Longitudinal and Related Methodological Issues in the Swedish Twin Registry

K. Warner Schaie, Ph. D.

The Pennsylvania State University

University Park, PA, USA

Paper prepared for a seminar of the

Swedish Planning Council

Stockholm Sweden

May 8-9. 2000

[illegible]

[illegible]

The Swedish twin registry has been the vehicle for obtaining a unique data base which has been widely exploited for a vast array of epidemiological and behavior genetic studies that have made important contributions to the research literature on human aging. The registry has at times been described as a longitudinal study that provides prospective data which can be related to intra-individual events and condition occurring at later points in the lives of the study participants. However, it should be made clear from the outset that the registry was not explicitly designed as a longitudinal study. Except for relatively recent (though extremely important) work associated with the SATSA study it seems to have engendered studies that have been primarily epidemiological in focus. Nevertheless, it is clear that one of the major future contributions of the registry (and associated data) will be in the assessment of questions which are relevant to normal and pathological development, and which require strong longitudinal focus.

I really do not think it is necessary or appropriate here for me to justify why the longitudinal approach is mandatory for an effort such as the registry. What I will concentrate on rather is the fact that because of its longitudinal nature the continuation of this study has to be concerned with all of the typical threats that can impair the validity of longitudinal studies. These threats include issues of experimental mortality (attrition), reactivity (practice effects), instrumentation effects, regression, and most importantly the impact of secular change on successive cohorts entering the registry.

The registry was designed to cover significant portions of the human life span, and it s now proposed that it be extended to become a major resource for life-span studies of twins. It is therefore important to attend to the issue of construct validity over time by addressing the stability of measurements both for the directly observed indicators of participants' behaviors, health status and socio-demographic attributes as well as for the latent constructs formed from these observations that are of particular interest for the study of long-term relationships. Hence, a discussion is required of the manner in which the broad array of interdisciplinary data available in the registry has been or might best be organized to yield a smaller set of more parsimonious latent constructs. At issue also are issues of demonstrating factorial equivalence of these latent constructs across time and different groups included in the registry.

In addition to the issues raised above, which affect any longitudinal study, it must also be stressed that the study of twin and/or family data raise special methodological problems because the members of a twin pair are not randomly selected, but rather represent a dyad whose data will be correlated due to heritability and/or temporally shared environment. Hence, the usual design consideration for longitudinal studies must be extended to consider the components of growth curves that are correlated within dyads. Distinctions must therefore be made between the developmental and causal patterns that obtain for the twin dyads and their individual members.

In this presentation, I will first consider the potential threats the internal and external validity of studies utilizing the registry the status of the Swedish Twin Registry as a series of longitudinal studies. Second, I will try to explicate the design implicit in the presently available data archives as well as the plans for their augmentation. and will make some recommendations as to how the archives might be organized to assess the magnitude and significance of the validity threats. I will thirdly address the issue of how the latent constructs underlying the data collected in the registry have been formulated and make some suggestions on possible alternatives. Fourth, I will consider the issue of factorial invariance of latent constructs and make some suggestions as to how this issue might be addressed for the registry. Finally, I will briefly comment on the additional complexities of longitudinal designs that involve dyadic data sets, although I will gladly defer to the opinions of my colleagues with a stronger behavior genetic expertise than I would claim to possess.

Let me begin by quickly reviewing the threats to the internal and external validity of longitudinal studies and remind us of the design requirements that must be met to address these threats. I can then comment on the extent to which these requirements are now met by the registry as well as to recommend what could be done to remedy current limitations.

[illegible]

Campbell and Stanley (1967) described eight different threats to the internal validity of quasi-experiments such as longitudinal studies: Maturation, effects of history, reactivity, instrumentation, statistical regression, experimental mortality, selection and the selection-maturation interaction. The first two of these, history and maturation, have special meaning for scientists studying individual development. *Maturation*, quite obviously, is not a threat to the validity of developmentally oriented longitudinal studies, but rather is the specific effect of primary interest to the investigator. Presumably, this is the normal developmental course of individuals over their life span, given their genetic predispositions and the characteristic demands of the culture and environment within such maturation occurs.

By contrast, *historical* effects are indeed the primary internal validity problem of longitudinal studies. History is directly involved in both cohort and time-of-measurement (period) effects. However, cohort effects represent the impact of historical effects upon a group of individuals who share similar environmental circumstances at equivalent temporal points in their life course. But, time-of-measurement effects represent those events that impact all members of the population, regardless of cohort membership, that experience common historical exposures. The specific threat to longitudinal studies is that historical effects may threaten the internal validity of designs that attempt to measure the effect of maturation (aging effects). The implication here is that effects thought to be age-dependent must be carefully disaggregated from those due historically limited environmental impacts. To do so, it is necessary to follow a minimum of two cohorts over similar age ranges (Schaie, 1977, 1988).

Longitudinal studies, such as those represented by some of the data collections in the registry are effected also by the other six threats to internal validity described by Campbell and Stanley. *Reactivity* may simply involve practice effects on performance measures to the extent that study participants spend less time figuring out problems previously solved and therefore improve their performance because of previous expose to procedures that are part of the experimental protocol. On the other hand, longitudinal study participants might also respond on subsequent test occasions very differently than would be the case if they had not been previously tested; a behavior change that could be confused with the effects of maturation. Methods for assessing practice effects are available when at least two sub-samples are available at different levels of measurement exposure (cf. Schaie, 1988).

The internal validity threat of *instrumentation* refers to differences in measurement techniques that covary with measurement occasions. In long-term longitudinal studies, such differences are likely to occur when study personnel changes, or when records regarding study protocol on previous occasions have been lost and slight variations in protocol are introduced inadvertently. Such effects could lead to the erroneous inference of having demonstrated maturational trends or the impact of societal interventions. I suspect that any long-term data collection such as the registry is likely to be plagued by this problem. The equivalence of data collections should be fully documented and statistical adjustments made where necessary.

Statistical regression is the tendency of variables containing measurement error to regress towards the population mean from one occasion to the next. This problem is of particularly important in sub-sets of data for which only two data points are available (see Baltes, Nesselroade, Schaie, and Labouvie [1972] and Schaie and Willis [1986] for examples of applications of the time-reversal method, that can be used to test for the effect of regression in such studies). It has been shown, however, that regression effects do not necessarily cumulate over extended longitudinal series (Nesselroade, Stigler, & Baltes, 1980). If evidence for statistical regression is found, one can either adjust for reliability of the base line scores, or model change at the latent construct level, thus permitting better control of error variance.

Members of longitudinal panels obviously cannot be forced to continue their participation. Consequently, another serious threat to the internal validity of longitudinal studies is that of *experimental mortality*. This term describes the attrition of participants from a sample between measurement occasions, whether such attrition is due to biological mortality, morbidity, or simply experimenter ineptness in maintaining good relations with his/her panel members. Most empirical studies of experimental mortality suggest that attrition is non-random at least between the first and second measurement occasion (Cooney, Schaie, & Willis, 1989; Schaie, 1988, 1996b). It is important to make distinctions between "natural" mortality; i.e., attrition caused by death or disability, from attrition caused by refusal to continue participation, or experimenters' failure to locate or access participants for logistic reasons. Such data should be provided by age/cohort groups.

Selection refers to the process of obtaining a sample from the population such that the observed effect is a function of the specific sample characteristics rather than of the maturational effect we wish to estimate (cf. Nesselroade, 1988). This issue is of particular importance in twin studies, if findings are to be generalized to other populations.The *selection-maturation interaction*, of course, refers to the case where maturational effects may be found in some samples but not in others. It would be a particularly importance service if differences between the registry samples and data from the general Swedish population could be described for the age/cohorts represented in the registry.

None of the internal validity threats can be controlled for or measured in single cohort longitudinal studies. When multiple data sets are available, however, the magnitude and significance of these effects can be estimated, and appropriate corrections applied in the substantive studies. Specific designs for appropriate analyses have been presented by Schaie (1977, 1988, 1996b).

□□□□□□□□□□□□□□□□□□□

Longitudinal studies also share certain limitations with respect the generalizability of their findings (cf. Cook & Campbell, 1975). Four major issues can be identified in this respect: The first concerns *experimental units*, that is the extent to which longitudinal data collected on one sample can permit inference to other populations. This issue is a particularly sensitive one in twin studies, where one must expect that twins represent a non-random sample of a broader reference population (see comments on selection, above). The second involves *experimental settings* or the extent to which findings from a given study would apply in different contextual circumstances. This is a concern that may apply as data acquisitions expand to younger individuals than have previously been surveyed or assessed. The third is concerned with *treatment variables*, that is the limitations imposed by specific settings or environmental exposures that might be limited to the experience of twins (e.g., greater temporal overlap in shared environment than for non-twins, The adoption data can obviously illuminate this issue. Finally, external validity may be threatened by certain aspects intrinsic to the measurement variables, to the extent to which task characteristics remain appropriate at different developmental stages as a longitudinal study progresses. This. of course is the problem of construct validity and factorial invariance to which we will return later (also cf. Maitland, Intrieri, Schaie, & Willis, 2000; Meredith, 1993; Schaie, Willis, Jay, & Chipuer, 1989; Schaie, Maitland, Willis, & Intrieri, 1998).

[illegible]

I will next briefly describe the data sets available in the registry. I will not attend to the substantive content, but rather restrict myself to the formal design attributes. I will then identify the design problems associated with the current data organization and suggest possible alternative schemes that would allow investigating the internal and external validity threats I will discuss later on.

□□□□□□□□□□□□□□□□□□□□

The Swedish twin registry was begun in 1959 (Cedrlöf and Lorich, 1978). Three cohorts have been investigated: The first ("old") cohort represents multiple birth dyads known to be alive in 1959, who were born between 1886 and 1925; N = 12,889).. This represents a cohort bandwidth of 40 years. In 1960-61 a questionnaire was sent to like-sexed twins. Those who responded constitute the initial data set for this cohort (N = 10,945). Additional data points on this cohort were collected in 1963 (N = 9.139), and 1967 to 1970 (N = 8,375). Minimal information has also been accessed for unlike-sexed twins from this cohort (N = 11,500 ±). Another data point is currently being collected on these subjects.

A second cohort ("the new registry" of twins born from 1926 to1958 (bandwidth 23 years, N = ±14,000 twin pairs), was compiled in 1970, A questionnaire similar to that used for the "old" cohort was sent out in 1972-73. (I did not have numbers of returns available to me). Follow-up data on this cohort are currently being collected

A third cohort born from 1969 to 1990 (bandwidth 22 years) has been identified. From this cohort only parents of twins born in 1985 and 1986 were contacted (numbers not knows).

As currently available to the registry (and including the data collections currently in progress there are actually only two cohorts available for longitudinal analysis. Except for the current data collection the times-of-measurement for the two cohorts are non-overlapping.

□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

As already mentioned, it is apparent that the twin registry was designed for epidemiological studies of twin populations that at their outset did not include explicit provisions for a systematic design of longitudinal follow-ups. As a consequence we are currently presented with two single-cohort studies which are not directly comparable. However, the current data collections and simple organizational restructuring of the data sets may offer some opportunities for introducing more efficient multi-cohort designs. In addition, estimation of factor scores that are properly weighted for differences in the regression of latent constructs on the observed variables may permit greater comparability across the two cohorts. I will discuss this issue later in more detail.

One of the major problems of the present organizational structure is that the bandwidth of the two cohorts with respect to year of birth differs, and in any event is too broad for both cohorts. Given the large sample size it should be possible to disaggregate the data base for both cohorts into narrower width cohorts so as to take advantage of multiple cohort designs that can be used more efficiently to control and/or estimate the effects of threats to the internal validity of studies using the registry. Such disaggregation would result in designs which I have previously referred to as cross-sequential strategies (cf. Schaie, 1975, 1977; Schaie & Willis, 1996). Some attempts at implementing cohort-sequential designs have been made for the SATSA sub-samples (cf. Finkel et al., 1995, 1998) since multiple data points are available here. However, extension of these designs to the broader array of registry variables still need to be accomplished.

A related problem is the fact that the observational intervals are not comparable for the two studies. Again, it may be possible to structure the available data in a manner that might allow statistical estimates (imputations) of changes occurring across standard intervals that might provide for better comparisons of the data sets contained in the registry.

The first cohort represented an age range in 1960/61 from 45 to 74 years. These data could, for example, be diaggregated into 3-year age cohort groups, to be able to estimate short-term longitudinal changes within groups for comparison with the 3 year time interval between the first two data collection. Alternatively, a five-year cohort interval might be desirable if the focus is to be on understanding short-term longitudinal change from the second to the third assessment interval. Change from the last to the current data acquisition could be estimated for five-year cohort groups and proportionalized to a five-year interval. Data would then be available for multiple-cohort analyses that could address the validity threats discussed above.

[illegible]

We now turn to the organization of the data structure to consider the latent variables that have thus far been investigated, and how one might proceed to enhance the parsimonious manner of organizing the structure of the twin registry. has had substantial discussion in the aging literature in recent years. Let me explain:

When we wish to compare observations across period of time (age) within individuals, or when we want to compare the status of two or more distinct population subsets (e.g. gender, urban/rural, or socio-economic status) we make the implicit assumption that the observations have the same relation to the underlying hypothetical construct of interest. This relationship is expressed technically as the equivalence of the factor loadings of the observed variables on the latent constructs.Only when the invariance of this relationship can be shown to hold can meaningful inferences be drawn.

Horn, McArdle and Mason (1983) drew attention to an important distinction between two levels of invariance in factor loadings (a distinction first introduced by Thurstone [1947, pp. 360-369]) that may have different implications for age change and age difference research: *configural invariance* and *metric invariance.* Meredith (1993) has spelled out in greater detail what are considered to be necessary conditions to satisfy this factorial invariance at different levels of stringency.

We would need to show at a minimum that the factor pattern across groups or time display *configural invariance.* In this case, all measures marking the factors (latent constructs) have their primary non-zero loading on the *same* ability construct across test occasions or groups. They must also have zero loadings on the same measures for all factor dimensions.

A second (more desirable) level of factorial invariance (termed *weak factorial invariance b*y Meredith) requires that the unstandardized factor pattern weights (factor loadings) can be constrained equal across groups or time. The technical and substantive considerations for this level of factorial invariance have found extensive discussion in the literature (cf. Horn, 1991; Horn & McArdle, 1992; Jöreskog & Sörbom, 1979; Meredith, 1993 ;Schaie & Hertzog, 1985; Sörbom, 1974; Thurstone, 1947). If this level of invariance can be accepted than it becomes possible to test hypotheses about the equivalence of factor means. More importantly, for our purposes, one can then test further hypotheses about the latent factor variances and covariances.

However, we should stress that it is probably questionable whether even the assumptions of weak factorial invariance can be met in complex empirical data sets such as is found in many aging studies. In fact, Horn, McArdle and Mason (1983) early on argued that configural invariance is likely to the best solution that can be obtained. Nevertheless, it should be possible to demonstrate more stringent levels of invariance for sub-systems across some ages and cohorts. Byrne, Shavelson and Muthén (1989) have proposed therefore that one should also test for partial measurement invariance. This proposition has been received with much controversy in the factor-analytic literature. However, it seems that testing for partial invariance is quite reasonable from the point of view of the substantively oriented scientist because of the undue sensitivity of most SEM estimates to local disturbances of model fit.

In any event, it should be evident that for both cross-sectional and longitudinal studies, configural invariance remains a minimal requirement, while demonstration of some form of metric invariance is essential before valid comparisons of latent factor scores can be made. This issue requires serious attention in further work on the registry data.

[illegible]

It seems reasonable to suppose, given the demonstration of configural invariance, that developmental processes or differential cohort experiences can lead to changes or differences in the magnitude of the regression of the latent constructs on the observed variables. Even though a particular test may measure the same latent construct over different life stages, or population sub-sets, it may do so with different degrees of efficiency. Before proposing to use structural equation modeling of a causal nature, or before using techniques such as growth curve modeling my recommendation would be that it is necessary to proceed.as follows:

1. Test the least restricted acceptable model, configural invariance.
    a. Constrain all non-salient factor loading to zero
    b. Estimate all other loadings for each group/time
    c. Estimate factor variances/covariances for each group time

2. Test the *weak invariance* model

   a. Constrain all factor loading to be equal across groups/time

   b. Estimate factor variances/covariances for each group time

3. If necessary, test partial invariance model

   a. Examine modification indices and/or standard errors of measurement for factor loadings to determine the partial invariance model

   b. Constrain all factor loading to be equal across groups/time, except those determined to be freed up in step a.

   c. Estimate all other loadings for each group/time

   d . Estimate factor variances/covariances for each group tim

A related issue here is to give consideration to changes in the factor space of the domains covered by the registry. There has been considerable attention recently to observations that behavioral performance and measures of sensory capabilities converge in advanced old age (cf. Baltes & Lindenberger, 1997). One could interpret this phenomenon to suggest that physiological processes should be given priority as outcomes in old age, even though the behaviors might be of greater interest as indices of life quality. But the literature is not yet clear whether the so-called dedifferentiation phenomenon (Werner, 1948) can be demonstrated to hold across all domains of behavior. Again latent factor analysis allows a formal test of this hypothesis.

Once the most stringent invariance model permitted by the data is accepted one can then proceed to test differentiation-dedifferentiation hypotheses as follows:

1. Constrain factor variances/covariances equal across all groups/time.

If the fit for this model in terms of Delta Chi-Square is not significantly worse than the accepted invariance model, it can be concluded that the hypothesis has been falsified.

2. If the fit is significantly worse then modification indices and/or standard errors of measurement for the variance/covariance matrices are inspected to determine a partial invariance model. The results can then be interpreted as a partial confirmation of the hypothesis.

Improvement of fit would generally be examined in terms of Delta Chi-Square, but other fit indices can obviously be used as well, although their distributional characteristics are not as well established for the purpose of model comparison (cf. Browne & Cudeck, 1993).

[illegible]

As far as I can tell from scanning some of the products of studies using registry data, there have been a number of efforts at structural modeling for purposes of some of the sub-projects (e.g., in the SATSA studies. Pedersen and Reynolds, 1998). Whenever possible, however, the definition of latent constructs and their observed markers is best conducted with the largest possible data sets, so that it becomes possible to use random sub-sets for initial model fitting and other random sets for establishing the stability of the accepted model.

In twin studies it would also seem important to study the invariance of the obtained dimensionalizations of the available data within monozygotic, heterozygotic and mixed gender twin sets. These in turn, need to be crossed to study invariance over time and age. Given the magnitude of the data set missing data imputation algorithms could also be used. But before such efforts can be employed, it would seem that a it is necessary to concatenate data arising from the several sub-studies to the extent that these are actually available to the registry data management. Such an effort would also be important for conducting future survival analysis that may depend upon the inclusions data only available for sub-sets of study participants.

[illegible]

I assume that the methodological problems related to behavior genetic issues are dealt with in Dr. DeFries' report and will therefore only mention some of the relevant longitudinal issues. Appraisal of longitudinal change in work with dyadic sets requires additional analyses as follows:

1. *Studies of invariance (Stability).* To determine whether there differential change within dyads, it is necessary to conduct such analyses for stability of inter-individual differences between sets of dyads, but also for stability of within dyad differences across time. The appropriate base model for the analysis of longitudinal invariance would be configural invariance design in which both members of the dyad and all times of measurement are seen as replications. In that case model fit can be improved by allowing the error covariance within dyads and across time to be estimated rather than fixed as would be true in between group designs.

2. *Analyses of level changes. (*M)ANOVAs used for this purpose need to treat time and dyadic membership as within group variables. For variables with high stability it might be desirable to enter change scores for the dependent variable. If possible estimation of means for latent variables should be done in conjunction with the invariance testing.

3. *Survival analysis designs.* It is suggested that for studies of human

aging, it makes more sense to determine log age rather than log time to index the occurrence of the out come events (cf. Schaie, 1989).. In dyadic studies, the two members of the dyad could be treated as competing outcomes.

□□□□□□□□□□□

The Swedish Twin Registry is a unique data base that deserves to be continued and augmented. In particular, it would be desirable to extend the study over the entire life span, with schedule assessments of a minimal core battery over regular time intervals. The registry has provided the basis for a large number of studies that have enriched the scientific literature and has much promise to continue to do so. This data set, properly handled, is not only relevant for health and social policy decision making in Sweden, but also represents a resource of international quality and importance for basic sciences research on human behavior genetics and aging.

The longitudinal design of the registry suffers from unequal bandwidth in terms of entry into the various cohorts, as well as for unequal spacing of observations over time. This creates difficulty in analyzing the various validity threats common to all longitudinal studies, and which in my judgment have not been sufficiently addressed in the past (although these issues are recognized and partially addressed in recent publications from the SATSA study).

In this report I have therefore recommended that the data be restructured so as to disaggregate narrower sub-sets that can be better assessed longitudinally and that can be compared across the two study cohorts by more closely matching equivalent time spans for different sub-sets.

I am also suggesting that the vast array of data now available could be better dimensionalized and studied at the latent construct level, once invariance of constructs with and across the dyadic pairs have been established for the various sub-set of interest. I am also recommending that data from the various sub-studies be concatenated in such a way as to facilitate missing data imputation for those participants on whom only limited data have actually been observed.

Having myself been involved in the acquisition and management of long-term longitudinal archives, I would also like to recommend that the utility of the registry data acquisition could be enhanced and preservation of data be ensured by making provisions for resources that would allow scanning of all data now available only in hard copy format. Once the current data acquisition is completed, it would also be desirable to investigate making cleaned and suitably protected data sets available via an appropriate web site.

It has been a privilege to review this important efforts, and I hope it can be continued and strengthened

□□□□□□□□□□

Baltes, P. B., Nesselroade, J. R., Schaie, K. W., & Labouvie, E. W. (1972). On the dilemma of regression effects in examining ability level-related differentials in ontogenetic patterns of adult intelligence. *Developmental Psychology, 6,* 78-84.

Browne, M. W., & Cudeck, R. (1993). Alternate ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models.* (pp. 136-162). Newbury Park, CA: Sage.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456-466.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105,* 456-466.

Campbell, D.T., & Stanley, J.C. (1963). Experimental and quasi- experimental designs for research in teaching. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 171-246). Chicago: Rand McNally.

Cederlöf, R., & Lorich, U. (1978). The Swedish Twin Registry. In *Twin research: Biology and epidemiology* (pp. 189-195). New York: Alan R. Liss.

Cook, T. C., & Campbell, D. T. (1975). The design and conduct of quasi-experiments and true experiments in field settings. In M. D. Dunette (Ed.), *Handbook of industrial and organizational psychology.* Skokie, IL: Rand McNally.

Cooney, T. M., Schaie, K. W., & Willis, S. L. (1988). The relationship between prior functioning of cognitive and personality dimensions and subject attrition in longitudinal research. *Journal of Gerontology: Psychological Sciences, 43,* P12-P17.

Finkel, D., Pedersen, N. L., McClearn, G. E., Plomin, R., & Berg, S. (1996). Cross-sequential analysis of genetic influences on cognitive ability in the Swedish Adoption/Twin Study of Aging. *Aging Neuropsychology & Cognition, 3,* 84-99.

Finkel, D., Pedersen, N. L., Plomin, R., & McClearn, G. E. (1998). Longitudinal and cross-sectional twin data on cognitive abilities in adulthood: The Swedish Adoption?Twin Study of Aging. *Developmental Psychology, 34,* 1400-1413.

Horn, J. L. (1991). Comments on issues in factorial invariance. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change* (pp. 114-125). Washington, DC: American Psychological Association.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research. 18,* 117-144.

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

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.

Maitland, S. B., Intrieri, R. C., Schaie, K. W., & Willis, S. L. (2000). Gender differences in cognitive abilities: Invariance of covariance and latent mean structure. *Aging, Neuropsychology & Cognition, 7*, 1-22.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58,* 525-543

Nesselroade, J. R. (1988). Sampling and generalizability: Adult development and aging research issues examined within the general methodological framework of selection. In K. W. Schaie, R. T. Campbell, W. Meredith, & S. W. Rawlings (Eds.), *Methodological issues in aging research*, (pp. 13-42). New York: Springer Publishing Co.

Pedersen, N. L., & Reynolds, C. A. (1998). Stability and change in adult personality: Genetic and environmental components. *European Journal of Personality, 12,* 365-386.

Schaie, K. W. (1975). Research strategy in developmental human behavior genetics. In K. W. Schaie, E. V. Anderson, G. E. McClearn, & J. Money (Eds.), *Developmental human behavior genetics*, (pp. 205-220). Lexington, MA: D. C. Heath.

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*, (pp. 39-58). New York: Van Nostrand Reinhold.

Schaie, K. W. (1988). 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. (1989). The hazards of cognitive aging. *Gerontologist, 29*, 484-493.

Schaie, K. W. (1996). *Intellectual development in adulthood: The Seattle Longitudinal Study.* New York: Cambridge University 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. 61-92). New York: Van Nostrand Reinhold.

Schaie, K. W., Maitland, S. B., Willis, S. L., & Intrieri, R. L. (1998). Longitudinal invariance of adult psychometric ability factor structures across seven years. *Psychology and Aging, 13*, 8-20.

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

Schaie, K. W., & Willis, S. L. (1996). *Adult development and aging* (4th ed.). New York: HarperCollins.

Schaie, K. W., Willis, S. L., Jay, G., & Chipuer, H. (1989). Structural invariance of cognitive abilities across the adult life span: A cross-sectional study. *Developmental Psychology, 25,* 652-662.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structures between groups. *British Journal of Mathematical and Statistical Psychology, 27*, 229-239.

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

Werner, H. (1948). *Comparative psychology of mental development.* New York: International Universities Press

. .