Psychology and Aging
2007, Vol. 22, No. 3, 525–545

0882-7974/07/$12.00 DOI: 10.1037/0882-7974.22.3.525

Latent Variable Analyses of Age Trends of Cognition in the Health and Retirement Study, 1992–2004

John J. McArdle
University of Southern California

Gwenith G. Fisher
University of Michigan

Kelly M. Kadlec
University of Southern California

The present study was conducted to better describe age trends in cognition among older adults in the longitudinal Health and Retirement Study (HRS) from 1992 to 2004 ($N > 17{,}000$). The authors used contemporary latent variable models to organize this information in terms of both cross-sectional and longitudinal inferences about age and cognition. Common factor analysis results yielded evidence for at least 2 common factors, labeled Episodic Memory and Mental Status, largely separable from vocabulary. Latent path models with these common factors were based on demographic characteristics. Multilevel models of factorial invariance over age indicated that at least 2 common factors were needed. Latent curve models of episodic memory were based on age at testing and showed substantial age differences and age changes, including impacts due to retesting as well as several time-invariant and time-varying predictors.

Keywords: longitudinal data, cognitive aging, structural factor analysis, latent variable structural equation modeling, latent growth-decline curve models

Classical research in gerontology has sought to determine the nature of adult age-related changes in cognitive functioning (e.g., Baltes & Schaie, 1976; Bayley, 1966; Botwinick, 1977; Bradway & Thompson, 1962; Horn, 1991; Jones & Conrad, 1933; Park, 2000; Schaie, 1996). Recent methodological studies have followed the same principles but have examined both age and cohort effects (Donaldson & Horn, 1992; McArdle & Anderson, 1990) and the influences of practice or retesting effects (McArdle, Hamagami, Meredith, & Bradway, 2001; Rabbitt, Diggle, Holland, & McInnes, 2004). Researchers have also been interested in the observation of large positive shifts in cognitive abilities among more recent birth cohorts (e.g., the Flynn effect; Flynn, 1984; Neisser, 1998). In research on older populations, some investigators have found a consistent cohort-related trend of decreased decline in cognitive functioning (Freedman & Martin, 1998, 2000). In contrast, Rodgers, Ofstedal, and Herzog (2003) recently refuted these claims on the basis of large-scale data analysis from the Health and Retirement Study (HRS) and the Asset and Health Dynamics of the Oldest Old Study (AHEAD; see Juster & Suzman, 1995). The purpose of the present study is to examine key questions related to age trends in cognition among older adults using contemporary techniques in latent structure modeling (e.g., McDonald, 1999; Meredith & Tisak, 1990; Park et al., 2002).

Previous Cognitive Research on the HRS

The HRS/AHEAD studies provide an excellent source of data for examining age trends in cognitive ability in the older U.S. population. The HRS and AHEAD studies began in 1992 and 1993, respectively, and in 1998 were combined into one study that attempts to be nationally representative of Americans over 50 years of age. The studies use a panel design in which the same respondents are interviewed every 2 years, and new respondents are added to the sample every 6 years to replenish the sample to adjust for aging and attrition (see Heeringa & Connor, 1996; Leacock, 2006; http://hrsonline.isr.umich.edu).

Herzog and Wallace (1997) conducted the most complete study to date examining the quality of the HRS cognitive measures in the HRS samples. They provided results for simple internal consistency reliability (using alpha indexes), exploratory factor analysis, and regression with demographic variables. Some of these findings show the limitations of the HRS cognitive battery, including lower than expected discriminations and reliabilities, and suggest a rel-

John J. McArdle and Kelly M. Kadlec, Department of Psychology, University of Southern California; Gwenith G. Fisher, Institute for Social Research, University of Michigan.

This research was supported by National Institute on Aging Grant AG-007137 to John J. McArdle and by National Institute on Aging Grant AG-027010, which provided additional Health and Retirement Study Aging, Demographics, and Memory Study support.

We thank our colleagues Fumiaki Hamagami, Kevin Grimm, John R. Nesselroade, and the other coprincipal investigators on the Health and Retirement Study for their helpful comments on drafts of this article. We thank John L. Horn for his inspiring efforts to understand and improve cognition measurement in aging research.

Computer scripts for all analyses are available on the University of Southern California National Growth and Change Study Web site (http://kiptron.usc.edu) under the subheading HRS_AgeTrends_2007.

Correspondence concerning this article should be addressed to John J. McArdle, Department of Psychology, University of Southern California, National Growth and Change Study Laboratory, SGM 826, Los Angeles, CA 90089. E-mail: jmcardle@usc.edu

atively complex two-factor structure. Other findings show that some of the cognitive tasks are clearly related to age, education, and health.

Differences due to testing modality—face-to-face versus telephone interviews—were studied in the recent work of Herzog and Rodgers (1999). Using the AHEAD data, they highlighted important issues in survey design requirements for working with an older sample. They pointed out the potential limitations of population estimates of measured cognition when only noninstitutionalized adults are used, demonstrated changes in nonresponse from one wave to the next, and studied differences between spouses. They also evaluated an HRS/AHEAD randomized experiment on administering the HRS cognitive test battery in a telephone versus face-to-face mode and found negligible differences in performance on the cognitive measures due specifically to mode of testing.

Other studies have used the HRS cognition measures as the main outcome of interest, as a key predictor of some other outcome, or as a control variable. One study examined the impact of education and wealth on cognitive function (Cagney & Lauderdale, 2002), another used HRS cognition measures to estimate the association of life expectancy and cognitive impairment (Suthers, Kim, & Crimmins, 2003), and a third examined lifestyle predictors of cognitive function and change in cognitive function (Ofstedal & Herzog, 2004). Another set of studies focused on cognition as a key predictor of health care utilization and behavioral outcomes, including physical functioning (Blaum, Ofstedal, & Liang, 2002; Fultz, Ofstedal, Herzog, & Wallace, 2003; Herzog & Wallace, 1997), informal caregiving (Langa et al., 2001), nursing home admission (Langa et al., 2004), and cohort trends in severe cognitive impairment (Freedman, Aykan, & Martin, 2001, 2002). The recent studies by Adams, Hurd, McFadden, Merrill, and Ribeiro (2003); Zelinski, Crimmins, Reynolds, and Seeman (1998); and Zelinski and Gilewski (2003) offered multivariate analyses of the relationships among multiple measures of physical health and HRS cognitive functioning.

Two recent analyses of HRS data examined cohort trends in cognitive functioning among older Americans. For example, Freedman, Aykan, and Martin (2001) investigated aggregate trends of cognitive impairment in a cohort of older Americans between 1993 and 1998 using two waves of data from HRS/AHEAD. They concluded that there was a decline in severe cognitive impairment between the two waves (6.1% in 1993 and 3.6% in 1998) that was not fully explained by differences in demographic, socioeconomic, and health characteristics between the two waves. In a related study, Rodgers et al. (2003) extended Freedman et al.'s (2001) work by examining cohort-level trends in cognitive scores across four waves of HRS/AHEAD (1993, 1995, 1998, and 2000). They conducted four sets of analyses to examine the trends in cognitive scores across the four waves of data. First they examined unadjusted scores and looked at changes from one wave of testing to another. In the second set of analyses they used multivariate regression to evaluate whether study design and demographic characteristics (including age) accounted for differences across waves. In a third set of analyses, the authors used similar regression models to adjust for additional respondent demographic characteristics related to cognitive scores, including gender, race, ethnicity, and educational attainment. Finally, Rodgers et al. (2003) addressed mode effects (telephone vs. face-to-face administration of the interview) and sample selection effects due to attrition and addition of new sample members in 1998 by adding specific contrast codes to this model.

Rodgers et al. (2003) concluded that there was little improvement in cognitive functioning across the cohorts. They suggested that the differences in the results between the two studies could be explained by statistical adjustments for methodological factors in the complex panel study of the HRS. They also examined the impacts of (a) prior exposure (practice), (b) mode of testing, and (c) alternative test forms but found only small differences due to these design features. In particular, they found that scores on the two word recall tasks were consistently higher after one exposure, whereas scores on the Serial 7s task (see below) were consistently lower. Another result was that cognitive scores of those who were interviewed by telephone were substantially higher on all four of the cognitive tests. However, because the in-home interview mode might have been chosen because of previously low cognitive scores, this was not resolved in their statistical analysis. Finally, Rodgers et al. also studied whether the alternative forms (i.e., four word lists) were equivalent to one another, and they found that one of the word lists was more difficult than the others.

The Current HRS Research

The analyses presented in this article differ in both form and purpose from those presented in any previous work with the HRS. To start, we describe the combination of all the available cognitive data from the 1992 to 2004 HRS/AHEAD surveys, using the final release version of the 2004 data (available as of August 2006). Most previous HRS research has used multiple regression analyses to understand changes across waves of testing, including specific demographic differences and design confounds, and cognition scores have been combined into a single composite score. Our new analyses add to the prior HRS analyses in several ways. First, because age is the major substantive issue of interest, we highlight age as the major methodological frame of reference, and we use contemporary versions of latent structure analysis to deal with key questions about the age-related structure and age-based dynamics of the cognitive variables measured in the HRS. We use latent structure models in the determination of common factors, latent path models to deal with demographic impacts on common factors, latent multilevel models to evaluate factorial invariance over time, and latent growth–decline curve models to deal with age-based changes in cognition. We deal with other key demographics, such as gender and birth cohort, as second-order effects in the latent growth–decline models (Jones & Meredith, 2000; McArdle & Anderson, 1990).

In general, we use these new latent structure analyses to evaluate the importance of age effects and the role of several key covariates, including the potential for test-practice effects, differences due to educational level or birth cohort, and differences for gender and marital status. All latent variable analyses presented in this article rely on common statistical assumptions about the reasons for incomplete data, and we discuss but do not elaborate on these assumptions (but see Ferrer, Salthouse, McArdle, Stewart, & Schwartz, 2005; McArdle, Ferrer-Caja, Hamagami, & Woodcock, 2002; McArdle & Woodcock, 1997). The analyses presented in this article raise issues that should be considered in future work and set the stage for necessary and potentially complex analyses of the role of specific cognitive factors in the HRS and other longitudinal cohort sequential or panel studies.

Method

Participants in HRS/AHEAD

The HRS is a nationally representative longitudinal study sponsored by the National Institute on Aging and conducted by the University of Michigan. The HRS researchers targeted community-dwelling adults in the contiguous United States who were 51 to 61 years old in 1992, when the baseline interview was conducted. Blacks, Hispanics, and Florida residents were oversampled (for details, see Heeringa & Connor, 1996). The initial HRS in-home interviews were conducted in over 7,700 households (with an 82% response rate), yielding more than 9,800 participants between the ages of 51 and 61. Within each household, the spouse or domestic partner of the sampled respondent was also interviewed, regardless of age. Follow-up interviews were conducted every 2 years. In 1993 and 1995, the AHEAD study was conducted among a national sample of adults age 70 or older. In 1998, the HRS and AHEAD studies merged, both assuming the name HRS, and two new cohorts were added to the HRS sampling frame. As a result, in 1998 the overall HRS sample of self-respondents was representative of the U.S. population of adults born in 1947 or earlier who resided in the 48 contiguous states. Data are now available from nine waves of testing: HRS 1992, AHEAD 1993, HRS 1994, AHEAD 1995, HRS 1996, HRS 1998, HRS 2000, HRS 2002, and HRS 2004. These data include information on more than 30,000 people in more than 17,000 families measured on over 100,000 interviews.

In the HRS, the respondents can be interviewed at their convenience. Interviewers attempt to set up appointments with spouses or partners on the same day. If respondents do not want the interviewer to come to their home, the interviews are carried out at a convenient location. The HRS longitudinal assessments currently use two modes of data collection. Baseline interviews and those with respondents age 80 or older are conducted in person, but most of the longitudinal follow-ups are administered over the telephone. When two people are living as a couple, they are often tested at the same occasion. Sample weights were designed to reflect the national representation of the sample by accounting for the oversampling of Blacks, Hispanics, and households in the state of Florida; compensating for unequal selection probabilities in geographical areas; and adjusting for geographic and race group differences in response rates. In addition, the weights were poststratified to match U.S. Census population totals (for details, see Heeringa & Connor, 1996). In most analyses to follow, we use the HRS respondent sampling weights, so the sample statistics and parameter estimates should reflect about 64 million people in the U.S. over the age of 50.

HRS Participants Selected for the Current Analyses

The HRS data used in the current analyses were selected and organized in a fashion similar to Rodgers et al. (2003). We started this analysis with the current HRS database of all people at all waves of testing and then eliminated any person who (a) was not a primary respondent, (b) had no data on gender or age, (c) had a sampling weight of zero, or (d) was under age 50. In addition, (e) in cases in which there were two persons per family, we randomly picked only one member of the family. Statistical information of the resulting subset of data for 17,355 self-respondents is described in Table 1. The demographic variables presented in this table include (a) chronological age at baseline testing, (b) years of formal education, (c) gender, (d) birth year (cohort), (e) number of waves of participation, (f) whether the participant was one member of a couple, and (g) whether the individual was still a participant in the most recent 2004 testing. The basic statistics and coding schemes used are described in Table 1, and the pairwise correlation coefficients among these same variables are described in Table 2. These correlations describe the overall characteristics of the first time of testing, and we refer to this as participants' *baseline interview.* Using this selection of data, we find a negative relationship of age and education ($r = -.29$) and only a small differentiation of age and birth cohort ($r = -.95$).

The simple correlation just presented does not fully define the entire longitudinal pattern of participation in the HRS, but Table 3 lists longitudinal features in more detail. In this table we show how the 17,355 participants provided 69,496 interviews by isolating features of 13 groups of people on the basis of the patterns of complete and incomplete data. We include the frequency (and percentage) of all patterns of participation across all interviews, and we summarize age, education, and gender within each group as well. This information is presented separately for our selected HRS participants who were measured in the most recent 2004 wave of data collection and for those who provided some data but

Table 1
Descriptive Statistics for the Entire Sample at Initial Testing ($N = 17{,}355$)

Variable	*M*	*SD*	Minimum	Maximum
Age in years[a]	61.88	10.99	50.00	103.25
Years of education[a]	12.37	3.27	0.00	17.00
Gender effect coded[b]	0.57	0.497	−0.5 = male	0.5 = female
Living as a couple effect coded[c]	−0.033	0.499	−0.5 = living alone	0.5 = living in a couple
Birth cohort[d]	1,935	13.33	1,890	1,953
No. waves tested[d]	3.57	2.14	1	7
Tested in 2004 dummy coded[e]	0.678	0.467	0 = not tested in 2004	1 = tested in 2004

Note. The statistical information presented used Health and Retirement Study respondent-level sampling weights.
[a] Measured at Test 1. [b] Of the total sample, 55.7% was female. [c] Of the total sample, 53.2% was living as a couple. [d] Data include the Health and Retirement Study and Asset and Health Dynamics of the Oldest Old study members. [e] Of the total sample, 67.8% was tested in 2004.

Table 2
Pearson Correlation Coefficients for the Entire Sample at the First Time of Testing for Each Person

Variable	1	2	3	4	5	6	7
1. Age	—						
2. Years of education	−.265	—					
3. Gender effect coded	.167	−.098	—				
4. Couple effect coded	−.283	.151	−.253	—			
5. Year born	−.952	.284	−.171	.248	—		
6. No. waves tested	.017	−.004	.093	.058	−.235	—	
7. Tested in 2004	−.467	.214	−.020	.141	.513	.290	—

Note. $N = 17{,}355$, so correlations $|r| > .01$ are significant at the $p < .0001$ level.

dropped out for any reason sometime before 2004. Within each of these two broad classes, we further organize individuals into groups defined by the number of waves of participation. For example, in our subsample of the HRS data, 1,948 people were tested in 2004 for the first time, but 3,465 were tested in 2004 for the seventh time. In contrast, 1,762 people were tested once and not tested again, and 249 people were tested six times and then were not tested in 2004. The dropout by 2004 occurred at different times and could be due to many reasons (mortality, lack of interest, etc.), but we used all data from all participants in our analyses.

The difference between the initial cross-sectional data and the subsequent longitudinal data is a key feature of the HRS, and one aspect of this issue is illustrated in Figures 1A and 1B . Figure 1A is a frequency distribution of the ages at the first time of testing, and this shows the clear separation of the HRS (younger) and AHEAD (older) study samples. Figure 1B is a frequency distribution of the ages at all times of testing and is based on every interview for every person selected, so this represents the full age distribution available and the current overlap of the HRS study samples.

Cognitive Measures in HRS/AHEAD

The cognitive performance measures in the HRS/AHEAD studies are used in this article, and these are outlined in Table 4. A more detailed description of these measures appears in Brandt, Spencer, and Folstein (1988) and Ofstedal, Fisher, and Herzog (2005). Specific cognitive measures included performance on (a) immediate and delayed free recall of a list of 10 nouns (for a possible score of 0–10 on each measure); (b) Serial 7s, a working memory and mental processing task in which respondents counted backward from 100 by 7s for a total of five trials (for a possible score of 0–5); and (c) mental status measures (with a possible combined score of 0–10), including counting backward from 20, naming the U.S. president and vice president by last name, naming two objects (scissors and cactus) on the basis of a brief description, and providing the date (month, day, year, and day of week) for an assessment of time orientation. Among many others, Rodgers et al. (2003) used a composite score they created by summing across the three scores (range from 0 to 35), but this is not used in the present

Table 3
Patterns of Longitudinal Participation and Frequency of Observations in the Health and Retirement Study for Demographic Information (Unweighted Counts and Weighted Means)

No. tests taken	Frequency of people	% of people	Frequency of data	% of data	Mean age at first test	Education in years	% female
			Measured in 2004				
1	1,948	11.2	1,948	2.8	53.4	13.5	46.4
2	92	0.5	184	0.3	60.1	11.2	42.7
3	281	1.6	843	1.2	58.7	11.9	43.2
4	2,268	13.1	9,072	13.1	58.9	12.9	55.6
5	389	2.2	1,945	2.8	64.8	11.6	57.2
6	2,088	12.0	12,528	18.0	70.1	12.0	69.0
7	3,465	20.0	24,255	34.9	55.8	12.7	58.9
			Not measured in 2004				
1	1,762	10.2	1,762	2.5	70.8	11.1	55.1
2	1,622	9.4	3,244	4.7	69.7	11.2	56.2
3	1,373	7.9	4,119	5.9	68.9	11.5	58.1
4	988	5.7	3,952	5.7	68.9	11.4	60.2
5	830	4.8	4,150	6.0	68.5	11.8	60.4
6	249	1.4	1,494	2.2	56.6	11.4	50.0
Total	17,355	100	69,496	100	61.9	12.4	55.7

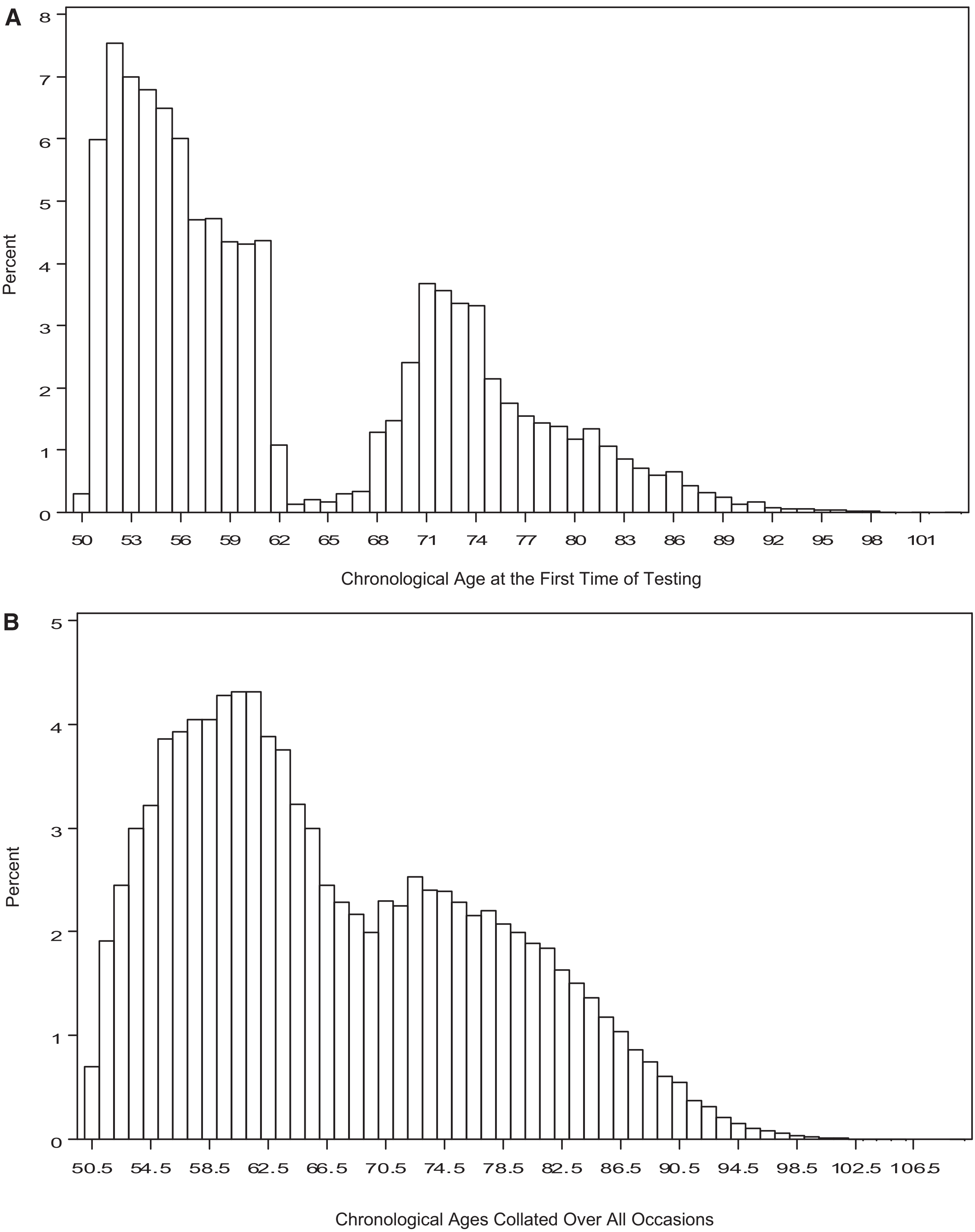


Figure 1. Frequency distributions for all chronological ages in the current Health and Retirement Study. A: Chronological age at the first time of testing (N = 17,355). B: Chronological age collated over all occasions (D = 69,496).

Table 4
A Description of Cognitive Measures in the Health and Retirement Study (HRS) and Asset and Health Dynamics of the Oldest Old Study (AHEAD) Collections

Measure	Typical features	Typical administration and scoring	Test waves administered	Testing notes
IR	10 words from four different lists	Typically given near end of survey; scores 0–10	All nine	Up to 20 items used in HRS 1992 and 1994
DR	10 words	Given 5 min after the IR	All nine	Five-min delay between IR and DR
Serial 7s	Subtract 7 from 100 for five trials	Administered between IR and DR; scores 0–5	All nine	
Backward Counting	Beginning with 20, count backward for 10 continuous numbers	Administered between IR and DR; scores 0 = incorrect, 1 = correct	Not used in HRS 1992 or 1994	In AHEAD 1995 and later waves, participants were instructed to count as quickly as possible. In AHEAD 1995, HRS 1996, and HRS 1998, backward counting from 86 was added
Dates	Today's date including month, day, year, and day of week	AHEAD 1995 and HRS 1996 administered between IR and DR; HRS 1998, 2000, 2002, and 2004 administered after DR; scores 0–4	Not used in HRS 1992 or 1994	In HRS 1998 and later waves, this question was only asked at baseline interview or of respondents 65 years of age or older
Names	Object naming (scissors and cactus); U.S. president and vice president naming by last name	AHEAD 1995 and HRS 1996 administered between IR and DR; HRS 1998, 2000, 2002, and 2004 administered after DR; scores 0–2	Not used in HRS 1992 or 1994	In HRS 1998 and later waves, this question was only asked at baseline interview or of respondents 65 years of age or older
Vocabulary	Five words in one of two lists from the WAIS-R	AHEAD 1995 and HRS 1996 administered after DR; HRS 1998, 2000, 2002, and 2004 administered after DR; scores 0–5	Used in AHEAD 1995 and later	In HRS 1998 and later waves, this question only asked at baseline interview or of respondents 65 years of age or older
Similarities	Seven word pairs as modified from WAIS-R	HRS 1992 administered between IR and DR; administered in modules in AHEAD 1993 and AHEAD 1994. Scores 0–7	HRS 1992 (all) and 1994 (module), AHEAD 1993 (module)	In AHEAD 1993, six word pairs were used instead of seven

Note. IR = Immediate Recall; DR = Delayed Recall; WAIS–R = Wechsler Adult Intelligence Scale—Revised.

work. The last two scales are based on shortened versions of Wechsler Adult Intelligence Scale—Revised subscales (Wechsler, 1981): (a) Similarities, which was measured on a subset of people before 1996, and (b) Vocabulary (five items), which was measured on everyone in HRS at least one time starting in 1996. All the Vocabulary data are used, but, because of the limited data on Similarities (less than 10%), they are not used in further analyses.

Summary statistics on all cognitive data for all occasions of measurement are presented in Table 5. This information is based on the seven cognitive variables at the first time of testing or the first time the participant had all seven tests (e.g., 1998 for AHEAD). To provide comparability across all scales and to simplify this measurement for further statistical analysis, we rescaled each variable into a percentage correct score (i.e., based on division by the maximum score and multiplication by 100). These statistics show that the subscales of Immediate Recall, Delayed Recall, and Vocabulary all had averages near 50% and had nearly normal distributions (skewness and kurtosis nearly zero). The Serial 7s scale was somewhat easier ($M = 70.5$), but it had a nearly normal distribution as well. However, the other three subscales, Backward Counting, Dates, and Names, had more than a 90% correct response rate and were negatively skewed.

Table 5
Descriptive Statistics for All Persons From the First Testing With the Most Cognitive Variables

Statistic	IR	DR	S7	BC	DA	NA	VO	Age	Educ
M	55.79	43.93	70.46	95.23	94.16	91.32	56.42	64.60	12.35
SD	18.52	22.30	44.07	21.09	14.27	16.71	21.20	11.45	3.28
Skewness	−0.15	−0.08	−0.68	−3.73	−2.98	−1.75	−0.37	0.53	−0.82
Kurtosis	0.01	−0.41	−0.96	12.01	10.75	2.61	−0.09	−0.87	0.98

Note. $N = 17{,}355$. This subsample was selected so there would be only one person per family, and respondent-level sampling weights were used to adjust to a U.S. national norm. IR = Immediate Recall; DR = Delayed Recall; S7 = Serial 7s; BC = Backward Counting; DA = Dates; NA = Names; VO = Vocabulary; Educ = education.

In Table 6 we present the pairwise correlations among the cognitive variables. Because there were some incomplete data on all measures (1%–5%) and about 20% of the people were never administered the Vocabulary measure (i.e., AHEAD 1995), these correlations were estimated with an incomplete data algorithm based on maximum-likelihood estimation (MLE-MAR in Mplus 4.0; Muthén & Muthén, 2006). The estimates presented in this table do not have any correction for nonnormality, so they are probably lower bound estimates. Nevertheless, the correlations among the seven cognitive measures exhibited a positive manifold, had a clear structure (i.e., were positive definite), and had negative correlations with age and positive correlations with education.

A summary description of all available longitudinal data from Table 3 is decidedly more complex, but some of the available information about the scores over age is presented in the plots of Figures 2A and 2B. In these figures we have plotted data from a randomly selected 1% of the selected people, and both the incomplete and the complete trajectories of data points are presented. The data for Figure 2A show the scores for Immediate Recall (y-axis) and the age at testing (x-axis), and the longitudinal data for each person are connected by a solid line. Figure 2B is the same kind of trajectory plot for Vocabulary scores, and here the incomplete data are clear. The straight lines in these figures represent the expectation of declining scores over age, and their derivation (from latent curve models) is discussed in detail in our later analyses.

There are many ways to give information about all the longitudinal cognitive data, but the complexity of the HRS data collection limits the simplicity of presentation. In subsequent analyses we characterize both longitudinal stability and fluctuations, so we use this approach for the initial description as well. Table 7 is a statistical summary of these kinds of associations among all the available longitudinal data from Table 2 (D = 69,496), with the total variance (σ_t^2) of the scores separated into between-persons variance (averages over time, σ_b^2) and within-person variances (deviations over time, σ_w^2). These terms provide initial estimates of longitudinal stability and changes for each variable by (a) collating the individual data on any score for each person ($Y[t]_n$) and (b) separating the individual average of the scores over time ($\mathbf{Y}_n$) from (c) the deviations around this average ($Y[t]_n - \mathbf{Y}_n$). In the typical analysis of persons in families (Nagoshi & Johnson, 1987) or in classrooms (Muthén, 1991), this is a simple way to separate the group differences from the individual differences within these groups. In the longitudinal context, however, this is a simple way to separate the stable average of a person's scores from the unstable deviations around the person's average. In this standard calculation, the unstable part includes the systematic individual-specific changes as well as all the time-varying measurement errors. It follows that the further calculation of longitudinal intraclass correlations (i.e., eta-squared) for each variable can be used to compare relative stability and change across the cognitive variables, but it should not be interpreted as a standard reliability coefficient (see McArdle & Woodcock, 1997). In these data, we can easily distinguish the most stable variables ($\eta^2 \sim .63$ for Backward Counting) from those that fluctuate most (η^2s $\sim .22$ for Serial 7s and Dates).

In Table 8 we present the same description for each pair of variables, and this is an initial description of the pattern of correlations of the stable and fluctuating components. The between-persons correlations (below the diagonal) indicate the similarities of the stable parts of the variables, and most of these are $\rho_b > 0.50$ (e.g., the highest $\rho_b = 0.96$ for Immediate Recall and Delayed Recall). The corresponding within-person (above the diagonal) correlations indicate the similarities among the systematic changes in a pair of variables over time (i.e., without measurement error), and most of these were much lower, around $\rho_w < 0.20$ (e.g., the highest $\rho_w = 0.61$ for Immediate Recall and Delayed Recall). The relative pattern of these between- and within-correlation matrices in Table 8 (in covariance form) is difficult to discern from this table, but it is used as the basis of longitudinal factor invariance analyses.

Latent Variable Modeling Analyses

The analyses we use are designed to characterize the basic structure and age changes in the cognitive variables measured in the current HRS. These questions can be answered via many different statistical methods, but we used a sequence of four contemporary forms of latent variable structural equation modeling:

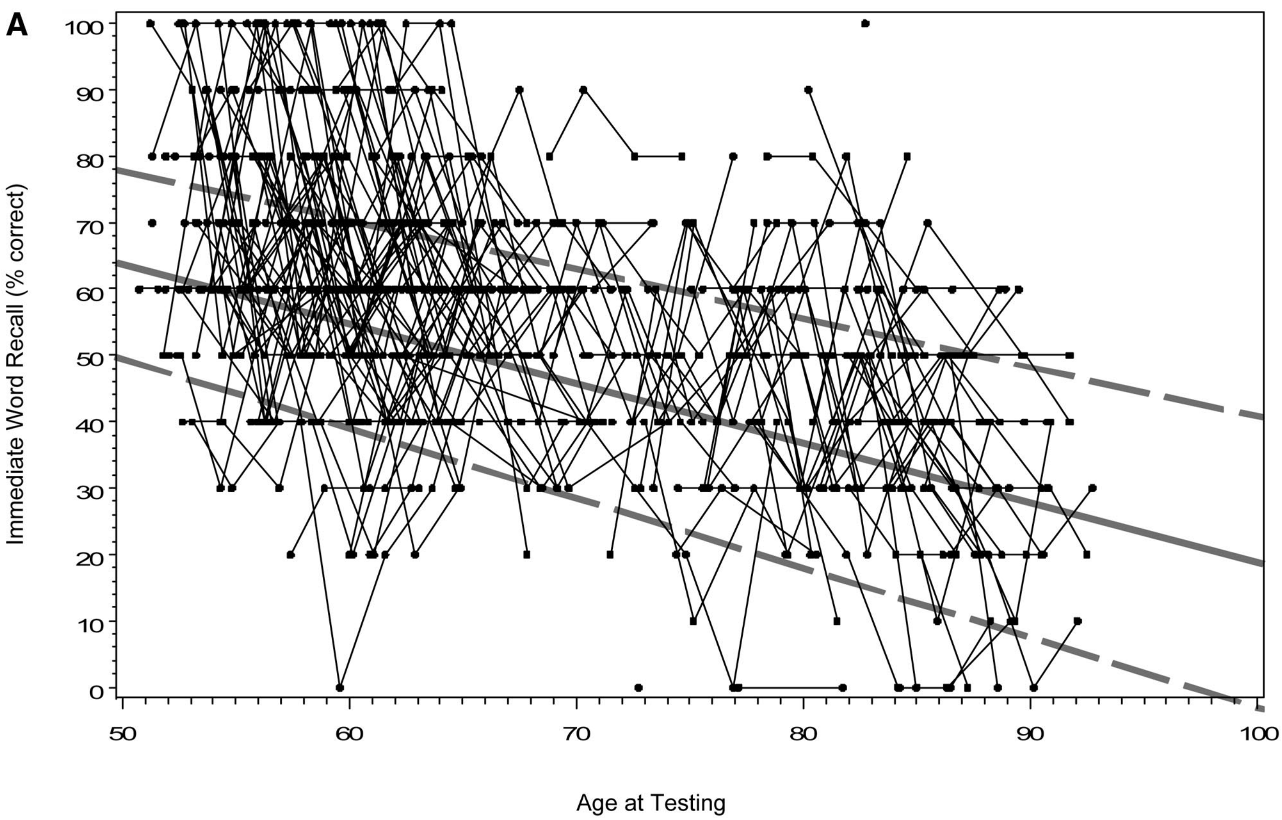
1. latent common factor models—unrestricted or exploratory factor analysis to define the basic factorial structure of the HRS cognitive measures,

2. latent path models—restricted or confirmatory factor analysis techniques to define the cross-sectional relation-

Table 6
Maximum Likelihood Estimate Missing at Random Estimates of Pairwise Correlation Coefficients for All Persons From the First Testing With the Most Cognitive Variables

Variable	1	2	3	4	5	6	7	8	9
1. IR	—								
2. DR	.772	—							
3. S7	.367	.356	—						
4. BC	.187	.169	.226	—					
5. DA	.279	.278	.254	.200	—				
6. NA	.357	.343	.379	.220	.307	—			
7. VO	.373	.341	.388	.182	.193	.391	—		
8. Age	−.401	−.403	−.215	−.099	−.202	−.176	−.158	—	
9. Education	.390	.359	.438	.186	.213	.402	.474	−.273	—

Note. Eigenvalues (7 × 7 submatrix) Λ = {2.97, 0.99, 0.84, 0.79, 0.62, 0.57, 0.23}. Overall −2LL = −500,638. IR = Immediate Recall; DR = Delayed Recall; S7 = Serial 7s; BC = Backward Counting; DA = Dates; NA = Names; VO = Vocabulary.



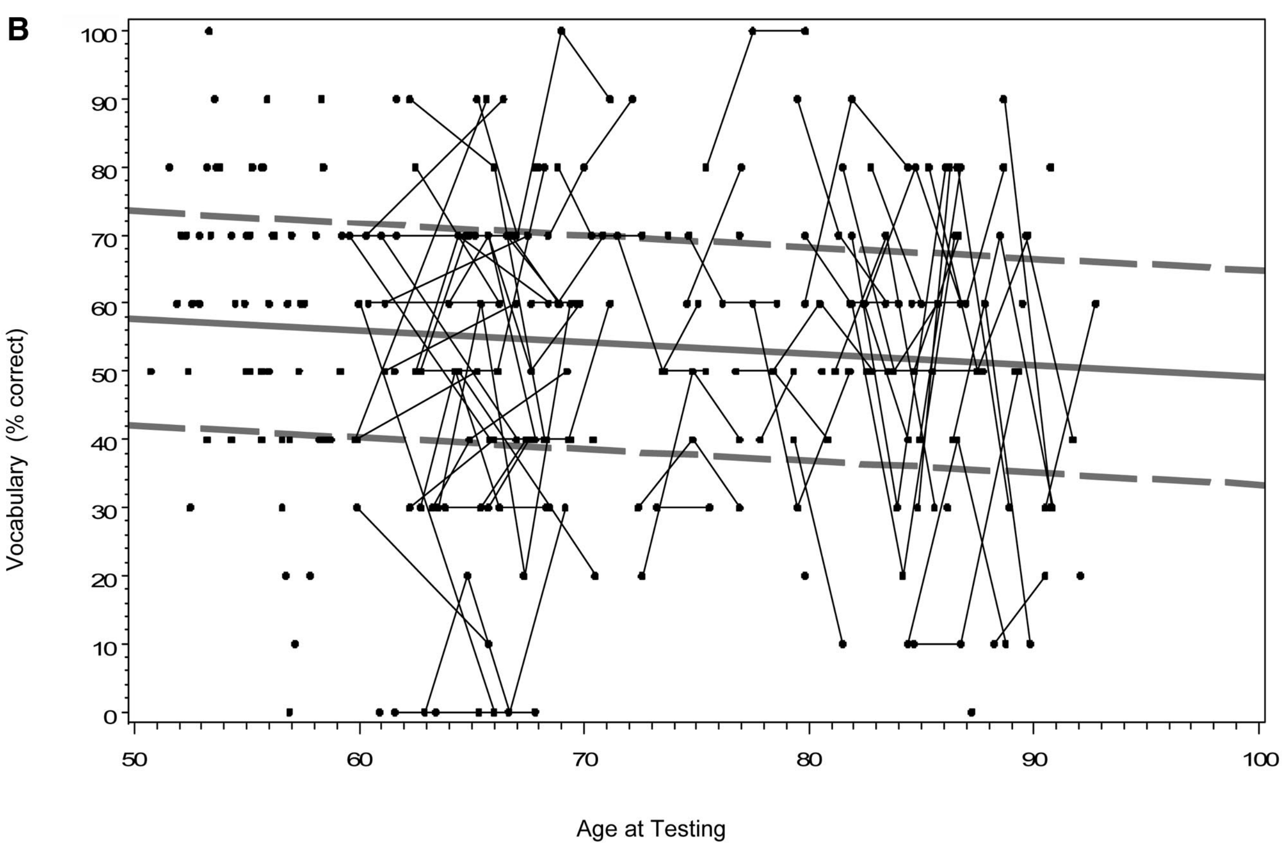


Figure 2. Longitudinal trajectories of cognitive scores as a function of age at testing for a random sample of Health and Retirement Study participants (1%; n = 174). Solid lines are mixed-model-based mean expectations, and dashed lines are plus or minus one standard deviation. A: Longitudinal trajectories of Immediate Recall scores. B: Longitudinal trajectories of Vocabulary scores.

Table 7
Maximum Likelihood Estimate Missing at Random for Between- and Within-Person Variability

Statistic	IR	DR	S7	BC	DA	NA	VO
Between-persons σ_b^2	198.5	282.5	102.2	749.4	54.4	148.8	228.8
Within-person σ_w^2	271.6	357.4	438.2	183.3	146.8	220.9	208.4
Intraclass correlation η^2 ($\eta^2 = \sigma_b^2/[\sigma_b^2 + \sigma_w^2]$)	.488	.510	.222	.631	.229	.504	.509

Note. $N = 17{,}355$. Only one person was included per family, and respondent-level sampling were weights used. $N_{\text{between}} = 17{,}355$; $N_{\text{within}} = 52{,}141$. IR = Immediate Recall; DR = Delayed Recall; S7 = Serial 7s; BC = Backward Counting; DA = Dates; NA = Names; VO = Vocabulary.

ships between the HRS cognitive factors and the demographic variables,

3. latent multilevel models—restricted or confirmatory factor analysis techniques to evaluate the longitudinal factorial invariance of the HRS common factors, and

4. latent curve models—mixed-effect modeling to define the growth–decline functions that characterize all longitudinal data as well as time-invariant and time-varying covariates.

Decisions about the accuracy, direction, and size of effects were made on a statistical basis. Of course, traditional criteria for statistical significance of fit and effects are virtually guaranteed with a sample size of more than 17,000. In this study, we routinely present overall chi-squares but rely on the root-mean-square error of approximation test statistic for the assessment of good fit ($\varepsilon_a < .05$; Browne & Cudeck, 1993). We also present individual Z values (parameters/standard errors) and use the $\alpha = .00001$ significance test level (e.g., $Z > 4.3$) as a screening criterion for further discussion.

Modeling With Incomplete Data

The previous HRS analyses (e.g., Rodgers et al., 2003) described the use of a multiple imputation procedure for handling incomplete data (using IVEware). This approach is reasonable for data that are missing within or over time, and it can be used when the incompleteness is due to attrition and other factors and the data are considered missing at random (MAR; after Little & Rubin, 1987). The same assumptions underlie analyses based on any latent variable structural equation model that includes all the data—not simply the complete cases (e.g., Horn & McArdle, 1980; McArdle, 1994). We expect nonrandom attrition, but our goal is to include all the longitudinal and cross-sectional data to provide the best estimate of the parameters of change as if everyone had continued to participate (Diggle, Liang, & Zeger, 1994; Little, 1995; McArdle & Anderson, 1990; McArdle & Bell, 2000; McArdle & Hamagami, 1991; McArdle, Prescott, Hamagami, & Horn, 1998). In computational terms, the available information for any participant on any data point (i.e., any variable at any occasion) is used to build up maximum likelihood estimates that optimize the model parameters with respect to any available data.

These incomplete data techniques are available in many current computer programs, and we use both Mplus 4.0 (Muthén & Muthén, 2006; see Ferrer, Hamagami, & McArdle, 2004) and SAS PROC MIXED (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006; Verbeke, Molenberghs, Krickeberg, & Fienberg, 2000). The Mplus program is advantageous because it also allows us to deal with (a) survey sampling weights, (b) categorical measurement models, (c) multilevel models, and (d) a random-slopes approach to latent curve models. We assess the goodness of fit of each model using classical statistical principles about the model likelihood (f_{MLE}) and change in fit (chi-square). In most models to follow, we use the MAR assumption to deal with incomplete longitudinal records, and we discuss these assumptions later (e.g., Cnaan, Laird, & Slasor, 1997; Little, 1995).

Table 8
Maximum Likelihood Estimate Missing at Random for Between-Persons and Within-Person Correlations

Test	1	2	3	4	5	6	7
1. IR	—	.605	.062	.102	.146	.133	.089
2. DR	.958	—	.052	.108	.138	.124	.060
3. S7	.462	.432	—	.050	.098	.065	.023
4. BC	.570	.543	.573	—	.118	.103	.022
5. DA	.657	.663	.589	.557	—	.208	.059
6. NA	.611	.600	.546	.615	.636	—	.066
7. VO	.624	.601	.494	.671	.459	.697	—

Note. $N = 17{,}355$. Only one person was included per family, and respondent-level sampling weights were used. Between-persons correlations are below the diagonal, and within-person correlations are above. IR = Immediate Recall; DR = Delayed Recall; S7 = Serial 7s; BC = Backward Counting; DA = Dates; NA = Names; VO = Vocabulary.

Results

Latent Common Factor Modeling

The first set of results is based on an exploratory common factor analysis of the raw data used to form the correlation matrix of Table 6. This is an unrestricted analysis because we used minimal identification constraints and rotated the factors using an oblique (PROMAX) solution (see McArdle, 2007; McDonald, 1999). These are standard exploratory factor analysis techniques, except that we accounted for both incomplete data (by MAR, with 36 patterns of incomplete data) and survey sampling weights (see Asparouhov, 2005; Pfeffermann, Skinner, Holmes, Goldstein, & Rabash, 1998; Stapleton, 2002). The results of a sequence of factor analysis models are presented in Table 9.

The first set of columns (Model 1) shows the result for a one-factor model, with dominant loadings on the Immediate Recall ($\lambda = .88$) and Delayed Recall ($\lambda = .86$) subscales, but with significant loadings on all seven measures. However, the goodness of fit, $\chi^2(14) = 3{,}991$ ($\varepsilon_a = .128$), was not adequate (i.e., $p\{\varepsilon_a <$

Table 9
Results From Unrestricted/Exploratory Common Factor Analyses of the Seven Health and Retirement Study Cognitive Measurements at First Testing Occasion With Continuous Variables and Sample Weights

	Model 1: One common factor $\chi^2(14) = 3{,}991$ ($\varepsilon_a = .128$)		Model 2: Two common factors $\chi^2(8) = 258$ ($\varepsilon_a = .042$)			Model 3: Three common factors $\chi^2(3) = 24$ ($\varepsilon_a = .020$)			
Measure	λ_1	ψ^2	λ_1	λ_2	ψ^2	λ_1	λ_2	λ_3	ψ^2
IR	*.88*	.23	*.79*	.13	.24	*.70*	.10	.11	.32
DR	*.86*	.27	*.85*	.06	.21	*.94*	.04	−.01	.09
S7	.46	.79	.04	*.59*	.63	.04	.30	.34	.64
BC	.24	.94	−.04	.38	.87	−.05	.38	.04	.85
DA	.35	.87	.06	.40	.81	.03	*.61*	−*.13*	.70
NA	.45	.80	−.03	*.67*	.58	−.02	.42	.30	.59
VO	.45	.80	.04	*.58*	.63	−.00	.00	*.72*	.48
Fac ρ			.58			.51	.65	.52	

Note. $N = 17{,}355$. Results are from the Mplus 4.0 EFA MISSING option with frequency weights and maximum likelihood with robust standard errors followed by Promax rotation. Italics are used to isolate the significant loadings. IR = Immediate Recall; DR = Delayed Recall; S7 = Serial 7s; BC = Backward Counting; DA = Dates; NA = Names; VO = Vocabulary; Fac ρ = factor intercorrelation.

.05} > .0001). The second set of columns (Model 2) shows the result for a two-factor model, and here the loadings were all relatively strong, and the goodness of fit, $\chi^2(8) = 258$ ($\varepsilon_a = .042$), was very good (i.e., $p\{\varepsilon_a < .05\} > .80$). The first factor now had significant loadings only on the Immediate Recall ($\lambda = .79$) and Delayed Recall ($\lambda = .85$) subscales, the second factor had high loadings on Names ($\lambda = .67$) and Serial 7s ($\lambda = .56$), and these two factors were positively correlated ($\rho_{12} = .58$). The first factor was the most obvious one, and we labeled it HRS Episodic Memory. The second factor was a bit more difficult to interpret, but the common features all represented very simple mental processing capacity and alertness, so we labeled it HRS Mental Status (after Herzog & Wallace, 1997). The two-factor model fitted very well, but the three-factor model was informative. The third set of columns of Table 9 shows a solution in which the fit was nearly perfect ($\varepsilon_a = .020$). In this model, the first two factors were basically the same as before, but the third factor was dominated by Vocabulary ($\lambda = 0.72$). If we isolated the Vocabulary scores, the second factor loaded only on the four original Mental Status variables.

The previous factor analytic results might be confounded by the fact that we were mixing three normally distributed continuous variables (Immediate Recall, Delayed Recall, and Vocabulary) with four categorical variables with substantial skewness (Backward Counting, Names, Dates, and even Serial 7s). To investigate any biases due to these standard assumptions, we next replicated the earlier sequence of factor analytic models using new methods that allowed a mixture of continuous and categorical variables in the same model (after Muthén & Muthén, 2006). This alternative approach to estimation (weighted least squares with minimum variance) started with the calculation of thresholds ($k - 1$) for each of the categorical variables, followed by a reestimation of the correlations (of Table 6), followed by the same factor model extraction and rotation. The results are presented in Table 10. As expected, the use of thresholds yielded higher positive correlations for relationships with the categorical variables and better model fit in every case. However, the results for the factor patterns were virtually identical to the previous results from Table 9. That is, the one-factor model was a poor fit, the two-factor model was very good and separated Episodic Memory from Mental Status variables, and the three-factor model was a perfect fit, suggesting the useful isolation of Vocabulary. This second approach, although not necessary in the overall sample, might be very useful in subsets of this population with less variation on some of the categorical variables, so we explore this again at various points in further analyses.

Latent Factor Path Modeling

The next set of analyses added a restricted set of factor loadings (e.g., as in confirmatory factor analysis) combined with selected sets of latent variable regressions (e.g., McArdle & Prescott, 1992; Park et al., 2002; Tulsky & Price, 2003). That is, we extended the previous common factor model with three continuous and four categorical variables (Table 10, Model 3) by adding (a) fixed zero loadings and (b) prediction of the common factor scores from key demographic variables. Although these were not strictly confirmatory models (i.e., with a priori hypotheses), the restricted set of loadings allowed us to examine other model features with additional precision.

The results of Table 11 represent a single latent variable path model that fitted fairly well ($\varepsilon_a = .030$). This model included the over-restricted factor pattern, in which the first factor was only the Episodic Memory variables (Immediate Recall and Delayed Recall), the second factor was only the four Mental Status variables (Serial 7s, Backward Counting, Dates, and Names), and the third factor was only Vocabulary. This measurement part of the model fitted the seven variable correlations nearly perfectly, with strong positive loadings and reasonably small uniquenesses. Of course, this is not surprising, as we based this pattern on the prior results of Table 10. The equations for the influences of the predictors of these two factor scores and Vocabulary are in the second set of columns in Table 11. Here we included age, education, gender, cohort, whether the participant was coupled, and mode of testing (i.e., telephone or face to face), with coding previously defined (Table 1). Raw estimates, standardized estimates, and Z values are included.

Table 10
Results From Exploratory/Unrestricted Common Factor Analysis With Three Continuous and Four Categorical Measures

Measure	Model 1: One common factor $\chi^2(12) = 1,861$ ($\varepsilon_a = .094$)		Model 2: Two common factors $\chi^2(7) = 171$ ($\varepsilon_a = .037$)			Model 3: Three common factors $\chi^2(2) = 30$ ($\varepsilon_a = .028$)			
	λ_1	ψ^2	λ_1	λ_2	ψ^2	λ_1	λ_2	λ_3	ψ^2
IR	*.74*	.46	*.72*	.17	.31	*.81*	.08	.04	.24
DR	*.71*	.49	*.93*	.01	.12	*.86*	.06	−.02	.22
S7	*.61*	.63[a]	−.01	*.67*	.56[a]	.02	*.59*	.11	.57[a]
BC	*.57*	.67[a]	−.03	*.65*	.61[a]	−.05	*.73*	−.06	.55[a]
DA	*.52*	.73[a]	.10	.45	.73[a]	.08	*.57*	−.11	.67[a]
NA	*.66*	.56[a]	.00	*.73*	.47[a]	−.03	*.59*	.17	.50[a]
VO	*.53*	.72	.02	*.55*	.68	−.01	.07	*.82*	.27
Fac ρ			.59			.38	.52	.58	

Note. $N = 17,355$. Results are from the Mplus 4.0 EFA MISSING option with frequency weights based on weighted least squares with minimum variance followed by oblique Promax rotation. The exact change in fit required additional rescaling proportions. Italics are used to isolate the significant loadings. IR = Immediate Recall; DR = Delayed Recall; S7 = Serial 7s; BC = Backward Counting; DA = Dates; NA = Names; VO = Vocabulary, Fac ρ = factor intercorrelation.
[a] A uniqueness parameter (γ) based on estimation of a set of thresholds for S7, BC, DA, and NA.

The results for the Memory factor include independent effects, which were negative for age (−9.9 decline per decade) and cohort (−3.7 decline per decade) and positive for education (1.8 per year), gender (5.5 higher for women), mode of testing (3.4 better on the telephone), and whether the participant was coupled (2.0 higher for a person in a couple). The impacts of education and age were the strongest effects, and the raw coefficients could be translated into the number of words lost or gained over specific years (i.e., $B_0 = 55.7$ at age 65 implies about 5.5 words remembered at age 65, $B_a = -9.9$ means a loss of a full word every decade, and $B_e = 1.8$ implies that 5 years of education add up to about one word recalled). The explained variance for the Memory factor was $R^2 = .35$.

The results for the Mental Status factor were not the same, and the raw weights were not in percentage units because of the rescaling of the common factor in categorical estimation, so we focus on standardized effects. We found significant effects only for education ($\beta = .56$), couple status ($\beta = .11$), and gender ($\beta = -.08$). The Vocabulary variance was strongly related to education ($B = 3.0$ per year), with smaller negative influences of age ($B = -3.9$ per decade) and cohort ($B = -3.5$ per decade). The removal of three coefficients for each variable produced a simpler model ($df = 3$) but led to significant misfit: education, $\chi^2(3) = 2,567$; age, $\chi^2(3) = 567$; gender, $\chi^2(3) = 556$; cohort, $\chi^2(3) = 520$; couple status, $\chi^2(3) = 520$; and telephone, $\chi^2(3) = 520$. The explained variance was $R^2 = .45$ for the Mental Status factor and $R^2 = .22$ for the observed Vocabulary factor. The largest effects (where $Z > 5$) are highlighted in the latent variable path diagram of Figure 3.

We considered the possibility of more complex age interactions in this latent path, and Table 12 is a list of the additional results. We started with the same path model but added inter-

Table 11
Restricted/Confirmatory Latent Path Results With Three Continuous and Four Categorical Measures, Two Common Factors, and Six Main Effects

	Model 1: Restricted factor loadings				Model 2: Latent variable regression weights			
Measure	Memory factor	Status factor	VO score	Unique variance	Latent equation	Memory factor	Status factor	VO score
IR	1.00/.93[a]	0.00	0	46/.13 (16)	Constant at 65	55.7/0.0 (329)	0.00/0.0[a]	56.9/0.0 (178)
DR	1.04/.83 (66)	0.00	0	150/.31 (43)	Age	−9.9/−.65 (20)	0.03/.05 (<1)	−3.9/−.22 (5)
S7	0.00	1.00/.67[a]	0	.67[b]	Educ	1.8/.33 (41)	0.13/.56 (46)	3.0/.47 (60)
BC	0.00	0.82/.57 (27)	0	.79[b]	Gender	5.5/.52 (19)	−0.12/−.08 (7)	1.5/.04 (4)
DA	0.00	0.67/.48 (31)	0	.86[b]	Cohort	−3.7/−.28 (8)	0.11/.19 (<4)	−3.5/−.22 (5)
NA	0.00	1.10/.73 (45)	0	.62[b]	Coupled	2.0/.06 (7)	0.16/.11 (10)	.06/.00 (< 1)
VO	0.00	0.00	1[a]	333/.78 (33)	Tele-FTF	3.4/.10 (13)	0.12/.08 (7)	−1.3/−.03 (3)
					Latent variable R^2	.347	.448	.220

Note. $N = 17,355$. Estimates are raw values from Mplus 4.0 with MEANSTRUCTURE MISSING model and weighted least squares with minimum variance missing at random estimates. Standardized values are listed after the slash, and Z values are in parentheses. Fit $\chi^2(24) = 405$ ($\varepsilon_a = .030$). IR = Immediate Recall; DR = Delayed Recall; S7 = Serial 7s; BC = Backward Counting; DA = Dates; NA = Names; VO = Vocabulary; Educ = education; Tele = telephone; FTF = face to face.
[a] Fixed nonzero value. [b] Uniqueness parameter from a threshold.

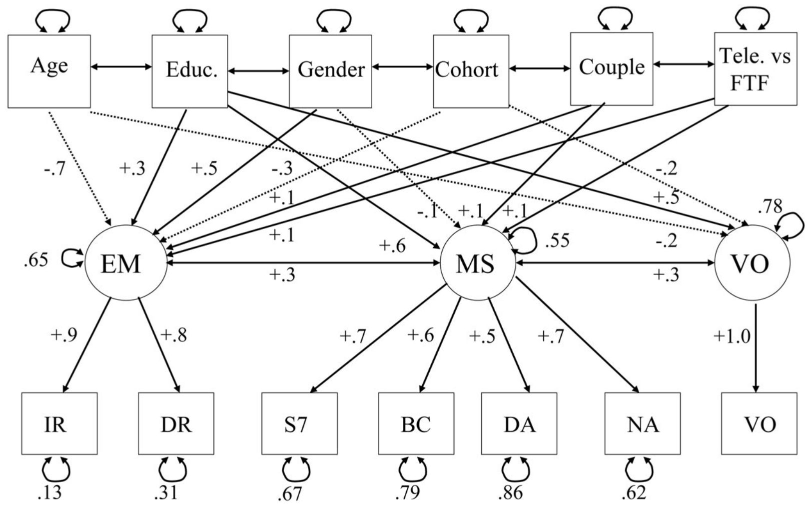


Figure 3. A cross-sectional latent path model for the seven Health and Retirement Study cognitive measures. This model includes three cognitive factors (Episodic Memory [EM], Mental Status [MS], and Vocabulary [VO]) and six second-order demographic predictors. On the basis of all cross-sectional data, $N = 17{,}355$. Only standardized parameters with Z values greater than 5 are included. The demographic variables are all assumed to be correlated, but these correlations are not included in the figure. Educ. = education; Tele. = telephone; FTF = face to face; IR = Immediate Recall; DR = Delayed Recall; S7 = Serial 7s; BC = Backward Counting; DA = Dates; NA = Names.

actions of all six predictors with age, and this model also fitted very well ($\varepsilon_a = .027$). The resulting factor loadings were virtually identical (to Table 11), so they are not listed again, and the latent variable explained variances did not change much either. However, the removal of the interaction coefficients for the latent variable outcomes led to significant misfits, so some interactions described above were needed for a complete explanation of the factor scores. The key results for the Memory factor included an Age × Age quadratic effect (age, $B = -6.6$; Age × Age, $B = 11.2$), and the previous cohort effects interacted with age (cohort, $B = 0.6$; Age × Cohort, $B = 6.5$), but there were no other notable effects (i.e., not Age × Education or Age × Gender). The Mental Status factor equation changed very little, with only a small Cohort × Age interaction. The equation for the Vocabulary score required no additional age interactions.

Table 12
Latent Path Results With Six Main Effects and Six Age Interactions

Latent equation	Memory factor	Status factor	Vocabulary score
	Model 1: Latent variable regression weights for main effects		
Constant at 65	56.3/0.0 (263)	0.00/0.0[a]	56.6/0.0 (135)
Age	−6.6/−.44 (13)	0.11/.16 (<3)	−3.9/−.22 (4)
Educ	1.8/.33 (41)	0.13/.57 (46)	3.0/.47 (61)
Gender	5.4/.15 (19)	−0.13/−.07 (8)	1.6/.04 (4)
Cohort	0.6/.05 (< 1)	0.18/.31 (6)	−3.1/−.20 (2)
Coupled	1.5/.04 (5)	0.15/.10 (9)	.20/0.0 (<1)
Tele-FTF	2.3/.07 (8)	0.06/.04 (3)	−1.2/−.03 (3)
	Model 2: Latent variable regression for age interactions		
Age × Age	11.2/.43 (13)	0.18/.16 (3)	−0.5/−.02 (< 1)
Educ × Age	−0.1/−.02 (< 3)	0.00/−.02 (<2)	0.0/−.01 (< 1)
Gender × Age	−.20/−.01 (< 1)	−0.03/−.02 (2)	0.2/0.00 (< 1)
Cohort × Age	6.5/.52 (16)	0.15/.28 (6)	0.9/.06 (< 1)
Coupled × Age	−1.0/−.03 (4)	0.15/.28 (6)	0.1/0.00 (< 1)
Tele-FTF × Age	0.4/.01 (< 2)	0.03/.02 (<2)	0.0/0.00 (< 1)
Latent variable R^2	.365	.458	.233

Note. $N = 17{,}355$. Fit $\chi^2(33) = 451$ ($\varepsilon_a = .027$). Standardized values are listed after the slash, and z values are in parentheses. Educ = education; Tele = telephone; FTF = face to face.
[a] Fixed nonzero value.

Latent Multilevel Longitudinal Invariance Modeling

One of the key assumptions in studies of longitudinal changes in score levels is that the same constructs are indicated by the same scores over time. Thus, as a prelude to further analyses of longitudinal changes, we examined the longitudinal invariance of the factorial structure just described in the first two sections of the Results. The principles of factorial invariance we use here are well known and have been discussed by many others (e.g., Horn and McArdle, 1992; Meredith & Horn, 2001). However, in this case we focused our studies on the longitudinal scores over time and age and under various procedures of measurement. Using a classic longitudinal approach, we could analyze the seven cognitive variables measured over all nine times and base this model fitting on a 63 $\times$ 63 covariance matrix (with 63 means). This multiple-wave approach would allow the inclusion of incomplete data patterns, provide a standard test of the invariance of the factor loadings over wave, and allow the common and unique factor variances (and means) to change over wave (see Meredith & Horn, 2001). However, as soon as we considered age as a key categorization of people, we created a much larger multiple-groups problem (i.e., longitudinal matrices of scores for each age group; see Horn & McArdle, 1980), so we needed a more concise computational approach.

Prior work in longitudinal factor modeling has shown how the principles of invariance over time can be applied to sums and differences (e.g., Hultsch, Hertzog, Dixon, & Small, 1998; McArdle & Nesselroade, 1994; Nesselroade & Cable, 1974). The key result of this prior work is that two occasion scores can be transformed into sums and differences without any loss of information, so the model of invariant factors can be evaluated the same way with original or transformed scores. We extended these principles to fit factor patterns to the between-persons and within-person matrices calculated over multiple occasions (as in Tables 7 and 8). In this approach, we examined the key questions of factorial invariance over time in a global way by examining the degree to which the same factor pattern could be used to fit the between-persons covariance matrix and the within-person covariance matrix (for further details, see McArdle, 2007). This global approach is used to indicate the overall invariance properties of each variable and each common factor. Model fitting follows the logic of analyzing persons in families (Nagoshi & Johnson, 1987) or in classrooms (Muthén, 1991) and uses recent techniques for multilevel invariance (see Lubke, Dolan, Kelderman, & Mellenburgh, 2003). We recognize that lack of fit will not (a) tell us about which particular occasions are different from one another, (b) separate out specific factor covariance over time, or (c) provide information about possible subgroups of persons (e.g., men and women) who have different factor patterns.

The specific approach we used was to fit all longitudinal data on all persons (D = 69,496) by first creating estimates of the 7 $\times$ 7 between- and within-person covariance matrices presented earlier (e.g., Table 7 and 8). Using these estimates as a saturated model, we then fitted and compared models of (a) configural invariance (same nonzeros in the between- and within-person loadings) versus (b) the same pattern with metric invariance (i.e., exact values in the between- and within-person loadings). The models fitted here included (a) a one-factor model for all seven cognitive measures, (b) a two-factor restricted model for all seven cognitive measures, and (c) a two-factor restricted model for six cognitive measures (i.e., without Vocabulary). In all models fitted here, the common and unique factor variances were allowed to vary for the between-persons and within-person matrices. These alternative models were easily fitted via a multilevel modeling approach (Mplus 4.0), with the individual as a cluster and sampling weights applied to all the data on the person. We did not estimate category thresholds because of the size of the computational problem. The overall estimates and goodness-of-fit statistics are described in Table 13 and these results showed a remarkable degree of consistency across time and among people. The comparison of models for the one-factor solution (a vs. b) yielded η^2 = .65 for the one metrically invariant factor but a relatively large misfit, normal $\chi^2(6)$ = 2,616. The comparison of models for the two-factor solution (c vs. d) yielded η^2s = .61 and .81 for the metrically invariant factors and a much smaller misfit, $\chi^2(5)$ = 651. The comparison of models for the six-variable solution (e vs. f) again yielded η^2s = .61 and .82 for the metrically invariant factors, and the misfit of the metric invariant model was the best found, $\chi^2(4)$ = 399. These intraclass estimates for the common factors are informative because they indicate a high degree of stability at the common factor level, especially in the second Mental Status factor.

Table 13
Goodness of Fit for Longitudinal Multifactor Multilevel Models for All Seven Cognitive Scores Over All Longitudinal Occasions

Goodness of fit	Model A: One-factor configural	Model B: One-factor metric	Model C: Restricted two-factor configural	Model D: Restricted two-factor metric	Model E: Restricted two-factor configural—VO	Model F: Restricted two-factor metric—VO
Variance between persons σ_b^2	185.3	209.1	196.9/448.5	198.7/398.3	195.5/428.0	197.1/350.0
Variance within person σ_w^2	139.2	113.7	127.0/29.4	125.8/64.9	129.0/30.9	125.4/78.1
Intraclass corr. η^2		.648		.612/.810		.611/.818
χ^2/df	4,617/28	7,629/34	577/26	1,347/31	237/16	699/20
MLR scaling	1.454	1.535	1.420	1.483	1.442	1.512
ε_a	.049	.057	.017	.025	.014	.022
ML χ^2/df		4,990/6		1,178/5		715/4
Normal χ^2/df		2,616/6		651/5		399/4

Note. See the matrices in Tables 7 and 8. Models were fitted with the Mplus 4.0 TWOLEVEL MISSING option with N = 17,355 and D = 69,496 via MLR/missing at random estimation. The Mplus 4.0 MLR correction factor was used to adjust the chi-square difference. Attempts to use categorical thresholds with weighted least squares with minimum variance estimation were limited by the large sample size. VO = Vocabulary; MLR = maximum likelihood with robust standard errors; ML = maximum likelihood.

Age-Based Latent Curve Models of Longitudinal Age Changes

The final set of analyses presented here is based on statistical research termed *multi-level, random coefficients,* or *latent curve models* (Bryk & Raudenbush, 1992; Jones & Meredith, 2000; McArdle & Hamagami, 1996, 2001; McArdle & Nesselroade, 2003; Snijders & Bosker, 1999; Verbeke et al., 2000). These statistical procedures are used to fit a model directly to the observed scores on the basis of various kinds of mathematical forms of curve or decline. The most common form of a latent curve model is based on a statistical model for a trajectory over time. This model typically includes latent scores representing (a) a latent intercept or initial level, (b) latent slopes representing the change over time, and (c) unobserved unique scores within each time. In most variations of this latent curve model, we added (d) a set of group weights or basis coefficients describing the timing of the observations (e.g., $A[t] = 0$ or 1 or $A[t] = \text{Age}[t]$). The error terms are assumed to be normally distributed with mean zero and variance (σ_u^2) and are presumably uncorrelated with all other components (as in McArdle & Hamagami, 2001; McArdle & Nesselroade, 2003). As usual, the initial level and slopes are often assumed to be latent random variables with fixed means (μ_o, μ_1) but random variances (σ_0^2, σ_1^2) and correlations (ρ_{01}).

In this latent curve model, the timing of changes was initially based on the age at testing (as in Figures 2A and 2B). As mentioned earlier, this approach differs from the more typical HRS over-time regression analyses (e.g., Rodgers et al., 2003). In this model, the age at any particular time was not considered as a random variable—instead, the changes over age in the scores were considered as the random intercepts and slopes, and the average group changes were described by the fixed basis parameters ($A[t] = \{\text{Age}[t] - 65\}/10$). The use of age as a basis represents the fitting of a linear curve model through the longitudinal trajectory data presented in Figures 2A and 2B, once again considering both the incomplete and the complete trajectories of data points. Of course, for this age-basis model to be viable, we need to assume that the untestable MAR assumptions apply to this age dimension. The score measured on a person at each age gives us some indication of his or her likely scores at the ages not measured, and persons measured at specific ages represent the age-based scores for anyone. MAR in this incomplete data design problem includes accounting for age changes in different cohorts (e.g., McArdle & Anderson, 1990; McArdle & Bell, 2000; Meredith & Tisak, 1990; Miyazaki & Raudenbush, 2000).

All models presented up to now were fitted with sample survey weights and all longitudinal data on all persons (via Mplus 4.0 MLR-MAR; see Asparouhov, 2005). This approach was reasonably easy for the normally distributed variables but computationally difficult for the skewed categorical outcomes. Therefore, to simplify this analysis, we now focus on the episodic memory data in the HRS; the initial latent curve results are presented in Table 14.

The first age-based model was fitted to Immediate Recall scores, and this linear age-change model fit was a dramatic improvement over a no-change or intercept-only model, corrected $\chi^2(3) = 7{,}966$. This model was fitted with chronological age scaled so the results included the three time-constant parameters at age 65 ($\mu_0 = 47.7$; $\sigma_0^2 = 119$; Immediate Recall, $\psi^2 = 183$) plus three time-dependent slope parameters indicating decreases in means ($\mu_1 = -8.73$) over every decade of age, with increases in variances and covariances ($\sigma_1^2 = 21$; $\sigma_{01} = 6$) over each decade. In terms of Figure 2A, this result represents a relatively large average impact compared with the apparent variation in the individual decline curves (i.e., -8.73 points lost per decade), with some systematic increasing individual differences. These changes in individual differences came from both variances and covariances, so for ease of interpretation we also added four rows representing the orthogonal decomposition of each term (e.g., the total slope variance was 26.9, or 8% of the total). The bold lines in Figure 2A are the expected values of age-based declines from this mixed-effects model of Immediate Recall scores, and the dashed lines represent plus or minus one standard deviation around the expected latent curve. Similarly, the lines in Figure 2B are the expected values of far smaller age-based declines from the latent curve estimated from the available Vocabulary scores ($\mu_0 = 55.0$; $\mu_1 = -1.7$).

The second column of results gives the same parameters and decomposition for the Delayed Recall scores on the same people over the same ages, and, although the improvement in fit was also large, corrected $\chi^2(3) = 7{,}071$, the numerical results were not exactly the same. Although the mean decline was approximately the same ($\mu_1 = -8.47$), the expected score at age 65 was much lower ($\mu_0 = 31.7$), indicating that Delayed Recall was a generally harder task. More important, the decomposition of the latent variable variance showed no systematic linear age variance, and the improvement in fit was largely based on the average decline. These results may change with a different model for the age trajectory, and we explore this in more complex models to follow.

The third column gives results attained when we fitted a latent curve model to a common factor formed from the two previous indicators. This model (a) used all longitudinal data from both Immediate Recall and Delayed Recall, (b) defined an invariant measurement model for Immediate Recall and Delayed Recall, and (c) allowed all age changes to be captured by the single common factor of Episodic Memory (for details, see alternative models in McArdle, 2007; McArdle et al., 2002). The results for this model were fairly good by all previous criteria—the measurement model fit was balanced across both variables (i.e., Immediate Recall, $\lambda = 1.00$; Delayed Recall, $\lambda = 1.15$) and reflected a good fit to the raw data, the age changes in the factor score represented a substantial improvement over the no-change baseline, and there was a small but potentially important age-change variation (1%). The good fit of this measurement + curve model (Model C in Table 14) added developmental evidence about the construct validity of this factor.

To deal with complex numerical problems of the simultaneous estimation of latent factor and slope parameters, we approached the same problem from an alternative perspective—we formed a simple composite from the unweighted average of the Immediate Recall and Delayed Recall scores. The results of the age-based latent growth model fitted to these data are presented in the last column of Table 14. This model was very different from its no-change baseline, $\chi^2(3) = 8{,}168$. The mean (fixed) parameters for these results looked very much like all previous results, with the age decline of -8.77 per decade (-1 word), but the other age-related variation was negligible. This implies that the Episodic Memory composite score can be used to carry much of the information in the common factor of Memory but also that these are not exactly the same.

Table 14
Numerical Results for Longitudinal Latent Curve Models With a Linear Age Basis for Four Alternative Forms of Health and Retirement Study Episodic Memory Scores

Parameters and fits	Model A: IR	Model B: DR	Model C: Common factor of memory	Model D: Composite score of memory
Fixed parameter				
Age 65 intercept μ_0	47.68 (342)	31.73 (170)	56.33 (460)	50.26 (388)
Age slope/decade μ_1	−8.73 (87)	−8.47 (77)	−8.29 (89)	−8.77 (89)
Factor loading IR[t] λ_{IR}	1.0[a]	0.0[a]	1.0[a]	1.0[a]
Factor loading DR[t] λ_{DR}	0.0[a]	1.0[a]	1.146 (996)	1.0[a]
Unique intercept μ_{DR}			−21.04 (69)	
Random parameter				
Intercept at 65 variance $\sigma_o{}^2$	119.9 (43)	161.4 (40)	132.0 (54)	153.3 (55)
Age slope/decade $\sigma_1{}^2$	20.9 (12)	4.9 (9)	13.5 (10)	12.6 (9)
Covariance intercept and slope	6.0 (< 4)	−4.8 (< 4)	−11.7 (10)	−13.3 (10)
Unique variance $\psi_u{}^2$			103.1 (74)	171.1 (110)
Unique variance $\psi_{IR}{}^2$	182.8 (109)		83.7 (69)	
Unique variance $\psi_{DR}{}^2$		225.8 (107)	129.0 (63)	
Combined random parameters (% variance)				
Intercept at 65 variance $\sigma_o{}^2$	125.9 (38)	156.5 (41)	120.3 (53)	140.0 (45)
Age slope/decade $\sigma_1{}^2$	26.9 (8)	0.1 (0)	1.8 (1)	−0.7 (0)
Unique variance $\sigma_u{}^2$	182.8 (54)	225.8 (59)	103.1 (46)	171.1 (55)
Total variance $\sigma_y{}^2$	335.6 (100)	382.5 (100)	225.2 (100)	310.4 (100)
Goodness of fit				
No change baseline—*f*/MLR correction	264,876/1.276	270,768/1.353	514,673/1.633	263,751/1.295
Linear age model/MLR correction	260,145/1.232	267,639/1.119	510,056/1.491	259,238/1.200
Improvement χ^2/parameters	9,462/6	6,258/6	9,234/9	9,026/6
Corrected χ^2/df	7,966/3	7,071/3	6,845/3	8,168/3

Note. Models were fitted with Mplus 4.0 RANDOM MISSING options with $N = 17{,}355$ and $D = 69{,}496$ via MLR/missing at random estimation. Values in parentheses are *Z* values to indicate parameters that were significant at various alpha test levels. Age was individually coded for each occasion as {Age[t] − 65}/10, so the intercept was at Age = 65, and the Age slope represented 1 decade. IR = Immediate Recall; DR = Delayed Recall; MLR = maximum likelihood with robust standard errors.
[a] Parameter is fixed at the current value.

Extended Latent Curve Models of Longitudinal Age Changes

In further analyses based on the same logic, we again faced computational problems using current computer software (i.e., Mplus 4.0). That is, fitting large-scale longitudinal models with sample weights and incomplete data is computationally demanding, and more complex latent curve models were not easy to fit the same way. To explore these models, we used standard mixed-model software without adjustment by sampling weights (i.e., SAS 9.1, PROC MIXED; Littell et al., 2006). A selected set of new results for Episodic Memory is presented in Table 15.

The first column gives new estimates for the linear age-based model with incomplete data (with maximum likelihood estimation MAR assumptions) for the HRS Memory Composite score. Similar to the previous Mplus estimates (Model D in Table 14), the linear age-change model provided (a) a dramatic improvement in fit over a no-change or intercept-only model and (b) significant time-constant parameters at age 65 plus time-dependent slope parameters indicating decreases in means over every decade of age, with increases in variances and covariances over each decade. These unweighted results yielded strong expectations about age-related declines in episodic memory.

The linear results naturally led to considerations of nonlinearity of the changes over age. The most typical way to do this is to add more complexity to the basis (as in McArdle & Nesselroade, 2003). The linear mixed model can be compared with the well-known use of a quadratic change model (Bryk & Raudenbush, 1992). Although the quadratic model offered a slight improvement in fit, $\chi^2(3) = 308$, this overall model was not well behaved for this overall sample. Unfortunately, as soon as we introduced these kinds of more advanced models into the current analyses of all HRS data, we experienced serious practical problems (e.g., negative variances and convergence problems in the quadratic; see Table 15, Model B). An alternative variation on nonlinear age relations is a two-part linear spline model fitted by first defining an age of turning or knot point (e.g., $\tau = 65$; cf. Cudeck & Klebe, 2002; Hall, Lipton, Sliwinski, & Stewart, 2000). These two-part models did not improve the fit at all—for example, two-part splines, $\chi^2(5) = 80$. Although our results do not suggest anything new, we do expect nonlinear age functions to be reasonable for some selections of HRS longitudinal data.

Table 15

Numerical Results for Health and Retirement Study Episodic Memory Composite Score in Alternative Longitudinal Latent Curve Models Based on Age, Retest, Time-Invariant Covariates, and Time-Varying Covariates

Parameters and fits	Estimates from alternative latent growth models: Model A: Linear age basis model	Model B: Quadratic age basis model	Model C: Linear age with retest basis	Model D: Adding time-invariant covariates	Model E: Adding time-varying covariates
Fixed parameter					
Age 65 intercept μ_0	50.60 (416)	50.89 (369)	49.60 (331)	51.77 (293)	51.91 (288)
Age slope/decade μ_1	−9.59 (104)	−9.17 (83)	−10.06 (104)	−14.00 (78)	−12.39 (63)
Age change in slope μ_2		−0.065 (5)			
Retest (t > 1) shift μ_r			1.79 (12)	3.10 (19)	2.02 (11)
Time-invariant educ					
Educ × Intercept at 65				2.06 (50)	2.04 (50)
Educ × Age Slope/Decade				−0.28 (11)	−0.28 (11)
Educ × Retest t > 1				−0.01 (< 1)	−0.02 (< 1)
Time-invariant gender					
Gender × Intercept at 65				4.78 (17)	5.22 (18)
Gender × Age Slope/Decade				−1.01 (6)	−0.81 (4)
Gender × Retest t > 1				1.31 (5)	1.08 (< 4)
Time-invariant cohort					
Cohort × Intercept at 65				−6.25 (33)	−4.99 (24)
Cohort × Age Slope/Decade				1.64 (20)	1.77 (21)
Cohort × Retest t > 1				−0.97 (7)	−1.24 (9)
Time-varying dyad					
Dyad × Intercept at *t*					1.61 (9)
Time-varying tele-FTF					
Tele-FTF × Intercept at *t*					0.24 (< 2)
Combined random parameters (% variance)					
Intercept at 65 variance σ_o^2	169.4 (46)	173.5 (43)	199.4 (46)	144.6 (38)	156.5 (40)
Age slope/decade σ_1^2	17.5 (5)	77.0 (19)	13.3 (3)	12.4 (3)	8.7 (2)
Age change in slope σ_2^2		−14.7 (−4)			
Retest variance σ_r^2			56.9 (13)	56.1 (15)	70.5 (40)
Unique variance σ_u^2	178.0 (49)	171.8 (42)	166.5 (38)	165.2 (44)	156.5 (18)
Total variance σ_y^2	364.9 (100)	407.5 (100)	456.0 (100)	436.0 (100)	392.2 (100)
Goodness of fit					
Improvement χ^2/parameters	11,065/6	11,373/9	11,567/9	16,727/18	41,921/20
Change χ^2/df	11,065/3	308/3	498/3	5,164/9	25,194/2

Note. Models were fitted with SAS 9.1 PROC MIXED without sampling weights but with N = 17,355 and D = 69,496 (observed mean = 50.13, variance = 437.4). The baseline model of equal means, deviations, and correlations yielded a fit of −2LL = 597,874, with σ_u^2 = 202.6, and intraclass correlation η^2 = .549. Values in parentheses are Z values to indicate parameters that were significant at various alpha test levels. Coding schemes follow Table 1: (a) Age was individually coded as {Age[t] − 65}/10, so the intercept was at age = 65 and the age slope represented one decade. (b) Retest was individually dummy coded so 0 = first testing and 1 = retesting. (c) Gender was half effect coded, with men = −0.5 and women = 0.5. (d) Education was coded as years −12. (e) Birth cohort was coded as cohort = (birth year − 1930)/10 to be approximately centered, so change represents a decade. (f) Dyad at testing was half effect coded for each wave, with not tested in dyad = −0.5 and tested in dyad = 0.5. (g) Telephone-face to face was half effect coded for each wave, with face-to-face testing = −0.5 and telephone testing = 0.5. Educ = education; Tele = telephone; FTF = face to face.

The latent curve model can also be expanded to account for multiple basis variables and multiple predictor variables (e.g., McArdle & Anderson, 1990). This leads to a simple way to consider practice effects in terms of components of change. Various representations of an age-based curve plus practice effects model have recently been discussed by other researchers, especially to eliminate confounds due to test practice or to model other time courses (e.g., Ferrer et al., 2005; Lövdén, Ghisletta, & Lindenberger, 2004; Rabbitt et al., 2004; Rabbitt, Diggle, Smith, Holland, & McInnes, 2001; Sliwinski, Hofer, & Hall, 2003). McArdle and Woodcock (1997) used a model with two basis coefficients describing the observations at different ages and occasions of retest. In particular, in the models to follow, we fixed A[t] = {Age [t] − 65}/10 and also fixed R[t] = {0 *if* t = 1, and 1 *if* t > 1}. This assumed that retest can be described as a function that has no impact at Time 1 and then has a constant impact thereafter (i.e., an initial retest effect that persists, after McArdle et al., 1998). In contrast to other treatments of retest effects in which patterns of retest basis (R[t]) were considered (e.g., Lövdén et al., 2004; Rabbitt et al., 2001, 2004), this retest model allows the estimation of a retest mean and variance (μ_r and σ_r^2) as well as correlations with other latent components.

The results of the addition of a retest basis to an age basis for the HRS Memory Composite are presented in the third column of Table 15 (Model C). The overall change in fit was reasonably large, $\chi^2(3) = 498$, and the resulting parameters seemed sensible—a larger age-based decline ($\mu_1 = -10.1$ per decade) combined with a positive retest effect ($\mu_r = 1.8$ after the first test). The overall variance components (accounting for both variance and covariance) showed that the age variance per decade ($\sigma_1^2 = 13.3$) was substantially less than the variance associated with the retest ($\sigma_r^2 = 56.9$). Of course, more complex alternative hypotheses about the nature of the retest basis can be fitted as well (as in McArdle & Woodcock, 1997). The only alternative we examined in this work was an increasing linear practice effect (R[t] = {0,1,2,3 . . . T}), and this alternative did not improve the fit substantially, $\chi^2(3) = 18$.

We extended the previous model with an age basis and a retest basis by adding three key demographic variables as time-invariant predictors—education, gender, and cohort. These variables were known to be correlated (see Table 2), so the results of separate demographic influences were likely to be confounded. To examine the independent impacts of these variables, we fitted a set of more complex mixed-effects models in which all variables and all their possible interactions were included. The numerical results in Model D of Table 15 show that the addition of these three variables as predictors of intercepts and slopes made a large difference in fit, $\chi^2(9) = 5{,}164$. The increases in educational level had a large positive impact at age 65 (2.0) and a small negative impact on the age slope (e.g., higher education with slightly more decline) but no impact on the retest. The women's significantly higher scores at age 65 (by 4.8, or half a word) declined at a slightly more rapid rate (−1.0) and had a slightly higher retest benefit (1.3). The cohort effect at age 65 was large and negative (−6.3 per decade) but was positive on the age slope (1.6 per decade) and negative on the retest (−1.0). This implies that significant differences between people born in different years were seen to be relatively large at age 65 but that, possibly more important, the age declines were significantly slower per decade for more recent birth cohorts. The overall reduction in the estimates of the latent variable variance was a key estimate of the explained variance due to these effects, and this reduction was relatively large for the intercept (from 199 to 145) but small for the age (13 to 12) and retest (57 to 56) slopes (see Xu, 2003).

The final model fitted here (Model E in Table 15) used an age basis and a retest basis, the three key demographic variables as time-invariant predictors, and two additional variables as time-varying covariates—whether the person was coupled at the time of testing and whether the test was administered by telephone or in face-to-face mode. The results showed that the addition of these two variables at each occasion as predictors of the outcome at the same occasion made a large difference in fit, $\chi^2(2) = 25{,}194$. The final results seemed to suggest that there was a significant positive impact on episodic memory for people in a current couple (1.6) but that there was no positive difference in memory scores for telephone versus face-to-face testing. The previous demographic effects above did not change direction, but some increased in size (age, $\beta = -12.9$), whereas others diminished (i.e., cohort, $\beta = -5.0$; retest, $\beta = 2.0$).

In a last effort to simplify this final model, we reevaluated the goodness of fit, leaving out each of the demographic predictors. The reduction in the overall likelihood showed the relative importance of each variable (education, cohort, and gender). The significance of these misfits leads us to conclude that all independent demographic effects were needed in the final model. The largest results from these models are presented in the mixed effects path diagram of Figure 4 (for diagram details, see McArdle & Hamagami, 1996).

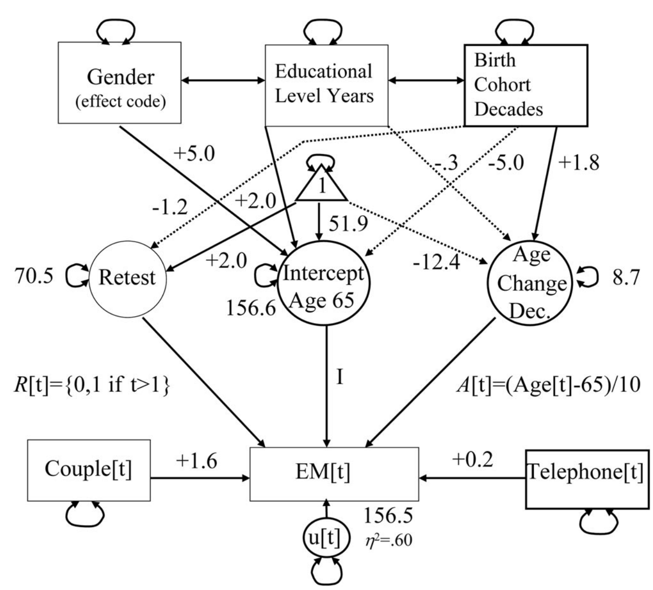


Figure 4. A longitudinal mixed-effects path diagram (after McArdle & Hamagami, 1996) for the Health and Retirement Study Episodic Memory (EM) composite scores. This model includes three latent components (intercept, age changes, retest), three demographic predictors (gender, education, cohort), and two time-dependent predictors (couple status, telephone). Results are based on all longitudinal data (D = 69,496). Values are raw maximum likelihood estimates with Z values greater than 5. Age Change Dec. = per decade age change; [t] = dependent on time.

Discussion

This research presents a set of new results on the longitudinal cognitive data from all available waves of HRS/AHEAD (1992–2004). The primary purpose of this research is to describe age trends of cognition among older adults in the HRS. We selected a representative national sample of more than 17,000 individuals on the basis of data from more than 69,000 interviews in the HRS. We used a set of latent variable models to organize this information in terms of both cross-sectional and longitudinal inferences about age and cognition and examined how additional variables explain individual differences in these constructs. These results are important for further HRS research and, more generally, for further research on age and cognition.

Although other researchers have used the HRS cognitive data by constructing a single composite score, the latent variable factor models presented in this article strongly suggest that a single factor of cognition was not measured in the HRS interviews. The factorial structure of the HRS cognitive measures was first investigated by Herzog and Wallace (1997) via a subset of the current data and classical methods. The new results we present are largely consistent with Herzog and Wallace (1997) and suggest at least one factor related to episodic memory (Immediate Word Recall and Delayed Word Recall) and a second factor related to the four other HRS tasks (Serial 7s, Backward Counting, Names, and Dates). The second factor was labeled Mental Status, and, although this is not a traditional common factor in multifactor theories of intelligence (e.g., Horn, 1991), it may be quite useful in identifying serious deficiencies that make it impossible to further measure a person on standard cognitive tests (e.g., Herzog & Wallace, 1997). Our inclusion of Vocabulary scores and the subsequent isolation of a Vocabulary variable into a third dimension make it likely that vocabulary is the only indicator of crystallized knowledge (after Horn, 1991) in the HRS. Perhaps more important, the common factor models we present point to the fact that the cognitive measures in the current HRS do not measure a number of other important factors in aging, such as Working Memory, General Speed, or Fluid Reasoning (e.g., Horn, 1988; Lachman & Spiro, 2002). It follows that measures of these and other important cognitive factors may need to be added to the HRS battery (e.g., McArdle, Rodgers, Fisher, Woodcock, & Horn, 2006).

The latent variable path model we present shows different prediction patterns among these common factor scores, and this adds additional external validity for their factorial separation (e.g., McArdle & Prescott, 1992; Park et al., 2002; Tulsky & Price, 2003). The most pervasive age and cohort declines were found for the Episodic Memory factor, and these declines were consistent with prior work in cognitive aging (e.g., Cerella & Hale, 1994; Hultsch et al., 1998; Park, 2000; Zacks & Hasher, 2006). The broad and representative sampling of people in the HRS makes this result more potent and allows us to examine other effects related to decline. In these path models, we found that, given the same cohort and the same age, there was less memory decline among women, those in a couple, and those who took the test over the telephone. In addition, there were several nontrivial age interactions. All of these effects require far more explanation than can be offered from these analyses, but they should be considered in future work.

The current HRS data collection also allowed us to reexamine the inferences about age differences from cross-sectional information in terms of inferences about age changes within a person from longitudinal data. Our first step was to examine the factorial invariance of the longitudinal measurement models in terms of multilevel factor analyses. These contemporary statistical models, combined with our large sample size, provided a powerful way to reject the hypothesis of metric invariance. However, the results we present here clearly suggest that, as long as we use two or three common factors instead of one common factor, we can reasonably fit a metric invariant model. These analyses suggest that the seven cognitive variables can be used to measure the same two or three common factors at each occasion (i.e., measurement invariance over wave of testing). As used in our work, this result provides a necessary psychometric platform for carrying out further longitudinal analyses. However, we recognize that there was much more between-persons information than within-person information, so this global approach was only capable of providing a first look at one limited view of factorial invariance over time. Additional invariance analyses over specific occasions, ages, and other groupings (e.g., gender, ethnicity, telephone vs. face-to-face interview) should follow and may uncover important qualitative differences in the sources and structure of these cognitive measurements.

Use of a latent curve model allowed us to organize age-related declines in terms of both group means and individual differences around these means. In the HRS, as in many other studies, the participants were not all measured at the same initial ages, so wave of measurement was not equivalent to age of measurement. This longitudinal modeling approach also allowed a statistical separation of the impact of age differences for participants who grew up in different birth cohorts. Although they are not emphasized in this article, differences in results and inferences can occur because different model-based organizations of the longitudinal data do not always yield the same results (even when age is included as a covariate; e.g., Sliwinski & Buschke, 1999; cf. McArdle et al., 2002). Of course, this means that age is not always the most important dimension of the changes. In research in which the participants are all measured at the same initial ages or after a specific incident (e.g., recovery time from surgery), the time passed between tests (e.g., later time lag) is often used as the basis of the trajectory (e.g., McArdle & Woodcock, 1997; Snijders & Bosker, 1999; Verbeke et al., 2000). Other more informative bases for time may be time since an illness or time before death (see Alwin & McCammon, 2001; Lindenberger, Singer, & Baltes, 2002; Singer, Verhaeghen, Ghisletta, Lindenberger, & Baltes, 2003).

The age-based approach applied to longitudinal scores clearly showed linear declines over age and did not suggest that more complex, nonlinear functional forms were needed. A simple formulation of retest effects was represented and showed small but positive increases in scores from first to subsequent testings, so models without retest effects tend to underestimate the basic aging declines in cognition (Ferrer et al., 2005; Lövdén et al., 2004; Rabbitt et al., 2004; Roediger & Karkicke, 2006). We used additional variables to understand some of the individual differences in age changes, and education, cohort, and gender had powerful independent effects. These demographic effects were largely descriptive features, and any shift in individual expectations required all demographic information. The time-varying information was similar—the positive effect of being in a couple was consistent with the positive cross-sectional effect, but the lack of an effect for being tested on the telephone was not consistent. These and other methodological features of the HRS require further analyses (see Park, 1999).

These results set the stage for many other HRS analyses not presented here. For example, we have not fully studied the impact of participant attrition in the HRS. We attempted to account for this nonrandom attrition by including all longitudinal and cross-sectional data in the models, but we recognize this potential confound, especially as it is related to other key variables, such as age selectivity of cohort selection (e.g., Alwin & McCammon, 2001; McArdle & Anderson, 1990; Miyazaki & Raudenbush, 2000). In other longitudinal studies, the persons who drop out are somewhat lower in cognitive performance at baseline than those who participate in the follow-up, and it is possible to introduce pattern-mixture assumptions to improve the accuracy of the model parameters (e.g., Hedeker & Gibbons, 1997; Little, 1995). One key concern for latent curve analyses with an older population is a statistical accounting for the relations of mortality status to cognition (Singer et al., 2003), and new statistical models merge latent curve analysis with survival or frailty analysis (e.g., Ghisletta, McArdle, & Lindenberger, 2006; Guo & Carlin, 2004; McArdle, Small, Backman, & Fratiglioni, 2005). More appropriate age-based models can be fitted with latent class mixture models for subgroups with different trajectories (Muthén & Masyn, 2005). After age-based confounds are considered, the HRS cognition data can be used in the investigation of causal models of health and well being (e.g., Adams et al., 2003; Blanchard-Fields, 2005; Zelinski et al., 1998; Zelinski & Lewis, 2003). Models using latent dynamic variables seem appropriate for the HRS cognitive measures within spouse pairs (Ghisletta & Lindenberger, 2003; McArdle, 2001; McArdle et al., 2001, 2004; Sliwinski et al., 2003).

The HRS/AHEAD data offer an opportunity for understanding cognitive functioning and age. The HRS longitudinal data studied in this article have features common to many data sets on aging (Hauser & Willis, 2005). The HRS data were not collected in a randomized controlled pretest–posttest design, and, as a result, independent causal impacts among variables are confounded and difficult to isolate. Also, there are large differences in initial cross-sectional ages and smaller differences in longitudinal time lags. Given these kinds of limitations, the HRS longitudinal data are unique in many respects, especially regarding the collection of a large representative sample of the U.S. older adult population and the attempt to measure several aspects of cognition (see McArdle et al., 2006). As we have tried to demonstrate, the HRS cognitive data offer many future opportunities for research on both the predictors and the outcomes of cognitive aging.

References

Adams, P., Hurd, M. D., McFadden, D., Merrill, A., & Ribeiro, T. (2003). Healthy, wealthy, and wise? Tests for direct causal paths between health and socioeconomic status. *Journal of Econometrics, 112,* 3–56.

Alwin, D. F., & McCammon, R. J. (2001). Aging, cohorts, and verbal ability. *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 65,* S1–S11.

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling, 12,* 411–434.

Baltes, P. B., & Schaie, K. W. (1976). On the plasticity of intelligence in adulthood and old age: Where Horn and Donaldson fail. *American Psychologist, 31,* 720–725.

Bayley, N. (1966). Learning in adulthood: The role of intelligence. In H. J. Klausmeier & C. W. Harris (Eds.), *Analyses of concept learning* (pp. 117–138). New York: Academic Press.

Blanchard-Fields, F. (2005). Introduction to the special section on emotion–cognition interactions and the aging mind. *Psychology and Aging, 20,* 539–541.

Blaum, C., Ofstedal, M. B., & Liang, J. (2002). The relationship of low cognitive performance to task specific disability: Findings from a population-based survey. *Journals of Gerontology, Series A: Biological Sciences and Medical Sciences, 57,* M523–M531.

Botwinick, J. (1977). Intellectual abilities. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (1st ed., pp. 580–605). New York: Van Nostrand Reinhold.

Bradway, K. P., & Thompson, C. W. (1962). Intelligence at adulthood: A twenty-five year follow-up. *Journal of Educational Psychology, 53,* 1–14.

Brandt, J., Spencer, M., & Folstein, M. (1988). The telephone interview for cognitive status. *Neuropsychiatry, Neuropsychology, and Behavioral Neurology, 1,* 111–117.

Browne, M., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen & S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Beverly Hills, CA: Sage.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Cagney, K. A., & Lauderdale, D. S. (2002). Education, wealth and cognitive function in later life. *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 57,* P163–P172.

Cerella, J., & Hale, S. (1994). The rise and fall in information-processing rates over the life span. *Acta Psychologica, 86,* 109–197.

Cnaan, A., Laird, N. M., & Slasor, P. (1997). Using the general linear mixed model to analyze unbalanced repeated measures and longitudinal data. *Statistics in Medicine, 16,* 2349–2380.

Cudeck, R., & Klebe, K. J. (2002). Multiphase mixed-effects models for repeated measures data. *Psychological Methods, 7,* 41–62.

Diggle, P. J., Liang, K.-Y., & Zeger, S. L. (1994). *Analysis of longitudinal data.* New York: Oxford University Press.

Donaldson, G., & Horn, J. L. (1992). Age, cohort, and time developmental muddles: Easy in practice, hard in theory. *Experimental Aging Research, 18,* 213–222.

Ferrer, E., Hamagami, F., & McArdle, J. J. (2004). Modeling latent growth curves with incomplete data using different types of structural equation modeling and multilevel software. *Structural Equation Modeling, 11,* 452–483.

Ferrer, E., Salthouse, T. A., McArdle, J. J., Stewart, W. F., & Schwartz, B. S. (2005). Multivariate modeling of age and retest in longitudinal studies of cognitive abilities. *Psychology and Aging, 20,* 412–422.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95,* 29–51.

Freedman, V. A., Aykan, H., & Martin, L. G. (2001). Aggregate changes in severe cognitive impairment among older Americans: 1993 and 1998. *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 56,* P100–P111.

Freedman, V. A., Aykan, H., & Martin, L. G. (2002). Another look at aggregate changes in severe cognitive impairment: Cumulative effects of three survey design issues. *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 57,* S126–S131.

Freedman, V. A., & Martin, L. G. (1998). Understanding trends in functional limitations among older Americans. *American Journal of Public Health, 88,* 1457–1462.

Freedman, V. A., & Martin, L. G. (2000). Contribution of chronic conditions to aggregate changes in old-age functioning. *American Journal of Public Health, 90,* 1755–1760.

Fultz, N., Ofstedal, M. B., Herzog, A. R., & Wallace, R. B. (2003). Additive and interactive effects of comorbid physical and mental conditions on functional health. *Journal of Aging and Health, 15,* 465–481.

Ghisletta, P., & Lindenberger, U. (2003). Age-based structural dynamics between perceptual speed and knowledge in the Berlin Aging Study: Direct evidence for ability dedifferentiation in old age. *Psychology and Aging, 18,* 696–713.

Ghisletta, P., McArdle, J. J., & Lindenberger, U. (2006). Longitudinal cognition-survival relations in old and very old age: 13-year data from the Berlin Aging Study. *European Psychologist, 11,* 204–223.

Guo, X., & Carlin, B. P. (2004). Separate and joint modeling of longitudinal and event time data using standard computer packages. *American Statistician, 58*(1), 1–9.

Hall, C. B., Lipton, R. B., Sliwinski, M., & Stewart, W. F. (2000). A change point model for estimating the onset of cognitive decline in preclinical Alzheimer's disease. *Statistics in Medicine, 19,* 1555–1566.

Hauser, R. M., & Willis, R. J. (2005). Survey design and methodology in the Health and Retirement Study and the Wisconsin Longitudinal Study. In L. J. Waite (Ed.), *Aging, health, and public policy: Demographic and economic perspectives* (pp. 209–235). New York: Population Council.

Hedeker, D., & Gibbons, R. D. (1997). Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods, 2*(1), 64–78.

Heeringa, S. G., & Connor, J. H. (1996). *Technical description of the Health and Retirement Study sample design* (Institute for Social Research Pub. DR-002). Ann Arbor: University of Michigan.

Herzog, A. R., & Rodgers, W. L. (1999). Cognitive performance measures in survey research on older adults. In N. Schwarz, D. Park, B. Knauper, & S. Sudman (Eds.), *Aging, cognition, and self-reports* (pp. 327–340). Philadelphia: Psychology Press.

Herzog, A. R., & Wallace, R. B. (1997). Measures of cognitive functioning in the AHEAD study. *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 52,* P37–P48.

Horn, J. L. (1988). *Thinking about human abilities*. New York: Plenum Press.

Horn, J. L. (1991). Measurement of intellectual capabilities: A review of theory. In K. S. McGrew, J. K. Werder, & R. W. Woodcock (Eds.), *Woodcock–Johnson technical manual* (pp. 197–246). Allen, TX: DLM Teaching Resources.

Horn, J. L., & McArdle, J. J. (1980). Perspectives on mathematical and statistical model building (MASMOB) in research on aging. In L. 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 guide to measurement invariance in research on aging. *Experimental Aging Research, 18,* 117–144.

Hultsch, D. F., Hertzog, C., Dixon, R. A., & Small, B. J. (1998). *Memory change in the aged.* New York: Cambridge University Press.

Jones, C. J., & Meredith, W. (2000). Developmental paths of psychological health from early adolescence to later adulthood. *Psychology and Aging, 15*(2), 351–360.

Jones, H. E., & Conrad, H. S. (1933). The growth and decline of intelligence: A study of a homogeneous group between the ages of ten and sixty. *Genetic Psychology Monographs, 13,* 223–275.

Juster, F. T., & Suzman, R. (1995). An overview of the Health and Retirement Study. *Journal of Human Resources, 30*(Suppl. 1995), S7–S56.

Lachman, M. E., & Spiro, A. (2002). *Critique of cognitive measures in the Health and Retirement Study (HRS) and the Asset and Health Dynamics Among the Oldest Old (AHEAD) Study* (HRS Data Monitoring Committee Report). Bethesda, MD: National Institute on Aging.

Langa, K. M., Chernew, M., Kabeto, M., Herzon, A. R., Ofstedal, M. B., Willis, R., et al. (2001). National estimates of the quantity and cost of informal caregiving for the elderly with dementia. *Journal of General Internal Medicine, 16,* 770–778.

Langa, K. M., Larson, E. B., Wallace, R. B., Fendrick, A. M., Foster, N. L., Kabeto, M. U., et al. (2004). Out-of-pocket health care expenditures among older Americans with dementia. *Alzheimer Disease and Associated Disorders, 18,* 90–98.

Leacock, C. (Ed.). (2006). *Getting started with the Health and Retirement Study, Version 1.0.* Ann Arbor: Survey Research Center, Institute of Social Research, University of Michigan.

Lindenberger, U., Singer, T., & Baltes, P. B. (2002). Longitudinal selectivity in aging populations: Separating mortality-associated versus experimental components in the Berlin Aging Study (BASE). *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 57,* P474–P482.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS system for mixed models* (2nd ed.). Cary, NC: SAS Institute.

Little, R. J. A. (1995). Modeling the dropout mechanism in repeated-measures studies. *Journal of the American Statistical Association, 90,* 1112–1121.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Lövdén, M., Ghisletta, P., & Lindenberger, U. (2004). Cognition in the Berlin Aging Study (BASE): The first 10 years. *Aging, Neuropsychology, and Cognition, 11,* 104–133.

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003). On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. *Intelligence, 31*(6), 543–566.

McArdle, J. J. (1994). Structural factor analysis experiments with incomplete data. *Multivariate Behavioral Research, 29,* 409–454.

McArdle, J. J. (2001). A latent difference score approach to longitudinal dynamic structural analyses. In R. Cudeck, S. du Toit, & D. Sorbom (Eds.), *Structural equation modeling: Present and future* (pp. 342–380). Lincolnwood, IL: Scientific Software International.

McArdle, J. J. (2007). Factor analyses of longitudinal data. In R. Cudeck & R. MacCallum (Eds.), *Factor analysis at 100* (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.), *The handbook of the psychology of aging* (pp. 21–43). New York: Plenum Press.

McArdle, J. J., & Bell, R. Q. (2000). *An introduction to latent growth models for developmental data analysis*. Mahwah, NJ: Erlbaum.

McArdle, J. J., Ferrer-Caja, E., Hamagami, F., & Woodcock, R. W. (2002). Comparative longitudinal multilevel 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. (1991). Modeling incomplete longitudinal and cross-sectional data using latent growth structural models. *Experimental Aging Research, 18*(1), 145–166.

McArdle, J. J., & Hamagami, F. (1996). Multilevel models from a multiple group structural equation perspective. In G. Marcoulides & R. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 89–124). Hillsdale, NJ: Erlbaum.

McArdle, J. J., & Hamagami, F. (2001). Linear dynamic analyses of incomplete longitudinal data. In L. Collins & A. Sayer (Eds.), *Methods for the analysis of change* (pp. 137–176). Washington, DC: American Psychological Association.

McArdle, J. J., Hamagami, F., Jones, K., Jolesz, F., Kikinis, R., Spiro, A., et al. (2004). Modeling memory and brain structure dynamic changes using longitudinal data from the Massachusetts General Hospital Aging Study. *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 59,* P294–P304.

McArdle, J. J., Hamagami, F., Meredith, W., & Bradway, K. P. (2001). Modeling the dynamic hypotheses of Gf-Gc theory using longitudinal life-span data. *Learning and Individual Differences, 12,* 53–79.

McArdle, J. J., & Nesselroade, J. R. (1994). Structuring data to study development and change. In S. H. Cohen & H. W. Reese (Eds.), *Life-span developmental psychology: Methodological innovations* (pp. 223–268). Hillsdale, NJ: Erlbaum.

McArdle, J. J., & Nesselroade, J. R. (2003). Growth curve analyses in contemporary psychological research. In J. Schinka & W. Velicer (Eds.), *Comprehensive handbook of psychology: Vol. 2. Research methods in psychology* (pp. 447–480). New York: Pergamon Press.

McArdle, J. J., & Prescott, C. A. (1992). Age-based construct validation using structural equation modeling. *Experimental Aging Research, 18,* 87–115.

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,* 69–114.

McArdle, J. J., Rodgers, W. L., Fisher, G. G., Woodcock, R., & Horn, J. L. (2006). *An experiment in measuring fluid intelligence over the telephone in the Health and Retirement Study (HRS)*. Unpublished manuscript, University of Southern California.

McArdle, J. J., Small, B. J., Backman, L., & Fratiglioni, L. (2005). Longitudinal models of growth and survival applied to the early detection of Alzheimer's disease. *Journal of Geriatric Psychiatry and Neurology, 18,* 234–241.

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.

Meredith, W., & Horn, J. L. (2001). The role of factorial invariance in modeling growth and change. In L. Collins & A. Sayer (Eds.), *Methods for the analysis of change* (pp. 201–240). Washington, DC: American Psychological Association.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55,* 107–122.

Miyazaki, Y., & Raudenbush, S. W. (2000). Tests for linkage of multiple cohorts in an accelerated longitudinal design. *Psychological Methods, 5,* 24–63.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*(4), 338–354.

Muthén, B., & Masyn, K. (2005). Discrete-time survival mixture analysis. *Journal of Education and Behavioral Statistics, 30,* 27–58.

Muthén, L. K., & Muthén, B. O. (2006). *Mplus user's guide* (4th ed.). Los Angeles: Muthén & Muthén.

Nagoshi, C. T., & Johnson, R. C. (1987). Between- vs. within-family analyses of the correlation of height and intelligence. *Social Biology, 34*(1–2), 110–113.

Neisser, U. (Ed.). (1998). *The rising curve: Long term gains in IQ and related measures.* Washington, DC: American Psychological Association.

Nesselroade, J. R., & Cable, D. G. (1974). "Sometimes, it's okay to factor differences scores"—the separation of state and trait anxiety. *Multivariate Behavioral Research, 9*(3), 273–281.

Ofstedal, M. B., Fisher, G. G., & Herzog, A. R. (2005). *Documentation of cognitive functioning measures in the Health and Retirement Study* (HRS/AHEAD Documentation Report DR-006). Ann Arbor: University of Michigan.

Ofstedal, M. B., & Herzog, A. R. (2004). *The effect of lifestyle on change in cognitive functioning among the oldest-old* (HRS research memo). Manuscript in preparation.

Park, D. C. (1999). Cognitive aging, processing resources, and self-report. In N. Schwarz, D. Park, B. Knauper, & S. Sudman (Eds.), *Aging, cognition, and self-reports* (pp. 45–70). Philadelphia: Psychology Press.

Park, D. C. (2000). The basic mechanisms accounting for age-related decline in cognitive function. In D. Park & N. Schwarz (Eds.), *Cognitive aging: A primer* (pp. 3–21). Philadelphia: Psychology Press.

Park, D. C., Hedden, T., Davidson, N. S., Lautenschlager, G., Smith, A. D., & Smith, P. K. (2002). Models of visuospatial and verbal memory across the adult life span. *Psychology and Aging, 17,* 299–320.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society: Series B, 60,* 23–40.

Rabbitt, P., Diggle, P., Holland, F., & McInnes, L. (2004). Practice and drop-out effects during a 17-year longitudinal study of cognitive aging. *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 59,* P84–P87.

Rabbitt, P., Diggle, P., Smith, D., Holland, F., & McInnes, L. (2001). Identifying and separating the effects of practice and of cognitive ageing during a large longitudinal study of elderly community residents. *Neuropsychologia, 39,* 532–543.

Rodgers, W. L., Ofstedal, M. B., & Herzog, A. R. (2003). Trends in scores on tests of cognitive ability in the elderly U.S. population, 1993–2000. *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 58,* S338–S346.

Roediger, H. L., & Karkicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17,* 249–255.

Schaie, K. W. (1996). *Intellectual development in adulthood: The Seattle Longitudinal Study.* New York: Cambridge University Press.

Singer, T., Verhaeghen, P., Ghisletta, P., Lindenberger, U., & Baltes, P. B. (2003). The fate of cognition in very old age (BASE): Six-year longitudinal findings and methodological implications of age-associated selectivity. *Psychology and Aging, 18,* 318–331.

Sliwinski, M., & Buschke, H. (1999). Cross-sectional and longitudinal relationships among age, cognition, and processing speed. *Psychology and Aging, 14,* 18–33.

Sliwinski, M. J., Hofer, S. M., & Hall, C. (2003). Correlated and coupled cognitive change in older adults with and without preclinical dementia. *Psychology and Aging, 18,* 672–683.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling.* London: Sage.

Stapleton, L. M. (2002). The incorporation of sample weights into multilevel structural equation models. *Structural Equation Modeling, 9,* 475–502.

Suthers, K., Kim, J. K., & Crimmins, E. (2003). Life expectancy with cognitive impairment in the older population of the United States. *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences, 58,* S179–S186.

Tulsky, D. S., & Price, L. R. (2003). The joint WAIS-III and WMS-III factor structure: Development and cross-validation of a six-factor model of cognitive functioning. *Psychological Assessment, 15,* 149–162.

Verbeke, G., Molenberghs, G., Krickeberg, K., & Fienberg, S. (Eds.). (2000). *Linear mixed models for longitudinal data.* New York: Springer.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale–Revised*. New York: Psychological Corporation.

Xu, R. (2003). Measuring explained variance in linear mixed effects models. *Statistics in Medicine, 22,* 3527–3541.

Zacks, R. T., & Hasher, L. (2006). Aging and long-term memory: Deficits are not inevitable. In E. Bialystok & F. I. M. Craik (Eds.), *Lifespan cognition: Mechanisms of change* (pp. 162–177). New York: Oxford University Press.

Zelinski, E. M., Crimmins, E., Reynolds, S., & Seeman, T. (1998). Do medical conditions affect cognition in older adults? *Health Psychology, 17,* 504–512.

Zelinski, E. M., & Gilewski, M. J. (2003). Effects of demographic and health variables on Rasch scaled cognitive scores. *Journal of Aging and Health, 15,* 435–464.

Zelinski, E. M., & Lewis, K. L. (2003). Adult age differences in multiple cognitive functions: Differentiation, dedifferentiation, or process-specific change? *Psychology and Aging, 18,* 727–745.

Received February 7, 2005
Revision received February 15, 2007
Accepted February 16, 2007 ■