Psychology and Aging
1987, Vol. 2, No. 3, 233-242

0882-7974/87/$00.75

Effects of Cognitive Training on Primary Mental Ability Structure

K. Warner Schaie, Sherry L. Willis, Christopher Hertzog, and John E. Schulenberg
Department of Individual and Family Studies, Pennsylvania State University

We report results of the first empirical test, as far as we know, of the assumption of structural invariance of latent constructs from pretest to posttest in cognitive training research on the elderly. In all, 401 participants in the Seattle Longitudinal Study, over 62 years old, received a 5-hr test battery at pre- and posttest that included 16 ability tests, marking the five primary abilities of Spatial Orientation, Inductive Reasoning, Numerical Ability, Verbal Ability, and Perceptual Speed. A total of 229 of our subjects received 5 hr of individual training on either Spatial Orientation or Inductive Reasoning. Restricted factor analysis with the LISREL algorithm tested the hypothesis of measurement equivalence across test occasions, separately for the control subjects and for each of the training groups. When ability-specific cognitive training intervenes, no structural change is observed for abilities not subject to intervention. However, slight shifts occurred in the optimal regression weights for the different markers for the training target abilities.

During the past few years there has been a growing interest among researchers to determine whether the cognitive performance of older adults can be improved by means of training interventions. Their overarching goal has been to demonstrate training gain on a latent construct (i.e., an intellectual ability factor), rather than to demonstrate that it is possible to improve subjects' performance on a specific measure by "teaching the test." In practice, however, training researchers have typically examined pretest–posttest performance gains for individual tests, rather than examine them at the construct level, and have reported significant training effects for one or more measures believed to represent the ability construct (Baltes & Willis, 1982, Sterns & Sanders, 1980).

These test-specific analyses have led some investigators (Donaldson, 1981; Birren, Cunningham, & Yamamoto, 1983) to criticize training effects as being narrow and limited. However, in recent studies we (Willis & Schaie, 1986) have reported significant training effects for two primary abilities at the latent construct level, providing further evidence that training improvement extends beyond merely teaching the test. In these studies we were able to test training effects at the construct level by assessing each construct with multiple marker tests. Moreover, the strategies that we used were relevant to the construct rather than to the specific features of the marker tests used to assess the construct. Previous assessments of training effects at both the test-specific and the construct level, however, have focused almost exclusively on *quantitative* change in subjects' performance.

A critical assumption that underlies evaluation of quantitative change in training performance—whether at the level of the construct or a test—is that the relation between the ability constructs and measures of these constructs (psychometric tests) in the assessment battery remains invariant from pretest to posttest. That is, quantitative comparisons are meaningful *only* if there is qualitative invariance (cf. Baltes & Nesselroade, 1973). For example, Donaldson (1981) has suggested that educational training procedures are crystallized in nature and may alter the character of the fluid abilities that were the target of training, such that fluid measures after training become more representative of the crystallized rather than the fluid intelligence domain. We are not explicitly testing the fluid/crystallized model. However, if Donaldson was correct, then we can deduce the following: Training on a primary ability should change the primary ability factor structure. Therefore, if we can show that training does not alter the factor structure, then we can reject Donaldson's interpretation of cognitive training effects. It may be argued that if Donaldson's hypothesis were shown to be correct, such results would seriously question the traitlike character of the fluid and crystallized intelligence domains (cf. Willis & Baltes, 1981). Nevertheless, we agree that assessment of structural change in the ability constructs is needed in training research.

How would such qualitative change in factor structure be manifested? The literature on comparative factor analysis indicates that the quintessential evidence for factorial invariance is the equality of unstandardized factor pattern weights (factor loadings; see Hertzog & Schaie, 1986; Meredith, 1964; Schaie & Hertzog, 1985). Following the terminology of Horn, McArdle, and Mason (1983), we can distinguish two levels of invariance in factor loadings (with different implications for training research): *configural invariance* and *metric invariance*.

Configural invariance requires that marker measures have their primary loading on the same ability constructs (i.e., ability factors) across occasions. If configural invariance is not main-

Research reported in this article was supported by Grant #R01 AG03544 from the National Institute on Aging.

We gratefully acknowledge the enthusiastic cooperation of members and staff of the Group Health Cooperative of Puget Sound.

Christopher Hertzog is now at the School of Psychology, Georgia Institute of Technology.

Correspondence concerning this article should be addressed to K. Warner Schaie, Department of Individual and Family Studies, S-110 Human Development Building, Pennsylvania State University, University Park, Pennsylvania 16802.

tained after training, then it is likely that the intervention may have produced qualitative changes in ability structure, perhaps indicating that the attributes measured by the tests have changed as a function of training. If this were the case, interpretation of quantitative training effects would then be ambiguous.

Metric invariance requires not only that markers have their primary loading on the same ability construct, but also that the magnitude of the loadings can be constrained to be equal between pretest and posttest. It seems reasonable to hypothesize, even if configural invariance after training is maintained, that training could cause pretest–posttest changes in the magnitude of the factor loadings for measures of abilities that had been trained. That is, it may not be possible to obtain complete metric invariance due to pretest–posttest shifts in the magnitude of the factor loadings for Tests A and B, even though the tests mark the same ability factor at both occasions. Such a lack of metric invariance would suggest problems for the interpretation of quantitative changes in individual tests, but such problems could readily be surmounted provided quantitative change can be assessed at the level of factor scores rather than observed scores. We hypothesize, therefore, that training will change the magnitude of factor loadings only for the abilities targeted for training, given the transfer of training literature that shows limited transfer of training gains to abilities not specifically trained (Baltes & Willis, 1982; Sterns & Sanders, 1980; Willis & Schaie, 1981). In other words, there should be no shifts in magnitude of factor loadings for measures of other ability constructs that were not the target of training.

A second important issue in training research focuses on the stability of individual differences across training intervention. Previous descriptive research findings indicate that adult individual differences in ability performance are highly stable. For example, Hertzog and Schaie (1986) found longitudinal correlations of a general ability factor with itself to exceed .9 over a 14-year interval. However, to the extent that the training intervention is differentially effective, then one would expect lower pretest–posttest stability in a training group than in a corresponding no-training control group. Conversely, stability of individual rankings across the intervention would indicate uniform, rather than differential training effects, given that significant training gain at the mean level has been shown (Willis & Schaie, 1986).

Data reported in this study are from a cognitive intervention study conducted with elderly subjects from the Seattle Longitudinal Study (SLS; Schaie, 1983), who received cognitive training on one of two primary mental abilities, Inductive Reasoning or Spatial Orientation. The study involved a pretest–posttest control group design, with subjects assessed at pretest and posttest on a broad measurement battery marking the five primary mental abilities of Inductive Reasoning, Spatial Orientation, Verbal, Number, and Perceptual Speed. Prior data analyses indicated significant training effects at the factorial level for both the inductive reasoning and spatial orientation training groups (Willis & Schaie, 1986).

In this article, we report the application of a repeated measures factor model to assess two issues: (a) the pretest–posttest factorial invariance of the ability battery, and (b) the stability from pretest to posttest of individual differences within each training group with respect to the target ability.

With respect to the first issue, the repeated measures factor analysis permits tests of invariance in the factor structure of the ability battery from pretest to posttest. These analyses have been conducted using the LISREL approach outlined by Jöreskog (1979; see also Hertzog, in press; Schaie & Hertzog, 1985). As discussed by Schaie and Hertzog (1985; see also Hertzog & Schaie, 1986), the critical test of change in the measurement properties of repeated measures data involves the test of invariance over time in the (unstandardized) regressions of variables on factors (i.e., the metric invariance in factor pattern loadings). With respect to changes in factor structure, we hypothesized that (a) the control group would show no changes between pretest and posttest in the unstandardized factor pattern loadings (regression coefficients relating tests to ability factors), (b) that the training groups would show no change in factor pattern loadings for nontarget abilities (i.e., Verbal, Number, Perceptual Speed), and (c) that any changes in the magnitude of factor loadings would be restricted to the ability trained in each training group (e.g., inductive reasoning ability in the induction training group).

With respect to the second issue, we hypothesized that pretest–posttest correlations of ability would be virtually perfect (i.e., stability of individual differences) except for the effects of training. With respect to differential pretest–posttest change, we hypothesized that (a) individual differences in primary ability factors would be highly stable (correlations above .9 in untrained control subjects), (b) individual differences on untrained abilities would be similarly stable in the training groups, but that (c) to the extent that training effects are differential, autocorrelations will be lower for trained abilities (i.e., for Spatial Orientation in the spatial training group, for Inductive Reasoning in the induction training group). The advantage of the LISREL model for this analysis is the fact that estimates of stability of individual differences are made at the level of the ability factor and are purged of the attenuating influences of measurement error on the correlations (Jöreskog & Sörbom, 1977; Schaie & Hertzog, 1985).

Method

Subjects

Our sample consisted of 401 participants (224 women and 177 men) over the age of 62 years from the Seattle metropolitan area, who had been participants in the Seattle Longitudinal Study (SLS; Schaie, 1983) since 1975 or earlier. This study includes longitudinal samples of subjects initially tested at three measurement periods (in 1956, in 1963, and in 1970). In addition, samples initially drawn in 1974 and 1975 for specialized research questions regarding the population served as a no-treatment control. All subjects are, or had been, members of the Group Health Cooperative of Puget Sound, a health maintenance organization.

Identical recruitment procedures have been used in all of the samples, that is, definition of a sampling frame by random draws from the health maintenance organization, followed by mail solicitation (see Schaie, 1983, for additional details on recruitment procedures in the SLS). Mean age of the total sample was 72.5 years (range = 64–95, SD = 6.41). Mean educational level was 13.9 years (range = 6–20, SD = 2.98). There were no sex differences in age or educational level. Mean income level was $19,879 (range = $1,000–$33,000, SD = $8,520). All of the subjects were community dwelling and most were White.

Design and Procedure

Training paradigm. All of the subjects received a 5-hr ability test battery at pretest and posttest. Training subjects (n = 229) received 5 hr of individual cognitive training on either Spatial Ability (n = 118) or Inductive Reasoning (n = 111). The remaining 172 subjects were used as a no-treatment control group, receiving only the pretest and posttest. Subjects were assigned to training groups on the basis of previous longitudinal patterns of change in Spatial Ability and Inductive Reasoning (see ahead).

Classification of participants. Training subjects' test performances on the Thurstone (1948) Primary Mental Ability (PMA) Reasoning and Spatial Orientation measures were classified as having remained stable or having declined over the prior 14-year interval (1970–1984). Because subjects entered the study at different points in time (1956–1970), performance in 1970 was used as a common baseline. Subjects were first classified by placing a 1-*SEM* confidence interval about their observed base score (cf. Dudek, 1979). If their 1984 score fell below this interval, they were considered to have declined, otherwise to be stable. No-treatment control subjects were initially tested in 1974–1975. They were consequently classified on the basis of performance changes from 1974–1975 to 1984. The desired assignment on the basis of past longitudinal patterns of change made random assignment to treatment and control groups impractical. The design therefore represents a nonequivalent groups design with a no-treatment control group (see Cook & Campbell, 1979). Across the treatment and control groups there were 170 subjects (42.4% of the sample) who were classified as having remained stable on the training target abilities, whereas 231 subjects (57.6%) had declined on one or both of the abilities.

Assignment of subjects. Subjects were assigned to either reasoning or space training programs on the basis of their past performance on these variables. Subjects who had declined on Inductive Reasoning, but not on Spatial Ability, or vice versa, were assigned to the training program for the ability exhibiting decline. Subjects who had remained stable on both abilities or had shown decline on both abilities were randomly assigned to one of the training programs.

Procedure. The study involved a pretest–treatment–posttest control-group design. In addition to the testing-only control group, the reasoning training group served as a treatment control for the space training group, and vise versa. The test battery was administered in two 2½-hr sessions conducted in small groups. Training involved five 1-hr, individually conducted training sessions. The majority of subjects were trained in their homes. Two middle-aged trainers, with prior educational experience in working with adults, served as trainers. Following training, subjects were assessed on a posttest battery involving the same measures administered at pretest. An interval of approximately one month separated pretest from posttest in both training and control groups. Subjects were paid $100 for participation in the study.

Measures

The test battery included psychometric measures representing five primary mental abilities. The battery included the Thurstone (1948) PMA measures, administered at previous SLS assessments. Additional measures were selected from other sources, principally the Educational Testing Service (ETS) reference kit (Ekstrom, French, Harman, & Derman, 1976) or the Adult Development and Enrichment Program (ADEPT) training battery (Baltes & Willis, 1982). Tests were selected on the basis of empirical evidence (e.g., Baltes, Cornelius, Spiro, Nesselroade, & Willis, 1980; Ekstrom et al., 1976) indicating that these tests would be relatively pure markers of the targeted ability factors. Each ability was represented by three to four markers (see Table 1). All of the tests are administered under time limits and are slightly speeded.

Table 1 also reports the test–retest correlations of these indicators in the control group. Under the assumption of perfect stability of individual differences in the true scores, these correlations estimate the reliability of the tests (Schaie & Hertzog, 1985). To the extent that individual differences are not perfectly stable, these correlations underestimate the markers' reliability. The results to be reported suggest that individual differences on the latent factors are almost perfectly correlated over time, so the degree of bias in these reliability estimates should be low. The correlations are all greater than .8, indicating satisfactory reliability for all instruments.

Table 1
Intellectual Abilities Measurement Battery

Primary ability and test	Source	Test–retest correlation
Inductive Reasoning		
PMA Reasoning	Thurstone, 1948	.884
ADEPT Letter Series (Form A)	Blieszner, Willis, & Baltes, 1981	.839
Word Series	Schaie, 1985	.852
Number Series	Ekstrom, French, Harman, & Derman, 1976	.833
Spatial Orientation		
PMA Space	Thurstone, 1948	.817
Object Rotation	Schaie, 1985	.861
Alphanumeric Rotation	Willis & Schaie, 1983	.820
Perceptual Speed		
Finding A's	Ekstrom et al., 1976	.814
Number Comparison	Ekstrom et al., 1976	.860
Identical Pictures	Ekstrom et al., 1976	.865
Numeric		
PMA Number	Thurstone, 1948	.875
Addition	Ekstrom et al., 1976	.937
Subtraction & Multiplication	Ekstrom et al., 1976	.943
Verbal		
PMA Verbal	Thurstone, 1948	.890
Vocabulary II	Ekstrom et al., 1976	.828
Vocabulary IV	Ekstrom et al., 1976	.954

Note. PMA = Primary Mental Ability; ADEPT = Adult Development and Enrichment Program.

Spatial Orientation. All of these tests (PMA Space, Object Rotation, Alphanumeric Rotation) are multiple response measures of two-dimensional mental rotation ability. The subject is shown a model line drawing and asked to identify which of six choices shows the model drawn in different spatial orientations. There are two or three correct responses possible for each test item. The Object Rotation test (Schaie, 1985) and the Alphanumeric test were constructed such that the angle of rotation in each answer choice is identical with the angle used in the PMA Spatial Orientation test (Thurstone, 1948). The three tests vary in item content. Stimuli for the PMA test are abstract figures; the Object Rotation test involves drawings of familiar objects, and the Alphanumeric test contains letters and numbers.

Inductive Reasoning. The PMA Reasoning measure (Thurstone, 1948) assesses inductive reasoning ability via letter series problems. The subject is shown a series of letters and must select the next letter in the series from five letter choices. The ADEPT Letter Series test (Blieszner, Willis, & Baltes, 1981) also contains letter series problems; however, some of the problems involve pattern description rules other than those found on the PMA measure. The Word Series test (Schaie, 1985) parallels the PMA measure in that the same pattern description rule is used for each item; however, the test stimuli are days of the week or months of the year, rather than letters. The Number Series test (Ekstrom et al., 1976) involves series of numbers rather than letters and involves differ-

ent types of pattern description rules involving mathematical computations.

Perceptual Speed. All Perceptual Speed measures come from the ETS factor reference kit (Ekstrom et al., 1976). Finding A's involves the cancellation of the letter *a* in columns of words, about half of which contain that letter. Picture Identification requires the subject to find the match among five simple test figures to a stimulus figure. Number Comparison involves comparing two sets of eight-digit numbers and marking those pairs that are not identical.

Numerical Ability. The first measure of Numerical Ability was the PMA Number test, which involves the checking of simple addition problems (Thurstone, 1948). The Addition test (Ekstrom et al., 1976) involves calculating the sums of four two-digit numbers. The Subtraction and Multiplication test (Ekstrom et al., 1976) requires calculating the sums and products for alternate rows of simple subtraction and multiplication problems.

Verbal Ability. All measures of Verbal Ability are multiple choice tests that require selecting a synonym for a stimulus word from four alternatives. The first measure is the PMA Verbal Meaning test (Thurstone, 1948). The other two measures are Levels 2 and 4, respectively, from the ETS factor reference kit (Ekstrom et al., 1976).

Training Programs

The focus of the training was on facilitating the subject's use of effective cognitive strategies identified in previous research on the respective abilities. A content task analysis was conducted on the two PMA measures representing the training target abilities. For each item of the PMA Reasoning test, the pattern description rule(s) used in problem solution were identified. Practice problems and exercises were developed on the basis of these pattern description rules. Subjects were taught through modeling, feedback, and practice procedures to identify the pattern description rules. A content task analysis of the PMA Space test was conducted to identify the angle of rotation for each answer choice. Practice problems were developed to represent the angle rotations identified in the task analysis (45°, 90°, 135°, and 180°). Cognitive strategies to facilitate mental rotation that were focused on in training included (a) development of concrete terms for various angles, (b) practice with manual rotation of figures prior to mental rotation, (c) practice with rotation of drawings of concrete, familiar objects prior to introduction of abstract figures, (d) subject-generated names for abstract figures, and (e) having the subject focus on two or more features of the figure during rotation. Further details of the training procedures are reported in Schaie and Willis (1986).

Statistical Procedure

The evaluation of equivalence in the factor structure of the psychometric battery in the different training groups was conducted by using LISREL VI (Jöreskog & Sörbom, 1984) to perform confirmatory factor analysis (see Jöreskog, 1971, Jöreskog & Sörbom, 1977, and Schaie & Hertzog, 1985, for further discussions of the technique). The analyses reported in this article used only one of LISREL's two factor-analysis measurement models. In LISREL notation, the measurement model may be specified as

$$y = \Lambda\eta + \epsilon, \tag{1}$$

which in matrix form specifies a p-order vector of observed variables, y, as a function of their regression on m latent variables (factors) in η, with regression residuals ϵ. The $p \times m$ matrix Λ contains the regression coefficients (factor loadings). Equation 1 implies that the covariance matrix of the observed variables in the populations, Σ, may be expressed as

$$\Sigma = \Lambda\Phi\Lambda' + \Theta, \tag{2}$$

where Λ is as before, Φ is the covariance matrix of the η, and Θ is the covariance matrix of the ϵs. Equation 2 should be recognized as a restricted factor analysis model that can be generalized to a multiple group model (Jöreskog & Sörbom, 1984).

The parameters of LISREL's restricted factor analysis model are estimated by the method of maximum likelihood, provided that a unique solution to the parameters has been defined by placing a sufficient number of restrictions in the second equation to identify the remaining unknowns. Restrictions are specified by either (*i*) fixing parameters to a known value a priori (e.g., requiring that a variable is unrelated to a factor by fixing its regression in Λ to 0) or (*ii*) constraining a set of two or more parameters to be equal. Overidentified models (which have more restrictions than are necessary to identify the model parameters) place restrictions on the hypothesized form of Σ, which may be used to test the goodness of fit of the model to the data using the likelihood chi-square test statistic. Differences in chi-square between *nested* models (models that have the same specification, with additional restrictions in one model) may be used to test the null hypothesis that the restrictions are true in the population. For example, a more restrictive model (i.e., with more restrictions placed on the model parameters) that is nested within a less restrictive model would be accepted over the less restrictive model if the difference in chi-square between the two models is not significant. Conversely, if the difference in chi-square is significant, then the less restrictive model would be accepted.

Another index of model fit reported is the LISREL goodness-of-fit (relative fit) index (GFI; Jöreskog & Sörbom, 1984). This index would be 1.0 if a perfect fit of the model to the data were obtained. The advantage of such relative fit indices is that they are less influenced by sample size than is the chi-square fit statistic, (e.g., Bentler & Bonett, 1980), although the interpretation of such statistics is a matter of ongoing evolution and controversy. Factor models with fits in the .8 to .9 range, as the ones we report ahead, may generally be considered useful approximations to the underlying "true" model, even though they do not account for all bivariate covariances in the data, provided that alternative specifications have been evaluated and ruled out. In initial model development, we carefully evaluated diagnostic fit indices such as LISREL modification indices and residual correlations before proceeding to address the group differences in factor structure.

In multiple groups analysis it is necessary to estimate factor models using covariance metric and sample covariance matrices rather than to analyze separately standardized correlation matrices. Standardization could obscure invariant relationships because of group differences in observed variances (Jöreskog, 1971). The covariance metric approach requires estimation of factor variances (rather than the traditional procedure of fixing these parameters to unity), identifying the metric of the factors by fixing a single regression in each column of Λ to the constant 1. The additional advantage for the covariance metric approach for longitudinal data is that changes in factor variances over time may be evaluated (Schaie & Hertzog, 1985). However, standardized factor loadings, factor correlations, and unique variances are easier to interpret. We generally report parameter estimates that have been rescaled to a standardized metric, using a SAS Proc Matrix Scaling program (Hertzog & Cannon, 1985). This program extends formulae supplied by Jöreskog (1971; see also Alwin & Jackson, 1981) to handle longitudinal designs (as is the case in our test–retest designs). The rescaling preserves group and pretest–posttest differences in variances but returns scaled values for factor loadings that are interpretable as standardized factor loadings. However, we also report maximum likelihood estimates and standard errors for certain parameters (e.g., ability factor variances). In general, parameters that exceed their standard errors by a ratio of 2:1 are reliably different from zero at approximately a 5% (per comparison) alpha level.

Preliminary Analyses

Before proceeding with training group comparisons, we first used the pretest data to select an appropriate factor model for the intelligence

battery described in Table 1. We hypothesized that five factors would be identified in the analysis: Induction, Space, Perceptual Speed, Number, and Verbal. The first examination of this hypothesis was done by inspecting the eigenvalues of the correlation matrix via a scree test. The pattern supported a five-factor representation of the data. Subsequent confirmatory factor analysis indicated that the five factors, as specified, did a relatively good job of accounting for the covariances among the psychometric tests.

Given that the training analysis classified groups by prior developmental history (i.e., stable levels vs. declining levels of ability; see Schaie & Willis, 1986), it was necessary to evaluate the metric invariance of the ability factor structure across the stable and decline groups. Models requiring metric invariance in the factor analysis parameter matrices across the two groups showed that the stable and unstable groups could be considered to have equivalent factor pattern weights and factor covariance matrices, $\chi^2(243, N = 401) = 463.17$, GFI stable = .847, GFI unstable = .892. This configuration suggests complete metric invariance of the solution in the common factor space for the stable and unstable groups, justifying pooling the data over the stable and unstable developmental pattern cases for training group comparisons. A similar analysis showed both configural and metric invariance across men and women in the samples, obviating any concern that training group differences in factor structure might represent confounded gender differences, $\chi^2(243, N = 401) = 466.22$, GFI men = .851, GFI women = .904. Finally, we specifically examined the invariance of the factor structure in the three training groups (space, inductive reasoning, control) at pretest. Given nonrandom assignment, one could argue that these groups might differ prior to training. Our test of metric invariance of factor pattern weights yielded an acceptable fit across groups, $\chi^2(305, N = 401) = 511.55$, GFI space = .783, GFI reasoning = .871, GFI control = .902, although the fit was slightly better for the reasoning training and control groups then for the space training group. An even better fit was obtained for a model testing configural invariance, $\chi^2(279, N = 401) = 452.58$, GFI space = .818, GFI reasoning = .884, GFI control = .913. We concluded therefore that there were no group differences in definitions of factors at pretest.

Results

The analysis consisted of a set of longitudinal factor analyses of the pretest–posttest data, separately in each of the training groups (control, inductive reasoning, and space).[1] We began with the control group analysis. The basic model extended the five-factor model developed for the pretest data to a repeated measures factor model for pretest–posttest data. It specified the same five factors at both pretest and posttest. As a result, a total of 10 factors is specified and the factor covariance matrix includes factor variances and covariances within each occasion (pretest or posttest) and covariances between pretest and posttest factors. The model also specified correlated residuals (specific components) to allow test-specific relations across time. These covariances are orthogonal to the covariances of the same factors over time and are needed to provide unbiased estimates of the stability of individual differences in the factors (see Hertzog & Schaie, 1986; Sörbom, 1975). For example, the model included a residual covariance of the residual for PMA Space at pretest and PMA Space at posttest.

The fit of the basic model was adequate, $\chi^2(399, N = 172) = 552.37$, GFI = .838. Inspection of the results, including fit diagnostics and parameter estimates, indicated that the 10-factor representation appeared to be a plausible representation of the data. The basic measurement model was then used as the basis for evaluating the hypothesis of longitudinal invariance in the factor pattern weights (Λ). A model constraining the corresponding loadings to be equal between pretest and posttest showed some indication of strain on the model, $\chi^2(412, N = 172) = 574.84$, GFI = .833. The change in fit was just significant at the .05 but not at the .01 level (change in $\chi^2(13, N = 172) = 22.47$, $p < .05$). The loss of fit was not large, but it was decided to provisionally treat the outcome as a rejection of the null hypothesis.[2] However, examination of the LISREL modification indices gave no indication of high stress on the constrained equal factor loadings. The indicator with the highest modification index, Word Series on the Induction factor, was next allowed to vary over occasion. This modification did not give a significant improvement in fit, change in $\chi^2(1, N = 172) = 2.92$, $p > .10$, nor did the LISREL GFI increase appreciably. We therefore concluded that the most parsimonious model was one that treats the factor pattern matrix as invariant between pretest and posttest. Table 2 provides the rescaled factor loadings, standardized unique variances and correlated errors, and factor intercorrelations for the accepted model.

Several features of the control group solution are noteworthy. First, the factor loadings of variables on their associated primary ability factors are generally high, and the proportions of unique variance are low. All of the factor loadings shown in Table 3 were significantly different from zero. The lowest loading involved Verbal Meaning on the Verbal Comprehension factor; indeed, the PMA Verbal Meaning test seems to be more closely related to Perceptual Speed than to Verbal Comprehension. In all other cases, however, the factor loadings exceed .6 and usually exceeded .8. Correspondingly, communalities of the variables were usually greater than .5 (greater than 50% of the variance determined by the factor). Second, the test–retest correlations of the latent variables were equal to or just less than a perfect 1.0 for all five abilities. Thus, individual differences on the abilities were almost perfectly preserved over the approximately one-month retest interval. Third, the autocorrelations of test-specific components were significant in all cases. Fourth, the control group displayed a tendency for increasing factor variances at posttest. This information is provided in Table 4.

The control group analysis provides a benchmark against which to evaluate the changes in factor structure and individual differences in the training groups.

Induction training group. The fit of the basic longitudinal factor model to the induction training group, compared to the fit of the model for the control group data, was not quite as good, $\chi^2(399, N = 111) = 574.43$, GFI = .774. The parameter estimates, however, were of similar magnitude. In testing the model requiring equivalence of the factor loadings between pretest and posttest, it was found that the fit decreased significantly, $\chi^2(412, N = 111) = 599.00$, GFI = .767; change in $\chi^2(13,$

[1] Originally, we attempted to use a simultaneous factor analysis in all three groups. This model contained too many free parameters and did not achieve a converged solution in over 600 CPU seconds! We therefore decided to estimate the model in each of the training condition groups separately.

[2] Given the nature of goodness-of-fit evaluation in structural models, the temptation is to *accept* the null hypothesis and argue for factorial invariance. A liberal Type I criterion is therefore advisable.

Table 2
Rescaled Solution for Accepted Pretest–Posttest Measurement Model for Control Group

Variable	Factor loadings[a]					Unique variance		Unique autocorrelation
	Induction	Space	Perceptual Speed	Number	Verbal	Pretest	Posttest	
PMA Reasoning	.939					.102	.132	.156
ADEPT Letter Series	.890					.255	.162	.332
Word Series	.894					.202	.200	.387
Number Series	.791					.390	.359	.573
PMA Space		.823				.322	.323	.557
Object Rotation		.861				.301	.218	.571
Alphanumeric Rotation		.859				.296	.229	.386
Finding A's			.606			.636	.629	.707
Number Comparison			.715	.144		.317	.332	.597
Identical Pictures			.832			.297	.316	.567
PMA Number				.912		.166	.172	.343
Addition				.943		.121	.100	.543
Subtraction & Multiplication				.901		.202	.177	.746
PMA Verbal Meaning			.631		.440	.205	.152	.379
Vocabulary II					.888	.286	.121	.170
Vocabulary IV					.910	.178	.167	.724

Note. PMA = Primary Mental Abilities; ADEPT = Adult Development and Enrichment Program.
[a] Because the accepted measurement model included the factor loadings being set equal over time, this matrix was identical for both pretest and posttest.

$N = 111) = 24.57, p < .05$. This statistically reliable difference was not surprising, given differences in the same models found in the control group. We hypothesized in advance that any shifts in factor pattern weights for the induction training group would be found primarily in the induction measures. A model constraining only the induction markers to be equal also fit significantly worse than the unconstrained measurement model, change in $\chi^2(3, N = 111) = 16.15, p < .001$. It appeared that most of the lack of fit (i.e., 16 of 25 chi-square units with only 3 of 13 *df*s) of the model specifying invariant pattern weights could thus be attributed to the Inductive Reasoning scales. In turn, the only significant difference in factor loadings among the Inductive Reasoning indicators involved the Word Series scale. Note that this was also the scale that showed the most stress in the constrained equal model for the control group. In contrast to the control group analysis, however, the 1-*df* test of the difference was significant in the induction group, $\chi^2(1, N = 111) = 12.42, p < .001$. The rescaled factor loading for Word Series on induction at pretest was 1.031, but decreased to .809 at posttest. What appears to be happening is a slight rotation of the Induction factor toward the letter and number series markers, with the communality of the latter showing modest increment. Changes from pretest to posttest appear to be no more than subtle changes in the relative values of factor loadings on the target ability; the factorial integrity of the factor model remains undisturbed.

Table 3
Factor Intercorrelations for the Accepted Measurement Model for Control Group

Factor	Pretest					Posttest				
	1	2	3	4	5	1	2	3	4	5
Pretest										
1. Induction	—									
2. Space	.675	—								
3. Perceptual Speed	.777	.736	—							
4. Number	.687	.584	.689	—						
5. Verbal	.631	.298	.381	.552	—					
Posttest										
1. Induction	.978	.662	.763	.724	.674	—				
2. Space	.631	.961	.752	.579	.286	.628	—			
3. Perceptual Speed	.820	.706	.994	.714	.456	.801	.729	—		
4. Number	.669	.540	.657	.987	.556	.717	.556	.704	—	
5. Verbal	.606	.284	.387	.510	1.001	.648	.279	.450	.572	—

Table 4
Estimated Factor Variances for the Three Training Conditions

Factor	Control		Induction		Space	
	Pretest	Posttest	Pretest	Posttest	Pretest	Posttest
Induction						
Variance	33.41	38.89	29.25	36.22	31.83	37.95
SE	3.96	4.67	4.66	5.98	4.61	5.43
Space						
Variance	66.24	77.29	66.99	62.09	60.47	70.36
SE	9.94	11.59	11.87	11.48	10.88	11.93
Perceptual Speed						
Variance	15.91	17.00	6.38	6.77	6.92	9.11
SE	3.46	3.73	2.39	2.53	2.06	2.68
Number						
Variance	85.58	93.95	97.68	107.75	68.88	72.27
SE	10.67	11.73	14.93	16.50	11.48	12.16
Verbal						
Variance	27.55	27.82	12.58	14.01	30.38	25.42
SE	3.59	3.48	2.08	2.41	4.79	4.01

Note. These values are from the accepted measurement model for the given training group.

The standardized pretest–posttest factor correlations for the induction training group are given in Table 5. As with the control group, all of the correlations are remarkably close to a perfect 1.0. Our interest centered on the hypothesis that individual differences in effects of Inductive Reasoning training would result in lower stability for the Induction factor, relative to the control group. To the contrary, the pretest–posttest correlation for the Induction factor was .986, indicating that training did not reduce the stability of individual differences in the target ability. As can be seen from Table 4, however, training did result in some increase in Induction factor variance. This pattern (increased variance, near-perfect stability) suggests that training gains were, if anything, nearly proportional to initial pretest differences.

Space training group. The basic longitudinal factor model did not fit the space-training data quite as well as it fit the data for the other two groups, $\chi^2(399, N = 118) = 685.07$, GFI = .746. There were, however, no clear indications from fit diagnostics or residual correlations of qualitative shifts in the factor structure at posttest. The test of invariant factor pattern weights over time resulted in a salient reduction in fit, $\chi^2(412, N = 118) = 707.61$, GFI = .740; change in $\chi^2(13, N = 118) = 22.54$, $p < .05$. Again, we tested the hypothesis that the stress on longitudinal invariance could be localized to the trained ability—spatial orientation. The analysis confirmed the hypothesis and revealed that the Object Rotation test was carrying most of the stress in the model with respect to invariant factor pattern weights. A model constraining the loadings of all tests except Object Rotation showed a significant 1-*df* difference between the constrained equal model, $\chi^2(411, N = 118) = 700.83$, GFI = .742, indicating significant change in the Object Rotation loading.

The remaining differences were not significant, as indicated by the nonsignificant difference between this model and the one allowing freely estimated factor loadings at pretest and posttest (change in chi-square compared to the model = 15.76, $df = 12$, $p < .20$). The rescaled loading of Object Rotation decreased from 1.013 to .846 at posttest. It remained, however, the best marker of spatial orientation, with communalities of .883 and .832 at pretest and posttest, respectively. The shift in factor loadings did tend to rotate the factor away from Object Rotation toward the other spatial indicators, as reflected in the increasing communalities for PMA Space (.636 to .737) and Alphanumeric Rotation (.679 to .758). As observed for Inductive Reasoning, some subtle shifts in factor structure also occurred for Spatial Orientation, as evidenced by quantitative shifts in factor loadings. Nevertheless, it is clear that the integrity of the Spatial Orientation factor was not greatly affected by training on that ability.

The stability of individual differences was high for all five ability factors in the space training group as well (see Table 5). As was the case in the induction group, we found little evidence that training on spatial ability decreased the stability of individual differences on that factor. As shown in Table 5, the pretest–posttest correlation was about .96 in both the control and space training groups. The estimated factor variance for Spatial Orientation also increased from pretest to posttest. However, a decrease was found for the Verbal Comprehension factor in this group, although the other three factors exhibited increases in variance comparable to those of the other groups.

Table 5
Correlations of Latent Variables From Pretest to Posttest for the Three Training Conditions

Factor	Control	Induction	Space
Induction	0.978	0.986	0.988
Space	0.961	0.933	0.958
Perceptual Speed	0.994	1.007	0.966
Number	0.987	0.995	1.001
Verbal	1.001	0.951	0.909

Discussion

Previous research on cognitive training in the elderly has largely been concerned with demonstrating the fact that a training regimen can help older subjects to improve their performance on selected tests of cognitive functioning (Willis, 1985). This line of research has been criticized on the grounds that intervention programs may do no more than teach the test, and do not necessarily provide convincing evidence that the trainees' standing on the underlying psychological construct has been affected. A further criticism has been the contention that educational training techniques when applied to tests of fluid abilities may in fact transform such tests into measures of crystallized ability. The present study directly addressed these criticisms by analyzing a data set that operationally defined abilities at the latent construct level by using multiple markers for each construct included in the study, and by using training paradigms that are directed toward improving performance on the latent construct, rather than being directed toward improvement at the level of a specific test.

A valid analysis of quantitative changes at the construct level, however, presupposes a demonstration that configural invariance has been maintained from pretest to posttest, and that any disturbance to metric invariance be confined to the markers for those constructs that were targets of training, but does not effect markers of constructs not subject to training. Tests of the hypothesis of structural invariance across cognitive training intervention were conducted by means of restricted factor analysis using the LISREL paradigm.

In view of the fact that we were interested in testing hypotheses about unobserved latent constructs, it was necessary to begin our inquiry by identifying a suitable measurement model that would relate the observed variables to the latent constructs of interest. This was accomplished by identifying a five-factor measurement model on the basis of the pretest data for all of our subjects. To increase the external validity of our findings, we next tested the equivalence of the obtained factor structure across subsets of subjects that had declined or remained stable, and across subsets aggregated by the sex of the subjects. In each case, the hypotheses of both configural and metric invariance could be accepted, and we were thus able to show that our measurement system appears to be appropriate for elderly persons, regardless of their sex or history of change in intellectual functioning.

In direct response to Donaldson's (1981) hypothesis that training might transform the character of the measures on which training occurred, we were next concerned with demonstrating that the integrity of the constructs targeted for training (Spatial Orientation and Inductive Reasoning) was maintained from pretest to posttest. The most stringent test of structural invariance from pretest to posttest involved the specification of an isolated stability model (i.e., one that permits only autocorrelated but no cross-lagged regression coefficients). It is possible to argue that even practice alone might result in shifts in the integrity of latent constructs. This proposition was examined first by fitting the data for our control group to the isolated stability model, and resulted in demonstrating both configural and metric invariance. Stability coefficients for the ability factors were above .90, and ability measures at pretest predicted approximately 97% of the individual differences variance at posttest. The remaining variance at posttest was accounted for by a slight increase in the concurrent correlation of two ability factors (Perceptual Speed and Numerical Ability) at posttest, most likely occurring as a consequence of shared mean increments due to strong practice effects on the marker tests defining these abilities. These findings of high stability across test occasions in the face of practice agree with earlier results from a repeated testing study by Hofland, Willis, and Baltes (1981).

The hypothesis of factorial integrity across training was next tested separately for each of the groups that received the training intervention. In each case, configural invariance was readily demonstrated. However, the isolated stability models did not obtain the optimal fit under either training conditions, but here too stability coefficients were in excess of .90. The stability coefficients from pretest to posttest were only slightly lower for the inductive reasoning and spatial orientation training groups. The perturbations in the projections of the observed variables on the latent ability factors introduced by training, moreover, seemed to be specific to that primary ability on which subjects had been trained; they were of small magnitude and did not substantially affect factor patterns for any of the observable–latent relations for nontrained abilities.

There was only one marker of each target ability for which metric invariance could not be demonstrated. For the induction training group, an improved fit could be obtained when the across-occasion constraint on the Word Series factor loadings was relaxed. For the space training group, similarly, an improved fit occurred when the across occasion constraint was relaxed for the Object Rotation factor loadings. These model modifications did not in any way impair those aspects of metric invariance that relate to the factorial integrity of the construct for which training occurred. That is, it was not necessary to allow any of the markers of the target abilities to load on any nontrained ability factor at posttest. Metric invariance was therefore demonstrated for all of the relations between observed variables marking the construct on which training occurred with all other constructs tested in our battery. This finding conclusively refutes Donaldson's (1981) hypothesis on the possible transformation of the construct relevance of measures after training.

In both instances in which significant change in optimal factor loading occurred at posttest in the training group (but not in the control group), the tests involved were the most concrete markers of the target ability. The magnitude of the factor loading of these markers decreased at posttest, whereas, at the same time, the magnitude of the least concrete markers increased. These shifts can be related to at least two distinct findings in prior research on information processing in older adults (Poon, 1985). First, it is well-known that older persons seem to be deficient in the spontaneous use of information-processing strategies (Labouvie-Vief & Gonda, 1976; Poon, Walsh-Sweeney, & Fozard, 1980). Second, older individuals are known to profit when familiar cues are provided that can be used in the processing of information (Smith, 1980). It follows that the test stimuli for the concrete markers could be processed more readily at pretest because of the availability of familiar cues (e.g., common household objects in the Object Rotation test, or over-

learned serial relations of weeks and months in the Word Series test).

The familiarity of the stimuli of the most concrete markers may have resulted in reduced individual differences on the mechanics of the test (processing of stimuli), and thereby maximized the individual differences contribution to the cognitive processes embodied in the latent construct (ie., mental rotation, abstraction of rules). The training paradigms involved strategies to facilitate processing of abstract stimuli, as well as strategies maximizing the cognitive processes. Hence, at posttest, variance components associated with the processing of the more abstract markers were reduced and their contribution to the construct-specific individual differences were enhanced.

Because of our finding of shifts in the latent–observable relationship for one of the markers in each of the training target abilities, we would caution investigators against using single markers in a training study, unless the factorial stability of such markers had been previously verified. Using a set of multiple indicators for a latent variable—such as were included in our study—on the other hand, makes it possible to identify training gain at the latent variable level, even if some of the indicators show shifts in factor loadings with training. In fact, such a design permitted us to show that (a) we have indeed trained on the latent variable, (b) we can unambiguously interpret individual differences and mean changes in the latent variable as a function of training, (c) we can identify the indicators that are reactive to training in terms of shifting measurement properties, and (d) the regression of observed marker variables on their latent ability factors is virtually undisturbed by test–retest effects over brief test intervals (2–4 weeks, in our case) when no ability-specific intervention occurs between test occasions.

It is noteworthy that both retest and training result in increased variability for the latent variables. In effect this means that practice and other interventions have increased rather than reduced individual differences in cognitive performance. As the analysis of changes in *level* of performance has shown (cf. Schaie & Willis, 1986), most subjects who declined or remained stable over the prior 14-year period gained at least somewhat from training, but there were wide individual differences in the magnitude of change. Nevertheless, changes in the subjects' relative position within their reference population were confined to a limited region within the distribution of individual differences, which as a whole tended to fan out somewhat at posttest. Such results would, of course, be expected if there was basic stability in the distribution of individual differences regardless of intervention. It is important to note, however, that the remarkable stability shown in our study may simply reflect that we were operating in one of the best-defined sectors of the ability domain, with measures having optimal psychometric characteristics. Other investigators should therefore be most cautious in *not* interpreting our findings as providing sufficient reassurance that they could safely ignore the need to apply procedures such as those described here in order to justify their own invariance assumptions.

References

Alwin, D. F., & Jackson, D. J. (1981). Applications of simultaneous factor analysis to issues of factorial invariance. In D. J. Jackson & E. F. Borgatta (Eds.), *Factor analysis and measurement* (pp. 249–278). London: Sage.

Baltes, P. B., Cornelius, S. W., Spiro, A., Nesselroade, J. R., & Willis, S. L. (1980). Integration versus differentiation of fluid/crystallized intelligence in old age. *Developmental Psychology, 16,* 625–635.

Baltes, P. B., & Nesselroade, J. R. (1973). The developmental analysis of individual differences on multiple measures. In J. R. Nesselroade & H. W. Reese (Eds.), *Life-span developmental psychology: Methodological issues* (pp. 219–252). New York: Academic Press.

Baltes, P. B., & Willis, S. L. (1982). Enhancement (plasticity) of intellectual functioning in old age: Penn State's Adult Development and Enrichment Project (ADEPT). In F. I. M. Craik & S. E. Trehub (Eds.), *Aging and cognitive processes* (pp. 353–389). New York: Plenum Press.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606.

Birren, J. E., Cunningham, W. R., Yamamoto, K. (1983). Psychology of adult development and aging. *Annual Review of Psychology, 34,* 543–575.

Blieszner, R., Willis, S. L., & Baltes, P. B. (1981). Training research in aging on the fluid ability of inductive reasoning. *Journal of Applied Developmental Psychology, 2,* 247–265.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand McNally.

Donaldson, G. (1981). Letter to the editor. *Journal of Gerontology, 36,* 634–636.

Dudek, F. J. (1979). The continuing misinterpretation of the standard error of measurement. *Psychological Bulletin, 86,* 335–337.

Ekstrom, R. B., French, J. W., Harman, H., & Derman, D. (1976). *Kit of factor-referenced cognitive tests* (Rev. ed.). Princeton, NJ: Educational Testing Service.

Hertzog, C. (in press). On the utility of structural regression models for developmental research. In P. B. Baltes, D. Featherman, & R. M. Lerner (Eds.), *Life-span development and behavior.* Hillsdale, NJ: Erlbaum.

Hertzog, C., & Cannon, C. (1985). *SAS Proc Matrix Scaling program.* Unpublished manuscript, Pennsylvania State University, University Park, PA.

Hertzog, C., & Schaie, K. W. (1986). Stability and change in adult intelligence: 1. Analysis of longitudinal covariance structures. *Psychology and Aging, 1,* 159–171.

Hofland, B. F., Willis, S. L., & Baltes, P. B. (1981). Fluid intelligence performance in the elderly: Intraindividual variability and conditions of assessment. *Journal of Educational Psychology, 73,* 573–586.

Horn, J. L., McArdle, J. J., & Mason, R. (1983). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. *Southern Psychologist, 1,* 179–188.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36,* 409–426.

Jöreskog, K. G. (1979). Statistical estimation of structural models in longitudinal developmental investigations. In J. R. Nesselroade & P. B. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 303–351). New York: Academic Press.

Jöreskog, K. G., & Sörbom, D. (1977). Statistical models and methods for analysis of longitudinal data. In D. J. Aigner & A. S. Goldberger (Eds.), *Latent variables in socioeconomic models* (pp. 285–325). Amsterdam: North Holland.

Jöreskog, K. G., & Sörbom, D. (1984). *LISREL VI user's guide.* Chicago: National Educational Resources.

Labouvie-Vief, G., & Gonda, J. N. (1976). Cognitive strategy training and intellectual performance in the elderly. *Journal of Gerontology, 31,* 327–332.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika, 29,* 177–185.

Poon, L. W. (1985). Differences in human memory with aging: Nature, causes, and clinical implications. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (2nd ed., pp. 427–462). New York: Van Nostrand Reinhold.

Poon, L. W., Walsh-Sweeney, L., & Fozard, J. L. (1980). Memory skill training for the elderly: Salient issues on the use of imagery mnemonics. In L. W. Poon, J. L. Fozard, L. S. Cermak, D. Arenberg & L. W. Thompson (Eds.), *New directions in memory and aging* (pp. 461–484). Hillsdale, NJ: Erlbaum.

Schaie, K. W. (1983). The Seattle Longitudinal Study: A 21-year exploration of psychometric intelligence in adulthood. In K. W. Schaie (Ed.), *Longitudinal studies of adult psychological development* (pp. 64–135). New York: Guilford.

Schaie, K. W. (1985). *Manual for the Schaie-Thurstone Adult Mental Abilities Test (STAMAT).* Palo Alto, CA: Consulting Psychologists Press.

Schaie, K. W., & Hertzog, C. (1985). Measurement in the psychology of adulthood and aging. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (2nd ed., pp. 61–94). New York: Van Nostrand Reinhold.

Schaie, K. W., & Willis, S. L. (1986). Can decline in adult intellectual functioning be reversed? *Developmental Psychology, 22,* 223–232.

Smith, A. D. (1980). Age differences in encoding, storage and retrieval. In L. W. Poon, J. L. Fozard, L. S. Cermak, D. Arenberg, & L. W. Thompson (Eds.), *New directions in memory and aging* (pp. 23–46). Hillsdale, NJ: Erlbaum.

Sörbom, D. (1975). Detection of correlated errors in longitudinal data. *British Journal of Mathematical and Statistical Psychology, 28,* 138–151.

Sterns, H. L., & Sanders, R. E. (1980). Training and education in the elderly. In R. E. Turner & H. W. Reese (Eds.), *Life-span developmental psychology: Intervention* (pp. 307–330). New York: Academic Press.

Thurstone, L. L. (1948). *Primary mental abilities.* Chicago: University of Chicago Press.

Willis, S. L. (1985). Towards an educational psychology of the adult learner: Cognitive and intellectual bases. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (2nd ed., pp. 818–847). New York: Van Nostrand Reinhold.

Willis, S. L., & Baltes, P. B. (1981). Letter to the editor. *Journal of Gerontology, 36,* 636–638.

Willis, S. L., & Schaie, K. W. (1981). Maintenance and decline of adult mental abilities: 2. Susceptibility to experimental manipulation. In F. Grote & R. Feringer (Eds.), *Adult learning and development* (pp. 40–57). Bellingham, WA: Western Washington University.

Willis, S. L., & Schaie, K. W. (1983). *Alphanumeric Rotation test.* Unpublished manuscript, Pennsylvania State University, University Park, PA.

Willis, S. L., & Schaie, K. W. (1986). Training the elderly on the ability factors of Spatial Orientation and Inductive Reasoning. *Psychology and Aging, 1,* 239–247.

Received November 13, 1985
Revision received November 11, 1986
Accepted November 12, 1986 ■