both unbalanced and incomplete with maximum likelihood estimation (MLE; Hedeker & Gibbons, 1997; Little & Rubin, 1987; McArdle & Anderson, 1990; McArdle & Bell, 2000; McArdle & Hamagami, 1992, 2001). These formal developments in LCM overlap with many recent developments of multilevel models (Bryk & Raudenbush, 1992; Goldstein, 1995; Snijders & Bosker, 1994; Woodhouse, Yang, Goldstein, & Rasbash, 1996) or mixed-effects models (Singer, 1998; Verbeke & Molenberghs, 2000). The important work by Browne and du Toit (1991) showed how the nonlinear dynamic models can also be considered in this same framework (see Cudeck & du Toit, 2003; McArdle & Hamagami, 1996, 2001; Pinherio & Bates, 2000).

These LCM developments also led to a revival of practical experimental variations based on planned incomplete data. For example, in both McArdle et al. (2002) and McArdle and Woodcock (1997), the cognitive test–retest data were collected by a design with varying intervals of time—a time-lag design. This time-lag design created subgroups of persons based on the time delay between testings, and a planned incomplete data approach was used to estimate parameters in a pooled analyses. Similar incomplete data latent growth models have also been used when the separate group data were neither planned nor randomly selected. In these cases, the available incomplete data pooling techniques could still be applied to describe a limited

mixture of age-based and time-based models with only two-time points of data collection, that is, an accelerated longitudinal design (Bell, 1953, 1954; McArdle & Anderson, 1990; McArdle & Bell, 2000; McArdle & Woodcock, 1997), and a multiple-group pooling approach proved especially useful in dealing with cohort-sequential data collections in studies across the life span (S. C. Duncan & Duncan, 1995; McArdle & Hamagami, 1992; cf. Swanson, 1999).

A seemingly separate literature has focused on the estimation of latent traits via IRT models (Bond & Fox, 2001; De Boeck & Wilson, 2004; Embretson & Reise, 2000; Fischer, 1995; Rost, 2002; van der Linden & Hambleton, 1997). IRT can be considered a collection of models designed to yield estimates of one or more latent traits based on responses to a set of individual items, whether binary (dichotomous) or multicategory (polytomous). The basic model of Rasch (1960, 1966) for dichotomous items was expanded by Fischer (1987, 1989) for use in the measurement of change over time (see also Fischer & Parzer, 1991; Fischer & Seliger, 1997). In one SEM-IRT-type integration, Hamagami (1998) showed how the longitudinal invariance of the factorial structure of a set of items could be evaluated through available SEM software (e.g., LISCOMP). This approach has been extended with general SEMs that show that all IRT models are latent variable SEMs with discrete indicators, so standard distinctions are artificial (e.g., Mc-Donald, 1999; Muthén, 2002; Skrondal & Rabe-Hesketh, 2004). Other aspects of longitudinal item response models are found in the work of Embretson (1991), Ferrando (2002), Mislevy and Wilson (1996), and Wilson (1989), and some key numerical issues have been raised by Feddag and Mesbah (2005), Fischer (2003), Pastor and Beretvas (2006), and Rijmen, De Boeck, and van der Maas (2005).

Changing Scales in Longitudinal Data

One of the most basic measurement requirements in all latent curve analyses is longitudinal measurement equivalence, in which the same attribute is measured on the same person in the same scale at every occasion. For practical reasons, many longitudinal researchers make sure to use the same tests (or items) at every repeated occasion. However, even when such precautions are taken, the meaning and function of the tests can change (Cattell, 1966; Horn & McArdle, 1992; McArdle & Cattell, 1994). When the same measures are repeated on the same persons, the assumption of *metric factorial invariance* over time can be examined with longitudinal SEMs (e.g., Hancock, Kuo, & Lawrence, 2001; Leite, 2007; McArdle, 1988, 2007; McArdle & Nesselroade, 1994, 2003; Sayer & Cumsille, 2001).

But another common scenario in longitudinal studies is when the primary measurements are not the same from one occasion to the next. Scales are altered for many good reasons, including age appropriateness, bad experiences in prior usage of some tests, and new and improved test batteries (see M. C. Jones, Bayley, Mcfarlane, & Honzik, 1971; Wohlwill, 1973). These considerations are sensible and practical, so it is likely that changing measurements of the same constructs will remain a feature of developmental research for many years to come. Unfortunately, researchers using growth curve methods cannot be sure how to separate differences in the scales over time from changes in the constructs over time. More specifically, if there are differences in test forms and test batteries, the difficulty of this problem increases. This problem is relevant in life-span developmental studies because researchers often include a wide range of ages or attempt to combine different studies based on different groups of persons measured on similar constructs.

Longitudinal researchers have approached these problems in several ways, through over-time prediction, within-occasion rescoring, absolute scaling, SEM with convergent factor patterns, and IRT linkage of common items.

Over-Time Prediction

A popular solution to this problem is simply to choose a mathematical-statistical model that does not require identical measurements. This choice is often made indirectly, such as when the analyst simply describes the correlations in scores over time and the constancy of individual differences (M. C. Jones et al., 1971). In a multiple regression prediction over time, earlier scores are used to predict later scores, and regression type models based on latent variables are popular (Jöreskog & Sörbom, 1979). Such an approach seems necessary in long-term longitudinal research in which the key constructs are considered to be different from one time to the next (i.e., in childhood vs. adulthood). However, these prediction models do not attempt to directly estimate change over time at the individual level (McArdle & Nesselroade, 2003; Nesselroade & Baltes, 1979).

Within-Occasion Rescoring

Another simple alternative is based on transforming the raw scores into z scores within a time point (Bayley, 1956; see Figure 1). Under the assumption that the same construct is measured by two scales, one variable is simply transformed into another scale by a regression calculation, usually with a focal occasion (e.g., 16 years old). This approach can lead to several potential problems. First, it removes the means and standard deviations within each time or age, so any systematic growth or change related to the these statistics cannot be easily identified. Second, the accuracy is limited by the size of the correlation among the observed scores, and the estimated scores will be attenuated unless a residual term is also imputed. Third, longitudinal data collections often have many cross-battery rescalings to con-

sider (e.g., for variables *W*, *X*, *Y*, and *Z*), so many regressions could be imputed with multiple pairs, or as triplets, or as an aggregate of several occasions of overlap. Finally, the measured sample at any time may be shifting because of attrition, yet each regression can be conducted only with a selected part of the available sample.

Absolute Scaling

A related longitudinal scaling method was proposed by Thurstone (1928) for cognitive scores. In this work, he first assumed that cognition over age could be represented as a linear growth model with linearly increasing means and linearly increasing standard deviations. Next, he assumed that any measurement of this construct at any age should be rescaled to fit this linear growth pattern. Under these assumptions, the group means and deviations of different tests at different ages should follow a linear pattern back to the starting point (i.e., the variance went to absolute zero). This creative approach used the latent curve parameters to create weights for a common metric. These linear assumptions seem unreasonable, so there seem to be very few applications of this scaling method (e.g., McArdle, 1988).

SEM With Convergent Factor Patterns

It is possible to consider a simultaneous structural model for the multiple scales in which some measures are missing at some occasions. The general problem of scale overlap can be seen in a common factor model of multivariate repeated measures in which the longitudinal data are accounted for by constant loadings but changing factor scores—a curveof-factor-scores model (Leite, 2007; McArdle, 1988, 1989). To the degree that multiple measurements are made at multiple occasions, this common factor hypothesis about the change pattern can be estimated and even rejected (McArdle et al., 2002; McArdle & Woodcock, 1997). It has been shown that the critical assumption of invariant loadings over time at the first order allows researchers to model the changes in the common factor scores in terms of a latent growth model at the second order (Hancock et al., 2001; Leite, 2007; McArdle, 1988, 1989; McArdle & Woodcock, 1997; Sayer & Cumsille, 2001). A first problem is that the scaling of the factor scores requires some fixed factor means, and this is typically done by assigning a zero factor mean at the first occasion (e.g., see Cattell, 1972; McArdle, 1988, 1989). With changing measures, we assume the same factor score can be measured by different variables at the different times, and the common factor equation is expressed for the possible patterns of measurement for the multiple measures. Although this is a compelling idea, it is important to consider how much observed information is needed to identify and estimate these model parameters, especially when using standard SEM computations based on high-dimensional integration. In general, the identification status and the ease of estimation of the covariance parameters depend on the number of observed measurements and occasions and on the overlap or coverage of multiple batteries within each time. If we have a large number of changing measurements, we can end up with low covariance coverage, and a common factor measurement model with or without invariance may not be identified (see McArdle, 1994; McArdle & Hamagami, 2004; McArdle & Woodcock, 1997).

IRT Linkage of Common Items

As a contemporary combination of the approaches listed above, we can try to estimate common scales of measurement using an IRT approach for the items in changing scales. In one specific form of this model, we can postulate Rasch-based restrictions—a single common factor for a large set of items, including equal loadings on the items. These simplifying assumptions allow planned missing items within occasion because of the experimental design (i.e., not because of the person's refusal; McArdle, 1994). This form of incomplete data IRT approach can be used to estimate a measurement model of the common traits over time, but unless some form of item overlap is present, it will not be possible to test the validity of the measurement invariance constraints (see Smith, 1996; Smith & Suh, 2003; van der Linden, 1994; Wilson, 2005; Wright & Linacre, 1994). In any case, further longitudinal modeling of the factor score estimates can be based on mixed-model latent growth curve analyses. Given enough information, the parameters of growth measurement models can be estimated simultaneously (Feddag & Mesbah, 2005; Fischer, 1989; Pastor & Beretvas, 2006; Rijmen et al., 2005). The strengths and weakness of each approach, as well the computational techniques available to carry out the estimation, have only recently been explored, but the general idea of fitting a longitudinal growth model together with an IRT model fits naturally into the contemporary multilevel IRT research (e.g., De Boeck & Wilson, 2004; Fox & Glas, 2001; Goldstein & Browne, 2002; Kamata, 2001).

Summary of the Current Approach

In the next section, we present details on real life-span longitudinal data in which, on at least one occasion, one or two out of a total of eight standardized tests have been administered. We describe eight cognitive scales in which the choice of the specific measurement for each occasion is slightly different but based on the most well regarded and age-appropriate tests available at the time. To limit these analyses, we focus on two key constructs in cognitive aging research: vocabulary and memory. Next, we define the features of multilevel latent curve models that allow us to describe the life-span features of the growth and changes of

these abilities. We also emphasize our assumptions about incomplete data. We propose a longitudinal growth model based on a longitudinal invariant Rasch test (LIRT), define the simplifying model assumptions, and show how this approach can be used to bring different measures of the same construct into a common scale. We consider several techniques for linkage across measurement scales and across multiple groups, and we fit a unidimensional Rasch model to item responses and a latent curve model together with changing latent scores over age and groups.

These LIRT models can be estimated either in multiple stages or by simultaneous methods, and we highlight the simultaneous estimation methods in this article (cf. McArdle & Nesselroade, 2003). In general, we emphasize how these kinds of LIRT analyses can provide a new and practically useful solution to the classic problem of changing measurement scales with different groups. We do not view LIRT as a methodological breakthrough but instead see the LIRT approach as a practical integration of important theoretical questions and contemporary modeling principles. We also show how LIRT leads to key issues for future longitudinal data collections and analyses.

Method

Participants

The data come from three long-term longitudinal studies of the growth (and decline) of a variety of cognitive abilities (see Table 1). The combination of these studies leads to N=419 persons measured from T=1 to 16 occasions on ages ranging from A=2 to 72 years (McArdle, Hamagami, Meredith, & Bradway, 2000). An overview of the cognitive testing in these three studies follows.

Berkeley Growth Study (BGS). The BGS (n = 75) was initiated by Nancy Bayley in 1928 to trace the normal intellectual, motor, and physical development through the first year (see M. C. Jones et al., 1971). The original participants of the BGS were selected for study as infants born in local hospitals in Berkeley, CA. Data collection continued through childhood and adolescence, with the sample taking an intelligence test every year until age 18. The sample was measured repeatedly on these kinds of cognitive tests at ages 21, 26, 36, 50, and 70. The two most recent measurement occasions included the spouses of the subjects, bringing the total sample size to 124. The sample is approximately half male (63) and half female (62).

Guidance–Control Study (GCS). The GCS (n=206) also began in 1928 by Jean Mcfarlane (also in M. C. Jones et al., 1971). The participants were selected from the population of every third infant born in Berkeley between January 1928 and June 1929. Half the GCS participants' mothers were offered guidance by the principal investigator about general issues of childhood behavior and develop-

Table 1
Summary of Available Data From Multiple Testing Occasions for Three Longitudinal Studies

Age	Berkeley Growth	Guidance-Control	Bradway-McArdle
(years)	(n = 61)	(n = 206)	(n = 111)
2-51/2			SB-L, SB-M (111)
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)	
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)		
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)
53	WAIS-R (41)	WAIS-R (118)	WAIS (53)
63			WAIS, WJ-R (48)
67			WAIS, WJ-R (33)
72	WAIS-R, WJ-R (31)		

Note. Available sample size for specific tests is contained in parentheses. SB-L, SB-M, SB-LM = Stanford—Binet Forms L, M, and LM; WB-I = Wechsler–Bellevue Intelligence Scale Form I; WAIS = Wechsler Adult Intelligence Scale; WAIS-R = Wechsler Adult Intelligence Scale—Revised; WJ-R = Woodcock–Johnson Psycho-Educational Battery–Revised.

ment (see C. J. Jones & Meredith, 2000). The participants whose mothers received guidance were the guidance group; the other participants were the control group. The participants from this study were repeatedly measured on various forms of intelligence tests almost every year from age 6 to age 15 and then at ages 18, 40, and 53. The final two measurement occasions also included measurements of their spouses.

Bradway–McArdle Longitudinal (BML) Study. The BML (n = 111) began in 1930 when the participants were tested in the Bay Area of California as part of the standardization sample of the Revised Stanford–Binet Intelligence Scale (see Bradway & Thompson, 1962). At the initiation of the study, the participants were between the ages of 2 and 5 years 6 months. Katherine Bradway retested the participants in 1941 for her doctoral thesis and continued the longitudinal study by testing the participants in 1957 and 1969. McArdle and colleagues followed up these participants in the 1980s and the 1990s (see McArdle et al., 2000). The Bradway-McArdle and both Berkeley samples are predominately Caucasian, from approximately the same cohort (circa 1928), and from the San Francisco area and have above-average social economic status.

Available Vocabulary and Memory Measurements

The raw data for all samples were recoded at the item level from the archival records of all three longitudinal studies. This analysis focuses on the Vocabulary and Memory Span items, and specific scoring details are presented in Tables 2 and 3. The Wechsler (1941) and Woodcock-Johnson scales (McArdle et al., 2002; Woodcock, 1990) have a specific set of items that are used to form the Vocabulary and Memory Span subscales. Adding complexity to this problem, the scoring procedures for these items change from one version to the next. The use of the Stanford-Binet scales (Becker, 2003; McNemar, 1942) is a more complex problem because the Vocabulary and Memory Span items are scored in a different way and are interspersed throughout the entire scale. In addition, the items representing the Vocabulary and Memory Span constructs are presented only if a person reaches a specific age level on the overall test.

It is useful to consider the multiple patterns of available test data. At each occasion of measurement, participants received only one or two of the eight cognitive test batteries. In all three studies combined, with eight scales, it is possible for many combinations of overlap or coverage among dif-

Table 2
Descriptions of Administration and Scoring of Alternative Scales: Scoring of Alternative Vocabulary Tests

Vocabulary test	Items	Scoring	Response type ^a	Stopping rule	Item overlap
1916 SB	50	0, ½, 1, converted to 0, 1, 2	Define	5 consecutive failures	34 with SB-L, SB-LM
SB Forms L and LM ^b	45	0, 1	Define	5 consecutive failures	34 with 1916 SB
WB Form I	42	0, 1, 2	Define	5 consecutive failures	
WAIS	40	0, 1, 2 ^c	Define	5 consecutive failures	33 with WAIS-R
WAIS-R	35	0, 1, 2	Define	5 consecutive failures	33 with WAIS
WJ-R Oral Vocabulary Synonyms	20	0, 1	Name Synonym	4 consecutive failures	1 with WAIS
WJ-R Oral Vocabulary Antonyms	24	0, 1	Name Antonym	4 consecutive failures	
WJ-R Picture Vocabulary	58	0, 1	Name Object	6 consecutive failures	

Note. SB = Stanford–Binet; SB-L, SB-LM = Stanford–Binet Forms L and LM; WB = Wechsler–Bellevue Intelligence Scale; WAIS = Wechsler Adult Intelligence Scale; WAIS-R = Wechsler Adult Intelligence Scale—Revised; WJ-R = Woodcock–Johnson Psycho-Educational Battery–Revised.

a Define = participant is asked to define the word presented; Name Synonym = participant is asked to name a word that has the same meaning as the word presented; Name Antonym = participant is asked to name a word that has the opposite meaning as the word presented; Name Object = participant is asked to name the object that is presented visually.

b Stanford–Binet Form M does not contain a vocabulary test.

c Except for the first three items, which are scored 0, 2.

ferent scales. However, as seen in Table 1, there are only six instances in which there was a direct administration of more than one scale at the same occasion (e.g., the Stanford–Binet Form L and the Stanford–Binet Form M were administered in the BML in 1931 at ages 2 years to 5 years 6 months).

Substantially more coverage was available at the item level. In addition to overlap because two scales were administered at the same testing occasion, item overlap occurred because the same item was administered on multiple test forms. For Vocabulary, 34 common items exist on the 1916 Stanford–Binet and the Stanford–Binet Form L, 45

items are common to the Stanford–Binet Forms L and LM, and 33 items are common to the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1955) with the WAIS-Revised (WAIS-R; Wechsler, 1981). For Memory, Digit Span items of a common length appear in the 1916 Stanford–Binet and the three revised editions of the Stanford–Binet (Forms L, M, and LM), and the Digit Span items of the Wechsler–Bellevue are the same as those of the WAIS and WAIS-R. (This assumes that Digit Span items of the same length are the same items even though the specific numerals the participants were asked to remember were different.)

Table 3

Descriptions of Administration and Scoring of Alternative Scales: Scoring of Alternative Memory Span Tests

Intelligence test	Item type	Attempts	Stopping rule
Stanford–Binet	Digits Forward 3 NA		NA
	Digits Backward	3	NA
	Memory for Words	3	NA
	Memory for Sentences	3	NA
Wechsler Memory	Digits Forward	2	2 incorrect on same length
	Digits Backward	2	2 incorrect on same length
Woodcock-Johnson-Revised	k–Johnson–Revised Memory for Words 1		4 consecutive failures
	Memory for Sentences	1	4 consecutive failures

Note. All items were scored 0 = incorrect, 1 = correct, except for the Woodcock–Johnson–Revised Memory for Sentences (0 = wrong, 1 = partial, 2 = right). The items in the Stanford–Binet are presented by difficulty, not by type of intelligence. Participants were not administered every Memory Span item on the Stanford–Binet. The items presented depended on each participant's individually assigned starting and stopping points for each test. NA = not applicable.

Another potentially important issue is that incomplete data are created within each test because of starting and stopping rules on each of the scales. Items were skipped when thought to be too easy (credit is given), and items the participant did not reach were assumed to be too hard (credit is not given). In all analyses presented here, we treated items that were not administered as missing. This was a conservative strategy, so we examined several other ways to score the items, including giving credit to items before the starting point (assumed correct), not giving credit to items after the stopping rule (assumed incorrect), and other combinations. A preliminary analysis of these methods, at the item level, produced no notable differences in the results from any model (i.e., the estimates across both coding schemes r(a,b) > .98), so we present only the conservative scoring strategy here.

This data collection raises another practical problem when there is a complete lack of item overlap. In these studies, the Wechsler-Bellevue was never administered at the same time as any of the other tests, and the Vocabulary items on the Wechsler-Bellevue were never given on another test. If we simply ignored these occasions, we would lose potentially valuable longitudinal data, so we explored several alternative solutions. We found that it was most useful to include items from the Information subscale, because both the Information and Vocabulary subscales are highly correlated (r(i,v) > .8), and because the Wechsler– Bellevue Intelligence Scale, WAIS, and WAIS-R share Information items (the Information subscale of the Wechsler-Bellevue shares 16 items with the WAIS and 9 items with the WAIS-R, and the WAIS and the WAIS-R information subscales share an additional 11 items, totaling 20 shared items between the two tests). Using this setup, we found that common persons and common item linking can be used to equate the Wechsler-Bellevue to the WAIS and WAIS-R, which in turn are linked to the other Vocabulary tests. We examined the impact of this use of shared information in several ways, but we could not find any notable impact on the results. For this reason, the only solutions presented here use this overlapping item collection.

Available Longitudinal Data

Given all these considerations, the net result is a large set of scores on cognitive test items. The Vocabulary scores are based on N=419 at an average of T=6 occasions with an average of I=34.9 specific items per testing occasion (i.e., D=2,507 individual records with 87,420 individual item scores). The Memory Span scores are based on N=416 participants at an average of T=7.5 occasions with an average of T=8.4 specific items per testing occasion (i.e., D=3,107 individual records and 25,943 individual items). However, even with all these longitudinal item data, the coverage of all possible items at all possible occasions is

only about 5%. That is, if all these persons had been measured on all possible items (I = 278 or I = 76) at all possible occasions (T = 16), the result would be almost 20 times more data than are currently available. It is clear that the historical choices to administer different tests at specific occasions have created a challenging problem for subsequent developmental analysis and inference.

Models

Latent Growth-Decline Curve Models

The overall goal of the current longitudinal analysis is to examine group and individual differences in the trajectory of people over the full life span, so we start with a focus on growth models. In the latent curve model used here, we assume we have observed the variable Y on persons (n = 1 to N) at multiple occasions (t = 1 to T), and we can write

$$Y[t]_n = g[t]_n + u[t]_n$$
 and $g[t]_n = g_{0n} + A[t]g_{1n},$ (1)

where we separate the within-time measurement equation from the over-time functional change equation. In the simple form of a measurement equation within each time, we separate the construct or trait score (g) from the time-specific unique scores (u). These unique scores are assumed to contribute variation to the observation at a given time but are independent of the trait score, and are independent of other unique scores across occasions of measurement. In this sense, they may be considered as unique factors with both specific (i.e., state) and random error components. For the purposes of all further analyses, these unique scores are distributed with a mean zero, a single unique variance (ψ^2) , and zero correlation with any of the other latent scores.

In the functional equation, g_{0n} is a latent score representing an individual's initial level (i.e., intercept), and g_{1n} represents the individual linear change over time (i.e., slopes). The set of A[t] are termed basis weights that define the timing or shape of the change over time for the group (e.g., age at testing), and we do not restrict the test occasion to be administered at a specific age. There is no additional residual in the functional equation, but this would be considered in multiple construct models (e.g., McArdle, 1988; McArdle & Woodcock, 1997). The latent components of this functional equation are assumed to have means, variances, and covariances and are related to other variables. In a multilevel form, we write these equations as

$$g_{0n} = \nu_{00} + \gamma_{01}X_n + d_{0n}$$
 and
 $g_{1n} = \nu_{01} + \gamma_{11}X_n + d_{1n}$, (2)

where the second-level scores are assumed to have means (μ_i) and covariances $(\phi_0^2, \phi_1^2, \phi_{01})$ and are accounted for by

using a regression based on external variables (X) with regression intercepts (ν), coefficients (γ), and disturbance terms (d).

Changes Implied From the Basis of the Latent Curve

The measurement equation above defines the separation of the g[t] from u[t] in the same way as a classical model of a true score separated from time-dependent error (e.g., Gullicksen, 1950). But in the functional equation, the constant part of the true score is the intercept or level score (g_0) , whereas the change in the true score from one time to another (g[t]-g[t+k]) is a function of the slope score (g_1) and the change in the time-based loadings $(\alpha[t]-\alpha[t+k])$. This interpretation is clarified when we write the first difference among successive latent scores as

$$\Delta g[t]_n = g[t]_n - g[t-1]_n$$

$$= (g_{0n} + A[t]g_{1n}) - (g_{0n} + A[t-1]g_{1n})$$

$$= (A[t]g_{1n}) - (A[t-1]g_{1n})$$

$$= (A[t] - A[t-1])g_{1n}$$
(3)

to isolate the interpretation of changes over time. This interpretation as a difference equation obviously requires the scaling of the latent scores g[t] to be identical at each occasion, and it is clear that substantial problems can arise if this assumption is not met (e.g., Cattell, 1966).

To the degree these scaling assumptions are met, we benefit from several other features of the latent curve models. The A[t] basis coefficients determine the metric or scaling and interpretation of these changes, so alterations of A[t] can lead to different curves. If we require all A[t] = 0, we effectively eliminate all slope parameters, whereas if we fix A[t] = t, we represent a straight line or linear growth curve. Alternatively, we can allow the latent basis to be estimated and take an optimal shape for the group curve (McArdle, 1986; Meredith & Tisak, 1990; Rao, 1958; Tucker, 1958). Restrictive nonlinear forms of the latent basis coefficients can be used to reflect specific growth hypotheses (McArdle & Nesselroade, 2003). Popular nonlinear models include polynomial models (quadratic, cubic) and exponential forms (e.g., Coleman, 1968; McArdle et al., 2002; McArdle & Hamagami, 1996). It is also possible to fit and compare a set of models in which the basis is written as Equation 1 with either

$$A[t] = 0,$$

 $A[t] = t,$
 $A[t] = Age[t],$
 $A[t] = \exp\{-\pi Age[t]\}, \text{ or }$
 $A[t] = \exp\{-\pi_g Age[t]\} - \exp\{-\pi_d Age[t]\}.$ (4)

The basis parameters represent a series of specific hypotheses to be tested. The first three models allow either no systematic change over time, linear change with time, or linear change with age at the time of testing. The distinctions among the first three models have been discussed in depth in previous literature (e.g., McArdle & Bell, 2000). The last two models have been used to represent either an exponential and nondecreasing change with age or a more complex, dual increasing-decreasing change over age. In this last model, the basis is formed as a difference between two exponential shapes with rates of growth (π_a) and decline (π_d) . Of course, this model is not novel in mathematics and statistics; it corresponds to a second-order constantcoefficient differential equation in continuous time, a second-order autoregression model in discrete time series, and a two-equation state-space model (McArdle & Nesselroade, 2003). This final dual exponential basis is of interest here when interpreted in terms of competing forces over age, and has been found to provide a reasonable fit with other lifespan abilities (McArdle et al., 2002). It is clear that the equivalence of measurement is required before we can consider any nonlinear form of latent changes (e.g., Carroll, Ruppert, & Stefanski, 1995).

Item Response Measurement Models

Another part of our approach follows the analysis of latent traits with IRT models (Bond & Fox, 2001; De Boeck & Wilson, 2004; Embretson & Reise, 2000; Fischer, 1995; van der Linden & Hambleton, 1997). IRT can be considered as a collection of models designed to yield estimates of one or more latent traits based on responses to a set of individual items, whether binary (dichotomous) or multicategory (polytomous). The basic model of Rasch (1960, 1966) for dichotomous items can be defined as

$$\ln\{Pr_{i,n}/(1 - Pr_{i,n})\} = g_n - \beta_i, \tag{5}$$

where $Pr_{i,n}$ is the probability of the correct response (Pr = 1) of person n to item i, the latent score g_n (usually termed θ_n) is the true score or ability (trait level) of person n, and β_i is the difficulty of item i. In this simple form, the log-odds of the correct response increases to the degree that the person's ability is higher than the item difficulty. A corresponding model of the probability of a correct response can be written as a standard exponential ratio (i.e., $\exp(Pr)/\{1 + \exp(Pr)\}$).

The item data we examine are collected at different occasions of measurement (t = 1 to T), so we extend this model by writing

$$ln\{Pr[t]_{i,n}/(1 - Pr[t]_{i,n})\} = g[t]_n - \beta_i, \tag{6}$$

where, at time t, the $Pr[t]_{i,n}$ is the probability of the correct response (Pr = 1) of person n to item i, the latent score $g[t]_n$

is the true score or ability (trait level) of person n at time t, and the difficulty of the item β_i does not change with time. This assumption of measurement invariance over time at the item level (i.e., β_i is constant over t=1 to T) is a testable hypothesis with complete longitudinal data at several occasions (McArdle & Nesselroade, 1994; McArdle, 2007). However, when the scales change from one time to the next, with only minimal overlap, the assumption of item invariance over time is not so easily tested. Nevertheless, given these assumptions and adequate fit to the overall measurement model, the Rasch estimated ability score $g[t]_n$ can be considered a valid interval scaled measure and substituted into Equation 1.

Many more complex variations of this basic model can be introduced. In the data presented here, some items used in these analyses can have graded outcome scores (i.e., 0, 1, or 2), so we can use a partial credit model (Glas & Verhelst, 1995; Masters, 1982; Wilson & Draney, 1997). The partial credit model can be written for any item as

$$\ln\{PrX[t]_{i,n}/(1 - PrX[t]_{i,n})\} = g[t]_n - \beta_i, \tag{7}$$

where, at time t, the $PrX[t]_{i,n} = Pr(X[t]_{i,n} = x|X[t]_{i,n} = x$ or $X[t]_{i,n} = x - 1$) is the probability the response of person n to item i is in category x, given that the response is either in category x or x - 1. This reduces to the longitudinal Rasch model of Equation 6 when dichotomous items are used.

In another alternative, we can write the classic twoparameter logistic model (Birnbaum, 1968) as

$$\ln\{Pr[t]_{i,n}/(1 - Pr[t]_{i,n})\} = \lambda_i(g[t]_n - \beta_I), \tag{8}$$

where, at each time t, the discrimination parameter λ_i modifies the difference and hence the probability of the correct response (Pr[t] = 1) of person n to item i at time t. Although there are now two independent characteristics of each item, the intercept β_I and slope λ_I , these are both assumed to be invariant over time. The generalized partial credit model (Muraki, 1992) combines the partial credit model and two-parameter logistic model by incorporating the slope parameter into the adjacent logits equation (see Equation 7), and this can be added to the longitudinal model (Equation 8). These models involving the slope parameter may be needed to achieve fit, and this raises a number of critical issues in measurement theory (see Andrich, 2002; Bock, 1997).

Estimation and Programming

Modeling Incomplete Longitudinal Curves

The longitudinal life-span data considered here are incomplete for many reasons: (a) The number of occasions was not the same across the three studies, (b) the individuals were not measured at the same ages, (c) there was some attrition because of death and relocation, (d)

the same scales were not used at every occasion, and (e) the same items were not presented to every individual even when the same test was administered. To deal with these issues, we estimated each of the models that follow using a variation of MLE based on high-dimensional integration under various assumptions about the reasons for the incomplete data (see Little & Rubin, 1987; McArdle, 1994).

In most cases here, the reason participants do not have a score on a specific item at a specific occasion was largely dependent on the investigators' plan. For this reason, this form of incomplete data is entirely unrelated to the score they would have received, and no statistical correction is needed (i.e., missing completely at random). In many cases, the reason the data are missing is directly related to the score, such as not being given certain test items that seemed too difficult, and corrections based on the data are needed (i.e., missing at random, or MAR). In some cases, the reason for incomplete data is less clear, such as when individuals dropped out of the study at some point in time (e.g., after age 29, the attrition was >50% in the Bradway-McArdle study), and other variables may be needed to account for selection effects. To deal with these problems, we fit all models with MLE-MAR estimation.

Initial Estimation of IRT Scores

In the LIRT approach described above, the IRT model (Equation 6) is used as the first order of measurement, and the longitudinal growth curve (Equations 1 and 4) can be used at the second order (i.e., the curve-of-factorscores model). One practical way to start this analysis is to estimate a Rasch scoring table that is invariant over time. As a simple first approximation, we can fit an IRT model to the available data at each occasion and then use the separate occasion item estimates to build up a scoring table for the conversion of the items to common composites. Joint maximum likelihood estimation of Rasch scores is implemented in Winsteps (Wright & Stone, 1979), and this creates estimates by iteratively alternating between item and person parameters, treating the other as fixed. This first stage may also be improved by using all the available data at all occasions for all persons (see Appendix, Script 1a).

It has recently been recognized that all IRT models can also be reformulated as nonlinear mixed-effects models (e.g., Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003). One common approach to fitting these models uses marginal maximum likelihood with classic multidimensional integration and typically relies on a design matrix to select the appropriate item for comparison with the latent trait in building up the likelihood function (e.g., De Boeck & Wilson, 2004; Sheu, Chen, Su, & Wang, 2005; Smits, De Boeck, &

Verhelst, 2003). However, the problems posed here are of substantially larger magnitude (e.g., 224 items), and alterations of these programs without design matrices are needed (see the NLMIXED program with ARRAYs in the Appendix, Script 1b).

This simple IRT approach with longitudinal data ignores several dependencies within persons over time. To account for these dependencies, we might create a separate dimension for each occasion, or facet, still requiring item invariance. The resulting person trait estimates can then be considered as observed data at each occasion for other longitudinal models (e.g., LGM-NLMIXED; see Appendix, Script 2). Of course, there are now a variety of elegant statistical and computational procedures for marginal maximum likelihood estimation of the parameters of latentgrowth mixed-effects multilevel models with MAR incomplete data (e.g., SAS MIXED and NLMIXED, Littell, Milliken, Stroup, & Wolfinger, 1996; Singer, 1998; Verbeke & Molenberghs, 2000; S-PLUS, Pinherio & Bates, 2000; MIXREG, Hedeker & Gibbons, 1997). These programs are practical in their ease of representing different models for the basis function (i.e., exact ages, nonlinear curves, free basis parameters; McArdle & Bell, 2000; Verbeke & Molenberghs, 2000).

Simultaneous Estimation With MLE and Markov Chain Monte Carlo (MCMC)

The LIRT model conceptualized here requires the assumptions that (a) the items are measuring a unidimensional construct, (b) the item difficulty is invariant with respect to time or age, and (c) the items are equally discriminating and this discrimination does not change across time or age. Violations of any assumptions could lead to fundamentally incorrect conclusions regarding the within-person changes and between-person differences in change. Thus, estimation in which we ignore the person dimension is likely to lead to poor estimates of growth and change.

To solve some of these problems, we considered the entire model, from Equation 1 to Equation 6, to be a unitary model for the longitudinal items and estimated the parameters simultaneously (i.e., a curve-of-factor-scores model). The primary advantage of a simultaneous analysis of the IRT-LCM is that we do not need to make a separate estimation of the LIRT score for each occasion, and this in turn should minimize the within-person dependence created by ignoring subjects in a two-stage approach. There is a possible gain in statistical efficiency from using the simultaneous approach, and the result should be more precise estimates and hypothesis tests. In practice, the first-stage estimates may best be used as starting values for the more comprehensive simultaneous solutions.

Other researchers have shown how to deal with the complexities of fitting a simultaneous model of this type using high-dimensional integration (e.g., Feddag & Mesbah, 2005; Ferrando, 2002), but the number of items and incompleteness we consider here far exceeds anything discussed in this previous literature. As a result, differences owing to the different computer programs, time intervals chosen, and MAR assumptions may be important considerations in further data analysis. Once again, we can employ the MLE-based programs, now with IRT and LCM combined, but once again the problems posed here are of substantially larger magnitude (e.g., up to 224 items at up to 16 occasions), and alterations of the typical programs are needed (NLMIXED plus ARRAYs; see Appendix, Script 3). Good initial starting values are crucial to the stability of this approach (McArdle & Wang, 2008).

Recent work has also shown how the latent curve models described here can be fitted with Bayesian inference and an MCMC algorithmic approach (e.g., WinBUGS, Congdon, 2003; Gelman & Meng, 2004; McArdle & Wang, 2008; Wang & McArdle, 2008; Zhang, McArdle, Wang, & Hamagami, 2008). One purpose of the MCMC approach is to approximate a marginal posterior estimate without using an intractable multidimensional numerical quadrature method (see Spiegelhalter, Thomas, Best, & Gilks, 1994). A Gibbs sampling algorithm is repeatedly used to generate a sequence of samples from the posterior joint probability distribution of model parameters. Good starting values are useful here as well because, at each MCMC iteration, the sampling of a parameter is generated given the previous posterior instance of all other model parameters and data. After a sequence of samples is generated for each parameter, an average of samples is obtained as the posterior estimate of the given model parameter. Numerous theoretical diagnostic techniques have been proposed in the past (Brooks & Gelman, 1997; Gelman & Rubin, 1992) for the convergence decision. In practice, one needs to visually inspect the trace of sequentially sampled posterior parameter estimates to examine whether the Markov chain reaches stability. A script for WinBUGS with a Bayesian approach to MCMC estimation is also presented in the Appendix (Script 4).

Results

Initial First-Stage Item Response Estimation

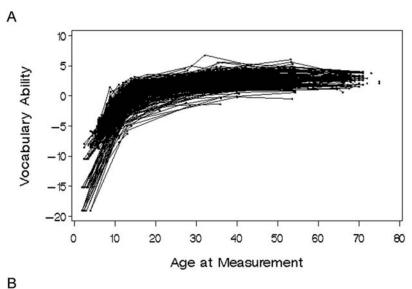
To initially represent the outcomes of these LIRT analyses, we estimated ability scores (g[t], using Winsteps and NLMIXED; see Appendix, Scripts 1a and 1b) as a simple function of the items administered at each occasion and the total number of correct and incorrect responses, ignoring the age at measurement. Although these initial Rasch estimates are likely to be biased away from the mean (i.e., joint maximum likelihood), these initial estimates raise a number of important empirical issues. For example, these theoretical scores can be scaled in many ways, and here we

chose a logit metric to reflect linear probability changes and defined it so the average of the item difficulties is zero (i.e., $\Sigma \beta_i = 0$).

After this calculation, we plotted the Vocabulary estimates for each person against the person's age at testing in the life-span trajectory plot, as displayed as Figure 2A. A corresponding plot of the Memory ability estimates against the participants' age at testing is shown in Figure 2B. In an important sense, these initial estimates are the first estimates of the key longitudinal data originally desired. That is, the life-span trajectories are now represented for each person, and the variable plotted, under the Rasch model, has the same interpretation at ages ranging from 2 to 75. These scores rise rapidly through childhood and adolescence and flatten out through adulthood, and there are substantial individual differences in both Vocabulary and Memory.

Simultaneous Longitudinal Item Response Models

We estimated a simultaneous LIRT solution for the models above. The results are listed in Table 4 for Vocabulary and Table 5 for Memory. It is known that this simultaneous model requires a numerical constraint typically placed on the mean of the intercept component in second-order growth curves (see Hancock et al., 2001; Leite, 2007; McArdle, 1988, 1989), and this constraint is imposed here. The marginal maximum likelihood estimation with NLMIXED (see Appendix, Script 3) took far longer than expected to carry out the numerical calculations, so these results are incomplete and are not presented. (As pointed out by a reviewer, this failure to converge requires further investigation.) In contrast, the MCMC estimation with WinBUGS (see Appendix, Script 4) was far more reasonable in terms of the



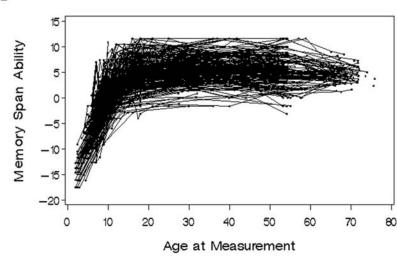


Figure 2. Longitudinal plots of Rasch estimated person abilities against age for (A) Vocabulary and (B) Memory abilities.

Table 4

Vocabulary Parameter Estimates and Fit Statistics for Age-Based Latent Growth Models From Simultaneous Bayesian

Estimation (Using WinBUGS)

Parameter	V0b: Level only baseline	V1b: Exponential growth	V2b: Dual exponential growth	V3b: Dual exponential with gender	V4b: Dual exponential with sample
Fixed effects					
Level mean v_{00}	= 0	= 0	= 0	= 0	= 0
Slope mean v_{10}	_	4.61 ^a (4.47, 4.78)	4.98 ^a (4.72, 5.25)	4.96 ^a (4.73, 5.19)	4.98 ^a (4.70, 5.30)
Growth rate π_{ρ}	_	.131 ^a (.127, .136)	.123 ^a (.117, .130)	.124 ^a (.119, .129)	.125 ^a (.120, .131)
Decline rate π_d	_	= 0	.001 ^a (.001, .002)	.001a (.0006, .002)	.001 ^a (.0003, .002)
Level on gender γ_{01}				.01(20,.21)	_
Slope on gender γ_{11}				11(28,.06)	_
Level on BML γ_{02}				_	$65^{a}(90,40)$
Slope on BML γ_{12}				_	08(34, .19)
Level on GCS γ_{03}				_	$47^{a}(66,26)$
Slope on GCS γ_{13}				_	10(33,.13)
Random effects					
Error deviation σ_e	2.34 ^a (2.22, 2.47)	.40 ^a (.36, .43)	$.40^{a}$ (.37, .43)	$.40^{a}$ (.37, .43)	$.40^{a}$ (.37, .43)
Level deviation σ_0	1.00 ^a (.85, 1.15)	.98 ^a (.90, 1.06)	.98 ^a (.90, 1.06)	.98 ^a (.90, 1.06)	.97 ^a (.89, 1.05)
Slope deviation σ_1	_	$.65^{a}$ (.59, .72)	$.70^{a}$ (.63, .78)	$.70^{a}$ (.63, .78)	$.69^{a}$ (.62, .78)
Correlation ρ ₀₁	_	$15^{a}(24,06)$	$15^{a}(25,06)$	$15^{a}(26,06)$	$16^{a}(26,07)$
Goodness of fit					
DIC	71660	70042	69998	70011	70012
Parameters	3	7	8	10	11

Note. Data D = 2,507 based on maximum N = 419, T = 13, I = 99. Cell entries include maximum likelihood estimation and 95% credible intervals in parentheses. Dashes indicate that a model does not include a parameter. Age was recentered at 10 years. Final information matrix was not positive definite in linear model. Quadratic and linear models did not converge. BML = Bradway–McArdle Longitudinal Study; GCS = Guidance–Control Study; DIC = deviance information criterion.

computer time required for each iteration. It was reasonable to run over 20,000 iterations for each model (i.e., 5,000 burn-in iterations, 10,000 runs, and an additional 5,000 to describe the converged estimates) and to start the estimation from three disparate starting points (even using different computers). The summary of numerical results with this Bayesian estimation approach is presented in the tables, where the summary information about each parameter is listed, including the mean of estimates, 95% credible interval, and deviance information criterion (DIC; see Congdon, 2003; Spiegelhalter et al., 1994).

Fitting Alternative Change Models

Several growth models (i.e., level only, exponential, dual exponential) were fit as the second-level LCM model in the simultaneous estimation of IRT-growth models. The parameters (p) estimated and fit statistics are presented in the first three columns of Table 4 for a series of simple growth models fit to the item-level vocabulary. The initial level-only model (V0) gives a baseline for evaluating fit (DIC = 71660) and includes two growth parameters ($\sigma_0 = 1.00$ and $\sigma_e = 2.34$). The mean of the latent level was fixed for identification purposes ($\mu_0 = 0$). The fitting of both linear models and quadratic models resulted in serious convergence problems. In contrast, the single exponential model

(V1) fit slightly better than the level-only model (DIC = 70042, $\Delta p = 4$). Here the slope mean ($\mu_1 = 4.61$) and the growth rate ($\pi_g = .13$) indicate a rapid early growth from childhood to early adulthood. The level and slope variances and their covariance ($\sigma_0 = .98$, $\sigma_1 = .65$, $\sigma_{01} = -.15$) show significant variation in vocabulary at age 10 and significant variation in individual changes. Furthermore, children who had a greater level of vocabulary ability at age 10 tended to have a slower rate of change. The dual exponential model (V2) fit slightly better than the single exponential model (DIC = 69998, $\Delta p = 1$), and the decline rate ($\pi_d = .001$) indicates a small but significant decline in ability through adulthood. As with the exponential model, the average rate of change was positive ($\mu_1 = 4.98$), and there was significant variation in the level ($\sigma_1 = .98$) and slope ($\sigma_1 = .70$) and their covariance ($\sigma_{01} = -.15$) was also significant. The expected mean (and deviations) of the age-based latent curve of Vocabulary is displayed in Figure 3A.

The same sequence of IRT-growth models was applied to the Memory Span items, and the dual exponential model also best represents the longitudinal data (see Table 5). The level-only model (DIC = 24090) provided a baseline for comparison and two growth parameters, the variation of the level ($\sigma_0 = .97$) and residual ($\sigma_e = .84$). The exponential model was an improvement over the level-only model (DIC = 20989,

^a Indicates a significant parameter at .05 level.

Table 5
Memory Span Parameter Estimates and Fit Statistics for Age-Based Latent Growth Models From Simultaneous Bayesian
Estimation (Using WinBUGS)

Parameter	M0b: Level only baseline	M1b: Exponential growth	M2b: Dual exponential growth	M3b: Dual exponential with gender	M4b: Dual exponential with sample
Fixed effects					
Level mean v_{00}	= 0	= 0	= 0	= 0	= 0
Slope mean v_{10}	_	3.42 ^a (3.22, 3.61)	4.18 ^a (3.91, 4.20)	4.19 ^a (3.77, 4.60)	4.20 ^a (3.86, 4.57)
Growth rate π_{g}	_	.213 ^a (.205, .225)	.186 ^a (.172, .197)	.186 ^a (.175, .201)	$.180^{a}$ (.170, .185)
Decline rate π_d		= 0	.007 ^a (.006, .008)	.007 ^a (.005, .009)	.007 ^a (.006, .009)
Level on gender γ_{01}				25(48,.02)	
Slope on gender γ_{11}				.13 (05, .33)	
Level on BML γ_{02}					-1.04^{a} (-1.43,66)
Slope on BML γ_{12}					13(46, .21)
Level on GCS γ_{03}					$52^{a}(75,28)$
Slope on GCS γ_{13}					.39 ^a (.13, .62)
Random effects					
Error deviation σ_e	.84 ^a (.77, .92)	$.38^{a}$ (.29, .46)	$.34^{a}$ (.25, .43)	.34 ^a (.24, .43)	.35° (.26, .45)
Level deviation σ_0	.97 ^a (.88, 1.06)	1.45 ^a (1.34, 1.57)	1.46 ^a (1.34, 1.58)	1.45 ^a (1.34, 1.57)	1.40 ^a (1.29, 1.51)
Slope deviation σ_1	_	.48 ^a (.41, .56)	.59 ^a (.50, .69)	.59 ^a (.49, .70)	.57 ^a (.47, .68)
Correlation ρ_{01}		$06^{a}(18,06)$	01(16, .14)	.01 (16, .17)	.01(14, .15)
Goodness of fit					
DIC	24090	20989	20905	20940	20904
Parameters	2	6	7	9	11

Note. Data D = 3,107 based on maximum N = 416, T = 16, I = 76. Cell entries include maximum likelihood estimation and 95% credible intervals in parentheses. Dashes indicate that a model does not include a parameter. Age was recentered at 10 years. BML = Bradway–McArdle Longitudinal Study; GCS = Guidance–Control Study; DIC = deviance information criterion.

^a Indicates a significant parameter at .05 level.

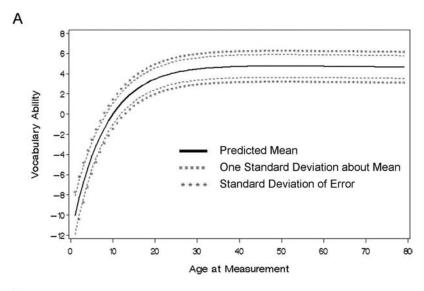
 $\Delta p=4$). The mean slope was positive ($\mu_1=3.42$), and the growth rate was positive ($\pi_g=.21$) and larger (i.e., faster) than that of Vocabulary. As with Vocabulary, there was significant variation in the level ($\sigma_0=1.45$) and slope ($\sigma_1=.48$), but their covariance ($\sigma_{01}=-.06$) was nonsignificant. Next, a dual exponential growth model was fit as the second-level growth model, and the fit was subsequently improved (DIC = 20905, $\Delta p=1$). The mean slope ($\mu_1=4.18$), growth rate ($\pi_g=.19$), and decline rate ($\pi_d=.007$) were all positive and significant, and this describes a latent curve with a fast rate of growth during childhood and adolescence and a marked decline through adulthood. There was significant variation in the level ($\sigma_0=1.46$) and slope ($\sigma_1=.598$), but their covariance ($\sigma_{01}=-.01$) was nonsignificant. The mean (and deviation) of the latent curve of Memory is displayed in Figure 3B.

These latent curve estimates of the dual exponential growth model obtained for Memory were quite different from those for Vocabulary. The dual exponential model has a marked acceleration through childhood and adolescence as in the Vocabulary model, but as young adulthood is reached, the function reaches its maximum and begins a decline. Although it is compelling to state that the decline in Memory is more pronounced than that of Vocabulary, we cannot compare them directly because they are not in the same scale of measurement.

Including Predictors of Change

In the next set of models (V3 and M3), we introduced sex differences in the LIRT levels and slopes. The IRT-growth models with a dual exponential basis were refit with Gender (coded -1/2 for female and 1/2 for male) as a predictor of the level and slope. For the Vocabulary and Memory span data, Gender was a nonsignificant predictor of the level and slope. To investigate the potential differences between the three groups, we compared the trajectories by including two dummy codes to compare the levels and slopes of the BML and GCS samples with those of the BGS sample (see Figure 4). The mean level (at age 10) of the BML and GCS samples was slightly lower than the mean level of the BGS sample on Vocabulary. For Memory Span, the BML and GCS samples had lower mean performances at age 10 compared with the BGS sample, but the GCS sample had a slightly greater rate of change than the BGS sample. It would be useful at this point to evaluate more fully the invariance of the measurement model over the multiple groups, but we recognize that these data are limited in this respect.

As these new growth charts show, the intellectual abilities underlying vocabulary and memory rise rapidly throughout childhood, peak in early adulthood, and decline at a very slow rate, if at all. The use of combined data increased the precision



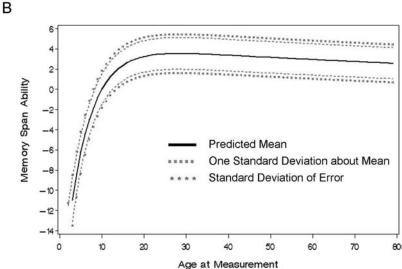


Figure 3. Latent growth curve model expected group trajectories based on the dual exponential model with one standard deviation around the mean for (A) Vocabulary and (B) Memory abilities.

of most tests (McArdle et al., 2000), and some significant but small differences were found between the separate studies (BGS, GCS, and BML). However, individual differences in the intercepts (at age 10) and subsequent changes were not related to group differences in gender. Of course, this sets the stage for the inclusion of other multilevel predictors, some of which differ over these samples (see Grimm & McArdle, 2007).

Discussion

Summary of Results

The basic requirements of meaningful and age-equivalent measurement models are a key problem in the behavioral sciences (see Burr & Nesselroade, 1990; Fi-

scher, 1995). The possibilities for a standard longitudinal measurement analysis were initially limited here by the complex longitudinal data collection (M. C. Jones et al., 1971). The classical solutions based on simple or complex rescoring were not used and have generally not been considered satisfactory to a wide research audience because of their ad hoc nature. Although not emphasized here, a full-information SEM approach based on using multiple indicators at the scale level failed to be estimable because of large amounts of incomplete data (see McArdle & Nesselroade, 2003). In contrast, the simultaneous estimation approach for a combined LIRT-growth model was able to account for the dependencies (longitudinal aspect) of the data. This method allowed us to

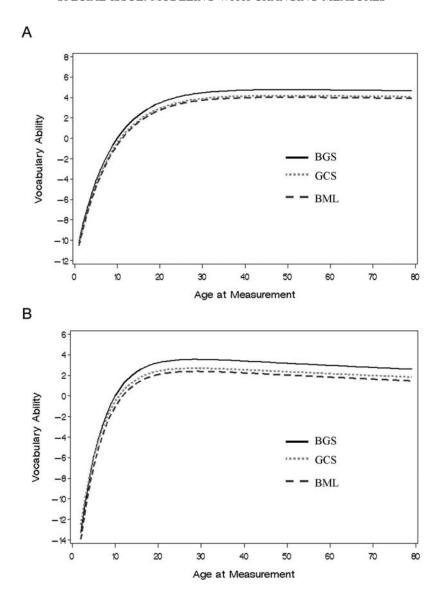


Figure 4. Expected group growth curves of (A) Vocabulary and (B) Memory abilities for the three independent study groups: the Berkeley Growth Study (BGS), the Guidance–Control Study (GCS), and the Bradway–McArdle Longitudinal (BML) Study.

consider changes over age in the same constructs for a longer span of time than any previous longitudinal study of cognitive aging.

To many substantive researchers, this kind of mixed-battery data collection is far better than a strict adherence to repeated measures because of age appropriateness, improved batteries, and so on. The subsequent IRT-growth analyses provided (a) a relative scaling of each item for Vocabulary and Memory as if everyone had taken all items on all scales, (b) ability estimates for participants at each occasion of measurement, and (c) parameters of average growth and of between-person differences in growth. Of course, it is also important to verify these results with other

data sets in which the scales are constant. For example, the loss of memory from adolescence to adulthood is coincident with the changing scales, so the accuracy of this decline requires further verification (see McArdle et al., 2002). The IRT-growth solution used here may reflect the best case scenario because it solves one of the key problems of changing measures of repeated constructs. An ancillary result of this analysis is the optimal selection of shortened tests (e.g., 24 items) with fairly well spaced item difficulty (available upon request). It is also important to consider checking the adequacy of the pooling of data by further sensitivity analyses (i.e., considering the results when one or more subsamples of people are treated as missing).

Simultaneous Versus Two-Stage Estimation

The simultaneous estimation approach used here is thought by many researchers to be the optimal way to model longitudinal item-level data in which systematic growth is expected. However, some viable alternative estimation approaches may be more practical and provide additional information not easily obtainable in the simultaneous estimation approach. One alternative we mentioned is a two-stage approach in which ability estimates for each person at each occasion are estimated with an item response model. In the second step, the ability estimates obtained from a first step can be used as observed data and modeled with growth curve analysis. This alternative twostage estimation was not pursued here because it was considered less optimal owing to the longitudinal dependencies within person. Some problems of factor score estimation (e.g., Croon, 2002) can be overcome simply by using a simultaneous estimation of the joint set of parameters (as in Tables 4 and 5). However, we do not want to overlook several practical advantages. First, the fit of items or people to the specific item response model can be evaluated in a standard IRT framework. Second, the ability estimates from the first step are easily plotted (e.g., as in Figure 2), and this can allow the researcher to check for outliers or unusual observations and consider the shape of development. Third, this simpler approach cuts down on the computational complexity and the amount of estimation time required. These are all practical issues worthy of further investigation.

Longitudinal Item Response Modeling Limitations

Of paramount importance here is our lack of ability to examine the assumption of metric factorial invariance over occasions (i.e., $\Lambda[t] = \Lambda[t + 1]$?; McArdle & Nesselroade, 1994), and we were limited in what we could accomplish. When this kind of restrictive model of changes in the factor scores among multiple variables provides a reasonable fit to the data, we have evidence for dynamic construct validity (McArdle, 2007; McArdle, Prescott, Hamagami, & Horn, 1998). Unfortunately, when the data are less than complete, or reflect nonoverlapping scales, we lose some or most of the statistical power of such tests. Because of the changing measures, we basically had to assume but not test invariance of the construct over time to proceed with our calculations. To make this a more reasonable analysis, we selected narrowly defined abilities of Vocabulary and Memory Span and considered these as part of a larger universe of items (Gullicksen, 1950). A more complete consideration of metric versus configural invariance with different loadings for different items would be possible in more carefully designed item selections (see Smith, 1996; Smith & Suh, 2003; Wilson, 2005). In general, we do expect that the LIRT method can be operationalized in studies with clearly defined constructs over repeated testings.

As we have shown using MCMC estimation, it is now possible to fit models with simultaneous estimation of item characteristics and higher order factors, including changes over time (e.g., Fox & Glas, 2001; Hamagami, 1998; Janssen, Tuerlinckx, Meulders, & De Boeck, 2004; Ram et al., 2005; Rijmen et al., 2003, 2005). It also follows that a simultaneous IRT-LCM model might increase accuracy from different stages of analysis. Unfortunately, this model could not be fitted with the standard MLE based on high-dimensional integration, and this may be due to the size of our problem and/or the lack of overlap in the items (e.g., Table 1). The large amount of incomplete information made it impossible to carry out SEM models at the scale level. The standard IRT-LCM calculation also made it difficult to fit a simultaneous growth-item model with standard MLE. In contrast, the MCMC approach to these problems used here highlights a practical solution that others may find useful when faced with these kinds of longitudinal models.

Issues for Future Longitudinal Studies

A set of theoretical and practical issues have emerged from these longitudinal analyses.

- 1. Contemporary data analysts need not simply rephrase substantive questions about development to deal with incommensurate measures (growth ~ regression). The models that are used for data analysis should drive the data collection, but this is not always the case. Although classic methods, such as factor-growth models as applied in standard SEM programs, were limited here by the lack of overlap of the scales, the item-intensive IRT-growth analyses were successful. The subsequent mixed-effects analyses demonstrate the possibility of measuring and evaluating growth and change in the same constructs over many ages with nonrepeated or changing measures.
- 2. In theory, it is not necessary or desirable for future longitudinal studies to require the same measures from one occasion to the next. As shown here, there are several contemporary techniques based on latent variables for dealing with repeated constructs without exactly repeated measures. This means a scale should not simply be repeated because it was given before. Instead, scale alteration over time should be designed to match reasonable substantive goals (i.e., age appropriateness) and not repeat items or scales that are irrelevant.
- Planning for overlapping scales or items within and between occasions is essential. It is clear that much can be accomplished by carrying over some scales

- or items from one occasion to the next to facilitate future analysis. If the practical problems of calibration at the scale and item level are recognized at the design stage, a variety of future analyses will be feasible. It is essential to study the linkage features of measurement in all longitudinal designs.
- Although not emphasized here, the IRT calibration does not require longitudinal data. Instead, measurement calibrations at the scale and item level can and should be completed in auxiliary studies outside the constraints of longitudinal studies. This encourages increased accuracy in scaling results based on much larger and wider range cross-sectional studies. But can a researcher just apply the scoring system from some larger IRT calibration studies and use it with new occasions and people? One unique aspect of the longitudinal data demonstrated here is the increased precision of the random effects from the simultaneous LIRT model fitting. However, we may not always be able to benefit from item model fitting and scale model fitting within the same study. Future longitudinal studies can benefit from considering many different approaches to cross-battery calibrations.
- 5. Pooling data with some overlapping measures is a powerful idea. The attempt to use all available information from any person measured on any of the variables of interest in a study can lead to increased multivariate power and precision (e.g., McArdle, 1994). Of course, biases can emerge when subgroups of persons within distinct groups are considered from the same population, and without overlapping information we may not realize these problems exist at all. We need to examine the assumptions of pooling group data whenever such data are available.

The results for LIRT multiple-group pooled data approximations presented here may be emphasized in different ways by different researchers. These differences represent potentially important theoretical and practical choices for future longitudinal researchers.

References

- Andrich, D. (2002). Understanding resistance to the data-model relationship in Rasch's paradigm: A reflection for the next generation. *Journal of Applied Measurement*, 3, 325–359.
- Bayley, N. (1956). Individual patterns of development. *Child Development*, 27, 45–74.
- Becker, K. A. (2003). *History of the Stanford–Binet Intelligence Scales: Content and psychometrics*. Itasca, IL: Riverside.

- Bell, R. Q. (1953). Convergence: An accelerated longitudinal approach. *Child Development*, 24, 145–222.
- Bell, R. Q. (1954). An experimental test of the accelerated longitudinal approach. *Child Development*, 25, 281–286.
- Birnbaum, A. (1968). Estimation of an ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 453–479). Reading, MA: Addison-Wesley.
- Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice, 16*, 21–32.
- Bond, T. G., & Fox, C. M. (2001). Applying the Rasch model: Fundamental measurement in the human sciences. Mahwah, NJ: Erlbaum.
- Bradway, K. P., & Thompson, C. W. (1962). Intelligence at adulthood: A twenty-five year follow-up. *Journal of Educa*tional Psychology, 53(1), 1–14.
- Brooks, S. P., & Gelman, A. (1997). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7, 434–455.
- Browne, M. W., & du Toit, S. H. T. (1991). Models for learning data. In L. M. Collins & J. L. Horn (Eds.), Best methods for the analysis of change: Recent advances, unanswered questions, future directions (pp. 47–68). Washington, DC: American Psychological Association.
- Bryk, A. S., & Raudenbush, S. W. (1992). Hierarchical linear models: Applications and data analysis methods. Thousand Oaks, CA: Sage.
- Burr, J., & Nesselroade, J. R. (1990). Change measurement. In A. von Eye (Ed.), *Statistical methods in longitudinal research: Vol. 1. Principles and structuring change* (pp. 3–34). New York: Academic Press.
- Carroll, R. J., Ruppert, D., & Stefanski, L. A. (1995). *Measure-ment error in nonlinear models*. London: Chapman and Hall.
- Cattell, R. B. (1941). Some theoretical issues in adult intelligence testing. *Psychological Bulletin*, *38*, 592.
- Cattell, R. B. (Ed.). (1966). *Handbook of multivariate experimental psychology*. Chicago: Rand-McNally.
- Cattell, R. B. (1972). *Real base, true zero factor analysis* (Multivariate Behavioral Research Monograph No. 72–1). Fort Worth: Texas Christian University Press.
- Cattell, R. B. (1998). Where is intelligence? Some answers from the triadic theory. In J. J. McArdle & R. W. Woodcock (Eds.), *Human cognitive abilities in theory and practice* (pp. 29–38). Mahwah, NJ: Erlbaum.
- Coleman, J. (1968). The mathematical study of change. In H. M. Blalock & A. B. Blalock (Eds.), *Methodology in social research* (pp. 428–475). New York: McGraw-Hill.
- Collins, L. M., & Sayer, A. G. (Eds.). (2001). New methods for the analysis of change. Washington, DC: American Psychological Association.
- Congdon, P. (2003). Applied Bayesian modelling. London: Wiley.Croon, M. (2002). Using predicted latent scores in general latent structure models. In G. A. Marcoulides & I. Moustaki (Eds.),Latent variable and latent structure models (pp. 195–223).Mahwah, NJ: Erlbaum.

- Cudeck, R., & du Toit, S. H. C. (2003). Nonlinear multilevel models for repeated measures data. In S. P. Reise & N. Duane (Eds.), *Multilevel modeling: Methodological advances, issues, and applications* (pp. 1–24). Mahwah, NJ: Erlbaum.
- De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.
- Duncan, D. R., & Barrett, A. M. (1961). A longitudinal comparison of intelligence involving the Wechsler–Bellevue I and WAIS. *Journal of Clinical Psychology*, *17*, 318–319.
- Duncan, S. C., & Duncan, T. E. (1995). Modeling the processes of development via latent variable growth curve methodology. *Structural Equation Modeling*, 2, 187–213.
- Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*, 495–515.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.
- Fan, X. (2003). Power of latent growth modeling for detecting group differences in linear growth trajectory parameters. *Structural Equation Modeling*, 10, 380–400.
- Feddag, M. L., & Mesbah, M. (2005). Generalized estimating equations for longitudinal mixed Rasch model. *Journal of Statistical Planning and Inference*, 129, 159–179.
- Ferrando, P. J. (2002). An IRT-based two-wave model for studying short-term stability in personality measurement. *Applied Psychological Measurement*, 26, 286–301.
- Fischer, G. H. (1987). Applying the principles of specific objectivity and generalizability to the measurement of change. *Psychometrika*, 52, 565–587.
- Fischer, G. H. (1989). An IRT-based model for dichotomous longitudinal data. *Psychometrika*, *54*, 599–624.
- Fischer, G. H. (1995). The linear logistic test model. In G. H. Fischer & I. W. Molenaar (Eds.), Rasch models: Foundations, recent developments, and applications (pp. 157–180). New York: Springer.
- Fischer, G. H. (2003). The precision of gain scores under an item response theory perspective: A comparison of asymptotic and exact conditional inference about change. *Applied Psychological Measurement*, 27, 3–26.
- Fischer, G. H., & Parzer, P. (1991). An extension of the rating scale model with an application to the measurement of change. *Psychometrika*, *56*, 637–651.
- Fischer, G. H., & Seliger, E. (1997). Multidimensional linear logistic models for change. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 323–346). New York: Springer.
- Fox, J.-P., & Glas, A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, 66, 271–288.
- Gelman, A., & Meng, X.-L. (Eds.). (2004). Applied Bayesian modeling and causal inference from incomplete-data perspectives. New York: Wiley.
- Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. Statistical Science, 7, 457–511.
- Glas, C. A. W., & Verhelst, N. D. (1995). Tests of fit for polyto-

- mous Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 325–352). New York: Springer.
- Goldstein, H. (1995). Multilevel statistical models (2nd ed.). New York: Oxford University Press.
- Goldstein, H., & Browne, W. (2002). Multilevel factor analysis modelling using Markov Chain Monte Carlo estimation. In G. A. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure models* (pp. 225–243). Mahwah, NJ: Erlbaum.
- Grimm, K. J., & McArdle, J. J. (2007). A dynamic structural analysis of the impacts of context on shifts in lifespan cognitive development. In T. D. Little, J. A. Bovaird, & N. A. Card (Eds.), Modeling contextual effects in longitudinal studies (pp. 363– 386). Mahwah, NJ: Erlbaum.
- Gullicksen, H. (1950). Theory of mental tests. New York: Wiley. Hamagami, F. (1998). A developmental-based item factor analysis. In J. J. McArdle & R. W. Woodcock (Eds.), Human cognitive abilities in theory and practice (pp. 231–246). Mahwah, NJ: Erlbaum.
- Hancock, G. R., Kuo, W., & Lawrence, F. R. (2001). An illustration of second-order latent growth models. *Structural Equation Modeling*, 8, 470–489.
- Harris, C. W. (Ed.). (1963). Problems in measuring change. Madison: University of Wisconsin Press.
- Hedeker, D., & Gibbons, R. D. (1997). Application of randomeffects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, 2, 64–78.
- Horn, J. L. (1988). Thinking about human abilities. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 645–685). New York: Plenum Press.
- Horn, J. L. (1998). A basis for research on age differences in cognitive capabilities. In J. J. McArdle & R. W. Woodcock (Eds.), *Human cognitive abilities in theory and practice* (pp. 57–91). Mahwah, NJ: Erlbaum.
- Horn, J. L., & McArdle, J. J. (1980). Perspectives on mathematical/statistical model building (MASMOB) in research on aging.
 In L. W. Poon (Ed.), *Aging in the 1980s: Psychological issues* (pp. 503–541). Washington, DC: American Psychological Association.
- Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimen*tal Aging Research, 18, 117–144.
- Janssen, R., Tuerlinckx, F., Meulders, M., & De Boeck, P. (2000).
 A hierarchical IRT model for criterion-referenced measurement.
 Journal of Educational and Behavioral Statistics, 25, 285–308.
- Jones, C. J., & Meredith, W. (2000). Developmental paths of psychological health from early adolescence to later adulthood. *Psychology and Aging*, *15*, 351–360.
- Jones, M. C., Bayley, N., Mcfarlane, J. W., & Honzik, M. P. (Eds.). (1971). The course of human development: Selected Papers from the Longitudinal Studies, Institute of Human Development, the University of California, Berkeley. Waltham, MA: Xerox.

- Jöreskog, K. G., & Sörbom, D. (1979). Advances in factor analysis and structural equation models (J. Magidson, Ed.). Cambridge, MA: Abt Books.
- Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Research*, *38*, 79–93.
- Leite, W. L. (2007). A comparison of latent growth models for constructs measured by multiple items. Structural Equation Modeling, 14, 581–610.
- Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). SAS system for mixed models. Cary, NC: SAS Institute.
- Little, R. T. A., & Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149–174.
- McArdle, J. J. (1986). Latent variable growth within behavior genetic models. *Behavior Genetics*, 16, 163–200.
- McArdle, J. J. (1988). Dynamic but structural equation models of repeated measures data. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 561–614). New York: Plenum Press.
- McArdle, J. J. (1989). Structural modeling experiments using multiple growth functions. In P. Ackerman, R. Kanfer, & R. Cudeck (Eds.), *Learning and individual differences: Abilities,* motivation, and methodology (pp. 71–117). Hillsdale, NJ: Erlbaum.
- McArdle, J. J. (1994). Structural factor analysis experiments with incomplete data. *Multivariate Behavioral Research*, 29, 409– 454.
- McArdle, J. J. (2007). Five steps in the structural factor analysis of longitudinal data. In R. Cudeck & R. C. MacCallum (Eds.), Factor Analysis at 100: Historical developments and future directions (pp. 99–130). Mahwah, NJ: Erlbaum.
- McArdle, J. J., & Anderson, E. (1990). Latent variable growth models for research on aging. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (3rd ed., pp. 21–44). San Diego, CA: Academic Press.
- McArdle, J. J., & Bell, R. Q. (2000). An introduction to latent growth models for developmental data analysis. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples* (pp. 69–107). Mahwah, NJ: Erlbaum.
- McArdle, J. J., & Cattell, R. B. (1994). Structural equation models of factorial invariance in parallel proportional profiles and oblique confactor problems. *Multivariate Behavioral Research*, 29, 63–113.
- McArdle, J. J., Ferrer-Caja, E., Hamagami, F., & Woodcock, R. W. (2002). Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span. *Developmental Psychology*, 38, 115–142.
- McArdle, J. J., & Hamagami, F. (1992). Modeling incomplete longitudinal and cross-sectional data using latent growth structural models. *Experimental Aging Research*, 18, 145–166.
- McArdle, J. J., & Hamagami, F. (1996). Multilevel models from a multiple group structural equation perspective. In G. Marcou-

- lides & R. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 89–124). Hillsdale, NJ: Erlbaum.
- McArdle, J. J., & Hamagami, F. (2001). Latent difference score structural models for linear dynamic analyses with incomplete longitudinal data. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 139–175). Washington, DC: American Psychological Association.
- McArdle, J. J., & Hamagami, F. (2004). Methods for dynamic change hypotheses. In K. van Montfort, J. Oud, & A. Satorra (Eds.), *Recent developments in structural equation models* (pp. 295–336). London: Kluwer Academic.
- McArdle, J. J., Hamagami, F., Meredith, W., & Bradway, K. P. (2000). Modeling the dynamic hypotheses of Gf-Gc theory using longitudinal life-span data. *Learning and Individual Dif*ferences, 12, 53–79.
- McArdle, J. J., & Nesselroade, J. R. (1994). Using multivariate data to structure developmental change. In S. H. Cohen & H. W. Reese (Eds.), *Life-span developmental psychology: Methodological contributions* (pp. 223–267). Hillsdale, NJ: Erlbaum.
- McArdle, J. J., & Nesselroade, J. R. (2003). Growth curve analysis in contemporary psychological research. In J. A. Schinka, W. F. Velicer, & I. B. Weiner (Eds.), *Handbook of psychology: Vol. 2. Research methods in psychology* (pp. 447–480). New York: Wiley.
- McArdle, J. J., Prescott, C. A., Hamagami, F., & Horn, J. L. (1998). A contemporary method for developmental-genetic analyses of age changes in intellectual abilities. *Developmental Neuropsychology*, *14*(1), 69–114.
- McArdle, J. J., & Wang, L. (2008). Modeling age-based turning points in longitudinal life-span growth curves of cognition. In P. Cohen (Ed.), *Applied data analytic techniques for turning points research* (pp. 105–128). New York: Routledge.
- McArdle, J. J., & Woodcock, R. W. (1997). Expanding test–retest designs to include developmental time-lag components. *Psychological Methods*, 2, 403–435.
- McDonald, R. P. (1999) *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.
- McNemar, Q. (1942) The revision of the Stanford–Binet Scale: An analysis of the standardization data. New York: Houghton Mifflin.
- Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, 55(1), 107–122.
- Metha, P. D., & West, S. G. (2000). Putting the individual back into individual growth curves. *Psychological Methods*, *5*, 23–43.
- Mislevy, R. J., & Wilson, M. (1996). Marginal maximum likelihood estimation for a psychometric model of discontinuous development. *Psychometrika*, *61*(1), 41–71.
- Muraki, E. (1992). A generalized partial credit model: Applications of an EM algorithm. *Applied Psychological Measurement*, *16*, 159–176.
- Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, 29, 81–117.

- Muthén, B. O., & Curran, P. J. (1997). General longitudinal modeling of individual differences in experimental designs: A latent variable framework for analysis and power estimation. *Psychological Methods*, 2, 371–402.
- Nesselroade J. R., & Baltes, P. B. (Eds.). (1979). *Longitudinal research in the study of behavior and development*. New York: Academic Press.
- Pastor, D. A., & Beretvas, S. N. (2006). Longitudinal Rasch modeling in the context of psychotherapy outcomes assessment. *Applied Psychological Measurement*, *30*, 100–120.
- Pinherio, J. C., & Bates, D. M. (2000). Mixed-effects models in S and S-PLUS. New York: Springer.
- Ram, N., Chow, S. M., Bowles, R. P., Wang, L., Grimm, K., Fujita, F., & Nesselroade, J. R. (2005). Examining interindividual differences in cyclicity of pleasant and unpleasant affect using spectral analysis and item response modeling. *Psychometrika*, 70, 773–790.
- Rao, C. R. (1958). Some statistical methods for comparison of growth curves. *Biometrics*, *14*, 1–17.
- Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.
- Rasch, G. (1966). An individualistic approach to item analysis. In P. F. Lazarsfeld & N. W. Henry (Eds.), *Readings in mathematical social science* (pp. 89–107). Chicago: Science Research.
- Rijmen, F., De Boeck, P., & van der Maas, H. L. J. (2005). An IRT model with a parameter-driven process for change. *Psychometrika*, 70(4), 651–669.
- Rijmen, F. Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, 8, 185–205.
- Rost, J. (2002). The growing family of Rasch models. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 2–42). New York: Springer.
- Sayer, A. G., & Cumsille, P. E. (2001). Second-order latent growth models. In L. M. Collins & A. G. Sayer (Eds.), *New methods for* the analysis of change (pp. 179–200). Washington, DC: American Psychological Association.
- Sheu, C.-F., Chen, C.-T., Su, T.-H., & Wang, W.-C. (2005). Using SAS PROC MIXED to fit item response models. *Behavior Research Methods, Instruments, & Computers, 37*(2), 208–218.
- Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, 23, 323–355.
- Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling. Multilevel, longitudinal, and structural equation models.* Boca Raton, FL: Chapman & Hall/CRC.
- Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modeling*, *3*, 25–40.
- Smith, R. M., & Suh, K. K. (2003). Rasch fit statistics as a test of the invariance of item parameter estimates. *Journal of Applied Measurement*, 4, 153–163.
- Smits, D. J. M., De Boeck, P., & Verhelst, N. D. (2003). Estimation of the MIRID: A program and a SAS-based approach.

- Behavior Research Methods, Instruments, & Computers, 35(4), 537–549.
- Snijders, T. A., & Bosker, R. J. (1994). Modeled variance in two-level models. Sociological Methods & Research, 22(3), 342–363.
- Sörbom, D. (1975). Detection of correlated errors in longitudinal data. *British Journal of Mathematical and Statistical Psychology*, 28, 138–151.
- Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. R. (1994). *BUGS: Bayesian Inference Using Gibbs Sampling, Version 0.30*. Cambridge, England: MRC Biostatistics Unit.
- Swanson, H. L. (1999). What develops in working memory? A life span perspective. *Developmental Psychology*, 35, 986–1000.
- Thurstone, L. L. (1928). The absolute zero in intelligence measurement. *Psychological Review*, *35*, 175–197.
- Tisak, J., & Tisak, M. S. (1996). Longitudinal models of reliability and validity: A latent curve approach. Applied Psychological Measurement, 20, 275–288.
- Tucker, L. R. (1958). Determination of parameters of a functional relation by factor analysis. *Psychometrika*, 23, 19–23.
- van der Linden, W. J. (1994). Fundamental measurement and the fundamentals of Rasch measurement. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 2, pp. 3–24). Norwood, NJ: Ablex.
- van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.
- Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer.
- Wang, L., & McArdle, J. J. (2008). A simulation study comparison of Bayesian estimation with conventional methods for estimating unknown change points. Structural Equation Modeling, 15, 52–74.
- Wechsler, D. (1941). *The measurement of adult intelligence* (2nd ed.). Baltimore: Williams & Wilkins.
- Wechsler, D. (1955). *Manual for the Wechsler Adult Intelligence Scale*. New York: Psychological Corporation.
- Wechsler, D. (1981). WAIS-R: Wechsler Adult Intelligence Scale Revised. New York: Harcourt Brace Jovanovich.
- Wiley, D. E., & Wiley, J. A. (1970). The identification of measurement errors in panel data. *American Sociological Review*, 35, 112–117.
- Willett, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin*, 116, 363–381.
- Wilson, M. (1989). Saltus: A psychometric model for discontinuity in cognitive development. *Psychological Bulletin*, 105, 276–289.
- Wilson, M. (2005). Constructing measures: An item response modeling approach. Mahwah, NJ: Erlbaum.
- Wilson, M., & Draney, K. (1997). Partial credit in a developmental context: The case for adopting a mixture model approach. In M. Wilson, G. Engelhard Jr., & K. Draney (Eds.), *Objective mea*surement: Theory into practice (Vol. 4, pp. 333–350). Greenwich, CT: Ablex.
- Wohlwill, J. F. (1973). The study of behavioral development. New York: Academic Press.

- Woodcock, R. W. (1990). Theoretical foundations of the WJ-R measures of cognitive ability. *Journal of Psychoeducational Assessment*, 8, 231–258.
- Woodhouse, G., Yang, M., Goldstein, H., & Rasbash, J. (1996).
 Adjusting for measurement error in multilevel analysis. *Journal of the Royal Statistical Society: Series A: Statistics in Society*, 159, 201–212.
- Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit statistics. *Rasch Measurement Transactions*, *8*, 370.
- Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: Mesa Press.
- Zhang, Z., McArdle, J. J., Wang, L., & Hamagami, F. (2008). A SAS interface for Bayesian analysis with WinBUGS. *Structural Equation Modeling*, 15, 705–728.

Appendix

Computer Scripts for Alternative Estimation of Longitudinal Invariant Rasch Tests

1a: WINSTEPS input script for initial Rasch model fitting

Winsteps Script for PCM Model &INST
TITLE='Vocabulary Analysis'
DATA=LIRT_Vocab_Items.txt
ITEM1=20
NI=278
PTBIS=Y
CODES = 012
GROUPS= 0
&END

1b: Alternative SAS PROC NLMIXED Script for Initial Rasch model fitting

```
TITLE2 'Initial Rasch Model for First Occasion';

PROC NLMIXED DATA = LIRT_Vocab_Items_224 (WHERE=(Time=1))

METHOD=GAUSS TECHNIQUE=NEWRAP NOAD QPOINTS=20;

ARRAY beta[224] beta1-beta224;

diff = gscore - beta[item_num];

p=1/(1+EXP(-diff));

MODEL item_mem ~ BINARY(p);

RANDOM gscore ~ NORMAL([0], [sigma_g*sigma_g]) SUBJECT = id;

PARMS s_g=1 beta1-beta224=.001;

ESTIMATE 'variance' sigma_g*sigma_g;

PREDICT p OUT=Vocab_Pred_Prob;

PREDICT gscore OUT=Vocab_Pred_Parm_Person;

ODS OUTPUT ParameterEstimates=Vocab_Pred_Parm_Item;

RUN;
```

2: SAS PROC NLMIXED Script for Dual Exponential Growth Model of estimated Rasch Scores

```
TITLE2 'Dual Exponential Model fitted to estimated scale scores';

PROC NLMIXED DATA = LIRT_Vocab_Scale;

Yt = g0 + g1 * At;

At=(EXP(-pi_d*(age-10)) - EXP(-pi_g*(age-10)));

MODEL measure ~ NORMAL(Yt, V_e);

RANDOM g0 g1 ~ NORMAL([nu_0, nu_1],

[V_0, C_01, V_1])

SUBJECT = id out=LIRT_Vocab_Growth_estimate;

PARMS nu_0=0 nu_1=1 pi_g=.10 pi_d=0.01

V_e=1 V_0=5 V_1=1 C_01=.01;

RUN;
```

3: SAS PROC NLMIXED Script for Simultaneous LIRT Approach

```
TITLE2 'Fitting the item-growth model to the longitudinal item data'; PROC NLMIXED DATA = LIRT_Vocab_Items_224 \,
      METHOD=GAUSS TECHNIQUE=NEWRAP NOAD QPOINTS=20;
       ARRAY beta[224] beta1-beta224;
       ARRAY g[16] g1-g16;
       ARRAY u[16] u1-u16;
         g[occ] = g_0 + g_1*At + u[occ];
At =EXP(-pi_d*(age-10)) - EXP(-pi_g*(age-10));
diff = g[occ] - beta[item_num];
         p=1/(1+EXP(-diff));
         v_0 = sigma_0*sigma_0;
         v_1 = sigma_1*sigma_1;
         v_u = sigma_u*sigma_u;
         c_01 = rho_01*sigma_0*sigma_1;
MODEL item_mem ~ BINARY(p);
RANDOM g_0 g_1 u1 u2 u3 u4 u5 u6 u7 u8 u9 u10 u11 u12 u13 u14 u15 u16 \sim
NORMAL([0,nu_1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0],
             [v_0,
              c_01, v_1,
           0, 0, v_u,
           0, 0, 0, v_u,
           0, 0, 0, 0, v_u,
           0, 0, 0, 0, 0, v_u,
           0, 0, 0, 0, 0, v_u,
           0, 0, 0, 0, 0, 0, v_u,
           0, 0, 0, 0, 0, 0, 0, 0, v_u,
0, 0, 0, 0, 0, 0, 0, 0, v_u,
           0, 0, 0, 0, 0, 0, 0, 0, 0, 0, v_u,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, v_u,
           SUBJECT = id;
PARMS nu_1=1 sigma_0=1 sigma_1=1 rho_01=.001 sigma_u=2
      beta1-beta224=1 pi_g=.15 pi_d=0.001;
ESTIMATE 'level variance' sigma_0*sigma_0;
ESTIMATE 'slope variance' sigma_1*sigma_1;
ESTIMATE 'leve-slope covariance' rho_01*sigma_0*sigma_1;
ESTIMATE 'unique variance' sigma_u*sigma_u;
PREDICT p OUT=Pred_Prob;
PREDICT gscore OUT=Vocab_Pred_Parm_Person;
ODS OUTPUT ParameterEstimates=Vocab_Pred_Parm_Item;
RIIN;
```

4: WinBUGS Script for Simultaneous One-Stage LIRT Approach

```
Model Dual Exponential with PCM
#variable
             specification
gscore[N,T] person trait levels
             item difficulties
beta[I]
             step difficulties
gamma[J]
tau
             precision for person distribution
sigma2
             variance of person distribution
p[N,I,T,J] category probabilities
x[N,I,T]
             item responses
z[N,I,T,J] working matrix #likelihood
\verb|model| \{
       for (n in 1:N) {
             for (t in 1:numobs[n])
                    for (i in 1:115) {
                           logit(p[n,t,i])<-gscore[n,t]-beta[i]
x[n,t,i] ~ dbern(p[n,t,i])
                    for (i in 116:224) {
                           z[n,t,i,1] < -1
                           pr[n,t,i,1] < -1/sum(z[n,t,i,])
                     log(z[n,t,i,2])<-gscore[n,t]-beta[i]-gamma[i-115]
                           pr[n,t,i,2] < -z[n,t,i,2]/sum(z[n,t,i,])
                           log(z[n,t,i,3])<-2*(gscore[n,t]-beta[i])
pr[n,t,i,3]<-z[n,t,i,3]/sum(z[n,t,i,])
                           x[n,t,i] \sim dcat(pr[n,t,i,])
                    gscore[n , t] ~ dnorm(mu[n , t],tauy)
mu[n,t] <- nu[n,1] +nu[n,2]* At</pre>
             At= (\exp(-1*pid*(age[n,t]-10)) - \exp(-1*pig*(age[n,t]-10)))
#priors
       for (i in 1:224) {
             beta[i] ~ dnorm(0,1.0E-6)
       for (i in 1:109) {
             gamma[i] ~ dnorm(0,1.0E-6)
       tauy ~ dgamma(0.001,0.001)
       for( n in 1 : N ) {
             nu[n,1:2]~dmnorm(munu[1:2],taunu[1:2,1:2])
       munu[1]<-0
       munu[2]~dnorm(0, 1.0E-6)
       taunu[1:2,1:2] \sim dwish(R[1:2, 1:2],2)
       sigma2nu[1:2, 1:2]<-inverse(taunu[1:2,1:2])
      sigmay <- 1 / sqrt(tauy)
pid~dnorm(0,1.0E-6)</pre>
       pig~dnorm(0,1.0E-6)
NOTE: Data entry in vector form needs to follow
```

Received December 31, 2005
Revision received February 11, 2009
Accepted February 11, 2009