Developmental Psychology
1989, Vol. 25, No. 4, 652-662

0012-1649/89/$00.75

Structural Invariance of Cognitive Abilities Across the Adult Life Span: A Cross-Sectional Study

K. Warner Schaie, Sherry L. Willis, Gina Jay, and Heather Chipuer
Pennsylvania State University

This study examined the hypothesis that psychometric tests retain equivalent factor structures across samples widely differing in age. We estimated a best-fitting measurement model for 17 psychometric tests covering the 5 primary abilities of Inductive Reasoning, Spatial Orientation, Verbal Ability, Numerical Ability, and Perceptual Speed, using a sample of 1,621 participants (ages 22 to 95) from the 5th wave of the Seattle Longitudinal Study. We disaggregated the participants into 9 subsets (M ages = 29, 39, 46, 53, 60, 67, 74, 81, and 90) and tested the fit of the accepted model for each subset. We confirmed configural invariance for all subsets, but could not establish either complete or incomplete metric invariance for any set. These results confirm the stability of factor patterns across age but indicate serious limitations for valid cross-age comparisons of individual markers of psychometric abilities in age-comparative studies.

Much of the literature on cognitive aging is concerned with comparing levels of performance of groups of individuals varying (often widely) in chronological age. The theoretical issues and evidence for the assumption of the internal validity of empirical studies, whether cross-sectional or longitudinal, that offer such comparisons have been discussed in considerable detail (Nesselroade & Labouvie, 1985; Schaie, 1973, 1977, 1988a). Another major assumption that has received only limited attention so far, however, has been the question of whether the factorial structure of assessment instruments remains equivalent both within subjects across time and between groups of subjects of different ages assessed at the same point in time (cf. Schaie & Hertzog, 1985). If satisfactory evidence of factorial invariance were lacking, it would be possible that the validity of quantitative comparisons might be impaired because of the occurrence of qualitative age changes or age differences among groups.

A critical assumption that underlies evaluation of quantitative change across age or differences between different age groups is that the relationship between the ability constructs and measures of these constructs (psychometric tests) in the assessment battery remains invariant across comparisons. That is, quantitative comparisons are meaningful *only* if there is qualitative invariance (cf. Baltes & Nesselroade, 1973).

The question arises, then, as to how such qualitative differences in factor structure would be manifested. The comparative factor analysis literature suggests that the required evidence for factorial invariance would be the demonstration of the equality of unstandardized factor pattern weights (factor loadings; see Hertzog & Schaie, 1986; Meredith, 1964; Schaie & Hertzog, 1985). Horn, McArdle, and Mason (1983) have recently focused attention on the distinction between two levels of invariance in factor loadings (with different implications for age change and age differences research) first introduced by Thurstone (1947, pp. 360–369): *configural invariance* and *metric invariance.*

Configural invariance requires that measures marking factors have their primary loading on the same ability constructs across occasions. If configural invariance is not maintained across time or between different cohort groupings, then it is likely that developmental processes or cohort effects may have produced qualitative changes in ability structure. If this were the case, interpretation of quantitative age changes or age differences 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 equally across time or between groups. It seems reasonable to hypothesize, even if configural invariance can be confirmed, that developmental processes or differential cohort experiences could cause changes or differences in the magnitude of the factor loadings for the ability measures. That is, it may not be possible to obtain complete metric invariance due to shifts or differences in the magnitude of the factor loadings for Tests A and B, even though the tests mark the same ability factor across time or for different cohorts. Finding a lack of metric invariance would raise problems for the interpretation of quantitative changes or differences in individual tests. Such problems could readily be surmounted, however, where quantitative change can be assessed at the level of *factor scores* rather than observed scores (cf. Hertzog & Schaie, 1988).

The research reported in this article was supported by Grant R01 AG4770 from the National Institute on Aging. We gratefully acknowledge the enthusiastic cooperation of the members and staff of the Group Health Cooperative of Puget Sound. This article was completed while K. Warner Schaie and Sherry L. Willis were visiting scientists at the Institute of Gerontology of the University of Michigan.

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

The issue of factorial invariance in longitudinal studies of intelligence has thus far been dealt with only for the relation of the first five primary mental abilities to a second-order *g* factor (Hertzog & Schaie, 1986). This study found highly stable individual differences in the projection of the primaries on the second-order factor over 14-year intervals in three samples that had mean ages of 37, 49, and 65, respectively, at the inception of these studies.[1]

Considerably more data exist on factorial invariance across adulthood from cross-sectional studies. The first major analyses of this kind were conducted by Cohen for the standardization samples of the Wechsler Adult Intelligence Scale (1957). Cohen concluded that what we would now call configural invariance was maintained from young adulthood to old age, but that factor loadings shifted substantially. Cunningham (1980, 1981) studied measures of speed and verbal ability from the ETS Kit of factor-referenced measures (Ekstrom, French, Harman, & Derman, 1976). He concluded for these studies that the factor space was maintained, with highly similar factor loadings, but that factor covariances tended to increase with age. Similar findings were obtained by Stricker and Rock (1987) in a study of the GRE that compared factor structures for the first three adult decades. However, in a more recent study of a battery involving speeded cognitive factors comparing two samples (one young adult and one young–old), White and Cunningham (1987) had to reject all simultaneous models and concluded that an additional factor was required to fit the data for their older group.

The present study addresses the issue of structural invariance across different age groups in a cross-sectional data set. It is more comprehensive then previous efforts, however, because it surveys a broader range of measures within the primary mental ability space and systematically covers nine age intervals across the adult life span from a mean age of 29 to a mean age of 90 years.

Data reported in this study are on participants in the Seattle Longitudinal Study (SLS), who were assessed on multiple measures of five primary mental abilities: Inductive Reasoning, Spatial Orientation, Perceptual Speed, Numerical Ability, and Verbal Ability. In this article, we report the application of restricted (confirmatory) factor analysis to assess the hypothesis of factorial invariance across different age/cohort groups assessed at the same point in time. These analyses have been conducted, using the LISREL approach outlined by Jöreskog (1979; also see Schaie & Hertzog, 1985). As discussed by Schaie and Hertzog (1985; see also Hertzog & Schaie, 1986), the critical test of differences in the measurement properties of separate data sets involves the test of invariance across age 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 test hypotheses at three levels of stringency: (a) *complete metric invariance,* which implies that there would be no difference between the best-fitting model for the total sample and each subset in the factor pattern loadings (regression coefficients relating tests to ability factors) and factor intercorrelations; (b) *incomplete metric invariance,* which implies no differences between the best-fitting model for the total sample and each subset in factor loading patterns, but allows for differences in the factor intercorrelations; and (c) *configural invariance,* which requires maintenance of factor patterns but allows for differences in factor loadings and factor intercorrelations.

Method

Subjects

Our sample consisted of 1,621 participants (738 men and 883 women), from the Seattle metropolitan area, who participated in the fifth wave (1984) of the Seattle Longitudinal Study (SLS; Schaie, 1983, 1988b; Schaie & Hertzog, 1986). The SLS includes subjects initially tested at five measurement points (1956, 1963, 1970, 1977, and 1984). Only data from the fifth wave are used in the present study because this was the first occasion for which multiple markers are available for the primary abilities of interest. Initially, all subjects represented random draws from the base population. That is, subjects are, or have previously been, members of the Group Health Cooperative of Puget Sound, a health maintenance organization. Possible effects of practice and attrition for those subjects who were repeatedly tested have been reported elsewhere (Schaie, 1988a); the effects of selection biases are relatively limited and would not seem to be important for purposes of structural analyses.

Identical recruitment procedures have been used in all samples: 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 59 years (range = 22–95; SD = 16.03). Mean educational level was 14.3 years (range = 1–20; SD = 3.06). There were no sex differences in age or educational level. Mean family income level was $23,200 (range = $1,000–$50,000 and over; SD = $9,606). All subjects were community dwelling, and most were Caucasian. Occupational levels were rated on a scale from 0 for *unskilled* to 9 for *professional* occupations. Those individuals who were gainfully employed at the time of assessment averaged an occupational level of 6.8 (SD = 1.87). The most frequent occupations represented involve skilled trades and clerical, sales, managerial, and semiprofessional jobs. All subjects were in good health at the time of testing; their records were prescreened by their attending physicians. Those potential participants who were acutely ill or had disabilities that would prevent them from participating in pencil-and-paper format assessment were eliminated. For the purposes of this study, the sample was disaggregated into nine subsets by date of birth (see Table 1).

Measures

The test battery included psychometric measures representing five primary mental abilities. The battery included the Thurstone Primary Mental Ability measures (Thurstone, 1948) administered at previous SLS assessments. Additional measures were selected from other sources

[1] The issue of construct equivalence of factor structure has also attained importance in the context of cognitive training research (cf. Willis, 1987). A short-term repeated measurement study of five ability factors by Schaie, Willis, Hertzog, and Schulenberg (1987) showed stability of factor structure in a sample of subjects ranging in age from 62 to 94 years over an interval of several weeks in control and training groups. Although there were sex differences in level, this study also showed that there were no significant sex differences in factor structure. That study provides the initial measurement model for the analyses presented in this article.

Table 1
Subsamples Entering the Structural Invariance Analyses

Group	SLS cohort	Year of birth	*N*	*M* age (in years)
1	1–2	1886–1899	39	90
2	3	1900–1906	136	81
3	4	1907–1913	260	74
4	5	1914–1920	291	67
5	6	1921–1927	260	60
6	7	1928–1934	193	53
7	8	1935–1941	154	46
8	9	1942–1948	124	39
9	10–11	1949–1962	164	29
Total sample			1621	59

Note. Following the convention used in all reports from the Seattle Longitudinal Study (SLS), lower cohort numbers represent earlier-born (older) subjects in all tabulations.

(principally the Educational Testing Service Reference Kit; Ekstrom et al., 1976) or the 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 2). All tests are administered under time limits and are slightly speeded.

Table 2 also reports the test–retest correlations of these indicators in a group of 172 participants who received these measures twice over a 2-week interval (Schaie, Willis, Hertzog, & Schulenberg, 1987). 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 actually underestimate the markers' reliability. The correlations are all greater than .8, indicating satisfactory reliability for all instruments.

Spatial orientation. Three 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 Rotation test (Willis & Schaie, 1983) were constructed so 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. The Cube Comparison test (Ekstrom et al., 1976) requires the matching of three-dimensional cubes upon mental rotation.

Inductive reasoning. The PMA Reasoning measure (Thurstone, 1948) assesses inductive reasoning ability by means of 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 different 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 As involves the cancellation of the letter *a* in columns of words of which about half 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

Table 2
Intellectual Abilities Measurement Battery

Primary ability	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
	Cube Comparison	Ekstrom et al., 1976	.951
Perceptual Speed	Finding As	Ekstrom et al., 1976	.814
	Number Comparison	Ekstrom et al., 1976	.860
	Identical Pictures	Ekstrom et al., 1976	.865
Numerical Ability	PMA Number	Thurstone, 1948	.875
	Addition	Ekstrom et al., 1976	.937
	Subtraction and Multiplication	Ekstrom et al., 1976	.943
Verbal Ability	PMA Verbal	Thurstone, 1948	.890
	Vocabulary II	Ekstrom et al., 1976	.828
	Vocabulary IV	Ekstrom et al., 1976	.954

problems (Thurstone, 1948). The Addition test (Ekstrom et al., 1976) involves calculating the sum 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 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).

Assessment Procedure

The measures described above were administered to small groups of subjects as part of a broader test battery that required approximately 5 hr, spread over two sessions. The tests were administered in a standard format and order by an examiner assisted by a proctor. Testing locations were at familiar sites, such as clinic conference rooms or church meeting rooms, close to the homes of our participants. A subject fee of $50 was provided on completion of both test sessions.

Statistical Procedure

The evaluation of equivalence in the factor structure of the psychometric battery in the different age/cohort groups was conducted by using LISREL VI (Jöreskog & Sörbom, 1984) to perform confirmatory factor analysis (see Alwin, 1988; Jöreskog, 1971; Jöreskog & Sörbom, 1977; and Schaie & Hertzog, 1985, for further discussions of the technique). The analyses reported in this paper 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 yields 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). The covariance matrix of the observed variables in the population, Σ, may then 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 represents a restricted factor analysis model that can then 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 on Equation 2 to identify the remaining unknowns. Restrictions are specified by fixing parameters to a known value a priori (e.g., requiring that a variable is unrelated to a factor by fixing its regression to 0).

Overidentified models (which have more restrictions than are necessary to identify the model parameters) place restrictions on the hypothesized form of Σ, which can 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) can be used to test the null hypothesis that the restrictions are true in the population. A more restrictive model (i.e., with more restrictions placed on the model parameters) that is nested within a less restrictive model will 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 will be accepted.

An additional index of model fit used in this study 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 may be less influenced by sample size than the chi-square fit statistic (e.g., Bentler & Bonett, 1980, but see Anderson & Gerbing, 1984; Bollen, 1986). Factor models with fits in the .8 to .9 range (such as those reported below) are generally considered to be useful approximations of 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. For our purposes, we will consider a GFI of .8 as an acceptable fit and a GFI in excess of .9 as an excellent fit. In initial model development, diagnostic fit indices such as LISREL modification indices and residual correlations were also carefully evaluated before proceeding to the analysis of group differences in factor structure.

Results

We conducted three distinct sets of confirmatory factor analyses to assess complete metric, incomplete metric, and configural invariance for each of the nine cohort groups. The initial structural model for these analyses was based on the factor structure derived in a prior analysis (Schaie et al., 1987) for a sample of 401 participants ranging in age from 62 to 94 years (M age = 72.1 years). The previous analysis confirmed a structure consisting of the following five factors: Inductive Reasoning, Spatial Orientation, Perceptual Speed, Numerical Ability, and Verbal Ability.

Estimation of the Measurement Model

Although it seemed reasonable to allow the results of the prior study of older people to provide the initial hypothesis for the factor pattern model, we felt that it was more desirable to conduct the tests of factorial invariance on the basis of a measurement model that used data collected more evenly across the entire adult life span. Therefore, the measurement model was estimated for the total data set (N = 1,621). All raw scores were rescaled into T-score form, on the basis of the entire SLS population at first test. Analyses were conducted on the variance–covariance matrix, with results rescaled into a correlation metric. The results of this analysis yielded a quite satisfactory fit, $\chi^2(107, N = 1{,}621) = 946.62$, $p < .001$ (GFI = .936). Table 3 shows the factor loadings and factor intercorrelations that enter subsequent analyses. In this model, each ability is marked by at least three operationally distinct tests. Each test marks only one ability, except for Number Comparison, which splits between Perceptual Speed and Number, and PMA Verbal Meaning, which splits between Perceptual Speed and Verbal Ability.

To examine the adequacy of the measurement model further, we tested a less restricted model. In this model, a marker variable was fixed to 1.0 for its salient factor loading and zero for each of the five factors, but all other values (lambda) were freely estimated. As would be expected, this relaxed model produced a significantly better fit, $\chi^2(6, N = 1621) = 229.63$, $p < .001$ (GFI = .983), by exhausting a larger proportion of common variance through allowing secondary loadings for all marker variables. However, the most salient loadings remained consistent with the hypothesis matrix for the more restricted model.

Table 3
Measurement Model for Total Data Set

Variable	Factor I	Sp	Ps	N	V	Unique variance
PMA Reasoning	.895*					.199
ADEPT Letter Series	.884*					.219
Word Series	.891*					.207
Number Series	.787*					.381
PMA Space		.831*				.309
Object Rotation		.877*				.231
Alphanumeric		.831*				.309
Cube Comparison		.594*				.647
Finding As			.524*			.725
Number Comparison			.576*	.270*		.424
Identical Pictures			.832*			.308
PMA Number				.838*		.297
Addition				.938*		.121
Subtraction and Multiplication				.865*		.252
PMA Verbal Meaning			.660*		.386*	.254
Vocabulary II					.897*	.195
Vocabulary IV					.893*	.203
Intercorrelations						
I	—					
Sp	.768	—				
Ps	.856	.765	—			
N	.562	.416	.555	—		
V	.484	.260	.317	.353	—	

Note. I = Inductive Reasoning; Sp = Spatial Orientation; Ps = Perceptual Speed; N = Numeric; V = Verbal. $\chi^2(107, N = 1621) = 946.62$. Goodness-of-Fit Index = .936.
* $p < .001$.

In the interest of examining age differences for a factor structure that is as parsimonious and conceptually simple as possible, we decided to proceed with the original, more restricted measurement model.

Complete Metric Invariance

In the first set of analyses, we tested the covariance matrices for each of the nine subsets to determine their fit to the measurement model determined for the entire sample (Table 3). For these analyses, the factor loading pattern was specified to be identical with that of the measurement model for the entire sample. The values of the factor loading matrix (lambda) and the factor variance–covariance matrix (phi) were fixed to those estimated in establishing the measurement model for the total data set. The values of the unique variance matrix (theta delta) remained free and were estimated. (The factor loadings and factor intercorrelations for these nine analyses are, of course, identical with those in Table 3 and have not been presented separately.) The chi-square statistics and the GFIs are presented in Table 4. Of course, these model fits are somewhat lower than for the total set but, except for Cohort 1–2 and Cohort 3 (the oldest cohorts), are still quite acceptable. Cohort 1–2 had the lowest GFI (.596), and Cohort 5 had the highest (.893). The small sample size of Cohort 1–2 (n = 39) probably accounts for its poorer fit.

Incomplete Metric Invariance

In the second set of analyses, we assessed the less stringent hypothesis of incomplete metric invariance for the nine cohort groups. For these analyses, the factor loading pattern remained the same and was set to that of the measurement model. The values of the factor loading matrix (lambda) remained fixed at those estimated for the total group. However, the values of the factor variance–covariance matrix (phi) and the unique variance matrix (theta delta) were allowed to be free and were estimated. (The factor loadings are identical with those in Table 3.) The factor variances and standardized covariances for the nine cohorts are presented in Table 5, and the chi-square statistics and GFIs are given in Table 4. The GFIs for this moderately constrained model yielded an improvement over those for the more constrained model (complete metric invariance). Similar to the complete metric invariance model, the poorest fit was again found for Cohort 1–2 (.669), and the best fit was found for Cohort 4 (.913). Substantial differences in factor variances are found across cohorts. Variances increase systematically until the sixties and, then, decrease again. Covariances also show substantial increment across cohorts with increasing age.

Configural Invariance

For the third set of analyses, the factor loading pattern was set to that of the previously determined model. However, the

Table 4
Chi-Square Values and Goodness-of-Fit (GFI) Indices for the Three Models

	Configural		Incomplete metric		Complete metric		
Cohort	x^2	GFI	x^2	GFI	x^2	GFI	*n*
1–2	161.59	.697	186.72	.664	247.15	.596	39
3	176.78	.869	269.07	.811	385.98	.789	136
4	169.36	.930	240.09	.904	347.71	.882	260
5	209.69	.925	242.39	.913	321.98	.893	291
6	215.12	.912	268.07	.889	361.60	.868	260
7	183.64	.907	225.85	.884	328.09	.856	193
8	133.17	.911	170.46	.888	260.03	.853	154
9	184.33	.848	216.64	.824	279.96	.801	124
10–11	271.86	.836	308.16	.823	448.89	.780	164
df	107		121		136		

values of the factor loading matrix (lambda), the factor variance–covariance matrix (phi), and the unique variance matrix (theta delta) were all allowed to be free and were estimated. We conducted analyses on the covariance matrices for each subset, with a metric being established by setting the largest loading on each factor to 1.0. Results were then standardized into a correlation metric to be comparable with the other analyses. Summaries of the factor loadings, factor variances, and standardized covariances across cohorts are presented in Tables 6 and 7, respectively. All chi-square statistics for these analyses were significant, and the goodness-of-fit indices ranged from a low of .697, for the oldest cohort, to a high of .930, for Cohort 4 (see

Table 5
Incomplete Metric Invariance Model Factor Variances and Standardized Covariances by Cohort Group

	Cohort								
Factor	1–2	3	4	5	6	7	8	9	10–11
				Factor variances					
Inductive Reasoning	10.44	16.04	22.17	27.23	25.26	21.28	17.59	16.02	18.06
Spatial Orientation	53.56	76.14	109.06	116.93	105.51	90.52	83.09	99.21	47.86
Perceptual Speed	24.06	32.56	26.86	27.68	20.42	15.23	16.52	12.93	13.38
Numerical Ability	140.02	235.32	194.56	264.62	249.69	246.27	263.96	188.24	164.00
Verbal Ability	58.28	53.10	37.80	33.14	31.43	29.04	22.93	24.91	18.61
				Standardized covariances					
				Inductive Reasoning with					
Spatial Orientation	.768	.696	.543	.631	.590	.542	.600	.503	.599
Perceptual Speed	.744	.726	.671	.767	.758	.644	.824	.697	.702
Numerical Ability	.614	.667	.597	.625	.501	.482	.587	.450	.516
Verbal Ability	.680	.626	.705	.638	.626	.512	.313	.415	.597
				Spatial Orientation with					
Perceptual Speed	.486	.725	.614	.590	.605	.416	.466	.443	.398
Numerical Ability	.471	.637	.360	.348	.364	.214	.323	.140	.265
Verbal Ability	.502	.434	.306	.330	.226	.229	.085	.027	.423
				Perceptual Speed with					
Numerical Ability	.802	.792	.610	.647	.524	.461	.720	.535	.405
Verbal Ability	.366	.449	.437	.390	.402	.418	.357	.246	.559
				Numerical Ability with					
Verbal Ability	.636	.537	.450	.410	.262	.172	.146	.216	.285

Table 6
Factor Loadings for the Configural Invariance Model

Cohort	Variable: PMA Reasoning	ADEPT Letter Series	Word Series	Number Series
	Inductive Reasoning			
1–2	.808*	.618*	.865*	.607*
3	.805*	.770*	.908*	.646*
4	.872*	.834*	.854*	.746*
5	.889*	.844*	.840*	.752*
6	.896*	.861*	.811*	.759*
7	.880*	.826*	.819*	.698*
8	.802*	.747*	.770*	.641*
9	.768*	.807*	.732*	.610*
10–11	.726*	.895*	.846*	.723*

Cohort	Variable: PMA Space	Object Rotation	Alphanumeric Rotation	Cube Comparison
	Spatial Orientation			
1–2	.879*	.772*	.629*	.274
3	.732*	.858*	.632*	−.044
4	.789*	.893*	.802*	.435*
5	.850*	.835*	.770*	.525*
6	.868*	.816*	.786*	.511*
7	.823*	.788*	.708*	.474*
8	.740*	.879*	.779*	.498*
9	.780*	.852*	.700*	.519*
10–11	.674*	.682*	.632*	.652*

Cohort	Variable: Finding As	Number Comparison	Identical Pictures	PMA Verbal Meaning
	Perceptual Speed			
1–2	.559*	.549	.700*	.807*
3	.650*	.890*	.780*	.681*
4	.549*	.530*	.700*	.687*
5	.572*	.547*	.702*	.527*
6	.526*	.447*	.576*	.475*
7	.457*	.477*	.521*	.398*
8	.600*	.373*	.641*	.325*
9	.503*	.474*	.620*	.212*
10–11	.443*	.453*	.355*	.368*

Cohort	Variable: Number Comparison	PMA Number	Addition	Subtraction/ Multiplication
	Numerical Ability			
1–2	.225	.690*	.808*	.832*
3	−.056	.840*	.936*	.877*
4	.299*	.833*	.894*	.884*
5	.250*	.870*	.947*	.897*
6	.254*	.840*	.951*	.836*
7	.182	.840*	.950*	.832*
8	.277*	.820*	.968*	.820*
9	.302*	.827*	.962*	.784*
10–11	.128	.742*	.957*	.816*

Cohort	Variable: PMA Verbal Meaning	Vocabulary II	Vocabulary IV
	Verbal Ability		
1–2	.123	.904*	.983*
3	.222*	.936*	.942*
4	.300*	.831*	.919*
5	.457*	.912*	.870*
6	.528*	.890*	.882*
7	.640*	.878*	.926*
8	.491*	.908*	.873*
9	.592*	.896*	.873*
10–11	.458*	.842*	.890*

* $p < .05$.

Table 4). Substantial differences were again found across cohorts for factor variances and covariances, with patterns similar to those shown in the incomplete metric invariance analysis.

Change in Chi-Square

To determine which model (configural, complete metric, or incomplete metric invariance) fit the data best, we assessed changes in chi-square statistics. The results of these analyses are presented in Table 8. As indicated in this table, there were no cohort groups for which the most constrained model (complete metric invariance) or the less constrained model (incomplete metric invariance) provided the best fit. The configural invariance model provided a significantly better fit than the more constrained models in all instances and must, therefore, be accepted as the most plausible description of the structure for our data set.

Differences in Factor Loadings

What is the nature of the cohort differences in factor loadings for the accepted configural invariance model? These differences appear to be ability-specific and located with respect to individual marker variables, rather than systematically related to overall differences across cohorts. We observed what seem to be only random sampling variations across cohorts for the Inductive Reasoning factor. For Spatial Orientation, however, Cube Com-

Table 7
Factor Variances and Standardized Covariances by Cohort Group for Configural Invariance Model

Factor	Cohort 1–2	3	4	5	6	7	8	9	10–11
				Factor variances					
Inductive Reasoning	10.69	14.15	21.15	26.47	26.19	15.38	18.55	14.08	15.88
Spatial Orientation	62.28	131.95	129.21	113.19	93.31	80.66	95.32	102.78	43.62
Perceptual Speed	15.92	27.22	17.56	20.14	13.04	10.08	15.57	15.53	5.20
Numerical Ability	154.58	221.19	184.87	276.68	223.19	208.71	235.62	155.23	163.84
Verbal Ability	60.27	51.75	40.94	31.86	32.46	31.59	26.13	27.48	19.55
				Standardized covariances					
				Inductive Reasoning with					
Spatial Orientation	.760	.673	.527	.628	.589	.539	.598	.511	.644
Perceptual Speed	.852	.773	.727	.784	.762	.632	.772	.636	.744
Numerical Ability	.664	.676	.594	.625	.494	.430	.579	.412	.515
Verbal Ability	.658	.646	.683	.636	.628	.514	.324	.446	.601
				Spatial Orientation with					
Perceptual Speed	.530	.699	.605	.575	.577	.400	.497	.467	.341
Numerical Ability	.465	.607	.343	.345	.356	.213	.321	.139	.257
Verbal Ability	.482	.416	.279	.331	.234	.245	.072	.025	.399
				Perceptual Speed with					
Numerical Ability	.873	.850	.650	.681	.548	.530	.683	.517	.545
Verbal Ability	.549	.543	.496	.384	.422	.298	.297	.236	.542
				Numerical Ability with					
Verbal Ability	.627	.517	.426	.404	.264	.166	.136	.213	.273

parison ceases to be a significant marker for the two oldest cohort groups. For Perceptual Speed, Finding As and Number Comparison show primarily random sampling variation across cohorts, but the loadings for Identical Pictures and Verbal Meaning appear to increase systematically for the older cohorts. On Numerical Ability, random sampling variability occurs for the primary markers, whereas the secondary marker of Number Comparison does not attain significance for four of the nine cohort groups. Finally, for Verbal Ability, the PMA Verbal Meaning test drops out as a significant loading for the oldest cohort and has substantially lower loadings for the two next oldest cohorts.

Discussion

As noted earlier, the validity of any age-comparative study of intelligence is directly based on the assumption that the measurement operations used in the study are comparable across age groups. That is, each observation is assumed to measure the same latent construct equally well, regardless of the age of the experimental subjects. Three levels of stringency of measurement equivalence were defined: complete metric invariance, incomplete metric invariance, and configural invariance. Demonstration of the most stringent requirement, complete metric invariance, would mean that measurement operations not only remain relevant to the same latent construct, but also that the correlations of the observable measures with the latent construct remain invariant across age and that the relationships among different constructs in a domain (factor intercorrelations) also remain invariant. In the most simple terms, such a demonstration would imply that inferences can validly be drawn from the results of an age-comparative study, both for age comparisons of directly observed means and of derived factor score means for the latent abilities being measured.

Next, we define a somewhat less stringent equivalence requirement, incomplete metric invariance, by allowing the unique variances and factor intercorrelations to vary across age groups but requiring the factor loadings to remain invariant. If our data permit acceptance of this relaxed requirement, we can still claim that our observations remain relevant to the same underlying latent constructs and that the relationship of the observations to these constructs remains invariant across age. However, the factor space (i.e., the factor variances and covariances among factors) cannot be claimed to remain invariant. In this case, it would still be permissible to draw inferences on comparisons of observed scores. Comparisons of factor scores, however, would be valid only if the changes in the factor space had been adequately modeled in the computational algorithm for such scores.

In the case of the least stringent requirement, configural invariance, we require that observables are relevant to the same

Table 8
Differences in Chi-Square Statistics Among the Three Models

Cohort	Model		
	Complete–incomplete[a]	Incomplete–configural[b]	Complete–configural[c]
1–2	60.43**	25.13	85.56**
3	116.91**	92.29**	209.20**
4	107.62**	70.73**	178.35**
5	79.59**	32.70*	112.29**
6	93.53**	52.95**	146.48**
7	102.24**	42.21**	144.45**
8	89.57**	37.29**	126.86**
9	63.32**	32.31*	95.63**
10–11	140.73**	36.30**	177.03**
df	15	14	29

[a] Difference between the complete metric invariance model and the incomplete metric invariance model. [b] Difference between the incomplete metric invariance model and the configural invariance model. [c] Difference between the complete metric invariance model and the configural invariance model.
* $p < .01$. ** $p < .001$.

latent construct across age. We do not insist, however, that the relationships among the latent constructs retain the same magnitudes, nor do we require that the correlations between observables and their underlying constructs remain invariant. Hence, a demonstration of configural invariance still allows the claim that the observable measures are relevant to the same constructs across age. However, it cannot be claimed, in this instance, that the observables measure the construct equally well at different ages.

In this study, we tested each of these requirements within the domain of psychometric intelligence, as sampled by multiple markers of the latent constructs of Inductive Reasoning, Spatial Orientation, Perceptual Speed, Numerical Ability, and Verbal Ability. We readily admit that this set of constructs is not a complete sampling of the domain of psychometric intelligence. Nor do we claim that each of the abilities samples has been as broadly marked as some might like. Our battery was constructed for the purposes of a substantive longitudinal study of adult development (Schaie, 1983, 1988b). Therefore, to obtain greatest possible construct stability across age and time, the conventional wisdom argued for marking constructs as narrowly as possible, without simply constructing parallel forms for each of the original tests. It would obviously have been possible to test alternative theoretical models, such as one derived from Horn's (1986) theory of fluid (Gf) and crystallized (Gc) intelligence. However, our past work, to which these analyses are closely linked, has been conducted within the primary mental abilities framework rather than in reference to second-order constructs such as the Gf–Gc model. Moreover, our prior work with this battery clearly suggests that a two-factor model does not sufficiently account for the common variance (Schaie et al., 1987), as was the case in the analysis of a similar battery by others (Baltes et al., 1980).

Although some might argue that our selection of measures has loaded the dice in favor of demonstrating construct equivalence across age and time, the results reported here do seem to lay a strong foundation for the validity of age-comparative studies with respect to the least stringent equivalence requirement. That is, if behavior is assessed across age with measures of satisfactory psychometric characteristics, it is most likely that such measures will indeed retain their conceptual position within the domain measured. Moreover, depending on the breadth or narrowness of the domain studied, the same number of factors may suffice to describe that domain (but see White and Cunningham, 1987, for possibly contradictory evidence as to number of factors).

We also recognize that this demonstration of stability of factor patterns may be ascribed in part to the nature of our subject population, which is somewhat skewed toward the upper end of the socioeconomic scale, both in level of education and intelligence. Thus, it might be argued that our results may be relevant primarily to those adults who lead somewhat cognitively challenging lives that may influence the stability of cognitive structures across cohorts (cf. Kohn & Schooler, 1973). Nevertheless, we must maintain that our sample represents at least as broad, and probably a broader, demographic spread than any other study of adults currently reported in the relevant research literature. These results, therefore, would seem to apply directly to most studies of adult cognitive functioning that use volunteer samples.

The demonstration of configural (factor pattern) invariance is initially reassuring to developmentalists in that it confirms our hope that it is realistic to track the same basic construct across age and cohorts in adulthood. Nevertheless, our study gives rise to serious cautions with respect to the adequacy of the construct equivalence of age-comparative studies. Indeed, here the relative narrowness of our battery causes us to be even more concerned as to what one would expect in less tightly constructed assessment programs. The second major finding of this study lies in the demonstration that neither complete nor incomplete metric invariance could be demonstrated with respect to a population-based measurement model for any age/cohort level. That is, given the conservative nature of our test, we must conclude that age-comparative studies at the single observable level are unlikely to provide completely equivalent estimates of latent constructs across age.

How serious is the divergence from complete metric equivalence? In part, shifts in the interrelation among ability constructs involve the increasing convergence of the ability factor space that has previously been associated with a differentiation–dedifferentiation hypothesis (Reinert, 1970). As this hypothesis predicts, factor covariances are lowest for our younger cohorts and increase with advancing age. There is also an increase in factor variances, at least until the seventies, when disproportionate drop-out of those at greatest risk once again increases sample homogeneity and reduces factor variances. Because our data set for the test of complete metric invariance was centered age-wise on late midlife, it is not surprising that the discrepancies in factor covariances when factor loadings are either constrained (Table 5) or unconstrained (Table 7) are confined primarily to the extremes of the age range studied. These shifts, consequently, would not seriously impair the validity of age-

comparisons using factor scores, except where extreme age ranges are involved.

Because we accept the configural invariance model as the most plausible description for the structure of our data set, we must also be concerned with shifts in the correlations of observable measures with the latent constructs. What is at issue here is that a particular observable measure may be a more or less efficient measure of a construct at one age than is true at other age levels. This shift could occur because of the influence of some extreme outliers in small samples (a possibility that we ruled out in this study by means of scatter plots for those cohorts showing deviations in factor loadings) or, more likely, by the attainment of floor effects in the older cohorts or ceiling effects in the younger cohorts for some variables. For example, with increasing age, the Cube Comparison test becomes a less effective marker of Spatial Orientation, whereas the PMA Space measure becomes a stronger marker. Similarly, the Number Comparison test (a marker of Perceptual Speed), which has a secondary loading on Numerical Ability in the general factor model, loses that secondary loading with increasing age. These findings suggest that age comparisons in performance level on some single markers of an ability may be confounded by the changing efficiency of that marker in making the desired assessment. Fortunately, in our case, the divergencies are typically quite local in nature. That is, for a particular ability, the optimal regression weights of observable measures on their factors may shift slightly, but there is no shift in the primary loading to another factor, and the structural relationships are well maintained across the entire age range sampled.

In sum, our results suggest that many markers of psychometric abilities are quite robust with respect to their construct equivalence across the adult age range. Consequently, our demonstration of invariant factor structures supports the validity of quantitative comparisons across different age levels provided that measures are used that have good reliability. However, because of the finding of age differences in the efficiency with which latent constructs are measured by individual markers, it also follows that inquiries of age differences, whenever possible, should use multiple markers so that inferences as to age differences can be made at the more stable construct level. Our findings, of course, are restricted to a limited part of the domain of psychometric intelligence in which a high degree of stability of these measures has been demonstrated in many investigations. It might be suggested, therefore, that the observed discrepancies from complete metric equivalence across age are likely to be far more serious in other substantive domains that involve less reliable measures.

References

Alwin, D. F. (1988). Structural equation models in research on human development and aging. In K. W. Schaie, R. T. Campbell, W. Meredith, & S. C. Rawlings (Eds.), *Methodological issues in aging research* (pp. 71–170). New York: Springer.

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49,* 155–173.

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: Pennsylvania State University 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.

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

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.

Bollen, K. A. (1986). Sample size and Bentler and Bonett's nonformed fit index. *Psychometrika, 51,* 375–377.

Cohen, J. (1957). The factorial structure of the WAIS between early adulthood and old age. *Journal of Consulting Psychology, 21,* 283–290.

Cunningham, W. R. (1980). Age comparative factor analysis of ability variables in adulthood and old age. *Intelligence, 4,* 133–149.

Cunningham, W. R. (1981). Ability factor structure differences in adulthood and old age. *Multivariate Behavioral Research, 16,* 3–22.

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

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.

Horn, J. L. (1986). Intellectual ability concepts. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 3, pp. 35–78). Hillside, NJ: Erlbaum.

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.

Kohn, M. L., & Schooler, C. (1973). Occupational experience and psychological functioning: An assessment of reciprocal effects. *American Sociological Review, 38,* 97–118.

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

Nesselroade, J. R., & Labouvie, E. W. (1985). Experimental design in research on aging. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (2nd ed., pp. 25–60). New York: Van Nostrand Reinhold.

Reinert, G. (1970). Comparative factor analytic studies of intelligence

throughout the human life-span. In L. R. Goulet & P. B. Baltes (Eds.), *Life-span developmental psychology* (pp. 467–484). New York: Academic Press.

Schaie, K. W. (1973). Methodological problems in descriptive developmental research on adulthood and aging. In J. R. Nesselroade & H. W. Reese (Eds.), *Life-span developmental psychology: Methodological issues* (pp. 253–280). New York: Academic Press.

Schaie, K. W. (1977). Quasi-experimental designs in the psychology of aging. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (2nd ed., pp. 39–58). New York: Van Nostrand Reinhold.

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. (1988a). Internal validity threats in studies of adult cognitive development. In M. L. Howe & C. J. Brainard (Eds.), *Cognitive development in adulthood: Progress in cognitive development research* (pp. 241–272). New York: Springer-Verlag.

Schaie, K. W. (1988b). Variability in cognitive function in the elderly: Implications for social participation. In A. D. Woodhead, M. A. Bender, & R. C. Leonard (Eds.), *Phenotypic variation in populations: Relevance to risk assessment* (pp. 191–212). New York: Plenum.

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., & Hertzog, C. (1986). Toward a comprehensive model of adult intellectual development. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 3, pp. 79–118). Hillsdale, NJ: Erlbaum.

Schaie, K. W., Willis, S. L., Hertzog, C., & Schulenberg, J. E. (1987). Effects of cognitive training upon primary mental ability structure. *Psychology and Aging, 2,* 233–242.

Stricker, L. J., & Rock, D. A. (1987). Factor structure of the GRE general test in young and middle adulthood. *Developmental Psychology, 23,* 526–536.

Thurstone, L. L. (1947). *Multiple factor analysis.* Chicago: University of Chicago Press.

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

White, N., & Cunningham, W. R. (1987). The age comparative construct validity of speeded cognitive factors. *Multivariate Behavioral Research, 22,* 249–265.

Willis, S. L. (1987). Cognitive training and everyday competence. In K. W. Schaie (Ed.), *Annual review of gerontology and geriatrics* (Vol. 7, pp. 159–188). New York: Springer.

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

Received October 21, 1987
Revision received August 11, 1988
Accepted December 14, 1988 ■