Estimating Age Change in List Recall in Asset and Health Dynamics of the Oldest-Old: The Effects of Attrition Bias and Missing Data Treatment

Robert F. Kennison and Elizabeth M. Zelinski
University of Southern California

Average change in list recall was evaluated as a function of missing data treatment (Study 1) and dropout status (Study 2) over ages 70 to 105 in Asset and Health Dynamics of the Oldest-Old data. In Study 1 the authors compared results of full-information maximum likelihood (FIML) and the multiple imputation (MI) missing-data treatments with and without independent predictors of missingness. Results showed declines in all treatments, but declines were larger for FIML and MI treatments when predictors were included in the treatment of missing data, indicating that attrition bias was reduced. In Study 2, models that included dropout status had better fits and reduced random variance compared with models without dropout status. The authors conclude that change estimates are most accurate when independent predictors of missingness are included in the treatment of missing data with either MI or FIML and when dropout effects are modeled.

Keywords: aging, memory, longitudinal, attrition, missing data

Attrition Bias and Missing-Data Treatments

A large literature on cross-sectional age comparisons of episodic memory shows that age-related declines in performance are ubiquitous (Zacks, Hasher, & Li, 2000). This general result has been substantiated in longitudinal comparisons (e.g., Hultsch, Hertzog, Dixon, & Small, 1998; Zelinski & Burnight, 1997), but recent work suggests that such findings are biased because cases with missing data are eliminated and attrition effects are not directly modeled (cf. Singer, Verhaeghen, Ghisletta, Lindenberger, & Baltes, 2003; Sliwinski, Hofer, Hall, Buschke, & Lipton, 2003). Attrition bias develops in longitudinal studies for a variety of causes, including death, ill health, and cognitive decline. The analytic strategy that produces the least bias is one that reasonably includes all available data (McArdle & Hamagami, 1991; see also Collins & Horn, 1991; Collins & Sayer, 2001), that is, the data of all participants, even if they have missing data at any point in the study. Within this full-information approach, various “proper” methods for treating missing data have been developed (cf. Graham & Hofer, 2000; R. J. A. Little & Rubin, 2002; Schafer, 1997; Wothke, 2000). Among them are MI and the FIML algorithm.

With the MI approach, more than one estimate of each missing observation is estimated and inserted into separate data sets that are analyzed individually with results collapsed, thereby creating variability in estimates (including both measurement and sampling error). By convention, this procedure uses independent predictors (e.g., sex, education) to improve accuracy. These predictors are included along with the extant longitudinal data in an estimation algorithm such as ordinary least squares (OLS) regression. More than one estimate is drawn from a posterior Bayesian distribution of possible estimates to reflect the uncertainty that arises from indeterminacy in estimates. This produces unbiased standard errors that result in proper variances, covariances, and means (Graham, & Hofer, 2000; Newman, 2003; Schafer, 1997). This is a two-step procedure: imputation followed by analysis.
The FIML approach involves a single step in which a model (e.g., structural equation or multilevel) is fit with the FIML algorithm (cf. Collins & Horn, 1991; Collins & Sayer, 2001; Graham & Hofer, 2000; T. D. Little, Schnabel, & Baumert, 2002; McArdle & Hamagami, 1991). The analyzed data include cases with missing values (i.e., full information). The algorithm does not actually estimate missing data; rather, it accounts for missing values in model parameter estimates. Thus, in a growth model, the intercept and slope parameters will reflect both extant and missing data. Because no special treatment of the data is required before modeling is done, FIML is straightforward. Like MI, it too has been shown to produce unbiased standard errors and proper parameter estimates (R. J. A. Little & Rubin, 2002). However, as discussed in detail later, the use of independent predictors in FIML has often been ignored (Graham & Hofer, 2000).

The validity of both approaches rests on the assumption that missing data are missing at random (MAR), that is, that the missing scores can be estimated from variables other than those being modeled (R. J. A. Little & Rubin, 1987). More formally stated, MAR occurs when missing data can be estimated from the observed data and is not dependent on what is missing after the observed data have been accounted for (see R. J. A. Little & Rubin, 1987, 2002; Schafer, 1997, for detailed discussions). Some diagnostics are available to examine whether the MAR assumption is supported for a given measure. Support for the MAR assumption, however, is predictive but not ensured. This is because it is impossible to know the true extent of attrition bias because the data necessary to make that determination are missing, and it is also unlikely that all of the sources that could account for attrition bias have, in fact, been used or measured. Logistic regression has been used as a diagnostic of the MAR assumption. Risk of dropout is predicted from the variable of interest (e.g., recall performance) and other independent predictors (e.g., age, gender). Reliable prediction of dropout status by independent predictors lends support for the assumption.

Because the MAR assumption is based, in part, on the presence of independent predictors, any missing-data treatment that does not include such predictors risks retaining bias. That is, the variables that predict missingness should also be used in the missing-data estimation procedure. There is mounting evidence that inclusion of such independent predictors reduces bias in simulation studies (Collins, Schafer, & Kam, 2001; Graham & Hofer, 2000; Meng, 1994; D. B. Rubin, 1996). However, Collins et al. (2001) demonstrated that including additional predictor variables does not fully reduce bias under all circumstances. They showed that the benefits of predictors are diminished when the amount of missing data is less than 25% of the sample or when a cause of missing data (e.g., death status) is at least moderately correlated ($r > .4$) with the variable containing the missing data (e.g., recall performance). Still, inclusion of predictors is recommended over their exclusion (Collins et al., 2001; Graham & Hofer, 2000) because some bias reduction is likely to be obtained.

Historically, the two missing-data treatments (MI and FIML) differ in their utilization of independent predictors. For MI, the recommended procedures strongly encourage inclusion of various predictors. For example, Schafer (1997) and others (Allison, 2002; R. J. A. Little & Rubin, 1987, 2002) state that any predictors that are to be included in later data analyses should be included in the imputation process. They also recommend adding other variables to improve prediction because the procedure has been shown to be robust to problems of overfitting.

In the case of the FIML approach, independent predictors have often been overlooked (e.g., McArdle, Ferrer-Caja, Hamagami, & Woodcock, 2002; McArdle, & Hamagami, 1991; Sliwinski et al., 2003). This is, in part, because clear guidelines for their inclusion have only recently been made available. For example, Graham et al. (Graham, 2003; Graham & Hofer, 2000) have demonstrated with their correlated-saturates approach that independent predictors can be incorporated into FIML models in the form of added predictors to regression-based structural equation models. The extension of this approach to structural equation or multilevel growth curve modeling can take various forms. The approach taken here is to include additional predictors of the latent intercept and slope parameters of growth curve models. In multilevel growth curve models, this is accomplished by including each predictor as a between-person conditional predictor of each of the growth parameters (i.e., intercept and slopes).

In the present study, we compared average estimated growth curves of recall performance over age (rather than time; see McArdle & Hamagami, 1991) for data samples treated in a manner consistent with either the MI or FIML approaches and either with or without the addition of predictors of missingness. With the age-based approach, the four-occasion (7-year) longitudinal sequences of each participant were analyzed over age, as though they were spread across the span of the observed ages (i.e., ages 70–105). For example, a 70-year-old person’s 7-year data would span ages 70 to 77, whereas a 75-year-old’s data would span ages 75 to 82. Thus, modeling change over age effectively spreads the data over a larger time span by creating additional missing data. Such age models rely on the convergence assumption: that a strong correspondence exists between cross-sectional and longitudinal data sequences (Bell, 1953; McArdle & Bell, 2000). The assumption asserts that cohort effects are zero. Violations of the assumption may produce inflated estimates of age change, especially if cohort effects are present and are positive for younger cohorts (e.g., McArdle et al., 2002).

Age-based modeling potentially involves two types of missing data: (a) missing data as a result of attrition from the time-based test occasions, or attrition-related missing data, and (b) missing data associated with analyzing the longitudinal sequences over age, or analysis-related missing data. The MI and FIML procedures are different in this respect. In the MI treatment, the missing longitudinal sequence data are estimated before modeling. Thus, in the analysis stage, the only missing data associated with the MI treatment are analysis related. In contrast, with the FIML approach, no treatment of missing data is applied before modeling so both forms of missing data exist (i.e., attrition-related and analysis-related missing data).

Despite these differences, both treatments, MI and FIML, should help to reduce attrition bias in predictable ways. As discussed in greater detail later, we expected that each data treatment would be associated with lower intercepts and greater declines in recall performance compared with models using only cases with complete data (list-wise deletion). Although demonstrated in simulation studies (Collins et al., 2001; Graham, 2003; Schafer, 1997) and in at least one study of real data (Allison, 2002), this is the first study to systematically compare bias reduction in the MI and FIML missing-data approaches with and without independent pre-
dictors of missingness in multilevel growth curve models and with real data.

Age, Death, and Dropout Effects in List Recall

In cross-sectional comparisons of memory performance, age differences have consistently been observed between college-age and older participants (cf. Zacks et al., 2000, for a review) and generally between young-old and old-old participants (e.g., Giambra, Arenberg, Zonderman, Kawas, & Costa, 1995; Small, Dixon, Hultsch, & Hertzog, 1999). Yet, in longitudinal studies, age-related declines in memory performance are not consistently found (cf. Haan, Shemanski, Jagust, Manolio, & Kuller, 1999; E. H. Rubin et al., 1998; Wilson et al., 2002), and outcomes disagree on when declines begin, the rate of decline, and whether declines are nonmonotonic (Colsher & Wallace, 1991; Giambra et al., 1995; Schaie, 1996; Singer et al., 2003). There are multiple reasons for such divergent results, among them practice effects, sample size, time lag durations, and choice of stimulus materials as well as death and dropout effects. Here we focus primarily on the latter: death and dropout effects. Longitudinal data gathered from participants just before death or dropout show greater declines than data gathered on similarly aged retestees (Rabbitt, Diggle, Holland, & McInnes, 2004; Rabbitt, Watson, Donlan, Bent, & McInnes, 1994; Rabbitt et al., 2002). Such dropout effects in longitudinal studies have been associated with increased age, poorer health, male gender, impaired cognitive performance, impaired episodic memory, depression, lower social status, low intelligence, and death (Baltes, Schaie, & Nardi, 1971; Hultsch et al., 1998; Johansson & Berg, 1989; McArdle, Hamagami, Elias, & Robbins, 1991; Rabbitt, Donlan, Bent, McInnes, & Abson, 1993; Rabbitt et al., 1994; Siegler & Botwinick, 1979; Sliwinski et al., 2003; Zelinski, Gilewski, & Schaie, 1993).

Selectivity analysis has been used to partition the variance attributable to total dropout, mortality-related dropout, and nonmortality-related dropout in intelligence measures. Lindenberger et al. (Lindenberger, Singer, & Baltes, 2002; see also Lindenberger et al., 1999) used the Pearson-Lawley selection formula to condition longitudinal parameter estimates with missing data by a set of baseline predictor variables. Selectivity in intelligence in the parent sample was associated with an effect size of nearly 0.6 SDs, and when age was partialed out the effect size was about 0.4 SDs, with substantial contributions from both mortality and nonmortality sources of dropout. They concluded that both mortality and other experimental causes of attrition contribute to bias in longitudinal assessments of intelligence, suggesting that mortality, although important, is only one of several sources of bias.

Dropout has been shown to be a stronger predictor of memory change than death. In pattern-mixture models, Sliwinski et al. (2003) found that time to dropout better explained age declines in memory than death. They suggested that time to death may represent effects of illness and medical interventions, whereas time to dropout involves additional attrition-related phenomena. Our focus is on dropout and its effect on estimates of change when dropout status is included in pattern-mixture growth curve models.

Thus, it is important to model attrition directly with dropout status indicators (Hedeker & Gibbons, 1997). Arguably, growth curve models that do not do this could retain attrition bias in their estimates of change (see Singer et al., 2003; Sliwinski et al., 2003). In Study 2, we examined change in pattern-mixture growth curve models that included dropout status to test hypotheses that partialing out attrition variance increases model fit, produces higher intercept values, and reduces average declines.

The Present Studies

In Study 1, the efficacy of MI and FIML to treat missing data and whether or not the addition of predictors of missingness aids bias reduction were examined by modeling change in AHEAD recall-data samples. Because MI is applied before analysis and FIML is applied during analysis, the application of the two produces what is essentially a $2 \times 4$ factorial of models. We delineate two stages of analysis: data treatment and data modeling. Four treatments were applied in the data-treatment stage, each resulting in a unique data set that was later analyzed in the data-modeling stage. First, the list-wise deletion (LD) treatment resulted in a complete-cases data set with a reduced sample size. All cases with any missing immediate-recall data were eliminated. Second, no treatment (NT) was applied, resulting in a full-information data set with both attrition-related and analysis-related missing data. The NT data relied solely on the FIML algorithm to deal with missing data and, therefore, represents the FIML missing-data treatment referred to earlier. Third, in the outcome-only multiple imputation (OOMI) treatment, missing longitudinal recall data were estimated using only the extant recall data. No additional predictors were added during the MI estimation procedure. In this treatment, as well as the next one (MI), missing data were estimated only for the four occasion (longitudinal) data. Because models were fit over age, new analysis-related missing data were introduced and treated by the FIML algorithm. Fourth, in the MI treatment, missing recall data were estimated from the extant recall data plus seven additional predictors.

The choice of the labels (NT and NT with predictors) to represent what are variants of the FIML treatment (discussed early in this article) was made because of ambiguity over the term FIML. The FIML algorithm is a model-fitting algorithm that is able to handle data sets with missing values, and because of its unique properties it has also been referred to as a missing-data treatment (Allison, 2002; Schafer, 1997; Wohtke, 2000). However, because the models we report all use the algorithm for model fitting, to treat analysis-related missing data we chose the NT label because no action was taken before model fitting (during the data-treatment stage).

In the data-modeling stage, multilevel growth curve models were fit to each of the four data sets (LD, NT, OOMI, and MI) either without or with additional independent predictors added to the intercept and slopes. Note that models run without predictors are usually referred to simply by their data treatment names without reference to the lack of predictors (e.g., NT rather than NT without predictors). The four levels of treatment, when crossed with two models (i.e., without or with predictors), yielded eight conditions: LD; LD with predictors; NT; NT with predictors; OOMI; OOMI with predictors; MI; and MI with predictors.

Although we report the results of all eight treatments, much of the discussion focuses on five of these: LD, NT, NT with predictors, OOMI, and MI. The LD treatment is used as a benchmark because it retains attrition bias. Both the NT with predictors and MI treatments are theoretically defensible missing-data treatments that fall in line with currently accepted methods. Their results are compared with the prediction that they should produce (a) poorer
performance at the intercept compared with LD (e.g., Hultsch et al., 1998); (b) steeper slopes of decline compared with LD; and (c) similar results when compared with each other. The second prediction has been supported by at least two studies of recall performance. Singer et al. (2003) showed steeper averaged age trajectories for the baseline cross-sectional sample than a longitudinal one for the Berlin Aging Study, suggesting that those who dropped out declined more rapidly. Sliwinski et al. (2003) modeled memory declines in dropouts versus continuers and decedents versus continuers in pattern-mixture models that included time-varying covariates of either years to dropout or years to death. They reported steeper memory declines in dropouts or decedents compared with retestees, but years to dropout accounted for more variance than years to death. We also confirm the second hypothesis in the AHEAD recall data by examining three-occasion change rates in those who either continued or dropped out at the fourth test occasion.

Comparisons of NT and OOMI are made to examine the effects of not including predictors in estimation of missing data. The results of these treatments are compared with the prediction that they should result in similar growth curves. In addition, these growth curves were expected to demonstrate less decline than NT with predictors and MI.

Study 2 was conducted with the goal of determining whether the inclusion of dropout status would increase model fit, reduce random variance, and be associated with larger intercepts and smaller declines than models that do not include such variables. Following Hedecker and Gibbons’s (1997) general approach, dropout status was added to the multilevel age models reported in Study 1, which included an intercept and linear and quadratic slopes.

**Study 1**

**Method**

**Participants.** The first four test occasions of the AHEAD survey were used for the data samples. Interviews were conducted either with the respondent or as a proxy interview with a spouse or relative. The present study deals only with respondent data. The four testings were conducted in 1993–1994, 1995–1996, 1998–1999, and 2000–2001. Test intervals were approximately 2 years, with the exception of the interval between the second and third test occasions, which was approximately 3 years.

Of the 8,222 individuals interviewed in 1993, data from 6,434 respondents were analyzed, after exclusion of proxy interviews \((n = 840)\), spouses younger than 70 \((n = 753)\), and cases with missing baseline data for recall \((n = 342)\), education \((n = 3)\), health \((n = 3)\), and depression \((n = 3)\). Note that some of the omitted cases had out-of-range or missing data on more than one of the measures. The mean age of the sample was 77.23 at baseline \((range = 70–103)\). Individual-level case weighting was used to accommodate intentional oversampling of Hispanic Americans, African Americans, and people living in Florida. Table 1 shows mean baseline values of age, education, health, and depression computed at each of the four test occasions on the available participants at that time. With each successive testing, the remaining sample tended to be younger, more educated, in better health, and less depressed on the baseline measures. Table 1 also shows the percentages of women, Whites, deceased participants, and number of people tested as a function of test occasion. There were increases in the percentages of women, Whites, decedents, and dropouts across test occasions.

**Measures.** The survey included measures of cognitive performance, health and health behaviors, economic status, and social network (cf. Soldo, Hurd, Rodgers, & Wallace, 1997). The interview, which was conducted either in person or by phone, lasted approximately 60 min; an average of 27 min was dedicated to the health and cognitive measures sections.

The immediate-recall test consisted of an orally presented list of 10 high-frequency English nouns, presented at a rate of 1 item every 2 s. After presentation, participants were prompted to orally recall as many of the items as possible in any order. Performance was assessed as the number of items correctly recalled out of 10. For the first test occasion, a single word list was presented. In subsequent testings, one of four word lists was randomly assigned to each respondent. None of the words used in the first list appeared in any of the later four lists. Small but reliable list effects have been reported (Herzog & Rogers, 1999). Measures of age, gender \((0 = \text{male}, 1 = \text{female})\), ethnicity \((\text{non-White} = 0, \text{White} = 1)\), years of education, self-rated health status on a 5-point scale \((5 = \text{poor health})\), depression, death status \((0 = \text{alive}, 1 = \text{dead})\), and dropout status \((0 = \text{continuer}, 1 = \text{dropout})\) were included as covariates of either years to dropout or years to death. They were increases in the percentages of women, Whites, decedents, and dropouts across test occasions.

Table 1

<table>
<thead>
<tr>
<th>Measure</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Recall</td>
<td>4.47</td>
<td>1.89</td>
<td>4.64</td>
<td>1.85</td>
</tr>
<tr>
<td>Age</td>
<td>77.23</td>
<td>5.59</td>
<td>76.71</td>
<td>5.31</td>
</tr>
<tr>
<td>Education</td>
<td>10.98</td>
<td>3.66</td>
<td>11.22</td>
<td>3.56</td>
</tr>
<tr>
<td>Health</td>
<td>2.99</td>
<td>1.16</td>
<td>2.90</td>
<td>1.13</td>
</tr>
<tr>
<td>Depression</td>
<td>1.66</td>
<td>1.97</td>
<td>1.53</td>
<td>1.89</td>
</tr>
</tbody>
</table>

**Percentages**

<table>
<thead>
<tr>
<th>Measure</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Females</td>
<td>62.43</td>
<td>62.79</td>
<td>63.68</td>
</tr>
<tr>
<td></td>
<td>Whites</td>
<td>84.86</td>
<td>86.11</td>
<td>86.46</td>
</tr>
<tr>
<td></td>
<td>Deceased</td>
<td>0.00</td>
<td>8.59</td>
<td>21.06</td>
</tr>
<tr>
<td></td>
<td>Retested</td>
<td>—</td>
<td>79.43</td>
<td>64.53</td>
</tr>
<tr>
<td></td>
<td>N</td>
<td>6,434</td>
<td>5,111</td>
<td>4,152</td>
</tr>
</tbody>
</table>

**Note.** Dashes indicate that participants could not be retested at Test Occasion 1.
Measures of missingness were added. In the MI treatment, the available longitudinal recall data and the predictor variables of age, gender, ethnicity, education, health, depression, and death status were used to estimate the missing values. Death status was used instead of dropout status because dropout status, when included in the MI procedure, violated model tolerance. Death strongly correlated with dropout ($r = .72$).

Five data sets were created for each MI treatment. Three to five imputations are usually sufficient for reliable estimates (Allison, 2002). In the present study, there was an average of $35.8\%$ missing data over the four measurement occasions. Thus, performing five imputations yielded results that are $93\%$ as efficient as an infinite number of imputations (Schafer, 1997). These data sets were then examined in multilevel models with output collapsed. The fixed-effect coefficients, random-effect variance components, chi-square statistics, and sigma-squared and its standard error were summarized as simple averages. The standard errors of the fixed effects were calculated with computations described in the HLM5 software manual (Raudenbush, Bryk, & Congdon, 2001; see also Allison, 2002).

Special equations were needed because the estimated standard errors are composed of two error components: sampling error and measurement error. These summary statistics are reported for the various MI treatments. In the MI treatment, the available recall data from the four test occasions was used to estimate the missing values. No additional predictors of missingness were included. In the MI treatment, the available longitudinal recall data and the predictor variables of age, gender, ethnicity, education, health, depression, and death status were used to estimate the missing values. Death status was used instead of dropout status because dropout status, when included in the MI procedure, violated model tolerance. Death strongly correlated with dropout ($r = .72$).

In the OOMI treatment, the available recall data from the four test occasions was used to estimate the missing values. No additional predictors of missingness were added. In the MI treatment, the available longitudinal recall data and the predictor variables of age, gender, ethnicity, education, health, depression, and death status were used to estimate the missing values. Death status was used instead of dropout status because dropout status, when included in the MI procedure, violated model tolerance. Death strongly correlated with dropout ($r = .72$).

Five data sets were created for each MI treatment. Three to five imputations are usually sufficient for reliable estimates (Allison, 2002). In the present study, there was an average of $35.8\%$ missing data over the four measurement occasions. Thus, performing five imputations yielded results that are $93\%$ as efficient as an infinite number of imputations (Schafer, 1997). These data sets were then examined in multilevel models with output collapsed. The fixed-effect coefficients, random-effect variance components, chi-square statistics, and sigma-squared and its standard error were summarized as simple averages. The standard errors of the fixed effects were calculated with computations described in the HLM5 software manual (Raudenbush, Bryk, & Congdon, 2001; see also Allison, 2002). Special equations were needed because the estimated standard errors are composed of two error components: sampling error and measurement error. These summary statistics are reported for the various MI treatments.

**Multilevel model estimation.** Multilevel models were fit to the various data sets. Multilevel modeling, as applied to longitudinal data, involves a process in which the repeated measures data are nested within individuals (Bryk & Raudenbush, 1992; MacCallum & Kim, 2000; Raudenbush & Bryk, 2002). Here, growth curves were fit to each individual’s data by means of random-effects regression. The multilevel regression equation for the within-persons age models of the present study can be summarized as follows:

$$y_{ij} = \pi_{0i} + \pi_{1i} \text{age}_{ij} + \pi_{2i} \text{age}_{ij}^2 + r_{ij},$$

where $y_{ij}$ is the recall score for individual $i$ measured at age $j$, which was equal to the intercept, $\pi_{0i}$, for individual $i$; the person’s linear slope of age; $\pi_{1i}$, the quadratic slope of age squared; $\pi_{2i}$, and error, $r_{ij}$. The between-persons model can be summarized as:

$$\pi_{0i} = \beta_{00} + U_{0i}$$
$$\pi_{1i} = \beta_{01} + U_{1i}$$
$$\pi_{2i} = \beta_{02} + U_{2i},$$

where $\beta$s were fixed effects and $U$s were random effects of the intercept ($\pi_{0i}$), linear slopes of age ($\pi_{1i}$), and quadratic slopes of age squared ($\pi_{2i}$). The independent predictors were included in the between-persons part. For example, in models that included independent predictors of the intercept and slopes, the between-persons equation for the slope of age became:

$$\pi_{1i} = \beta_{10} + \beta_{11}(\text{age}) + \beta_{12}(\text{gender}) + \beta_{13}(\text{ethnicity})$$
$$+ \beta_{14}(\text{education}) + \beta_{15}(\text{health}) + \beta_{16}(\text{depression})$$
$$+ \beta_{17}(\text{death status}) + U_{1i},$$

where the predictors of age, gender, ethnicity, education, health status, and death status were added to the slope of age. These predictors were added to the intercept and quadratic slope of age squared in a similar manner. Individual growth parameters (i.e., intercepts and slopes) were treated as independent variables at the between-persons level with both fixed and random effects. Fixed effects estimated the average variance that was systematically accounted for. Random effects estimated variance that could be attributed to individual differences in the fixed parameters.

HLM software, version 5.04 (Raudenbush et al., 2001), was used to fit the multilevel models. The software offers a choice between the FIML and restricted information maximum likelihood (REML) algorithms. The latter tends to generate somewhat larger standard errors and is considered more accurate when the sample size is small (Raudenbush & Bryk, 2002); however, because degrees of freedom are based on the number of random parameters included in the model, model fits can only be compared for models differing in their random part (cf. Hox, 2000; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). The FIML algorithm provides fit comparisons in models differing in the number of fixed parameters included. Results using both algorithms were compared. Our comparison showed similar coefficients, standard errors, and random variance estimates. In addition, the pattern of fixed and random effects was identical. Thus, because the models we set out to compare differed in the number of fixed parameters to be estimated, the large sample size of the AHEAD data set, and the similar results, we chose to use the FIML algorithm to fit the various models to all data sets.

Longitudinal change was assessed across the observed age range of 70 to 105, and the intercept was the grand mean centered for all models at 77.23 years, which was the mean age of the AHEAD sample at baseline ($n = 6,434$). Grand mean centering has been shown to minimize the correlation between the intercept and slopes and to reduce problems associated with multicollinearity (Hofmann & Gavin, 1998). In fitting the multilevel models, we initially compared the fits of an intercept-only model with an intercept and linear slopes-of-age model, and an intercept and linear and quadratic slopes-of-age model, for each of the data treatments. In most of the multilevel models that included predictors of the intercepts, linear age, and quadratic age terms, the measures of age, gender, ethnicity, education, health, depression, and death status were included. Because the number of living participants, death status was not included in its with-predictors models. These predictors were the same as those used in the MI treatment, and they were added as between-participants conditional predictors of the intercepts and linear and quadratic slopes of age. Each of the predictors was mean centered at entry into the model, which was done to preserve the metric of the parameter estimates (Hofmann & Gavin, 1998).

### Results and Discussion

**Patterns and predictors of missing data in recall.** Table 2 shows longitudinal recall performance and demographic information for participants as a function of the last testing completed. Note that the sample sizes vary within the categories of last test completed because, in some cases, respondents skipped a testing and then returned later (this also partly explains why the sample sizes do not always match those of Table 1). It is clear that those who dropped out after the first testing have the lowest Time 1 recall, poorest health rating, and highest depression; they are the...
The correlations, variances, and covariances of recall performance at each of the four test occasions and for the predictor variables appear in Table 3. The correlations for immediate recall over the four test occasions ranged from .44 to .53, suggesting that performance on a given test occasion could be predicted by performance on any other occasion. Age correlated significantly with recall performance at each occasion and also with each of the predictor variables except ethnicity. Both dropout and death status were negatively associated with baseline recall performance, such that those who died or dropped out tended to exhibit poorer recall. As shown in Table 3, dropout status correlated more strongly with recall at Test Occasions 1, 2, and 3 than death status, indicating that those who died or dropped out tended to exhibit poorer recall. The correlations, variances, and covariances of recall performance at each of the four test occasions and for the predictor variables appear in Table 3. The correlations for immediate recall over the four test occasions ranged from .44 to .53, suggesting that performance on a given test occasion could be predicted by performance on any other occasion. Age correlated significantly with recall performance at each occasion and also with each of the predictor variables except ethnicity. Both dropout and death status were negatively associated with baseline recall performance, such that those who died or dropped out tended to exhibit poorer recall. As shown in Table 3, dropout status correlated more strongly with recall at Test Occasions 1, 2, and 3 than death status, indicating that those who died or dropped out tended to exhibit poorer recall.
Table 3  
Correlations (Above Diagonal), Variances (on Diagonal), and Covariances (Below Diagonal) of Immediate Recall for Each Test Occasion, and Independent Predictors

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Recall 1</td>
<td>3.53</td>
<td>.51</td>
<td>.47</td>
<td>.44</td>
<td>-.31</td>
<td>.08</td>
<td>.21</td>
<td>.38</td>
<td>-.20</td>
<td>-.19</td>
<td>-.22</td>
<td>-.27</td>
</tr>
<tr>
<td>2. Recall 2</td>
<td>1.70</td>
<td>3.41</td>
<td>.53</td>
<td>.47</td>
<td>-.29</td>
<td>.10</td>
<td>.17</td>
<td>.32</td>
<td>-.22</td>
<td>-.17</td>
<td>-.40</td>
<td>-.45</td>
</tr>
<tr>
<td>3. Recall 3</td>
<td>1.58</td>
<td>1.74</td>
<td>3.58</td>
<td>.50</td>
<td>-.31</td>
<td>.11</td>
<td>.14</td>
<td>.33</td>
<td>-.21</td>
<td>-.17</td>
<td>-.60</td>
<td>-.69</td>
</tr>
<tr>
<td>4. Recall 4</td>
<td>1.37</td>
<td>1.41</td>
<td>1.57</td>
<td>3.10</td>
<td>-.29</td>
<td>.09</td>
<td>.14</td>
<td>.31</td>
<td>-.20</td>
<td>-.12</td>
<td>-.72</td>
<td>1.00</td>
</tr>
<tr>
<td>5. Age</td>
<td>-3.24</td>
<td>-2.86</td>
<td>-.89</td>
<td>-3.24</td>
<td>31.26</td>
<td>.05</td>
<td>-.01</td>
<td>-.12</td>
<td>.10</td>
<td>.11</td>
<td>.31</td>
<td>.32</td>
</tr>
<tr>
<td>6. Gender</td>
<td>0.07</td>
<td>0.09</td>
<td>0.10</td>
<td>0.08</td>
<td>.13</td>
<td>.24</td>
<td>-.00</td>
<td>-.15</td>
<td>.03</td>
<td>.11</td>
<td>-.09</td>
<td>-.04</td>
</tr>
<tr>
<td>7. Ethnicity</td>
<td>0.14</td>
<td>0.11</td>
<td>0.09</td>
<td>0.08</td>
<td>-.02</td>
<td>-.01</td>
<td>.13</td>
<td>.29</td>
<td>-.14</td>
<td>-.08</td>
<td>-.04</td>
<td>-.07</td>
</tr>
<tr>
<td>8. Education</td>
<td>2.60</td>
<td>2.10</td>
<td>2.18</td>
<td>1.85</td>
<td>-.53</td>
<td>-.05</td>
<td>.03</td>
<td>13.41</td>
<td>-.28</td>
<td>-.09</td>
<td>-.16</td>
<td></td>
</tr>
<tr>
<td>9. Health</td>
<td>-.045</td>
<td>-.47</td>
<td>-.44</td>
<td>-.37</td>
<td>.03</td>
<td>.02</td>
<td>-.06</td>
<td>-1.19</td>
<td>1.34</td>
<td>.41</td>
<td>.24</td>
<td>.24</td>
</tr>
<tr>
<td>10. Depression</td>
<td>-.072</td>
<td>-.58</td>
<td>-.59</td>
<td>-.36</td>
<td>1.21</td>
<td>.10</td>
<td>-.06</td>
<td>-.66</td>
<td>0.93</td>
<td>3.89</td>
<td>.16</td>
<td>.16</td>
</tr>
<tr>
<td>11. Death status</td>
<td>-.20</td>
<td>-.75</td>
<td>-.95</td>
<td>-1.25</td>
<td>0.81</td>
<td>-.02</td>
<td>-.01</td>
<td>-.16</td>
<td>0.13</td>
<td>0.15</td>
<td>.22</td>
<td>.72</td>
</tr>
<tr>
<td>12. Dropout status</td>
<td>-.26</td>
<td>-.93</td>
<td>-.14</td>
<td>-2.58</td>
<td>0.89</td>
<td>-.01</td>
<td>-.01</td>
<td>-.29</td>
<td>0.14</td>
<td>0.16</td>
<td>0.17</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Note. Gender: 0 = male, 1 = female; ethnicity: 0 = white, 1 = nonwhite; health rating on a scale from 1 = excellent and 5 = poor; Correlations with Recall 1 are based on 6,434 participants, Recall 2 on 5,111 participants, Recall 3 on 4,152 participants, and Recall 4 on 3,339 participants. All other correlations are based on 6,434 participants. Correlations over .03 are significant at p < .01.

Thus, although initial performance was a predictor of missing observations, it was only one of an array of significant and independent predictors. These results are consistent with the MAR assumption (R. J. A. Little & Rubin, 1987; Schafer, 1997), suggesting that the predictors examined may be useful in estimating missing recall data.

To judge objectively the appropriateness of each of the data treatments in simulating the missing recall data, a preliminary analysis examined whether greater declines were to be expected in those who dropped out versus those who remained in the study. A multilevel model was fit to the data of respondents with complete recall data on the first three test occasions, using dropout (coded 0 = continue, 1 = dropout) at the fourth test occasion as a predictor of estimated growth curve parameters. In that model, the intercept was 4.97, the linear slope of age was −0.616, and the slope of age squared was −0.149, all significant at p < .01. Dropout status was a significant predictor of the intercept (β = −0.650, p < .01) and slope of age (β = −0.321, p < .05) but not the slope of age squared (β = 0.100, p > .05). Because of a limited number of degrees of freedom, random effects were only estimated for the intercept and slope of age. These were associated with significant variance in the intercept (variance component = 1.288, p < .01) and the slope of age (variance component = 0.038, p < .05). Respondents who dropped out at the fourth test occasion had steeper linear slopes of decline than those who remained in the sample. It should be emphasized that this analysis was based on only those cases with complete longitudinal data for the first three time periods and is likely to be a conservative estimate of the differences in change to be expected between full-information data sets and those treated with LD (complete cases). These results support the prediction that individuals with missing longitudinal data will show lower intercepts and greater decline than those with complete data.

Growth curve models. The deviance statistic for each growth curve model, distributed as chi-square, appears in Table 4. Results indicated that the best fitting models for all treatments included the intercept plus linear and quadratic slopes of age (i.e., the second-order polynomial). These models each fit significantly better than the intercept and linear slope models (Δdf = 95.28, df = 4, p < .01; Δdf = 131.02, df = 4, p < .01; Δdf = 113.89, df = 4, p < .01; and Δdf = 221.18, df = 4, p < .01, for LD, NT, OOMI, and MI, respectively), which, in turn, had better fits than intercept-only models. This pattern of results also held for models with predictors of the intercepts and slopes: better fits of the quadratic models compared with linear change models (Δdf = 123.99, df = 4, p < .01; Δdf = 143.33, df = 4, p < .01; Δdf = 155.50, df = 4, p < .01; and Δdf = 211.36, df = 4, p < .01, for LD, NT, OOMI, and MI, respectively) and better fits of the linear change models than the intercept-only models. Thus, the best fitting models were consistently the second-order polynomial models, and at this juncture these models are referred to simply as age models because they reflect age-based change.

The fixed and random effects for the models of the four data treatments without and with predictors of the intercepts and slopes appear in Table 5. Figure 1 shows the mean growth curves for each model without predictors and, in addition, the NT with predictors model, the main treatments of interest. Figure 2 shows the 95% CIs around the intercepts, linear slopes, and quadratic slopes for each of the eight models.

As shown in Figure 2, the LD model had a higher intercept than the other without predictors models. This finding is consistent with the prediction that poorer performance at the intercept would be obtained for full-information treatments (e.g., NT and MI) compared with LD because those who drop out and are not included in the LD sample tend to be among the worst performers. One other difference in the intercepts surfaced: MI was associated with a lower intercept than NT.

The change parameters (fixed effects) in Table 5 indicate that some characteristics of the growth curves were similar across treatments. They each demonstrated a general pattern of decline, and, with the exception of the quadratic slope in the OOMI treatment, all of the change parameters were significantly greater than zero, indicating reliable linear and quadratic declines. Increasingly negative slopes of decline generally characterized change in recall performance.
Despite these similarities, there were also substantive differences. Estimated declines, as indexed by the linear and quadratic coefficients shown in Table 5, varied between models. A comparison of the 95% CIs of the linear and quadratic slopes suggests differences between models in the rates of decline. For the linear slopes, similar declines were observed for LD, NT, and OOMI. The MI models had a steeper decline than the others. For the quadratic slopes, greater negative curvatures were associated with LD, NT, and MI compared with OOMI. The interaction between the linear and quadratic change coefficients for the various models can be observed in the growth curves shown in Figure 1, where, compared with LD, OOMI demonstrated less decline, NT demonstrated similar decline, and MI demonstrated greater declines.

Standard errors associated with the various models reflect uncertainty resulting from both measurement and sampling errors. As noted, the main difficulty with many traditional missing-data treatments is that standard errors are underestimated because they only include measurement error. This leads to an increased possibility of Type I error. Although the magnitudes of standard errors for various missing-data treatments are sometimes compared with LD (Allison, 2002), with the prediction that standard errors of appropriate missing-data treatments should be similar to LD, we do not make these comparisons here because the LD model we tested contained missing analysis-related data and is, therefore, not a satisfactory benchmark.

The standard errors for the age models without and with predictors appear in Table 5. For the intercepts, the SEs of the four models ranged from 0.021 to 0.025, indicating similarity among the estimates. For the linear age slopes, the range in estimates was somewhat larger (0.034–0.045) such that the standard errors

<table>
<thead>
<tr>
<th>Model</th>
<th>LD</th>
<th>NT</th>
<th>OOMI</th>
<th>MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviance</td>
<td>df</td>
<td>Deviance</td>
<td>df</td>
<td>Deviance</td>
</tr>
<tr>
<td>Without predictors of intercepts and slopes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept model</td>
<td>46,241</td>
<td>3</td>
<td>73,052</td>
<td>3</td>
</tr>
<tr>
<td>Age model</td>
<td>45,563</td>
<td>6</td>
<td>71,833</td>
<td>6</td>
</tr>
<tr>
<td>Age² model</td>
<td>45,468</td>
<td>10</td>
<td>71,703</td>
<td>10</td>
</tr>
<tr>
<td>Dropout status model</td>
<td>—</td>
<td>—</td>
<td>71,229</td>
<td>11</td>
</tr>
<tr>
<td>With predictors of intercepts and slopes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept model</td>
<td>45,169</td>
<td>9</td>
<td>70,641</td>
<td>10</td>
</tr>
<tr>
<td>Age model</td>
<td>44,762</td>
<td>18</td>
<td>70,116</td>
<td>20</td>
</tr>
<tr>
<td>Age² model</td>
<td>44,638</td>
<td>28</td>
<td>69,973</td>
<td>31</td>
</tr>
<tr>
<td>Dropout status model</td>
<td>—</td>
<td>—</td>
<td>69,801</td>
<td>32</td>
</tr>
</tbody>
</table>

Note.  
LD = listwise deletion; NT = no treatment; OOMI = outcome only multiple imputation; MI = multiple imputation. Dashes indicate a model that was not examined.

Table 5

Fixed Effects (β), Standard Errors, and Random Variance for the Age Models, Without and With Additional Predictors of the Intercept and Slopes

<table>
<thead>
<tr>
<th>Model</th>
<th>β</th>
<th>SE</th>
<th>Random variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age models without predictors of intercept and slopes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>5.048**</td>
<td>4.678**</td>
<td>4.627**</td>
</tr>
<tr>
<td>Age slope</td>
<td>−0.736**</td>
<td>−0.780**</td>
<td>−0.845**</td>
</tr>
<tr>
<td>Age² slope</td>
<td>−0.219**</td>
<td>−0.225**</td>
<td>0.005</td>
</tr>
<tr>
<td>Within-subjects</td>
<td>1.539</td>
<td>1.563</td>
<td>1.579</td>
</tr>
<tr>
<td>Age models with predictors of intercept and slopes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>4.971**</td>
<td>4.504**</td>
<td>4.530**</td>
</tr>
<tr>
<td>Age slope</td>
<td>−0.481**</td>
<td>−0.427**</td>
<td>−0.592**</td>
</tr>
<tr>
<td>Age² slope</td>
<td>−0.845**</td>
<td>−0.685**</td>
<td>−0.238**</td>
</tr>
<tr>
<td>Within-subjects</td>
<td>1.527</td>
<td>1.563</td>
<td>1.570</td>
</tr>
</tbody>
</table>

Note.  
LD = listwise deletion; NT = no treatment; OOMI = outcome-only multiple imputation; MI = multiple imputation.

*p < .05.  **p < .01.
LD and MI appear to be larger than those of the NT and OOMI models. The quadratic slope standard errors also had a wider range of values (0.028–0.041), and they showed that the NT and MI models fell between the largest (LD) and the smallest (OOMI) estimates.

As shown in Figure 2, the addition of predictors to each missing-data treatment had little effect on the intercepts, with the exception of NT, which was associated with a lower intercept compared with its associated without-predictors model. This suggests further bias reduction in that intercept. Differences in the rate of declines were found with the addition of predictors. As shown in Figure 3, compared with their referent without-predictors models, estimated declines were larger for each model. However, Figure 2 suggests that for LD, NT, OOMI, and MI, the larger declines of the with-predictors treatments was the result of larger quadratic coefficients, whereas linear declines were either unchanged (MI) or smaller (LD, NT, OOMI). Still, the addition of predictors acted to increase overall decline estimates.

Of the with-predictors models, NT with predictors is of greatest interest because it is most similar to Graham’s FIML with saturated-correlates approach (Graham, 2003; Graham & Hofer, 2000). Figure 1 shows that average decline was greater for NT with predictors compared with LD. However, an examination of the 95% CIs in Figure 2 indicates that a larger quadratic function was primarily responsible for this difference. The finding that added predictors to the intercept and slopes increased decline in the estimated growth curves suggests that attrition bias was accounted for. This is important because it demonstrates real differences in growth curve models fit without and with predictors and suggests that models fit without predictors may be subject to greater attrition bias.

Although both NT with predictors and MI produced greater estimated declines than LD, a comparison of the two curves in Figure 1 suggests that NT with predictors had larger declines than MI, and this difference is the result of a larger quadratic coefficient (see Figure 2). This is evidence that the hypothesis that NT with predictors and MI would produce similar growth curves was not fully supported, although one could argue that there was a general consistency between the two treatments.

As shown in Table 5, the standard errors of the with-predictor models tended to be larger than those of the without-predictor models, suggesting greater uncertainty as a result of greater model complexity. All of the with-predictor models produced similar standard errors for the intercepts, and the NT and MI models produced similar standard errors for the linear and quadratic slopes.

Random variance is variance that is not accounted for by the fixed parameters. The random effects appear in Table 5 for each model. Each of the four without-predictors models was associated with significant estimates of random variance in all of the parameters estimated. Even when the predictors were added to each of the models, there remained significant random variance, although it was often reduced in comparison to the without-predictors models. Reductions in random variance were found in the intercepts and slopes of age of all models. This suggests that the predictors were accounting for attrition-related bias. However, random variance estimates were largely unchanged for the quadratic slopes in all models. Also the addition of predictors made only modest reductions in within-subjects random variance estimates (sigma-squared).

Evaluation of missing-data treatments. This discussion focuses on the main treatments of interest: NT, NT with predictors, OOMI, and MI. Because true average change in recall is unknowable when data are missing, one can only speculate about which curves (treatments) best represent change. That said, we suggested earlier that bias-reducing treatments should be associated with lower intercepts and greater change in recall performance than LD.

As noted, there were similarities among the four main treatments. They were each associated with lower intercepts compared with LD, increasingly negative slopes of decline in recall perfor-
mance, and significant random variance in the intercepts and slopes. Differences for models without predictors were found in the magnitude of decline observed, with MI showing greater declines than the other treatments, most notably LD, suggesting that the addition of independent predictors helped to reduce attrition-related bias in that treatment. Thus, the prediction that MI would demonstrate greater declines than LD was confirmed. Although NT was associated with a lower intercept and, therefore, poorer overall performance, the trajectory of change paralleled that of LD, suggesting that bias reduction was achieved only for the intercept. This implies that the simple application of a growth curve model without either performing MI before modeling or adding independent predictors to the model is ill advised.

The addition of predictors to the intercepts and slopes of the growth curve models increased estimated declines in all models tested, suggesting that variables added at the analysis stage reduced attrition bias. In particular, the addition of predictors to the NT model dramatically increased estimated declines differentiating that model from LD, thus confirming the hypothesis that the FIML missing-data approach would result in additional bias reduction when predictors were added.

Yet, although both proper treatments (NT with predictors and MI) met the prediction of greater declines than LD, NT with predictors was associated with steeper declines than MI. Thus, the suggestion that the two treatments would produce similar results was not fully met.

The implications here are unclear; however, we consider two possible explanations for the obtained differences. First, it is certainly the case that the reported differences are somewhat exaggerated by the large age range included, especially at the upper reaches of the life span for which data are scarce. Although larger quadratic coefficients (e.g., NT with predictors) demonstrate greater declines than models associated with smaller coefficients (e.g., MI), these differences often do not appear until late in the curve (life span). Indeed, a comparison of the right portion of the NT with predictors and MI curves in Figure 1 showed that the two curves begin to separate only in the early 90s. Thus, a reasonable conclusion is that the two treatments produce roughly similar estimates over much of the age range.

A second possibility is that NT with predictors relies more heavily on the convergence assumption than MI and is, therefore, more strongly an affected by cohort effect, which would be expected to increase declines. The MI treatment is less dependent on the convergence assumption because it involved imputation of missing data in the four-occasion extant data. Thus, attrition-associated missing data were dealt with in a manner that did not rely on the assumption. In contrast, all missing data of NT with
predictors were treated during the data-modeling stage, which was done in an age-based model that relied on the convergence assumption. The difficulty with this explanation is that it is not known whether cohort effects exist in age-based change estimates of recall performance in the AHEAD sample. This possibility could be examined by comparing trajectories of change in replacement participants who represent younger cohorts with that of the original sample.

The guiding principles applied to choose between appropriate and inappropriate treatments were that appropriate treatments should demonstrate lower intercepts and greater decline than LD. The NT with predictors, MI, and MI with predictors treatments each met these criteria. Thus, the two proper treatments (NT with predictors and MI) met our criteria, and we consider them to be appropriate missing-data interventions for these data. Our results suggest that one’s choice of which treatment to apply is largely a matter of preference. Certainly, NT with predictors is less labor intensive than MI because the data-treatment stage is eliminated (at least for attrition bias), and because multiple data sets do not have to be analyzed and results do not have to be collapsed. However, the NT with predictors approach has the disadvantage that attrition reduction is always model dependent, whereas once MI has been performed, various models can be examined without having to be concerned about attrition bias reduction for each model and possible differences between models.

The MI with predictors model met our initial criteria for appropriate missing-data treatments by having a lower intercept and greater decline than LD. In fact, the MI with predictors model is appropriate missing-data treatments by having a lower intercept and average decline compared with LD. In addition, the MI with predictors model is more appropriate than MI because the data-treatment stage is eliminated (at least for attrition bias), and because multiple data sets do not have to be analyzed and results do not have to be collapsed. However, the NT with predictors approach has the disadvantage that attrition reduction is always model dependent, whereas once MI has been performed, various models can be examined without having to be concerned about attrition bias reduction for each model and possible differences between models.

We initially predicted that similar results would be associated with the NT and OOMI models. Yet NT was associated with larger declines than OOMI, suggesting that simple application of the FIML algorithm to treat missing longitudinal data may be more effective than OOMI. However, neither of these treatments is recommended because they each produced growth curves and parameter estimates similar to LD.

Study 2

Study 2 was conducted with the goal of determining whether additional bias associated with attrition could be accounted for in pattern-mixture growth curve models (Singer et al., 2003; Sliwinski et al., 2003). Several researchers have argued that selectivity effects must be directly modeled when identifying age trajectories because it is not possible to accurately model the effects of aging on memory or cognition at the upper reaches of the life span without consideration of health or biological status (Rabbit et al., 2002; Singer et al., 2003; Sliwinski et al., 2003). Following Hedeker and Gibbons’s (1997) general approach, the effect of dropout status was evaluated to examine these hypotheses. If dropout status improves model fit relative to a model with parameters only associated with age change, it suggests that dropout is an important variable when modeling longitudinal change. It was expected that dropout models would produce higher average intercepts and smaller average declines compared with age models that do not include dropout status. It was also predicted that dropout status would be a significant predictor of performance: that its inclusion would improve model fits and reduce random variance estimates. In addition, in Study 2, we report separate growth curve models for dropouts and continuers, with the expectation that models associated with only dropouts will demonstrate greater declines than those of continuers.

Method

Participants, measures, and missing-data treatments. The models reported are extensions of the age models associated with the NT, OOMI, and MI missing-data treatments of Study 1. The LD sample and its associated models were not included in these analyses because the sample included no missing data. The participants, measures, and missing-data treatments applied in Study 2 are each described in detail in Study 1.

Pattern-mixture growth curve models. Pattern-mixture models (Hedeker & Gibbons, 1997) in this instance are multilevel growth curve models that include within-subjects predictors that have different configurations than those associated with age. That is, dropout status (continuer vs. dropout) was included in the within-person level equation. These new models fit each of the missing-data treatments had the form

\[ y_{ij} = \pi_0 + \pi_1 \text{Age}_i + \pi_2 \text{Age}^2_i + \pi_4 \text{dropout status}_i + \epsilon_{ij} \]

where \( y_{ij} \) was the recall score for individual \( i \) measured at age \( j \), which was equal to the intercept, \( \pi_0 \), for individual \( i \), plus that person’s slope of age, \( \pi_1 \), the slope of age squared, \( \pi_2 \), dropout status (continue or dropout), \( \pi_4 \), and error, \( \epsilon_{ij} \). The between-persons model included fixed (\( \beta_0 \)) and random (\( U \)) effects for the intercept (\( \pi_0 \)) and change parameters (\( \pi_1, \pi_2 \)) but only a fixed effect for dropout status (\( \pi_4 \)); that is:

\[ \pi_0 = \beta_0 + U_{0i} \]
\[ \pi_1 = \beta_1 + U_{1i} \]
\[ \pi_2 = \beta_2 + U_{2i} \]
\[ \pi_4 = \beta_4 \]

Random effects (\( U \)) were not calculated for dropout status because there were insufficient degrees of freedom available. By keeping the rest of the model identical to the age models, it allowed for unconfounded model comparisons with those models. Dropout status was mean centered for all models reported.

Results and Discussion

As shown in Table 4, the dropout status models were associated with improved fits compared with the age models of Study 1 for the without-predictors models (\( \Delta d = 473.02, df = 1, p < .01 \); \( \Delta d = 394.95, df = 1, p < .01 \); and \( \Delta d = 835.54, df = 1, p < .01 \)) for NT, OOMI, and MI, respectively) and for the with-predictors models (\( \Delta d = 171.36, df = 1, p < .01 \); \( \Delta d = 127.45, df = 1, p < .01 \); and \( \Delta d = 369.54, df = 1, p < .01 \)) for NT, OOMI, and MI, respectively. Thus, the addition of dropout status generally improved model fit.

The results of the dropout status models appear in Table 6 and the associated growth curves in Figure 3. These results are compared with the age models presented in Study 1 and shown in Table 5. The dropout status models, compared with the age mod-
Table 6: Fixed Effects (B) With 95% Confidence Interval (in Parentheses), Standard Errors, and Random Variance for the Dropout Status Models, Without and With Additional Predictors of the Intercept and Slopes

<table>
<thead>
<tr>
<th>Variable</th>
<th>NT</th>
<th>OOMI</th>
<th>MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.737** (4.697, 4.777)</td>
<td>4.930** (4.883, 4.977)</td>
<td>4.512** (4.469, 4.555)</td>
</tr>
<tr>
<td>Age slope</td>
<td>0.766** (0.836, 0.697)</td>
<td>0.767** (0.833, 0.702)</td>
<td>1.040, 0.876)</td>
</tr>
<tr>
<td>Age2 slope</td>
<td>0.172** (0.235, 0.108)</td>
<td>0.161** (0.230, 0.092)</td>
<td>0.255, 0.190</td>
</tr>
<tr>
<td>Dropout</td>
<td>0.813** (0.886, 0.740)</td>
<td>0.715** (0.796, 0.634)</td>
<td>0.634, 0.427)</td>
</tr>
</tbody>
</table>

**Note.** NT = no treatment; OOMI = multiple imputation. Dashes indicate parameters that were not estimated.

In sum, pattern-mixture growth models, which included dropout status, were associated with significant dropout status parameters and better fits but similar age change estimates compared with their associated age models. Dropouts recalled fewer items on average than continuers. The lack of strong effects on estimated average growth curves suggests that dropout effects may be small for age-based estimates, especially in models that include predictors of missingness (e.g., NT with predictors, MI) in treatment of missing data. However, the fact that dropout status coefficients were consistently significant across models suggests that dropout effects varied systematically with the type of treatment received.

For the dropout status with predictors models, declines, as estimated in the linear and quadratic slopes, were similar to those of the age with predictors models. Again, the only significant difference observed was in the intercept of the OOMI dropout model, which had a higher intercept than its associated age with predictors model. Only small effects on the age-based growth estimates were observed; however, dropout status coefficients were consistently significant (see Table 6) such that dropouts recalled fewer items on average than continuers. Here the advantage for continuers ranged across models from about one half of an item to nearly nine tenths of an item. Also, inclusion of dropout status generally increased the random variance accounted for. For NT, the dropout model had a decrease in variance in the intercept of 3.31%, a decrease in the slope of age of 2.07%, but an increase in variance in the slope of age squared of 1.37%. Compared with the age with predictors model, OOMI with dropout and predictors was associated with slightly smaller variance estimates in the intercept (3.48%), slope of age (0.40%), and slope of age squared (2.16%). In the MI treatment, the dropout status model was associated with less random variance in the intercept (7.83%), slope of age (2.62%), and slope of age squared (7.16%) than the age model. Although change estimates (fixed effects) were not significantly affected by the addition of dropout status, its inclusion differentiated groups of participants, increased the variance accounted for, and improved model fit.
effects were present in the age models. These results point to the importance of considering dropout effects. As Hedeker and Gibbons (1997) note, variables that reflect much more complex patterns of missingness are also possible. For example, Sliwinski et al. (2003) have demonstrated the effectiveness of including time since death or time since dropout as time-varying covariates in pattern-mixture models.

Our strategy thus far for studying age trajectories in recall has been to include data of both continuers and dropouts, making use of nearly all of the available information. An alternative strategy is to examine change in dropouts separately from continuers (i.e., LD). This approach shows the differences between groups.

We ran separate age models for dropouts with NT, NT with predictors, MI, or MI with predictors. Because dropout effects are known to negatively affect memory performance, we made the following predictions: Dropouts will be associated with lower intercepts than continuers (LD), and dropouts will be associated with greater declines than continuers. To this end, we analyzed and compared the growth curves of dropouts (N = 3,095). Models were compared with the LD model described in Study 1. The deviance scores for the data samples of the dropouts were 22,800; 22,249; 46,476; and 45,606; respectively, for NT, NT with predictors, MI, and MI with predictors. The addition of predictors improved model fits for NT (Δdf = 551, df = 18, p < .01) and for MI (Δdf = 870, df = 18, p < .01).

The fixed effects of the models examined here appear in Table 7, and these are compared with the LD model discussed previously in Study 1 with results presented in Table 5. The average growth curves associated with all of the relevant models appear in Figure 4. Figure 4 shows that all of the models were associated with lower intercepts compared with LD, and this observation was confirmed by the finding of nonoverlapping intercepts of the four dropout models compared with LD. For change estimates, however, greater declines were observed for NT with predictors, MI with predictors, and possibly MI compared with LD. NT demonstrated less decline than LD. However, only some of these observations were supported by comparing the growth curve coefficients with 95% CIs.

For the linear slopes, MI and MI with predictors both had steeper declines than LD. Neither NT nor NT with predictors demonstrated different linear declines than LD. Similarly, for the quadratic slopes, both MI and MI with predictors had larger negative coefficients than LD, whereas NT and NT with predictors were not different. Thus, the initial observation that NT showed a larger decline than LD was not confirmed by comparing the 95% CIs. This discrepancy may exist for two reasons. First, the changes demonstrated in the figure are the product of the combined linear and quadratic coefficients, which interact to create the curve. The coefficient comparisons do not capture this interaction. Second, the standard errors and, therefore, the 95% CIs associated with the NT models were large and larger than those of MI, reducing the likelihood of obtaining significant differences. Thus, strong inferences about the NT and NT with predictors models are not possible.

The prediction that data sets consisting of only dropouts’ data would have lower intercepts than LD was supported. This suggests that, at the intercept, average performance is lower for dropouts compared with continuers. Yet the prediction that dropout models would be associated with steeper declines than LD only received clear support from the MI models, whereas for the NT models support was tenuous. One methodological difference that may explain these discrepant results is that the MI dropouts were drawn from the previously imputed full-information data described in Study 1, whereas the NT models were associated with reduced sample size data sets. For the NT models, the reduced sample size and lack of imputed longitudinal data may have acted to increase standard errors and reduce statistical power.

General Discussion

Summary of Missing Data Treatments

The results of the two studies demonstrated similarities in the results of the four main missing-data treatments (NT, NT with predictors, OOMI, and MI): (a) lower intercepts compared with...
LD, (b) increasingly negative slopes of decline in recall performance, (c) significant random variance in the intercepts and slopes, and (d) decreased random variance when variance associated with dropout was accounted for. More important, both of the methodologically accepted treatments (NT with predictors and MI) demonstrated lower intercepts and steeper slopes of decline than LD, suggesting that attrition bias was reduced. In addition, both treatments were associated with significant random variance in their intercepts and change parameters. Despite the similarities between the two treatments, NT with predictors produced greater average declines than MI. The possible underlying reasons for this difference were discussed, and we concluded that either approach met our criteria of greater declines than LD, and neither one could be ruled out by other criteria. Thus, the results of Study 1 were that both NT with predictors (i.e., the FIML approach) and MI produced theoretically acceptable results. One main implication is that missing-data treatments, if they are to be effective, need to include additional predictors. When predictors were not included, as in the results of NT and OOMI, results did not meet criteria.

Generalization of our results to other measures, domains, and data samples should be applied with caution. We have only explored differences between missing-data treatments within the context of a single cognitive measure, a large representative sample, and with multilevel growth curve models. Yet we have provided some guidelines for how this might be done. Careful examination of the measures, assessment of missing-data patterns, and assessment of the predictors of missingness are important. Comparing the results of different missing-data treatments to discover which treatments provide theoretically acceptable estimates is also recommended. We have demonstrated the possible advantages of establishing a set of theoretically and empirically defensible assumptions, which can then be examined with various missing-data treatments.

Study 2 demonstrated some benefits of directly estimating dropout effects. Doing so provided better model fits because dropout status predicted performance, resulting in smaller random variance estimates. Unfortunately, the results of models that included sub-groupings of dropouts or continuers were only definitive at the intercept. Models of dropouts showed greater declines than continuers only for the MI models but not for the NT models, suggesting that estimation of missing data with more impoverished longitudinal data and smaller sample sizes may affect statistical power. Still, findings associated with the MI models suggest that dropouts decline faster than continuers.

These studies, taken together, suggest a two-pronged approach to dealing with attrition-related bias. First, proper treatment of missing data is crucial to resulting change estimates. This can take the form of MI or FIML (NT with predictors) treatments as long as predictors are included in the applied treatment. Second, dropout effects may also be modeled directly in pattern-mixture growth curve models. These approaches, applied together, should help to maximize reduction of attrition bias. However, further remedies are needed to reduce additional forms of bias. These include modeling practice effects, controlling for regression toward the mean, and greater reliance on nonlinear modeling.

Summary of Change in Recall Performance and the Effects of Death and Dropout

The current array of findings consistently showed that recall performance declines with age. Yet differences in the magnitude of change observed were found between missing-data treatments (Study 1) and between dropouts and continuers (Study 2). Although beyond the scope of the present study, further attention should be paid to the issues of missing-data treatment and death and dropout effects and possible interactions between the two. Longitudinal samples such as AHEAD may provide an opportunity to study these issues in greater depth if participants are monitored across the entire life span.
As seen in Tables 2 and 3, there are some important predictors of recall performance as well as dropout status in the demographic variables. Respondents who had complete data were substantially superior in initial recall, with mean recall differences at the first testing relative to those dropping out ranging from 1.36 to 0.96 words, or about 0.5 SD. One possibility for their superiority on recall is that they have not yet succumbed to the effects of poor health that often precede dropout and death. That is, they are predominantly the remaining portion of the original sample who are still in good health and functionally intact. However, approximately 20% can be expected to drop out by the fifth test occasion, and those should be more similar to other dropouts than to those remaining in the study. Data from those as yet unidentified dropouts should increase the observed declines in the fourth test occasion, but currently their data cannot be discriminated from those of participants who will continue to the fifth testing and beyond.

Some support for this was obtained by comparing the magnitude of annual declines associated with a respondent’s final test completed (i.e., the third test occasion for people who dropped out at the fourth test occasion and the fourth test occasion for people who dropped out at the fifth test occasion), but currently their data cannot be discriminated from those remaining in the study. Significant annual declines were observed in the last testing completed compared with the previous testing for those who dropped out at third test occasion (annual decline = 0.10 items), t(1000) = 3.53, p < .01, those who dropped out at fourth test occasion (annual decline = 0.19 items), t(889) = 9.71, p < .01, and those who remained in the study (annual decline = 0.20 items), t(3210) = 12.67, p < .01. The declines observed in the last testing completed were much steeper than those observed in the penultimate testing for those who dropped out at the third test occasion (0.07 annual decline in the second test occasion), F(1, 889) = 8.55, MSE = 0.791, p < .01, and those who remained in the study (0.08 annual decline in the third test occasion), F(1, 3122) = 80.15, MSE = 1.198, p < .01. This suggests the effects of poor health and other factors manifest as large declines as dropout or death approaches. As suggested, a way to confirm this would be to monitor the entire sample, or a large portion of the sample, through death and use dynamic models that allow for direct identification of differences between groups in change (McArdle & Hamagami, 2001).

The death rates of those who drop out may be a key underlying variable that separates the dropouts from retesters. More than 75% of those who only participated in the initial testing died over the 7-year period of the study. It suggests that studies comparing older cohorts with full participation in longitudinal retests may show effects that may not be as clearly associated with social and historical phenomena driving outcomes as they are with survival. That said, as survival rates increase in current cohorts, better performance may be observed as a cohort effect. Alternatively, it is possible that longer life spans in future cohorts may serve to reduce the effect of survival and to increase age declines (Singer et al., 2003).

Nevertheless, average age declines were observed in the present study in the majority of models examined. Across the age range of 70 through very old age, change was not uniform (linear) but rather reflected steeper slopes of decline with older age. Whether these effects occur with latent variables, with multiple measures of memory and other cognitive constructs, and in the longitudinal convenience samples that are the norm for memory research remains to be seen.

References


