Journal of Gerontology: PSYCHOLOGICAL SCIENCES
2004, Vol. 59B, No. 2, P84–P97

Practice and Drop-Out Effects During a 17-Year Longitudinal Study of Cognitive Aging

Patrick Rabbitt,[1] Peter Diggle,[2] Fiona Holland,[2] and Lynn McInnes[3]

[1]Age and Cognitive Performance Research Centre, University of Manchester, England.
[2]Department of Biomathematics and Statistics, University of Lancaster, England.
[3]Department of Psychology, University of Northumbria, England.

Interpretations of longitudinal studies of cognitive aging are misleading unless effects of practice and selective drop-out are considered. A random effects model taking practice and drop-out into account analyzed data from four successive presentations of each of two intelligence tests, two vocabulary tests, and two verbal memory tests during a 17-year longitudinal study of 5,899 community residents whose ages ranged from 49 to 92 years. On intelligence tests, substantial practice effects counteracted true declines observed over 3 to 5 years of aging and remained significant even with intervals of 7 years between successive assessments. Adjustment for practice and drop-out revealed accelerating declines in fluid intelligence and cumulative learning, linear declines in verbal free recall, and no substantial change in vocabulary. Socioeconomic status and basal levels of general fluid ability did not affect rates of decline. After further adjustment for demographics, variability between individuals was seen to increase as the sample aged.

MANY excellent longitudinal studies of cognitive change have had the same basic aims. The most general has been to determine the average form of the trajectory of age-related change and, in particular, whether or not the average rates of change accelerate in old age (e.g., Hertzog & Schaie, 1988; Rabbitt, 1993a; Schaie & Strother, 1968). A corollary aim has been to determine whether rates of change differ between different mental abilities or are similar for all (e.g., Arenberg 1974; Colsher & Wallace, 1991; Heron & Chown, 1967; Hertzog & Schaie, 1988; Hultsch, Hertzog, Small, McDonald-Miszczak, & Dixon 1992; Johansson, Zarit, & Berg, 1992; Lansen, 1997; Owens, 1953, 1966; Powell, 1994; Rabbitt, 1993a; Schaie, 1996; Schaie & Labouvie-Vief, 1974; Schaie & Strother, 1968; Schaie & Willis, 1993; Terman & Oden, 1947, 1959). A third aim has been to test how rates of cognitive change are affected by demographic factors such as educational and social advantage (e.g., Bosworth, Schaie, & Willis, 1999; Evans et al., 1993; Forner, 1972), by gender (e.g., Bosworth et al., 1999; Voitenko, & Tokar, 1983), by epidemiological factors such as general health (e.g., Bell, Rose, & Damon, 1972; Birren, Butler, Greenhouse, Sokoloff, & Yarrow, 1963; Costa & McCrae, 1980; McInnes & Rabbitt, 1997; Rabbitt, Bent, & McInnes, 1997), by specific pathologies (e.g., Hertzog, Schaie, & Gribbin, 1978), by maintenance of physical mobility and engagement in everyday physical activities (e.g., Clark, 1960; Clement, 1974; Dirken, 1972; McInnes & Rabbitt, 1997), or by genetic factors (e.g., Bank & Jarvik, 1978; Payton et al., 2003; Pendleton et al., 2002; Terman & Oden, 1947, 1959). This raises the general issue of the extent and etiology of individual differences in trajectories of aging. Prima facie, because individuals are affected in different ways and to different extents by their lifestyles, health histories, and genetic factors, we might expect that their trajectories of aging correspondingly diverge so that variance in performance between members of a sample will increase as the members age (Morse, 1993; Rabbitt, 1982, 1993a). It follows that individual differences in rates of change provide more information about the functional determinants of cognitive aging than do average trajectories of decline.

Achievement of these general aims has been frustrated by persistent methodological problems. One issue is that analyses have simply regressed performance data across successive time–assay points. Neglect of longitudinal correlations in the data can lead to incorrect inference. A second issue is that when participants are repeatedly assessed on the same or similar tasks, improvements with practice may lead to underestimates of true rates of change and, in particular, may disguise an age-related acceleration of rate of decline. Further, if participants improve more on some tasks than others, analyses may incorrectly conclude that the particular mental abilities that support these tasks decline at different rates. Useful discussions and some empirical investigations (e.g., Zelinski & Burnight, 1997; Zelinski, Gilewski, & Stewart, 1993; Zelinski & Stewart 1998) suggest that, because patterns of correlations between scores on different tests remain stable across successive assessments, improvements are similar across tasks and thus do not mimic or mask differences in the rates of decline of different mental abilities. This hope has been vitiated by cross-sectional studies showing that practice effects vary with complex interactions between individuals' overall levels of general fluid mental ability and the particular kinds of tasks on which they are compared. Less able and older individuals show greater initial and overall improvements on easy tasks (Rabbitt, 1993b), but on difficult tasks the more able and younger show much greater immediate and sustained gains (Rabbitt, Banerji, & Szemanski, 1989). Similar interactions among the effects of practice, of individual differences in ability, and of task difficulty in longitudinal data would conceal age-related declines on simple tasks on which less able older individuals show relatively greater improvements and exaggerate apparent declines on difficult tasks on which they show relatively smaller improvements. Unless analyses determine the relative sizes of

practice effects both for different tasks and for older and younger and more and less able individuals, they will incorrectly estimate true rates of overall age-related decline and may also misleadingly suggest that performance declines more rapidly on some tasks than on others.

A third, well-documented, but incompletely resolved methodological problem has been that older, frailer, and less able participants drop out of longitudinal studies earlier than the younger, healthier, and more able. Thus, successive data points reflect the performance of a progressively more elite subset of the original sample, and the true extent of cognitive changes is disguised (e.g., Baltes, 1968; Forner, 1972; Lachman, Lachman, & Taylor, 1982; Lindenberger, Singer, & Baltes, 2002; Mason & Mason, 1973; Nesselrode & Baltes, 1979; Palmore, 1978; Schaie, Labouvie, & Barrett, 1973; Schlesselman, 1973a, 1973b; Schulsinger, Knop, & Mednick, 1981).

Parenthetically, a typical practice in longitudinal investigations is not to recruit a single sample of participants who are thereafter followed until the study ends but rather to continue to recruit new waves of participants, at least throughout the early years. The possibility that the cohorts recruited in successive waves may differ from each other both in demographics and in overall levels of ability gives rise to a corollary, and largely unexplored, methodological problem that may be termed the "drop-in effect." Unless analyses take recruitment cohort differences into account, estimates of rates of cognitive decline will be misleading, especially if cohorts differ more on performance of some tasks than of others.

Apart from obvious age differences, participants who withdraw early from longitudinal studies tend to have poor levels of general health, education, and socioeconomic advantage. Men also tend to drop out earlier than women (Rabbitt, Watson, Donlan, Bent, & McInnes, 1994). Such trends can lead to complex misinterpretations. For example, because women tend to perform better than men on some verbal learning tasks (Rabbitt, Donlan, Watson, McInnes, & Bent, 1996; Rabbit et al., 2002), the rates at which verbal learning declines with age may be underestimated unless gender differences in drop-out are taken into consideration.

Some investigators have estimated the effects of selective drop-out by comparing patterns of differences between age groups observed in initial cross-sectional screenings of a volunteer population against the patterns of age-related changes that become apparent as longitudinal data are accumulated. Because patterns of age-related differences revealed by cross-sectional and longitudinal comparisons seem very similar, investigators have concluded that selective drop-out may not always lead to serious misinterpretations (e.g., Sliwinski & Bushke, 1999; Zelinski & Burnight, 1997; Zelinski et al., 1993; Zelinski & Stewart, 1998). The comparison of cross-sectional against longitudinal trends is a useful exploratory step that can tell us whether drop-out affects the relative amount of change in different cognitive abilities, but it does not reveal the extent to which drop-out has masked the actual amount of changes. In particular, such comparisons do not show whether substantial progressive increases in variability between members of an aging population have been masked by selective withdrawal of the oldest and less able. We believe that only longitudinal studies can properly address all of the effects of drop-out on population changes in performance over time. Thus, a main aim of the analyses described here was to model changes over time to take account of drop-out effects.

To consider how this may be done, we find it important to distinguish among three different drop-out scenarios (Lindenberger, 2002; Rubin, 1976). The first is the completely random drop-out: The drop-out process is independent of the measurement process. The second is the random drop-out: The drop-out process is dependent on the observed measurements prior to drop-out but is independent of the measurements that would have been observed had the participant not withdrawn. The third is the informative drop-out: The drop-out process is dependent on the measurements that would have been observed had the participant not dropped out. Not surprisingly, analyses made under the informative drop-out assumption are fraught with difficulty. The results of such analyses typically depend on modeling assumptions that are difficult or impossible to check from observed data. For example, in most observational studies it is extremely difficult even to identify the precise time at which a participant made a decision to drop out. In contrast, analyses under the assumption of completely random drop-out are generally straightforward because no distinction need be made between measurements that are unavailable because of drop-out and those that are unavailable because they were never intended. Put another way, completely random drop-out implies that the incomplete data can simply be treated as if from an unbalanced experimental design, with no commonality to the times at which measurements are made on different subjects.

The simplicity of analysis under the completely random drop-out assumption is bought at a price. If this assumption is invalid, then so may be the resulting inferences about the measurement process. However, if likelihood-based methods of inference are used, validity is retained under the weaker assumption of random drop-out. This is important because longitudinal data are typically correlated over time. This means that even when the true drop-out process is informative, the most recent measurements on a given subject before drop-out are partially predictive of the missing measurements after drop-out. By allowing for the effects of these measurements on drop-out (which is what the random drop-out assumption implies), we can partly compensate for the missing information (see, e.g., Scharfstein, Rotnitzky, & Robins 1999, and the associated discussion). To appreciate how the likelihood-based methods automatically make this kind of compensation, Diggle, Liang, and Zeger (1994, chap. 11) showed a simulated data set from a model in which the mean response is constant over time, but the probability of drop-out for any given subject at any given time is a decreasing function of that subject's most recent measurement. The effect of this random drop-out mechanism is that low-responding subjects progressively drop out, leading to an apparent rising trend in the mean response over time as the observed mean is calculated from the progressively more selective subpopulation of survivors. This rising trend is what would be estimated by a naive regression analysis of the data that ignores both the drop-out process and the longitudinal correlation in the data.

An important implication is that, for longitudinal data with drop-out, there is no reason why a fitted mean response curve should track the observed mean response trajectory of the survivors. In contrast, a model fitted by likelihood-based

Table 1. Schedule of Recruitment and Testing of Volunteers (Newcastle)

Wave	TB 1 Administered (1982–1985)	TB 1 (1985–1986)	Robbins & Sahakian Battery (1985–1986)	TB 2 (1985–1991)	TB 1 (1991–1992)	Currently Registered
Wave 1–2 (1982–1985)	2,052 (first test)	1,578 (retest 1)	828 (first test)	1,167 (first test)	977 (second test)	977
Wave 3 (1988)		629 (first test)		492 (first test)	425 (second test)	425
Wave 4 (1989–1990)				607 (first test)	601 (first test)	601
Wave 5 (1991–1992)					67 (first test)	67

Notes: TB = test battery. Total tested on at least one battery = Waves 1 and 2 (2,048) = Wave 3 (629) + Wave 4 (601) + Wave 5 (67) = 3,345. Total retested once on Battery 1 = Wave 1 (1,578) + Wave 2 (425) = 2,003. Total tested at least once on both Battery 1 and Battery 2 = Wave 1 (1,167) + Wave 2 (492) = 1,659. Total still registered on panel = 2,070.

methods under the assumption of random drop-outs estimates what the mean response would have been if it had been possible to follow up the entire study population. That is, the observed means are estimating the mean response conditional on not dropping out before the end of a determined census period. The unconditional and conditional means coincide only if the data are uncorrelated in time, or if the drop-out process is completely random. We would argue that neither of these assumptions is plausible for most data from longitudinal studies of cognitive aging. We therefore conclude that a likelihood-based analysis under a random drop-out assumption is a sensible analytic strategy because it focuses on the complete study population, rather than on the progressively self-selected subpopulation of subjects who do not drop out of the study. We used this analysis to examine the data set described in the paragraphs that follow.

The present study was made to investigate the extent to which improvements associated with repeated testing and by selective drop-out and drop-in effects have obscured answers to the basic questions about the nature of age-related cognitive changes. The analyses also addressed substantive hypotheses on the true relationships that would be revealed when practice effects have been identified and drop-out has been taken into consideration.

First, practice effects would be found to be substantial and large enough to disguise the true rates and forms of trajectories of longitudinal changes. Second, the sizes of practice effects would be found to differ both between different kinds of tasks and also between more and less intellectually able, and so implicitly younger and older, participants (as has been found in brief, cross-sectional laboratory studies by Rabbitt, 1993b, and Rabbitt et al., 1989). Third, when the true forms of trajectories of age-related changes can be established, rates of cognitive change will be found to accelerate with age and to differ with task demands. It will also be possible to more accurately determine the extents to which individual differences in rates of change vary with demographic variables such as gender and socioeconomic status and with individual differences in level of general fluid mental ability (gf). Fourth, after the effects of demographics, gender, and individual differences in ability have been taken into consideration, variance in cognitive performance between participants will be found to increase significantly as the sample ages.

To examine the effects of practice, we used a drop-out and drop-in random effects model to analyze data from successive presentations of the same battery of six different tasks during a 17-year longitudinal study of 5,899 healthy, community resident, older people. The model allowed us to determine the true sizes of practice effects, the extent to which practice effects differ between tasks, and the extent to which practice effects and Task × Practice interactions differed between individuals of different ages and levels of gf.

Methods

Participants

Details of recruitment of the sample are given elsewhere (Rabbitt, Donlan, Bent, McInnes, & Abson, 1993). The data analyzed here were obtained from 5,899 active community residents from Newcastle-upon-Tyne ($n = 3,261$) and Greater Manchester ($n = 2,638$) who ranged in age from 49 to 92 years on entry. Years of education varied from 5 to 21 ($M = 11.7$, $SD = 4.2$). Numbers of health complaints recorded on the Cornell Medical Index (Brodman, Erdman, & Wolff, 1949) varied from 2 to 18 ($M = 6.4$, $SD = 3.2$). Table 1 shows the time pattern of successive waves of recruitment and the subsequent retesting schedule by city of residence. Participants experienced the battery between one and four times, depending on when they were recruited and whether, and when, they dropped out. Tables 1 and 2 detail the patterns of successive waves of recruitment and the extent of drop-out in Newcastle and Manchester.

The index of socioeconomic advantage was the categories for Classification of Occupation published by the (U.K.) Office of Population Censuses & Surveys (1980).

Procedure

Volunteers traveled independently to laboratories in Manchester or Newcastle-upon-Tyne, where they were tested in groups of 10 to 15. Sessions were conducted in large quiet rooms by two experimenters who checked that participants with visual or auditory problems had brought their prescribed prostheses and were not inconvenienced. Sessions lasted, on average, for 90 min with 15-min tea and coffee breaks. Volunteers each received £5 (U.K.) for each session to cover their travel expenses. The tests were administered over two successive sessions within a period of 8 weeks.

Cognitive Tests

During the first testing session, volunteers completed the Heim (1970) AH4-1 and AH4-2 intelligence tests, and the Raven (1965) Mill Hill "A" and "B" (MHA and MHB) vocabulary tests. During the second session, they completed a cumulative verbal learning (CVL) task and a verbal free recall (VFR) task.

Both AH4 tests are well-standardized measures of gf and correlate strongly ($R = .7–.8$) with instruments such as the

Table 2. Schedule of Recruitment and Testing of Volunteers (Manchester)

Wave	TB 1 (1985–1988)	TB 2 (Oct. 1988–Sept. 1990)	TB 1 (Oct. 1990–Aug. 1992)	Currently Registered
Waves 1, 2, and 3 (recruited 1985–1988)	2,361	1,452	1,147	1,147
Wave 4 (recruited 1989–1990)		522 (first test)	256 (first test)	256
Wave 5 (recruited 1991–1992)			104 (first test)	104

Notes: TB = test battery; Oct. = October; Sept. = September; Aug. = August. Total tested on at least one battery = Waves 1–3 (2,361) + 522 (Wave 2) + 104 (Wave 3) = 2,992. Total retested once on Battery 1 (Waves 1–3) = 1,147. Total tested at least once on both Battery 1 and Battery 2 = Waves 1–3 (1,452) + Wave 4 (256) = 1,708. Total still registered on panel = Waves 1–3 (1,147) + Wave 4 (256) + Wave 5 (104) = 1,507.

Wechsler Adult Intelligence Scale battery (see Heim, 1970). In each test, volunteers are given untimed and unscored practice on 5 demonstration problems with provided solutions and then have 10 min to solve as many as possible out of a total of 65 problems. AH4-1 problems include equal numbers of logical reasoning tests, verbal comparisons, and arithmetic and number series problems. AH4-2 problems are nonverbal, involving addition and subtraction of complex shapes, completion of logical series of shapes, and matching by mental rotation of irregular shapes. For each AH4 test the scores analyzed are the numbers of problems correctly solved in 10 min.

The MHA recognition vocabulary test comprises a list of 33 different rare words, each accompanied by a set of 6 different words from which participants must select the most appropriate synonym. The MHB production vocabulary test comprises a list of 33 different rare words for each of which participants provide as accurate and precise a definition as possible (Raven, 1965). Both MH tests are untimed. Scores analyzed consist of the number of correct synonyms identified or definitions given.

In the CVL test, a list of 15 different three-syllable words matched for concreteness and for frequency (1/10,000, Kucera & Francis, 1967, norms) are projected, one item at a time, by a Kodak "Carousel" projector at a rate of 1.5/s. Words appear in Times Roman print, boldface, as a string that is 150 cm long and 15 cm high on a projection screen that is no further than 5 m from any member of the group tested. The room is blacked out to maximize visibility, and participants' results are not recorded unless the participants remembered to bring prescribed spectacles and have no difficulty reading the displays. After the first presentation of the 15 words, participants write down as many as they can recall, in any order they wish, on the first page of an answer booklet. They then turn to a new page and the 15 words are then presented in a new random order, and recall is again attempted. Scores analyzed are the total numbers of words correctly recalled over four such presentations.

In the VFR test, 30 words, selected and matched as for the CVL task, are presented once, in random order, at a rate of 1 item/1.5 s. Volunteers then immediately write down as many as they can recall in any order they wish. Scores analyzed are the total numbers of words correctly recalled.

Description of Test Data and Exploratory Analysis

Because the tests have total scores ranging from 30 to 65, each participant's raw scores were converted to a percentage correct to provide a common measurement scale. Mean scores at entry, broken down by the selected demographic groups, are shown in Table 3.

Choice of scores.—We made the decision to use percentage correct scores rather than standardized scores for the following reasons: Let Y_{ijk} denote the raw score on test i recorded by subject j on the kth testing occasion. Our models for these data take the generic form

$$Y_{ijk} = x'_{ijk}\beta + A_{ij} + B_{ij}t_{jk} + E_{ijk}, \quad (1)$$

in which x_{ijk} is a vector of explanatory variables, with associated regression parameters β; A_{ij} and B_{ij} are subject-specific random effects, assumed to be normally distributed and realized independently for different subjects, but possibly correlated within subjects; t_{jk} is the age of the jth subject on the kth testing occasion; and E_{ijk} are measurement errors, assumed to be mutually independent, and normally distributed.

Our substantive hypotheses concern claims that, first, particular elements of β are nonzero, reflecting real, group-level effects on test outcomes; second, particular terms in the variance matrix of the random effects are nonzero, reflecting dependencies between the different outcomes achieved by a given subject on different tests or between a given subject's general level of achievement and his or her rate of change with age (in both cases, measured relative to the subject's peer group as defined by the explanatory variables).

The normalized score corresponding to each Y_{ijk} is Y^*_{ijk},where

$$Y^*_{ijk} = (Y_{ijk} - m_i)/s_i \quad (2)$$

and m_i is the observed mean of Y_{ijk} for a given cognitive test, i; s_{ik} is the observed standard deviation of Y_{ijk} for a given cognitive test, i.

Combining Equations 1 and 2 leads to

$$Y^*_{ijk} = x'_{ijk}\beta^*_i + A^*_{ij} + B^*_{ij}t_{jk} + E^*_{ijk}. \quad (3)$$

Comparing Equations 1 and 3 with respect to the random effect terms A and B, we see that they are essentially identical. Specifically, if the A_{ij} and B_{ij} are multivariate normally distributed, then so are A^*_{ij} and B^*_{ij} and vice versa. Further, because the transformations from the variants with and without asterisks are made componentwise, then any two random effects that are uncorrelated before transformation remain so after transformation and vice versa. In particular, a test for whether any two "without asterisk" random effects are or are not

Table 3. Mean Percentage of Correct Scores at Initial Test by Selected Demographics

Demographic	Participants *n* (%)	AH4-1 *M* (*SD*)	AH4-2 *M* (*SD*)	MHA *M* (*SD*)	MHB *M* (*SD*)	CVL *M* (*SD*)	VFR *M* (*SD*)
Gender							
Female	4,176 (71)	47.5 (17.3)	42.9 (16.2)	66.2 (14.2)	48.5 (18.5)	70.4 (14.3)	28.0 (11.1)
Male	1,735 (29)	50.8 (17.8)	49.0 (17.0)	70.5 (14.1)	52.1 (18.5)	66.4 (15.0)	25.7 (11.1)
City							
Manchester	2,646 (45)	49.7 (17.8)	45.7 (16.9)	69.8 (13.8)	54.8 (17.9)	71.2 (14.0)	28.3 (11.6)
Newcastle	3,265 (55)	47.6 (17.3)	43.8 (16.4)	65.6 (14.5)	45.4 (18.0)	67.5 (14.9)	26.4 (10.8)
Social class							
Category 1, professional	266 (4.5)	61.5 (13.9)	59.5 (14.6)	79.0 (11.1)	63.5 (15.7)	71.8 (13.9)	29.5 (12.2)
Category 2, intermediate	1,876 (31.7)	56.4 (16.6)	51.0 (16.0)	74.1 (12.2)	58.3 (16.6)	73.0 (13.3)	30.3 (11.8)
Category 3 (N), nonmanual skilled	2,080 (35.2)	48.4 (14.9)	43.6 (14.3)	66.5 (12.4)	48.1 (16.3)	70.2 (13.6)	27.4 (10.4)
Category 3 (M), manual skilled	765 (12.9)	38.6 (15.4)	37.5 (15.3)	60.7 (13.3)	40.3 (16.7)	63.0 (14.6)	22.7 (9.5)
Category 4, partly skilled or Category 5, unskilled	488 (8.3)	34.7 (14.7)	33.5 (15.1)	55.9 (13.8)	34.9 (16.4)	61.3 (15.1)	22.0 (9.2)
Missing/uncoded	436 (7.4)	39.7 (17.9)	38.0 (17.4)	61.4 (16.5)	42.7 (19.9)	61.4 (17.9)	24.3 (10.8)
Entry age (years)							
49–59	1,418 (24)	55.2 (17.2)	53.0 (16.2)	68.3 (13.9)	51.4 (18.0)	75.0 (13.2)	31.6 (12.2)
60–69	2,910 (49)	49.4 (16.7)	45.3 (15.6)	67.3 (14.1)	49.4 (18.3)	69.8 (13.8)	27.6 (10.6)
70–79	1,431 (24)	41.6 (16.5)	36.6 (14.7)	67.2 (15.0)	48.3 (19.4)	64.1 (14.5)	23.5 (9.9)
80+	152 (3)	33.7 (15.6)	29.6 (12.8)	68.0 (15.3)	40.4 (19.7)	56.7 (17.2)	21.2 (10.9)

Notes: For participants, $N = 5{,}911$; MHA and MHB = Mill Hill A and Mill Hill B (tests). AH4-1 and AH4-2 are intelligence tests. CVL = cumulative verbal learning; VFR = verbal free recall.

correlated is equivalent to a test that the corresponding "with asterisk" random effects are or are not correlated.

With respect to the regression parameters, the important difference between Equations 1 and 3 is that the β* parameter has acquired a subscript *i*. This implies that the substantive meaning of a hypothesis involving interactions between explanatory variables and cognitive tests is indeed different on the raw and standardized scales (e.g., a hypothesis that the average effect of a 1-year increase in age is numerically the same across all cognitive tests). However, the substantive meaning of a hypothesis concerning main effects is unchanged (e.g., a hypothesis that an increase in age does or does not affect the average response to a particular cognitive test).

The overall conclusion is that although the numerical values of estimated parameters would be affected by a change from raw to standardized scores, the substantive conclusions sought from the analysis and claimed in this article are not.

Model specification.—The first methodological issue to which we have drawn attention is the need to take account of the fact that measurements taken over time on the same individual tend to be correlated. There are several approaches to model this correlation structure, and the choice depends on the main scientific questions of interest. Our present aims are (a) to identify factors predictive of cognitive decline at the population level and (b) to gain insight into individual differences relative to the population levels. This leads us naturally to the random effects model, which has two parts: a model for the average response over time for respondents with given values of all explanatory variables, and a model for the random variation about the mean response. For the second component, we postulate a set of latent variables, or "random effects," which represent deviations of individual respondents from the population average for some relevant features. This random variation occurs in addition to the residual variation.

To address the substantive hypotheses set out in the introduction, we need to describe for each cognitive test how the population-average scores depend on the following set of explanatory variables: age, gender, socioeconomic status (SES), city of origin (Manchester or Newcastle), wave of recruitment to the study (cohort), level of general intellectual ability (gf) as indexed by AH4 test scores, and whether tests are being taken for the first, second, third, or fourth time (practice effects). Accordingly, the analysis was made to address the combined effects of age, gender, SES, and practice effects, which are of substantive interest, whereas effects distinguishing cities and year-of-entry cohorts are included as a means of adjusting for unidentified confounding factors. The effects of differences in level of intellectual ability (I gf) are then also explored.

We now introduce the following notation. For a given response, let Y_{ij} denote the percent correct score for the *i*th subject on the *j*th occasion. Hence $i = 1{-}n$ (the number of subjects with at least one response measure) and $j = 1$–4. We use x_{ijk}, where $k = 1{-}p$, to denote the values of the set of p explanatory variables associated with each Y_{ij}. In particular, let x_{ij1} denote age (in years over 49, the minimum entry age) and x_{ij2} denote age squared. Improvement at the three repeat test occasions ($j \geq 2$) is modeled as a series of step functions. Then the mean value of Y_{ij} is μ_{ij} defined by

$$\mu_{ij} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \cdots + \beta_p x_{ijp},$$

and the measured value Y_{ij} is

$$Y_{ij} = \mu_{ij} + A_i + B_i x_{ij1} + E_{ij},$$

where A_i and B_i are subject-specific parameters giving the deviation of the *i*th subject's intercept and slope from the average

population response. We assume these to be random variables drawn from a bivariate normal distribution with mean zero and covariance matrix

$$\sum = \begin{bmatrix} \sigma_A^2 & \rho_{AB}\sigma_A\sigma_B \\ \rho_{AB}\sigma_A\sigma_B & \sigma_B^2 \end{bmatrix}$$

independent of E_{ij}, which follows a univariate normal distribution with mean zero and variance $\sigma^2{}_E$. Further random effects (such as quadratic age terms or practice effects) are not considered. As described in the paragraphs that follow, age appears to have little or no effect on the MH tests. For this pair of tasks, we therefore omitted the subject-specific random slope effects B_i from the model.

We used maximum likelihood, using the *mle*( ) function within the Splus software environment, for model estimation under the assumption that drop-out was random (i.e., the probability that a respondent drops out may depend on his or her observed measurement history, but not on the unobserved responses).

We fitted separate models to each response, adopting the same method for selecting the mean structure.

Results

Not all participants provided scores on all measures. Thus a total of 5,894–5,899 individuals with at least one response measure were included in the AH4 and MH analyses. As already mentioned, the CVL and VFR tests were performed on a second session. Because substantial numbers withdrew after the first testing session of the first data collection, only 5,254 participants also had a CVL or a VFR result.

Selection of the Mean Structure

Based on the empirical evidence of a declining age effect for the AH4-1, AH4-2, CVL, and the VFR tests, the following steps are used to derive a mean model for these responses.

Assuming that the relationship between cognitive decline and age is captured adequately by a linear and a quadratic term, we adopt the following step-down approach to test whether there is any evidence that the overall age trajectory differs with SES, gender, or practice. Based on participants from Newcastle who, overall, were followed up for longer than those from Manchester, a sequence of models is fitted. Under the assumption that the improvement at the second data-collection point adequately captures the Practice × Age interaction, we fit models allowing the quadratic curve to depend on any combination of SES, gender, and improvement. We then fit simpler interaction models, retaining the simpler model at each step if the generalized likelihood ratio test (LRT) for the additional terms is nonsignificant at the conventional 5% level. We do not separately test for an interaction with the linear or quadratic component of age. Hence, a test comprises $2(m - 1)$ degrees of freedom, where m is the number of levels of the factor in question.

There are no significant interactions with age for AH4-1 or VFR, whereas there is a significant single interaction for Improvements at Visit 2 × Age for AH4-2, and for Gender × Age for CVL. Repeating these steps but considering only main effect practice terms results in the same "final" models for AH4-1, CVL, and VFR, whereas for AH4-2 the age quadratic now depends on the Gender × SES combination. Although they are statistically significant, all these interactions have very small effects that we considered unimportant in substantive terms. We also noted that the covariance matrix estimates for the random effects were robust to the choice of mean model (i.e., main effects only or with age interactions). For these reasons we proceed to fit models without interactions to the combined data set, including a term for city of residence, and test the quadratic age trend against the linear term for these four responses. In all cases this was highly significant ($p < .0001$, 2 *df*).

For the vocabulary tests the linear age term based on the complete Manchester and Newcastle data set was statistically significant for MHB but not MHA scores, with estimated mean declines of −0.06 and −0.03 per annum, respectively. Although these trends are clearly too slight to be of substantive importance, predicting a decline of only 0.6% and 0.3% respectively over 10 years, the linear age term was retained in order to better estimate the covariance structure.

Before describing the mean estimates for each response pair and comparing the parameter estimates of substantial interest, namely the age trends and practice components across tests, we now give a brief explanation of the parameter estimates. In all models the first level of each factor is the reference group. Thus, the intercept parameter represents the percentage score for a respondent who is in socioeconomic category (C) 1; female; of age 49; and a resident in Manchester and taking the test for the first time in 1983. The improvement parameters measure the average step increase between successive testing occasions ($j \geq 2$). The effects for socioeconomic categories (Cs) from 2 to 4–5 represent the estimated difference in mean percent correct scores between socioeconomic groups 2 to 4–5 and socioeconomic group 1. The entry year values represent the difference in scores between 1984–1992 and 1983, and the city term gives the mean difference between Newcastle and Manchester participants. This allows us to test the substantive working hypotheses that, when practice and drop-out have been taken into consideration, tests will show accelerated age-related declines, and that rates of decline will be seen to differ between tests, and also between individuals of higher and lower gf.

AH4-1 and AH4-2 Scores

Tables 4 and 5 summarize the estimated mean effects for AH4-1 and AH4-2, respectively. The AH4-1 intercept of 66.0 is 4.9 points higher than the AH4-2 intercept. The fact that both quadratic coefficients are negative indicates that scores on both these tests show accelerated decline with age. The rates of decline are very similar. For example, consider the entry scores: On AH4-1, the average score for a 60-year-old would be 64.6, and for a 70-year-old, 58.6 (a 6-point drop), falling to 46.5 (a further 12-point fall) for an 80-year-old. On the AH4-2 task, the corresponding scores would be 58.2, 52.1 (6-point drop), and 38.6 (a further 13-point drop). This supports the first working hypothesis that, after practice and drop-out are taken into consideration, at least on tests of gf, the rates of decline are seen to accelerate with increasing age.

The next substantive hypothesis to be tested is that practice effects would differ between tests and age groups, and possibly also between sexes, between demographic groups and as a function of overall level of gf. Tables 4 and 5 show that the average practice step increase for AH4-2 on the second occasion of

Table 4. AH4-1 Model Estimates

Mean Response Model Parameter	Estimate	*SE*	*t* Value	*p* Value	95% CI
Intercept	66.03	1.53	43.08	<.001	63.03, 69.03
age − 49	0.04	0.05	0.74	.459	−0.07, 0.15
$(age - 49)^2$	−0.02	0.001	−17.57	<.001	−0.03, −0.02
Improvement					
$j \geq 2$	4.51	0.16	27.58	<.001	4.19, 4.83
$j \geq 3$	2.21	0.21	10.28	<.001	1.79, 2.63
$j \geq 4$	1.90	0.36	5.25	<.001	1.19, 2.61
Male vs. female	2.20	0.44	5.02	<.001	1.34, 3.07
Socioeconomic status					
C2 vs. C1	−3.77	0.96	−3.94	<.001	−5.64, −1.89
C3 (N)	−10.53	0.98	−10.72	<.001	−12.45, −8.60
C3 (M)	−19.94	1.04	−19.19	<.001	−21.98, −17.91
C4/5	−23.20	1.13	−20.56	<.001	−25.41, −20.99
Missing	−19.12	1.16	−16.44	<.001	−21.40, −16.84
Newcastle vs. Manchester	−3.20	1.07	−3.00	.003	−5.29, −1.11
Entry year					
1984 vs. 1983	0.71	0.64	1.12	.261	−0.53, 1.96
1985	−0.82	1.25	−0.65	.515	−3.27, 1.64
1986	−0.50	1.30	−0.39	.699	−3.05, 2.04
1987	−1.54	1.33	−1.16	.245	−4.14, 1.06
1988	1.97	0.74	2.65	.008	0.51, 3.43
1989	4.64	0.89	5.24	<.001	2.90, 6.37
1990	5.37	1.47	3.65	<.001	2.49, 8.25
1991	1.19	1.84	0.65	.516	−2.41, 4.79
1992	1.64	1.55	1.06	.289	−1.39, 4.68

Notes: CI = confidence interval; C = category; N = nonmanual skilled; M = manual skilled. Covariance and residual parameters: For σ_A, estimate = 15.43 and 95% CI = 14.71, 16.18; for σ_B, estimate = 0.41 and 95% CI = 0.36, 0.47; for ρ_{AB}, estimate = −0.51 and 95% CI = −0.58, −0.44; for σ_E, estimate = 5.71 and 95% CI = 5.58, 5.84.

Table 5. AH4-2 Model Estimates

Mean Response Model Parameter	Estimate	*SE*	*t* Value	*p* Value	95% CI
Intercept	61.12	1.44	42.57	<.001	58.31, 63.94
age − 49	−0.17	0.06	−3.01	.003	−0.28, −0.06
$(age - 49)^2$	−0.02	0.00	−15.57	<.001	−0.02, −0.02
Improvement					
$j \geq 2$	5.58	0.17	32.94	<.001	5.24, 5.91
$j \geq 3$	4.72	0.22	21.07	<.001	4.28, 5.15
$j \geq 4$	5.88	0.38	15.34	<.001	5.13, 6.64
Male vs. female	5.14	0.41	12.62	<.001	4.34, 5.93
Socioeconomic status					
C2 vs. C1	−5.96	0.89	−6.73	<.001	−7.70, −4.23
C3 (N)	−11.36	0.91	−12.50	<.001	−13.14, −9.58
C3 (M)	−17.58	0.96	−18.25	<.001	−19.47, −15.69
C4/5	−20.37	1.04	−19.50	<.001	−22.42, −18.33
Missing	−17.56	1.08	−16.22	<.001	−19.68, −15.44
Newcastle vs. Manchester	−0.57	0.99	−0.58	.564	−2.51, 1.37
Entry year					
1984 vs. 1983	−0.27	0.59	−0.46	.646	−1.42, 0.88
1985	2.22	1.16	1.91	.056	−0.05, 4.50
1986	2.58	1.21	2.14	.032	0.22, 4.94
1987	2.22	1.23	1.80	.071	−0.19, 4.63
1988	3.19	0.69	4.64	<.001	1.84, 4.54
1989	9.03	0.82	11.00	<.001	7.42, 10.64
1990	8.50	1.37	6.22	<.001	5.82, 11.17
1991	7.41	1.71	4.34	<.001	4.07, 10.75
1992	4.25	1.45	2.93	<.001	1.41, 7.08

Notes: CI = confidence interval; C = category; N = nonmanual skilled; M = manual skilled. Covariance and residual parameters: For σ_A, estimate = 14.63 and 95% CI = 13.94, 15.35; for σ_B, estimate = 0.43 and 95% CI = 0.38, 0.49; for ρ_{AB}, estimate = −0.58 and 95% CI = −0.63, −0.51; for σ_E, estimate = 6.22 and 95% CI = 6.09, 6.36.

testing are comparable (gains of 4.5 and 5.6 points, respectively). On the third and fourth occasions the level falls to under 2.5 on AH4-1, but it remains above 4.5 for AH4-2, corresponding to a cumulative gain of 5.6, 10.3, and 16.2. Thus practice effects differ, even between tests that putatively measure the same performance index, gf. Although initial improvements are similar on the two AH4 tests, in the longer term, AH4-2 scores show greater and more sustained gains.

On average, men performed significantly better than women on both of the AH4 tests. Tables 4 and 5 show that, after the other covariates in the model are adjusted for, scores on both AH4 tests markedly vary with SES category. Participants in C1 score at least 20 points more, on average, than participants in C4–5. Another demographic factor, city of residence, also affects AH4 test scores. On the AH4-1, Manchester residents score higher than Newcastle residents, but on the AH4-2, this city effect is not significant. Implications of this male advantage are discussed in the paragraphs that follow. The overall effects of SES are as expected.

There are also significant differences between waves of recruitment; that is, there are significant drop-in effects. Individuals who entered the study in 1989 and 1990 have markedly higher scores on both AH 4 tests than those who entered in the first year of the study, 1983. Note that a principal aim of the analysis is to examine the hypothesis that rates of change over time are affected both by demographic factors and by overall levels of gf, and this cannot be properly addressed if this drop-in effect is neglected.

Cumulative Verbal Learning and Verbal Free Recall Scores

Tables 6 and 7 summarize these results for CVL and VFR, respectively. Note that the intercept estimate of 75.6 for CVL is more than double the 35.5 point score on VFR, which is the lowest score overall.

As for the AH4 responses, the quadratic term is negative for both tests, indicating some acceleration of decline with age. However, the deviation from linearity is small. As an example, the average CVL entry score is 73.6 for a 60-year-old, 68.6 for a 70-year-old, and falls to 60.2 for an 80-year-old; that is, there are declines of 5.0 and 8.4 points. The corresponding VFR scores of 32.4, 28.6, and 23.8 points equate to declines of 3.8 and 4.8 points over the successive 10-year intervals. The average improvement on the second experience of CVL was 1.0. This, although statistically significant, is slight. The equivalent improvement for VFR was negative and nonsignificant. On the third and fourth testing occasions, an improvement followed by a decline occurred for both these tests. These fluctuations remain unexplained.

On both the CVL and VFR tests, women performed significantly better than men. On the CVL and VFR tests, the differ-

Table 6. CVL Model Estimates

Mean Response Model Parameter	Estimate	*SE*	*t* Value	*p* Value	95% CI
Intercept	75.57	1.37	55.14	<.001	72.89, 78.26
age − 49	0.07	0.06	1.06	.290	−0.06, 0.19
(age − 49)^2	−0.02	0.00	−13.36	<.001	−0.03, −0.02
Improvement					
$j \geq 2$	0.97	0.20	4.75	<.001	0.57, 1.37
$j \geq 3$	5.51	0.28	19.50	<.001	4.95, 6.06
$j \geq 4$	−3.16	0.48	−6.53	<.001	−4.11, −2.21
Male vs. female	−4.28	0.39	−11.01	<.001	−5.04, −3.52
Socioeconomic status					
C2 vs. C1	−0.58	0.82	−0.71	.478	−2.18, 1.02
C3 (N)	−3.76	0.84	−4.47	<.001	−5.40, −2.11
C3 (M)	−8.63	0.90	−9.61	<.001	−10.38, −6.87
C4/5	−10.88	0.98	−11.09	<.001	−12.80, −8.96
Missing	−9.12	1.16	−7.83	<.001	−11.40, −6.84
Newcastle vs. Manchester	−1.09	0.97	−1.12	.262	−2.98, 0.81
Entry year					
1984 vs. 1983	4.57	0.54	8.53	<.001	3.52, 5.62
1985	6.96	1.12	6.23	<.001	4.77, 9.14
1986	4.98	1.15	4.33	<.001	2.73, 7.23
1987	6.05	1.19	5.10	<.001	3.72, 8.37
1988	8.31	0.74	11.25	<.001	6.86, 9.75
1989	7.59	0.77	9.88	<.001	6.09, 9.10
1990	7.97	1.27	6.27	<.001	5.48, 10.46
1991	4.99	1.64	3.05	<.001	1.78, 8.19
1992	3.87	1.37	2.83	<.001	1.19, 6.55

Notes: CVL = cumulative verbal learning; CI = confidence interval; C = category; N = nonmanual skilled; M = manual skilled. Covariance and residual parameters: For σ_A, estimate = 10.24 and 95% CI = 9.28, 11.29; for σ_B, estimate = 0.49 and 95% CI = 0.44, 0.56; for ρ_{AB}, estimate = −0.45 and 95% CI = −0.57, −0.32; for σ_E, estimate = 7.59 and 95% CI = 7.43, 7.76.

Table 7. VFR Model Estimates

Mean Response Model Parameter	Estimate	*SE*	*t* Value	*p* Value	95% CI
Intercept	35.46	1.07	33.10	<.001	33.36, 37.55
age − 49	−0.23	0.05	−4.41	<.001	−0.34, −0.13
(age − 49)^2	−0.01	0.00	−4.01	<.001	−0.01, 0.00
Improvement					
$j \geq 2$	−0.16	0.18	−0.91	.365	−0.51, 0.19
$j \geq 3$	1.96	0.24	8.14	<.001	1.49, 2.44
$j \geq 4$	−0.93	0.40	−2.33	.020	−1.72, −0.15
Male vs. female	−2.07	0.29	−7.01	<.001	−2.65, −1.49
Socioeconomic status					
C2 vs. C1	0.05	0.62	0.07	.942	−1.18, 1.27
C3 (N)	−2.99	0.64	−4.65	<.001	−4.24, −1.73
C3 (M)	−5.83	0.68	−8.56	<.001	−7.17, −4.50
C4/5	−6.71	0.75	−8.99	<.001	−8.17, −5.25
Missing	−4.99	0.91	−5.48	<.001	−6.77, −3.21
Newcastle vs. Manchester	−1.21	0.72	−1.68	.092	−2.61, 0.20
Entry year					
1984 vs. 1983	1.21	0.41	2.94	<.001	0.41, 2.02
1985	1.35	0.84	1.61	.107	−0.29, 3.00
1986	0.52	0.86	0.61	.545	−1.17, 2.22
1987	1.06	0.89	1.18	.237	−0.69, 2.80
1988	2.30	0.56	4.09	<.001	1.20, 3.41
1989	1.77	0.58	3.03	.002	0.63, 2.92
1990	2.07	0.98	2.12	.034	0.15, 3.98
1991	2.11	1.26	1.67	.094	−0.36, 4.58
1992	0.11	1.05	0.10	.917	−1.95, 2.17

Notes: VFR = verbal free recall; CI = confidence interval; C = category; N = nonmanual skilled; M = manual skilled. Covariance and residual parameters: For σ_A, estimate = 9.67 and 95% CI = 8.95, 10.46; for σ_B, estimate = 0.23 and 95% CI = 0.16, 0.31; for ρ_{AB}, estimate = −0.72 and 95% CI = −0.78, −0.65; for σ_E, estimate = 7.00 and 95% CI = 6.87, 7.15.

ences in trends for SES Cs are much less marked than on the AH4 tests. Again, the largest difference is between C1 and C4–5 (10.9% and 6.7% points for CVL and VFR, respectively). On these tests, as on all others, average scores are higher for Manchester than for Newcastle residents, but in this case these differences are not significant. Interestingly, the mean CVL scores are from 3.9 to 8.3 points higher for participants who entered the study from 1984 onward, compared with those entering in 1983. VFR scores show a similar but less marked recruitment wave, or drop-in effect.

Mill Hill A and Mill Hill B Scores

Tables 8 and 9 show the results for MHA and MHB, respectively. The intercept estimate for the MHA test is 80.6, which is higher than for any other test and is 15.4 points higher than the intercept estimate of 65.1 for the MHB test.

Assuming that the age trajectories for MHA and MHB are described by a linear term, there is a nonsignificant decline of −0.03% points per year for MHA, and a significant decline of −0.06% points per year for MHB. Assuming a mean rate corresponding to the lower confidence interval for MHB, we find that this equates to a decline of only 1.2% over 10 years.

Tables 7 and 8 show negligible positive or negative changes of 1.5 or less on MHA at all repeat testings, and on MHB for the first and second repeat testings. The estimated improvement of 9% points on MHB on the last testing occasion is an unexplained anomaly. Scores on both the MHA and MHB tests are significantly higher for men than for women, and for Manchester residents than for Newcastle residents. The differences between SES Cs were substantial, and they were comparable in magnitude with the effect sizes for AH4. Note that, in contrast to all other tests, on MHA the average entry year scores for cohorts recruited from 1984 onward were lower, though not significantly so for those starting in 1983. There was no clear pattern for MHB. This illustration that successive recruitment cohorts may differ on some tests though not on others emphasizes that drop-in effects may be complex and must be taken into consideration when one analyzes longitudinal data.

Comparisons of Practice Effects Between Tasks

For AH4-1, AH4-2, and CVL, rates of decline are similar and accelerate with age. For VFR, the rate of decline is linear, and MH scores remain stable over time. The size of the average practice effect on the second occasion varies between tasks: Gains of over 4.5% points on the AH4 tests contrast with gains of 1% or less on the other tasks. On the third and fourth occasion, a gain of over 4.5 for AH4-2 and CVL contrasts with negative estimates for several other tasks. However, note that these differences may reflect a lack of fit between the quadratic

and improvement parameters at older ages for which relatively few data points are available.

A further question about practice effects is whether their sizes vary with the interval between successive repetitions of the tasks. The fact that some individuals missed particular retesting sessions but later returned to the study allowed us to make a secondary analysis to determine whether the substantial practice effect of over 4.5 points on the second occasion of taking the AH4-1 and AH4-2 tests varies with the duration of the interval between initial and second experiences. Restricting the analysis to individuals with scores at entry and at the second scheduled visit for each test, we found that the interval between these two time points ranged from 1 to 8 years. Because of small numbers in the lowest and highest categories, the categories we used were ≤ 2, 3, 4, 5, 6, or 7+ years. We fitted a model to each response, replacing the single-step function at the second occasion by a six-level term corresponding to the gap times. For the AH4-1 response, the mean practice effects are similar across the intervals, ranging from 3.5 to 5.2 points, with no clear trend over time. The estimates are slightly more variable for AH4-2, ranging from 3.8 to 7.9. For both responses, this model provides an improvement in fit compared with the simpler model with practice at the second visit coded as a two-level factor ($p = .006$ and $p < .0001$, 5 *df*). That is to say, the average sizes of improvement caused by a previous experience of either AH4 test remained the same over intervals of 2 to 7 years.

Table 8. MHA Model Estimates

Mean Response Model Parameter	Estimate	*SE*	*t* Value	*p* Value	95% CI
Intercept	80.56	1.33	60.61	<.001	77.96, 83.16
age − 49	−0.03	0.02	−1.41	.155	−0.07, 0.01
Improvement					
$j \geq 2$	1.07	0.16	6.57	<.001	0.75, 1.39
$j \geq 3$	−1.48	0.21	−7.02	<.001	−1.90, −1.07
$j \geq 4$	1.47	0.34	4.32	<.001	0.80, 2.13
Male vs. female	1.89	0.39	4.82	<.001	1.12, 2.66
Socioeconomic status					
C2 vs. C1	−3.18	0.85	−3.73	<.001	−4.85, −1.51
C3 (M)	−17.09	0.93	−18.41	<.001	−18.91, −15.27
C3 (N)	−10.35	0.88	−11.82	<.001	−12.06, −8.63
C4/5	−21.11	1.01	−20.95	<.001	−23.08, −19.14
Missing	−16.20	1.04	−15.61	<.001	−18.23, −14.16
Newcastle vs. Manchester	−4.21	0.96	−4.39	<.001	−6.08, −2.33
Entry year					
1984 vs. 1983	−0.44	0.57	−0.78	.433	−1.55, 0.66
1985	−1.44	1.12	−1.28	.200	−3.64, 0.76
1986	−1.72	1.16	−1.48	.139	−4.00, 0.56
1987	−2.15	1.19	−1.81	.071	−4.47, 0.18
1988	−0.18	0.66	−0.27	.790	−1.47, 1.12
1989	−1.24	0.79	−1.56	.120	−2.79, 0.32
1990	−1.14	1.32	−0.86	.388	−3.74, 1.45
1991	−2.56	1.65	−1.55	.121	−5.79, 0.68
1992	−2.57	1.40	−1.84	.066	−5.30, 0.17

Notes: MHA = Mill Hill A; CI = confidence interval; C = category; N = nonmanual skilled; M = manual skilled. Covariance and residual parameters: For σ_A, estimate = 11.84 and 95% CI = 11.59, 12.09; for σ_E, estimate = 6.07 and 95% CI = 5.97, 6.18.

Deviation Around the Mean Response: Random Effects

The final hypotheses examined are that participants' rates of cognitive decline vary with their overall levels of gf, and that variance in performance between participants increases as the study continues, and the mean age of the sample increases.

In these models, the A_is reflect the extent to which individuals deviate from the average response value, and the B_is measure their deviations in slope, that is, in rates of decline. The maximum likelihood estimates (and 95% confidence intervals, or CIs) of the standard deviations (σ_A and σ_B) and correlation (ρ_{AB}) of these random effects, assumed to be normally distributed with mean zero, are presented in the table notes of Tables 4 through 9. After adjustment for covariates, the individual estimates of the A_i and B_i are of interest as they can be used to predict intercepts and slopes of individual trajectories of change. They are usually estimated as the conditional expectation of the effects given the observed data, and they are sometimes termed empirical Bayes (EB) estimates. In addition, histograms and scatterplots can be used to detect unusual individuals.

From values of $\hat{A}_i$ and $\hat{B}_i$ estimates for individuals based on the AH4-1 model, we calculate sample standard deviation estimates of 12.84 and 0.15. These are considerably smaller than the *mle* estimates of 15.43 and 0.41. The estimated sample correlation is weakly negative ($\hat{\rho} = -0.29$), whereas the *mle* is strongly negative ($\hat{\rho}_{AB} = -0.51$, 95% CI of −0.58 to −0.44). Discrepancies of similar magnitudes are observed for the other responses. This suggests that the empirically observed estimates do indeed substantially underestimate the true variability in the random effects. Actually, for any linear combination of the random effects, the EB estimates are less than or equal to the true variability in the random effects (Verbeck & Molenberghs, 1997, chap. 3). This latter result provides theoretical support for our findings.

Within this model, the issue of whether individuals' trajectories of cognitive decline vary with their basal levels of mental ability (AH4 test scores) can be approached only if we make a strong assumption that there is a particular age before which decline proceeds at a constant rate. Given this assumption, we see that estimates $\hat{\rho}_{AB}$ from the models have substantive value. Assuming that this critical age is the lowest entry age in the sample, 49 years, we find that the outcome is that, among individuals who were aged 49 at entry, those with higher initial AH4-1 scores tended to show relatively more rapid cognitive decline on AH4-1 than did those with lower initial scores. As shown in Tables 4 through 7, the correlation estimates for AH4-1, AH4-2, CVL, and VFR are all significantly negative.

Nevertheless, it is important to note that these correlations are arbitrary and depend on the age used for "centering." For example, if age 65 is used instead of age 49, the correlation is approximately zero for the AH4-1 scores, indicating parity in rates of change for individuals at all levels of AH4-1 scores. They become increasingly positive as centering ages older than 65 are selected, indicating faster rates of decline for individuals with lower AH4-1 scores.

The question of whether variability between participants increases with sample age can be addressed in a similar way. For each response, the σ_A standard deviation estimate can be used to calculate the 95% expected range (the range over which 95% of

Table 9. MHB Model Estimates

Mean Response Model Parameter	Estimate	*SE*	*t* Value	*p* Value	95% CI
Intercept	65.13	1.59	40.91	<.001	62.01, 68.25
age − 49	−0.06	0.03	−2.35	.019	−0.12, −0.01
Improvement					
$j \geq 2$	0.97	0.23	4.28	<.001	0.53, 1.42
$j \geq 3$	−1.31	0.30	−4.39	<.001	−1.89, −0.76
$j \geq 4$	9.31	0.49	18.93	<.001	8.34, 10.27
Male vs. female	0.94	0.47	2.01	.045	0.02, 1.86
Socioeconomic status					
C2 vs. C1	−4.28	1.02	−4.21	<.001	−6.27, −2.29
C3 (N)	−13.60	1.04	−13.03	<.001	−15.64, −11.55
C3 (M)	−21.64	1.11	−19.55	<.001	−23.81, −19.48
C4/5	−26.26	1.20	−21.82	<.001	−28.62, −23.90
Missing	−19.78	1.25	−15.88	<.001	−22.22, −17.34
Newcastle vs. Manchester	−3.71	1.15	−3.23	<.001	−5.96, −1.46
Entry year					
1984 vs. 1983	−2.49	0.67	−3.71	<.001	−3.81, −1.18
1985	2.32	1.35	1.72	.085	−0.32, 4.96
1986	−0.04	1.39	−0.03	.976	−2.77, 2.69
1987	0.69	1.42	0.49	.627	−2.10, 3.48
1988	0.79	0.79	1.00	.318	−0.76, 2.33
1989	−1.01	0.95	−1.06	.287	−2.87, 0.85
1990	−2.42	1.59	−1.52	.128	−5.54, 0.70
1991	−3.70	1.98	−1.86	.062	−7.58, 0.19
1992	−2.01	1.68	−1.20	.230	−5.29, 1.27

Notes: MHB = Mill Hill B; CI = confidence interval; C = category; N = nonmanual skilled; M = manual skilled. Covariance and residual parameters: For σ_A, estimate = 13.53 and 95% CI = 13.22, 13.85; for σ_E, estimate = 9.00 and 95% CI = 8.84, 9.16.

the population values would fall) for the intercept. For example, the 95% expected ranges for the AH4-1 and AH4-2 mean entry levels are $66.0 + 1.96 \times 15.4 = [35.8–96.3]$ and [32.5–89.8], respectively. The expected ranges for the other tests are smaller because the $\hat{\sigma}_A$ are less. We can also examine the relative variability of the A_i with respect to the corresponding fixed effect estimate. Based on the parameter estimates presented in Table 4, the relative variability is $15.4/66.0 = 23\%$ for AH4-1. Estimates of similar size were obtained for AH4-2 (24%), VFR (27%), and MHB (21%), whereas CVL and MHA showed markedly less relative variability (14% and 15%).

The σ_B estimates ranged from 0.41 to 0.49 on the AH4 and CVL responses, whereas there was less variability in the individual slopes for the VFR response ($\hat{\sigma}_B = 0.23$). Approximately 95% of the individual B_i values lie within $\pm 1.96\ \hat{\sigma}_B$ of the zero mean. Although the σ_B estimates for these four responses appear small, they are amplified by the multiplication with age (in years over 49) in the model. For example, a difference of 0.45 (roughly 1 *SD* on the AH4 and CVL tests) in the slopes for any two participants with equal cognitive function on entry to the study would result in a difference of 4.5 and 9.0 percentage points after 10 and 20 years, respectively. Because participants have different rates of decline, this implies that between-individual variability in performance increases with sample age and also that this increase in variability is most marked for those tasks with large between-participant variability. It is important to note that, because the effects of covariates such as gender, SES, and city of residence were taken into consideration in computing this variance, they cannot provide functional explanations for it. Because age is modeled as a quadratic function in the mean part of the models, neither the expected ranges nor the relative variability of the B_i can be calculated. Finally, under the specified random effect models, the residual standard deviation estimates were similar between tests and ranged from 5.7 to 9.0 percentage points.

Discussion

It has long been recognized that longitudinal data may be seriously misinterpreted if participants in longitudinal studies improve because they repeatedly take the same tests. The current analyses tested the working hypothesis that practice effects do occur in a longitudinal study, that they are substantial enough to mask age-related declines, and that they vary between tasks and between individuals of different ages and levels of ability.

Practice Effects

Practice effects are significant and substantial on both AH4 tests and on the CVL test. For example, on the second occasion, gains of over 4.5 percentage points on the AH4 tests contrast with improvements of 1% or less on the other tasks. On the third and fourth occasions, an improvement of over 4.5 is predicted for AH4-2 and CVL, contrasting with negative estimates for several other tasks. Note that these gains from practice are comparable with the declines in average scores, after practice and drop-out have been taken into consideration, of 6 points between age 60 and 70 (64.6 and 58.6, respectively) on the AH4-1 and 6 points (58.2 to 52.1, respectively) on the AH4-2.

The sizes of practice effects do differ between tasks, and between older and younger individuals. On AH4-1 and AH4-2 tasks, practice effects were indeed markedly greater for older than for younger participants, with estimated improvements between first and second testing of 1.5–2.5% for a 49-year-old as against over 4.5% for a 70-year-old. On all other tasks they ranged from only 0.1 to 1.6 points, and they were independent of age. Neglect of practice effects leads to underestimation of the true extent of age-related changes and may disguise the fact that they are accelerated rather than linear. Further, marked differences in practice effects between tasks and age groups may be misinterpreted as evidence that brain aging affects performance on some tasks, and so some mental abilities, earlier and more severely than others.

Practice improvements were greatest between the first and second encounters with a task, and were thereafter modest. At first sight this seems paradoxical because considerable bodies of evidence, such as those reviewed by Kausler (1990), show that age slows the learning of novel tasks. This would lead us to expect that, the older individuals are, the less they should improve during a longitudinal study. One explanation for this counterintuitive finding is that older individuals perform poorly when they first encounter novel cognitive tests because they need longer to understand what the tests demand of them and to accommodate to an unfamiliar environment (Rabbit, 1993b). On this premise the large and long-lasting practice gains observed between the first and subsequent test sessions during this longitudinal study not only reflect specific task learning but also general familiarity with the testing environment and procedures. Note that this possibility carries the awkward

methodological implication that, even if particular tasks are not repeated, for example, by using "parallel forms," increasing familiarity with the general testing procedures may benefit older participants more than younger participants and so counteract age differences in rates of decline. We suggest that these findings also have theoretical implications. Difficulties in coping with task novelty, and marked gains once initial problems have been overcome, are characteristics of patients with focal prefrontal cortical damage (Burgess, 1997). In this context the present findings may be interpreted as further evidence for age-related declines in "executive" functions supported by the prefrontal cortex that enable us to cope with novel tasks (Burgess & Shallice, 1996; Lowe & Rabbitt, 1998; Shallice & Burgess, 1991) This behavioral evidence has been assumed to reflect neurophysiological findings that the prefrontal cortex suffers earlier and more rapid neurophysiological and cerebrovascular changes than other areas (Gur, Gur, Orbist, Skolnik, & Reivitch, 1987; Haugh & Eggers, 1991; Scheibel & Scheibel, 1975; Shaw et al., 1984). In this framework of interpretation, it is a surprising new finding that, once experienced, tasks and testing situations do not regain "novelty" through disuse, even over periods as long as 7 years.

Drop-Out

We have argued that these likelihood-based analyses under random drop-out assumptions allow good estimates of what actual trajectories of change would have been had drop-out not occurred and so permit more realistic estimates of how rates of age-related cognitive change differ between age groups, socioeconomic groups, and gender groups. Note, however, that these analyses adjust for, but do not give information about, drop-out effects. The relationship between volunteers' propensity to drop out and their cognitive measurement profiles, their gender, socioeconomic category, or general health status are different questions of substantive interest in their own right. We propose to investigate these relationships by using informative drop-out models. The results will be reported separately in due course.

The analysis also detected, and took into consideration, significant differences between the average levels of ability of cohorts recruited at different points during the study. These drop-in effects differed between tasks. On the AH4 and CVL tests, cohort recruitment differences were large enough so that interpretation of the data would have been affected if they had been neglected. In contrast, they were negligible on the MH vocabulary tests. This implies that analyses must not assume that cohort differences on any single "benchmark" test can be taken as representative of differences on all other tests.

The remaining working hypotheses were that, after practice and drop-out effects had been considered, it would be possible to more accurately determine actual rates of changes, and so to discover whether these are constant or are accelerated by increasing age, and whether they differ between different kinds of tasks, between more and less able individuals, and with demographic factors such as gender and socioeconomic advantage. Finally, it was predicted that after all of the aforementioned factors had been taken into consideration, variance in cognitive performance between members of a sample would be seen to significantly increase as the members age.

Does Rate of Cognitive Decline Accelerate With Sample Age?

After practice and drop-out effects were adjusted for, there was clear evidence that rates of decline accelerated with age on the two AH4 tests and the CVL task.

Do Scores on Different Cognitive Tests Decline at Different Rates?

On the AH4-1 and AH4-2 tests and on the CVL task, declines accelerated with age. Declines in VFR scores were less marked and were linear rather than accelerated. On the MHA and MHB vocabulary tests, there was little or no decline. This last finding agrees with the consensus of previous studies that declines in tasks that are assumed to be supported by gf contrasts with stability on tests such as the MHA and MHB vocabulary tests, in which performance is supported by "crystallized" knowledge acquired over a lifetime and maintained by practice in old age (Horn, 1982). The different trajectories of change for the CVL and VFR tests also provide a longitudinal confirmation of Horn's (1982) many cross-sectional demonstrations that age affects performance on some tests of fluid mental abilities more than on others.

How Are Rates of Decline Affected by Gender, by Level of Socioeconomic Advantage, and by Individual Differences in General Intellectual Ability?

On average, men performed better than women on the AH4-1 and AH4-2 and MHA and MHB, but women performed better than men on the CVL and VFR tasks. Superiority of men on the AH4 and MH tests may partly be explained by the fact that, for these generations of participants, women had much poorer educational and career opportunities, most especially in the industrial North of England. The finding of superiority of women on CVL and VFR tests confirms and extends cross-sectional comparisons within this sample by Rabbitt and colleagues (1996). The gender effect on CVL scores appears to be complex, because it also depends on age. The advantage in CVL scores for women is relatively small at young to middle ages and thereafter widens. One possible explanation for this might be that because women live longer they also retain mental competence later in life. However, this seems unlikely because there is no similar Gender $\times$ Age interaction on any other task. In our view, and in the absence of other evidence, the particular advantage for verbal learning (CVL) is as likely to reflect lifestyle factors as intrinsic differences in the level and the maintenance of particular mental abilities. These and other hints of interactions between differences in lifestyle and preservation of particular abilities in old age require further investigation.

There were marked differences in cognitive performance between socioeconomic categories on all tests. The mean difference between occupational groups C1 and C4–5 was over 20 percentage points on the AH4 and MH tasks and 7–11 percentage points on the CVL and VFR tasks. In spite of this clear evidence that SES affects overall levels of performance, there is no evidence that it differentially affects rates of decline. This is unexpected because SES is a good proxy for many factors that are known to slow biological decline, such as level of general health and of lifetime health care, level of educa-

tion, and exposure to toxicity (Kitagawa & Hauser, 1973). Socioeconomic disadvantage is also associated with higher and earlier mortality in later life, and there is robust evidence that approach to death reduces level of cognitive performance during longitudinal studies (Berkowitz, 1964; Bosworth et al., 1999; Botwinick, West, & Storandt, 1978; Jarvik & Blum, 1971; Johannsen & Berg, 1989; Lieberman, 1965; Rabbitt et al., 2002; Reimanis & Green, 1971; Riegel & Riegel, 1972; Riegel, Riegel, & Myer, 1967; Small & Backman, 1997). There is also evidence that socioeconomic disadvantage, and in particular lower educational attainment, is linked to the prevalence of Alzheimer's disease in old age (Bonaiuto, Rocca, & Lippi, 1990; Evans et al., 1993; Korczyn, Kahana, & Galper, 1991). Obviously, more detailed analyses exploring relationships among socioeconomic factors, age, and cognition in this particular population sample are required.

Even when effects of SES and gender are taken into account, Manchester residents perform significantly better than Newcastle residents on the AH4-1, MHA, and MHB tests. There is no evidence of any difference in performance between cities on CVL or VFR. These differences remain cryptic because the city term is likely to be a proxy for a variety of unidentified factors for which the modeling process could not control.

These analyses also suggest that the level of general intellectual ability of participants on entry to a longitudinal study may affect their rates of subsequent cognitive change, though not in the direction that previous research has led us to expect. If we make the reasonable assumption that, in members of this sample, cognitive decline can be dated from age 49 (the age of the youngest volunteers on entry to the study) and that individuals' rates of decline previous to age 49 had been constant, the analysis shows that after practice and drop-out effects had been considered, individuals who entered the study with higher overall levels of ability declined more rapidly than those who entered with lower levels of ability. This finding is inconsistent with previous suggestions that higher levels of performance in young adult life may be associated with longer retention of ability and with lower incidence of dementias and predementing conditions in old age (see, e.g., Snowden et al., 1996). It does, however, agree with an analysis of data from a subgroup of this sample by Rabbitt, Chetwynd, and McInnes (2003) based on the entirely different premise that, because individuals' scores on the MHA vocabulary test do not change with age, they can be used as proxies for their AH4 test scores in middle age and so can be compared against their current, observed AH4 test scores to estimate age-related losses.

Note, however, that the outcome of the present analysis depends on the age used for "centering" in the population. If age 65 is used for "centering," then rates of decline do not vary with levels of gf, and if ages older than 65 are used for "centering," then it appears that the less able decline more rapidly than the more able. The implications of these findings with regard to methods of analysis of individual differences in the forms of trajectories of cognitive change are currently being further explored.

As a Population Ages Does Variability Between Its Members Increase?

The standard deviation estimates provide useful insight into the amount of variability between individuals on each task. The estimated standard deviation for the linear rate of decline was similar for the AH4 and CVL tasks, ranging from 0.41 to 0.49, but for VFR it was only 0.23. The differences between the slopes for individuals give rise to increased variability in performance with age. For example, a pair of participants with equal cognitive function on entry to the study whose slopes differed by 0.4 would differ by 4 points after 10 years and by 8 points after 20 years. Differences of this size are of practical importance because they are large enough to provide useful insights into the functional causes of marked individual differences in rates of cognitive decline in old age.

There are two quite different reasons why, as the members of a sample age, they should increasingly diverge in terms of their levels of cognitive performance. One is that differing genetic legacies and lifetime health histories bring about differences in trajectories of biological aging, which will diverge over time (Rabbitt, 1982, 1993a). A second is that as people age and so become less able, their performance on any task on which they are tested varies more from moment to moment and, as a direct consequence, their average levels of performance also vary more from session to session and from day to day (Rabbitt, 1999; Rabbitt, Osman, Stollery, & Moore, 2001). As day-to-day variability increases for all members of a sample, so they will differ more with respect to each other when they are all tested on any single occasion. Thus, increasing variability between members of aging samples has at least two, functionally different, causes. The possibility of confounds between these effects means that any single, cross-sectional observation of members of a population at a particular time point will give us an inaccurate, and probably exaggerated, estimate of actual individual differences in trajectories of cognitive aging. For better estimates to be obtained, longitudinal data are essential; ideally, we also need to estimate, as far as possible, the effects of session-to-session or day-to-day variability by taking several samples of performance on each task at each successive longitudinal data point. Estimates of intrinsic within-participant variability obtained from these samples will allow long-term trends resulting from differences in trajectories of change to be more precisely determined. Such data will also be useful in showing the extent to which increases in the intrinsic variability of individuals' performance, as distinct from changes in their mean levels of performance, alter as they age.

Acknowledgment

Address correspondence to Patrick Rabbitt, Age and Cognitive Performance Research Centre, Zochonis Building, University of Manchester, Manchester, M13 9PL, United Kingdom. E-mail: rabbitt@psy.man.ac.uk

References

Arenberg, D. (1974). A longitudinal study of problem solving in adults. *Gerontology, 29*, 650–658.

Baltes, P. B. (1968). Longitudinal and cross-sectional sequences in the study of age and generation effects. *Human Development, 11*, 145–171.

Bank, L., & Jarvik, L. F. (1978). A longitudinal study of aging human twins. In Schneider, E. L. (Ed.), *The genetics of aging* (pp. 303–333). New York: Plenum Press.

Bell, B., Rose, C. L., & Damon, A. (1972). The normative aging study: Interdisciplinary and longitudinal study of health and aging. *Aging and Human Development, 3*, 5–17.

Berkowitz, B. (1964). Changes in intellect with age: IV. Changes in

achievement and survival. *Newsletter for Research in Psychology, 6*, 18–20.
Birren, J. E., Butler, R. N., Greenhouse, S. W., Sokoloff, L., & Yarrow, M. R. (1963). *Human aging* (Publication No. 986). Washington, DC: U.S. Government Printing Office.
Bonaiuto, S., Rocca, W. A., & Lippi, A. (1990). Impact of education and dementia: Clinical issues. In H. A. Whitaker (Ed.), *Neuropsychological studies of non-focal brain damage* (pp. 1–15). New York: Springer.
Bosworth, H. B., Schaie, K. W., & Willis, S. L. (1999). Cognitive and sociodemographic risk factors for mortality in the Seattle Longitudinal Study. *Journal of Gerontology: Psychological Sciences, 54B*, P273–P282.
Botwinick, J., West, R. L., & Storandt, M. (1978). Predicting death from behavioral test performance. *Journal of Gerontology, 33*, 755–762.
Brodman, E., Erdman, A. J., Jr., & Wolff, H. G. (1949). *Manual for the Cornell Medical Index Health Questionnaire*. New York: Cornell University Medical School.
Burgess, P. W. (1997). Theory and methodology in executive function research. In P. M. A. Rabbitt (Ed.), *Methodology of frontal and executive function* (pp. 81–116). Hove, England: Psychology Press.
Burgess, P. W., & Shallice, T. (1996). Bizarre responses, rule detection and frontal lobe lesions. *Cortex, 32*, 241–259.
Busse, E. W., Maddox, G. L., Buckley, C. E., III., Burger, P. C., George, L. K., March, G. R., et al. (1985). *The Duke Longitudinal Studies of Normal Aging: 1955–1980*. New York: Springer.
Clark, J. W. (1960). The ageing dimension. A factorial analysis of individual differences with age on psychological and physiological measurement. *Gerontology, 15*, 183–187.
Clement, F. J. (1974). Longitudinal and cross-sectional assessments of age-changes in physical strength as related to sex, social class and mental ability. *Gerontology, 15*, 94–116.
Colsher, P. L., & Wallace, R. B. (1991). Longitudinal application of cognitive measures in a defined population of community dwelling elders. *Annals of Epidemiology, 3*, 71–77.
Costa, P. T., & McCrae, R. R. (1980). Somatic complaints in males as a function of age and neuroticism: A longitudinal analysis. *Journal of Behavioral Medicine, 3*, 245–253.
Diggle, P. J., Liang, K. Y., & Zeger, S. L. (1994). *Analysis of longitudinal data*. Oxford, England: Oxford University Press.
Dirken, J. M. (1972). *Functional age of industrial workers*. Groningen, The Netherlands: Wolters-Noordhof.
Evans, D. A., Beckett, L. A., Albert, M. S., Hebert, L. E., Scherr, P. A., Funkenstein, H. H., et al. (1993). Level of education and change in cognitive function in a community population of older persons. *Annals of Epidemiology, 3*, 77–81.
Forner, A. (1972). *Aging and society* (Vol. III). New York: Sage.
Gur, R. C., Gur, R. E., Orbist, W. D., Skolnik, B. E., & Reivitch, M. (1987). Age and regional cerebral blood flow at rest and during cognitive activity. *Archives of General Psychiatry, 44*, 617–621.
Haugh, H., & Eggers, R. (1991). Morphometry of the human cortex cerebri and cortex striatum during aging. *Neurobiology of Aging, 12*, 336–338.
Heim, A. W. (1970). *The AH4 group tests of intelligence*. Windsor, England: NFER/Nelson.
Heron, A., & Chown, S. (1967). *Age and function*. London: Churchill.
Hertzog, C., & Schaie, K. W. (1988). Stability and change in adult intelligence: 2. Simultaneous analysis of longitudinal means and covariance structures. *Psychology and Aging, 3*, 122–130.
Hertzog, C., Schaie, K. W., & Gribbin, K. (1978). Cardiovascular disease and changes in intellectual function from middle to old age. *Journal of Gerontology, 33*, 872–883.
Horn, J. L. (1982). The theory of fluid and crystallized intelligence in relation to concepts of cognitive psychology and aging in adulthood. In F. I. M. Craik & S. Trehub (Eds.), *Aging and cognitive processes* (pp. 237–278). Boston: Plenum Press.
Hultsch, D. F., Hertzog, C., Small, B. J., McDonald-Miszczak, L., & Dixon, R. A. (1992). Short-term longitudinal change in cognitive performance in later life. *Psychology and Aging, 7*, 571–584.
Jarvik, L., & Blum, K. (1971). Cognitive declines as predictors of mortality in twin pairs. In E. Palmore & F. Jeffers (Eds.), *Prediction of life span* (pp. 46–73). Lexington, MA: Heath.
Johansson, B., & Berg, S. (1989). The robustness of the terminal decline phenomenon: Longitudinal data from the Digit-Span Memory Test. *Journal of Gerontology: Psychological Sciences, 44*, P184–P186.
Johansson, B., Zarit, S. H., & Berg, S. (1992). Changes in cognitive function of the oldest old. *Journal of Gerontology: Psychological Sciences, 47*, P75–P80.
Kausler, D. H. (1990). *Experimental psychology, cognition and human aging*. New York: Springer-Verlag.
Kitagawa, E. M., & Hauser, P. M. (1973). *Differential mortality in the United States: A study in socioeconomic epidemiology*. Cambridge, MA: Harvard University Press.
Korczyn, A. D., Kahana, E., & Galper, Y. (1991). Epidemiology of dementia in Ashkelon, Israel. *Neuroepidemiology, 10*, 100.
Kucera, H., & Francis, N. (1967). *Frequency and word association norms*. Providence, RI: Brown University Press.
Lachman, R., Lachman, J. L., & Taylor, D. W. (1982). Reallocation of mental resources over the productive lifespan: Assumptions and task analyses. In F. I. M. Craik & S. Trehub (Eds.), *Aging and the cognitive process* (pp. 227–252). New York: Plenum Press.
Lansen, P. (1997). The impact of aging on cognitive functions: An 11-year follow up study of four age cohorts. *Acta Neurologica Scandinavica Supplimentum, 96* (No. 172), 172–193.
Lieberman, M. A. (1965). Psychological correlates of impending death: Some preliminary observations. *Journal of Gerontology, 20*, 181–190.
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). *Journal of Gerontology: Psychological Sciences, 57B*, P474–P482.
Lowe, C., & Rabbitt, P. M. A. (1998). Test/re-test reliability of the CANTAB and ISPOCD neuropsychological batteries: Theoretical and practical issues. *Neuropsychologia, 36*, 915–923.
Mason, K., & Mason, W. (1973). Some methodological issues in cohort analysis of archival data. *American Sociological Review, 38*, 242–258.
McInnes, L., & Rabbitt, P. M. (1997). The relationship between functional ability and cognitive ability among elderly people. In *Facts and research in gerontology* (pp. 34–45). Paris: Serdi.
Moller, J. T., Cluitmans, P., Rasmussen, L. S., Houx, P., Canet, J., Rabbitt, P., et al. (1998). Long-term post-operative cognitive dysfunction in the elderly: ISPOCD1 study. *The Lancet, 351*, 857–861.
Morse, C. K. (1993). Does variability increase with age? An archival study of cognitive measures. *Psychology and Aging, 8*, 156–164.
Nesselrode, J. R., & Baltes, P. B. (1979). *Longitudinal research in the study of behaviour and development*. New York: Academic Press.
Office of Population Censuses & Surveys. (1980). *Classification of occupations*. London: Her Majesty's Stationary Office.
Owens, W. A. (1953). Age and mental abilities; a longitudinal study. *Genetic Psychological Monographs, 48*, 3–54.
Owens, W. A. (1966). Age and mental abilities: A second adult follow-up. *Journal of Educational Psychology, 57*, 311–325.
Palmore, E. (1978). When can age, period and cohort be separated? *Social Forces, 57*, 282–295.
Payton, A., Holland, F., Diggle, P., Rabbitt, P., Horan, M., Davidson, Y., et al. (2003). Cathepsin D exon 2 polymorphism associated with general intelligence in a healthy older population. *Molecular Psychiatry, 8*, 14–18.
Pendleton, N., Payton, A., van den Boogerd, E. H., Holland, F., Diggle, P., Rabbitt, P. M. A., et al. (2002). Apolipoprotein E genotype does not predict decline in intelligence in older adults. *Neurosciences Letters, 324*, 74–76.
Powell, D. H. (1994). *Profiles in cognitive aging*. Cambridge, MA: Harvard University Press.
Rabbitt, P. M. A. (1982). Cognitive psychology needs models for old age. In A. D. Baddeley & J. Long (Eds.), *Attention and performance IX* (pp. 142–165). Hove, England: Erlbaum.
Rabbitt, P. M. (1993a). Does it all go together when it goes? *Quarterly Journal of Experimental Psychology, 46(A)*, 385–433.
Rabbitt, P. M. A. (1993b). Crystal quest: An examination of the concepts of "fluid" and "crystallised" intelligence as explanations for cognitive changes in old age. In A. D. Baddeley & L. Weiskrantz (Eds.), *Attention, selection, awareness and control* (pp. 197–231). Oxford, England: Oxford University Press.
Rabbitt, P. M. A. (1999). Measurement indices, functional characteristics and psychometric constructs in cognitive ageing. In T. J. Perfect & E. A. Maylor (Eds.), *Models of cognitive aging* (pp. 160–187). Oxford, England: Oxford University Press.
Rabbitt, P. M. A., Banerji, N., & Szemanski, A. (1989). Space Fortress

as an IQ test? Predictions of learning and of practised performance in a complex video game. *Acta Psychologica, 71*, 243–257.
Rabbitt, P., Bent, N., & McInnes, L. (1997). Health, age and mental ability. *The Irish Journal of Psychology, 18*, 104–131.
Rabbitt, P. M., Chetwynd, A., & McInnes, L. (2003). Do clever brains age more slowly? Further exploration of a Nun result. *British Journal of Psychology, 94*, 63–71.
Rabbitt, P. M., Donlan, C., Bent, N., McInnes, L., & Abson, V. (1993). The University of Manchester Age and Cognitive Performance Research Centre and North East Age Research longitudinal programmes 1982 to 1997. *Zeitschrift Gerontology, 26*, 176–183.
Rabbitt, P. M. A., Donlan, C., Watson, P., McInnes, L., & Bent, N. (1996). Unique and interactive effects of depression, age socio-economic advantage and gender on cognitive performance of normal healthy older people. *Psychology & Aging, 10*, 221–235.
Rabbitt, P. M. A., Osman, P., Stollery, B., & Moore, B. (2001). There are stable individual differences in performance variability, both from moment to moment and from day to day. *Quarterly Journal of Experimental Psychology, 54A*, 981–1003.
Rabbitt, P. M., Watson, P., Donlan, C., Bent, L., & McInnes, L. (1994). Subject attrition in a longitudinal study of cognitive performance in community-based elderly people. In B. J. Vellas, J. L. Albarede, & P. J. Garry (Eds.), *Facts and research in gerontology* (pp. 203–207). Paris: Serdi.
Rabbitt, P., Watson, P., Donlan, C., McInnes, L., Horan, M., Pendleton, N., et al. (2002). Effects of death within 11 years on cognitive performance in old age. *Psychology and Aging, 17*, 1–14.
Raven, J. C. (1965). *The Mill Hill Vocabulary Scale*. London: Lewis.
Reimanis, G., & Green, R. (1971). Immanence of death and intellectual decrement in the aging. *Developmental Psychology, 5*, 270–272.
Riegel, K. F., & Riegel, R. M. (1972). Development, drop and death. *Developmental Psychology, 6*, 306–348.
Riegel, K. F., Riegel, R. M., & Myer, G. (1967). A study of the drop-out rates in longitudinal research on aging and the prediction of death. *Journal of Personality and Social Psychology, 5*, 342–348.
Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*, 581–592.
Schaie, K. W. (1996). *Intellectual development in adulthood: The Seattle Longitudinal Study*. New York: Cambridge University Press.
Schaie, K. W., & Labouvie-Vief, G. (1974). Generational versus ontogenetic components of change in adult cognitive functioning: A fourteen-year cross-sequential study. *Developmental Psychology, 10*, 305–320.
Schaie, K. W., Labouvie, G. V., & Barrett, T. J. (1973). Selective attrition effects in a fourteen-year study of adult intelligence. *Gerontology, 28*, 328–334.
Schaie, K. W., & Strother, C. R. (1968). A cross-sequential study of age changes in cognitive behaviour. *Psychological Bulletin, 70*, 671–680.
Schaie, K. W., & Willis, S. L. (1993). Age difference patterns of psychometric intelligence in adulthood: Generalizability within and across ability domains. *Psychology and Aging, 8*, 44–55.
Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for non-ignorable drop-out using semiparametric non-response models (with discussion). *Journal of the American Statistical Association, 94*, 1096–1120.
Scheibel, M. E., & Scheibel, A. B. (1975). Structural changes in the aging brain. In H. Brody, D. Harmon, & J. M. Ordy (Eds.), *Aging* (Vol. 1, pp. 11–37). New York: Raven Press.
Schlesselman, J. J. (1973a). Planning a longitudinal study: I. Sample size determination. *Journal of Chronic Diseases, 26*, 532–560.
Schlesselman, J. J. (1973b). Planning a longitudinal study: II. Frequency of measurement and study duration. *Journal of Chronic Diseases, 26*, 561–570.
Schulsinger, F., Knop, J., & Mednick, S. A. (1981). *Longitudinal research*. Lexington, MA: Hinghams.
Shallice, T., & Burgess, P. W. (1991). Higher-order cognitive impairments and frontal lobe lesions in man. In H. S. Levin, H. M. Eisenberg, & A. L. Benton (Eds.), *Frontal lobe function and dysfunction* (pp. 211–259). New York: Oxford University Press.
Shaw, T. G., Mortel, K. F., Meyer, J. S., Rogers, R. L., Hardenberg, J., & Cutaia, M. M. (1984). Cerebral blood flow changes in benign aging and cerebrovascular disease. *Neurology, 34*, 855–862.
Shock, N. W., Greulich, R. C., Andres, R., Arenberg, D., Costa, P. T., Jr., Lakatta, E. G., et al. (1984). *Normal human aging: The Baltimore Longitudinal Study of Aging* (NIH Publication No. 84-2450). Washington DC: U.S. Government Printing Office.
Sliwinski, M., & Buschke, M. (1999). Cross-sectional and longitudinal relationships among age, cognition and information processing speed. *Psychology and Aging, 14*, 18–33.
Small, B. J., & Backman, L. (1997). Cognitive correlates of mortality: Evidence from a population-based sample of very old adults. *Psychology and Aging, 12*, 309–313.
Snowden, D. A., Kemper, S. J., Mortimer, J. A., Greinier, L. H., Wekstein, D. R., & Markesbery, W. R. (1996). Linguistic ability in early life and cognitive function and Alzheimer's disease in later life. Findings from the Nun study. *Journal of the American Medical Association, 21*, 528–532.
Terman, L. M., & Oden, M. H. (1947). *Genetic studies of genius. Vol. 4: The gifted child grows up*. Stanford, CA: Stanford University Press.
Terman, L. M., & Oden, M. H. (1959). *Genetic studies of genius. Vol. 5: The gifted group at mid life*. Stanford, CA: Stanford University Press.
Verbeck, G., & Molenberghs, S. (1997). *Linear mixed models in practice. A SAS-oriented approach*. New York: Springer-Verlag.
Voitenko, V. P., & Tokar, A. V. (1983). The assessment of biological age and sex differences of human aging. *Experimental Aging Research, 9*, 239–244.
Zelinski, E. M., & Burnight, K. P. (1997). Sixteen-year longitudinal and time lag changes in memory and cognition in older adults. *Psychology and Aging, 12*, 503–513.
Zelinski, E. M., Gilewski, M. J., & Stewart, K. W. (1993). Individual differences in cross-sectional and 3 year longitudinal memory performance across the adult life span. *Psychology and Aging, 8*, 176–186.
Zelinski, E. M., & Stewart, S. T. (1998). Individual differences in 16-year memory changes. *Psychology and Aging, 13*, 622–630.

Received February 8, 2001
Accepted August 18, 2003
Decision Editor: Margie E. Lachman, PhD