GEORGIA INSTITUTE OF TECHNOLOGY — OFFICE OF CONTRACT ADMINISTRATION

SPONSORED PROJECT TERMINATION/CLOSEOUT SHEET

Date 9-1-88

Project No. G-42-631 School/~~Lab~~ Psych

Includes Subproject No.(s) N/A

Project Director(s) C. K. Hertzog GTRC/~~GIT~~

Sponsor DHHS/PHS/NIH/National Institute of Health

Title Individual Differences in Adult Cognitive Development

Effective Completion Date: 7-31-88 (Performance) N/A (Reports)

Grant/Contract Closeout Actions Remaining:

- [ ] None
- [x] Final Invoice or Copy of Last Invoice Serving as Final
- [ ] Release and Assignment
- [ ] Final Report of Inventions and/or Subcontract: Patent and Subcontract Questionnaire sent to Project Director [ ]
- [ ] Govt. Property Inventory & Related Certificate
- [ ] Classified Material Certificate
- [ ] Other _____

Continues Project No. _____ Continued by Project No. G-42-609

COPIES TO:

Project Director
Research Administrative Network
[redacted]
Accounting
Procurement/[redacted]
[redacted]
Reports Coordinator (OCA)
[redacted]
Contract Support Division

[redacted]
GTRC
Project File
Other _____

14:38:49 OCA PAD INITIATION - PROJECT HEADER INFORMATION 08/06/87

Project # G-42-631 MOD # REV # 0
Contract # 5 K04 AG00335-03 OCA file # Status A
Contract entity GTRC Prime contract #
PDPI HERTZOG C K (DR.
SSN - - Unit PSYCH Phone ( ) -
Project unit PSYCH Unit code 02.010.154
Sponsor/Division DHHS/PHS/NIH / NATL INSTITUTE OF HEALTH
Sponsor#/division # 108 / 001
Type of document GRANT
Award period: from 87 / 08 / 01 to 88 / 07 / 31 (perf) 88 / 10 / 31 (rpts)

Sponsor amount	New this change	Total to date
Contract value	$ 58249	58249
Funded	$ 58249	58249

Cost sharing # G-42-323 Cost sharing $ 2000
Does subcontracting plan apply? (Y/N) N
Title:
INDIVIDUAL DIFFERENCES IN ADULT COGNITIVE DEVELOPMENT

CTR project # R6002-2A0 CTR cost sharing # F6002-2A0

Are there existing subprojects? (Y/N) N
Is this a subproject? (Y/N) N Main project #
Continuation of project # G-42-619 Type of research RES

Coproject director name

SSN - - Unit

Coproject director name

SSN - - Unit

PROJECT ADMINISTRATION DATA

Administrative data OCA contact E. FAITH GLEASON PAD CO EFG 894-4820

Sponsor technical contact	Sponsor issuing office
MATILDA W. RILEY, D.SC.	CAROL L. TIPPERY
(301) 496 - 1472	(301) 496 - 1472
NATIONAL INSTITUTE ON AGING	NATIONAL INSTITUTE ON AGING
BETHESDA, MD 20892	BETHESDA, MD 20892

Security class (U,C,S,TS) ONR resident rep. is ACO (Y/N) N
Defense priority rating
supplemental sheet
Equipment title vests with Sponsor GIT X Comment follows -

Admin comments -
RESEARCH CAREER DEVELOPMENT AWARD

5-92-031

MEMORY PERCEPTIONS AND MEMORY PERFORMANCE IN ADULTHOOD AND AGING

David F. Hultsch, Ph.D.
Department of Psychology
University of Victoria
Victoria, British Columbia V8W 2Y2

Roger A. Dixon, Ph.D.
Max Planck Institute
for Human Development and Education
Lentzealle 94
1000 Berlin 33
West Germany

and

Christopher Hertzog, Ph.D.
Department of Individual and Family Studies
Pennsylvania State University
University Park, PA 16801

Running Head: Memory Perceptions and Performance

Abstract

Evidence is accumulating to suggest that memory perceptions are important for understanding memory functioning in later life. This paper reviews two questionnaires designed to measure such perceptions, and identifies several research questions requiring further study. In general, memory perceptions appear to be multidimensional and involve belief and affective components as well as knowledge components. In addition, the relationship between memory perceptions and memory performance appears to differ as a function of age and type of task. Salient issues for future research include the extent to which older adults' memory perceptions show systematic state-like fluctuations over time, and the extent to which these perceptions represent an accurate picture of individuals' memory abilities.

Footnote

The research conducted in our laboratories and the preparation of this report was supported by grants AG00910, AG00048, and AG05165 from The National Institute on Aging and by Grant 492-84-002 from the Social Sciences and Humanities Research Council of Canada.

Requests for reprints should be sent to David F. Hultsch, Department of Psychology, University of Victoria, Victoria, British Columbia, V8W 2Y2, Canada.

Individuals' understanding of their own cognitive functioning is currently receiving considerable renewed attention. An explicit example of this attention is the introduction of the concept of metamemory by Flavell (1971). He defined this construct as involving three classes of knowledge: (1) knowledge of the memory demand characteristics of particular tasks or situations, (2) knowledge of potentially employable strategies relevant to a given task or situation, and (3) memory-relevant characteristics of the person themselves (Flavell & Wellman, 1977). Recently, individuals' knowledge, beliefs, and affects about their own memory have become important foci of researchers attempting to understand apparent age-related declines in memory performance in later life (Dixon & Hertzog, 1984; Dixon & Hultsch, 1983a,b; Hulicka, 1982; Zelinski, Gilewski, & Thompson, 1980). The central assumptions underlying much of this work are that (1) individuals' knowledge, beliefs and affects about their own memory are important determinants of their behavior in memory demanding situations, and (2) that such memory self-perceptions become particularly salient in the latter half of life, contributing significantly to observed declines in performance.

These notions are derived from more general life-span developmental arguments emphasizing the importance of linkages among social, personality, and cognitive processes in development (e.g., Baltes, Dittman-Kohli, & Dixon, 1984; Hultsch & Pentz, 1980). More specifically, this view suggests that age-related changes in basic memory processes may be but one contributing factor in the typically observed decline in performance with increasing age. In particular, individuals' performance will be shaped not only by their actual skills, but also by their understanding of the cognitive demand characteristics of the situation, and their perceptions about the likely outcomes of their behaviors in such a situation. This perspective does not deny age-related changes in underlying memory processes. Such changes undoubtedly do exist. However, it does presume that observed age differences are influenced by factors other than those defining memory abilities per se. It argues that peoples' perceptions of their own memory may be important factors as well.

While preliminary work is available to support the basic idea that older adults' knowledge, beliefs, and affects with respect to memory-demanding situations are related to actual performance in these situations (e.g., Bruce, Coyne, & Botwinick, 1980; Dixon & Hultsch, 1983a; Zelinski, et al., 1980), several notable problems remain. First, one difficulty has been the measurement of individuals' knowledge and perceptions of their own memory. There has been confusion with regard to the scope and content of the construct, and many instruments and procedures designed to operationalize it have neglected issues of reliability and validity altogether (Cavanaugh & Perlmutter, 1981). Second, efforts to relate perceptions of memory to actual memory performance have met with some success (Dixon & Hultsch, 1983a; Zelinski, et al., 1980). However, the meaning of such relationships is unclear. For example, the correlation between perceived incidence of memory failures and performance in the elderly may reflect a veridical understanding of reduced memory capacities, or inaccurate perceptions (over- or under-estimates) of memory capacities that may influence performance because of their effect on effort or other variables.

Our own work in this area began with the development of a questionnaire designed to measure adults perceptions of their everyday memory functioning

(Dixon & Hultsch, 1983b, 1984). Originally, eight theoretically meaningful dimensions were defined and operationalized as summarized in Table 1. After content validity was established for a pool of 206 items, the instrument was administered sequentially to three separate samples of adults. Computation of internal consistency estimates and factorial validity resulted in a 120-item instrument. As summarized in Table 2, at least six of the subscales (Strategy,

---

Insert Tables 1 and 2 about here

Task, Change, Anxiety, Achievement, and Locus) appeared to be acceptably reliable and factorially valid. The Capacity subscale was internally consistent, but correlated highly with the Change subscale. Finally, the Activity subscale exhibited fairly modest levels of internal consistency . resulting concern about its adequacy as a subscale. We have since questioned the inclusion of the Activity subscale in the questionnaire for conceptual reasons as well. The other subscales all appear to index adults' knowledge and perceptions of their own memory. The items of the Activity subscale, on the other hand, index the nature and frequency of everyday activities that might support maintenance of effective memory functioning. We now suggest that such activities are representative of a different construct that might be labeled cognitive life style.

In addition to the Metamemory in Adulthood instrument summarized above, a number of other questionnaires have been developed to measure everyday memory functioning (Herrmann, 1982). Unfortunately, many of these have been developed without sufficient attention to their psychometric properties. However, in addition to the MIA, the Metamemory Questionnaire (MQ) developed by Zelinski, Gilewski, & Thompson (1980), appears to be particularly useful for examining memory perceptions in adulthood. This 92-item questionnaire contains nine a priori scales including: General Rating, Reliance on Memory, Retrospective Functioning, Frequency of Forgetting, Frequency of Forgetting While Reading, Remembering Past Events, Seriousness, Mnemonics Usage, and Effort Made to Remember. These are summarized in Table 3. A shorter (64-item) form has also

----

Insert Table 3 about here

----

been developed (Gilewski, Zelinski, Schaie, & Thompson, 1983). Like the MIA, the questionnaires developed by Zelinski and Gilewski have acceptable psychometric properties. Both questionnaires measure strategy use, perceived change in memory, and frequency of forgetting. In addition, the MIA taps memory knowledge, the affective aspects of memory perceptions, whereas the MQ examines the demands on memory, memory for past events, seriousness of memory failures, and overall judgements of memory adequacy.

In sum, we have made some progress in developing reliable and factorially-precise self-report measures of adults' everyday memory functioning. However, several issues remain. First, it is critically important to understand whether the specific domain of memory self-perceptions can be differentiated from more general self-perceptions. For example, it has been well established that memory complaints occur with greater frequency among adults suffering depressed affect even though their actual memory performance is no worse than that of non-depressed adults (Zarit, 1982). Similarly, it is possible that perceptions of reduced memory competence may be solely a function of age-related declines in feelings of self-esteem and personal control (Langer, 1981). Similar arguments apply to other dimensions such as anxiety, and achievement. Second, it is possible that previous work may have overdimensionalized the memory perception construct. This possibility is raised by the relatively substantial correlations found among some of the subscales of the MIA and MQ. We have hypothesized that memory perceptions may be best described by three dimensions: memory knowledge (i.e., knowledge and frequency of strategy use; knowledge of the demand characteristics of situations requiring memory); perceived self-efficacy of memory (i.e., perceptions of changes in one's memory capabilities; beliefs about the modifiability of one's memory); and affects associated with memory demanding situations (i.e., feelings of anxiety, achievement, and depression associated with one's memory).

Are there age-related differences in adults' perceptions of their memory system? Our own research suggests that there are (Dixon & Hultsch, 1983b). In particular, we found age differences on the Task, Capacity, Change, and Locus subscales. As summarized in Table 4, these age differences suggest that older adults evince less general knowledge about memory tasks, and perceive themselves as less efficatious in memory demanding situations than younger adults. Consistent results were found by Zelinski et al. (1980) who reported that older adults indicated that they experience more memory failures than younger adults.

Insert Table 4 about here

What is the relationship between individual's knowledge, beliefs, and affects about their memory and actual memory performance? Again, research suggests that such relationships exist, and that they may differ for different age groups. Table 4 also summarizes the results of regression analyses reported by Dixon and Hultsch (1983a). It appeared that Task and Strategy were the best overall predictors of performance for all adults in all samples. Certain age differences were evident however. Strategy (followed by Capacity and Task) were the best predictors of memory performance for the young. In contrast, Task (followed by Achievement and Locus) were the most important predictors for the old. Thus, younger adults' performance is predicted by what is known about retrieval strategies and physical reminders, what is believed about their capacity to perform on given tasks, and what is known about memory tasks and processes in general. Older adults' performance is predicted by what is known about memory tasks and processes in general, their level of motivation to achieve in memory demanding situations, and their belief in the degree of control they exercise over their memory functioning. These results suggest the possibility that the performance of older adults is more related to their feelings and beliefs about their memory than is the case for younger adults.

This possibility is important because it has become increasingly clear that individuals' self-reports of their memory capacities do not necessarily predict performance accurately. For example, Sunderland, Harris, and Baddeley (1983) recently examined the relationship of subjects' report of everyday memory problems and performance on several laboratory tasks in people suffering from severe closed head injury. Ratings of the patients' memory problems were obtained from the patients themselves and from their relatives. Patients' questionnaire responses showed only weak correlations with performance measures, whereas relatives' assessments showed strong relationships. Thus, patients with poor memories may be poor at recalling instances of memory failure resulting in an overestimation of their memory abilities.

Results such as these have led Herrmann (1982) to suggest that self-report questionnaires are not valid indicators of everyday memory functioning. However, other explanations of the lack of significant correlations between self-perceived memory and performance are possible. For example, it may be that self reports of memory capabilities do not predict performance on some laboratory tasks because the latter do not tap the skills evaluated on the former. This possibility is suggested by the finding that most of the correlations between self-perceptions and memory and performance have been obtained with fairly familiar tasks such as memory for meaningful texts. In contrast, when less familiar verbatim list tasks are used, the correlations are generally nonsignificant (Sunderland et al., 1983; Zelinski et al., 1980). Thus, memory self-perceptions may reflect accurate assessments of capability within a fairly narrow range of ecologically relevant memory situations.

Similarly, we have noted that memory perceptions, particularly those associated with memory demanding situations may be better predictors of performance for older adults than for younger adults (Dixon & Hultsch, 1983a). This suggests the possibility that what people believe and feel about their memory abilities may be as important as their actual memory abilities. Two possibilities are apparent. On the one hand, older adults may be accurately describing their deteriorating memory abilities that determine their poor performance. On the other hand, older adults' erroneous perceptions of declining memory may result in anxiety, feelings of loss of control, and decreased effort in memory demanding situations, which in turn, may produce poor performance. It is clear that longitudinal data will be required to determine whether changes in memory ability lead to changes in memory perceptions or vice versa.

Measures of intraindividual change in memory perceptions are important for another reason. We have generally approached the measurement of memory perceptions as though they were trait-like characteristics of individuals. That is, we have assumed they represent judgements that are stable across time. However, memory perceptions may be more state-like than trait-like (Dixon, 1985). The presence of such intraindividual variability, however, does not mean that the construct is useless. If such variability is systematically linked to other labile variables such as health, mood states, and energy levels, then it is possible that self-perceived memory may be an important antecedent predicting shifts in memory performance over time.

In sum, there is evidence accumulating to suggest that memory perceptions are important for understanding memory functioning in later life. In general, memory perceptions appear to be multidimensional and may be usefully conceptualized as containing belief and affective components as well as knowledge components. In addition, the relationship between memory perceptions and memory performance may differ for different types of tasks and for different age groups. Many questions remain, however. It is unclear to what extent older adults' self-reports of their memory knowledge, beliefs, and affects represents an accurate picture of their memory abilities. Similarly, we do not know to what extent individuals' memory perceptions show systematically state-like fluctuations over time. The perspective suggested here is that older adults memory perceptions may operate in fairly complex fashion. However, it may also be the case that such perceptions constitute an important class of variables that must be measured in order to understand adult age changes in memory functioning.

References

Baltes, P.B., Dittmann-Kohli, F., & Dixon, R.A. (1984). New perspectives on the development of intelligence in adulthood: Toward a dual-process conception and a model of selective optimization with compensation. In P.B. Baltes & O.G. Brim, Jr. (Eds.), Life-span development and behavior (vol. 6, pp. 33-76). New York: Academic Press.

Bruce, P.R., Coyne, A.C., & Botwinick, J. (1980). Adult age differences in metamemory. Journal of Gerontology, 37, 354-357.

Cavanaugh, J.C., & Perlmutter, M. (1982). Metamemory: A critical examination. Child Development, 53, 11-28.

Dixon, R.A. (1985). Metamemory and aging: issues of structure and function. Paper presented at Third George A. Talland Memorial Conference on Memory and Aging, Cape Cod.

Dixon R.A., & Hertzog, C. (1984). A functional approach to memory and metamemory development in adulthood. Paper presented at Conference on Memory Development and the Life Span, Munich, West Germany.

Dixon, R.A., & Hultsch, D.F. (1983a). Metamemory and memory for text relationships in adulthood: A cross-validation study. Journal of Gerontology, 38, 689-694.

Dixon, R.A., & Hultsch, D.F. (1983b). Structure and development of metamemory in adulthood. Journal of Gerontology, 38, 682-688.

Dixon, R.A., & Hultsch, D.F. (1984). The metamemory in adulthood (MIA) instrument. Psychological Documents, 14, 3.

Flavell, J.H. (1971). First discussant's comments: What is memory development the development of? Human Development, 14, 272-278.

Flavell, J.H., & Wellman, H.M. (1977). Metamemory. In R.V. Kail, Jr. & J.W. Hagen (Eds.), Perspectives on the development of memory and cognition. (pp. 3-34). Hillsdale N.J.: Lawrence Erlbaum.

Gilewski, M.J., Zelinski, E.M., Schaie, K.W., & Thompson, L.W. (1983). Abbreviating the metamemory questionnaire: Factor structure and norms for adults. Paper presented at the 91st Annual Meeting of the American Psychological Association, Anaheim, CA.

Herrmann, D.J. (1982). Know thy memory: The use of questionnaires to assess and study memory. Psychological Bulletin, 92, 434-452.

Hulicka, I.M. (1982). Memory functioning in late adulthood. In F.I.M. Craik and S. Trehub (Eds.), Aging and cognitive processes. (pp. 331-351). New York: Plenum.

Hultsch, D.F., & Pentz, C.A. (1980). Encoding, storage, and retrieval in adult memory: The role of model assumptions. In L.W. Poon, J.L. Fozard, L.S. Cermak, D. Arenberg, & L.W. Thompson (Eds.), New directions in memory and aging: Proceedings of the George A. Talland Memorial Conference. (pp. 73-94). Hillsdale, N.J.: Lawrence Erlbaum.

Langer, E.J. (1981). Old age: An artifact? In J. McGaugh & S. Kiesler (Eds.), Aging: Biology and behavior. (pp. 255-282). New York: Academic Press.

Sunderland, A., Harris, J.E., & Baddeley, A.D. (1983). Do laboratory tests predict everyday memory? A neuropsychological study. Journal of Verbal Learning and Verbal Behavior, 22, 341-357.

Zarit, S.H. (1982). Affective correlates of self-report about memory of older people. International Journal of Behavioral Geriatrics, 1, 25-34.

Zelinski, E.M., Gilewski, M.J., & Thompson, L.W. (1980). Do laboratory t sts relate to self-assessments of memory ability in the young and old? In L.W. Poon, J.L. Fozard, L.S. Cermak, D. Arenberg, & L.W. Thompson (Eds.), New directions in memory and aging: Proceedings of the George A. Talland Memorial Conference. (pp. 519-544). Hillsdale, N.J.: Lawrence Erlbaum.

Table 1

The Eight Dimensions of the Metamemory in Adulthood (MIA) Instrument

Dimension	Description	Sample Item
1. Strategy	Knowledge and use of information about one's remembering abilities such that performance in given instances is potentially improved (+ = high use)	Do you write appointments on a calendar to help you remember them?
2. Task	Knowledge of basic memory processes, especially as evidenced by how most people perform. (+ = high knowledge)	For most people, facts that are interesting are easier to remember than facts that are not.
3. Capacity	Perception of memory capacities as evidenced by predictive report of performance on given tasks. (+ = high capacity)	I am good at remembering names.
4. Change	Perception of memory abilities as generally stable or subject to long-term decline. (+ = stability)	The older I get the harder it is to remember things clearly.
5. Activity	Regularity with which respondent seeks and engages in activities that might support cognitive performance. (+ = high regularity)	How often do you read newspapers?
6. Anxiety	Feelings of stress related to memory performance. (+ = high anxiety)	I do not get flustered when I am put on the the spot to remember new things.
7. Achievement	Perceived importance of having a good memory and performing well on memory tasks. (+ = high achievement)	It is important that I am very accurate when remembering names of people.
8. Locus	Perceived personal control over remembering abilities. (+ = internality)	Even if I work on it my memory ability will go downhill.

Based on Dixon and Hultsch (1983b)

Table 2

Summary of Internal Consistency (Cronbach's alpha)
and Item Factor Loadings for Eight Metamemory in Adulthood Subscales

Subscale	Range of alphas Across 3 Samples	Factor Loadings Across 3 Samples
Strategy	.78 to .90	.32 to .75
Task	.74 to .87	.34 to .76
Capacity	.74 to .90	.30 to .67
Change	.82 to .92	.38 to .83
Activity	.28 to .76	.32 to .55
Anxiety	.78 to .87	.33 to .73
Achievement	.61 to .84	.30 to .64
Locus	.71 to .80	.31 to .67

Based on Dixon and Hultsch (1983b)

Table 3

Sample Items from the Metamemory Questionnaire (MQ) Instrument

Scale	Sample Item
1. General rating	1. How would you rate your memory in terms of the kinds of problems you have?
2. Reliance on memory	2. How often do you need to rely on your memory without the use of remembering techniques, such as making lists, when you are engaged in ...(a) social activities?
3. Retrospective functioning	3. How is your memory compared to what it was ... (b) one year ago?
4. Frequency of forgetting	4. How often do these present a memory problem for you ...(a) names?
5. Frequency of forgetting when reading	5. As you are reading a novel, how often do you have trouble remembering what you have read ...(a) in opening chapters once you have finished the book?
6. Remembering past events	6. How well do you remember things which occurred ...(a) last month?
7. Seriousness	7. When you actually forget in these situations, how serious of a problem do you consider the memory failure to be ...(a) names.
3. Mnemonics	9. How often do you use these techniques to remind yourself about things ...(a) keep an appointment book.
). Efforts made to remember	10. How much effort do you usually have to make to remember in these situations ...(a) names.

ote: These sample items are also on the shortened Memory Functioning Questionnaire MFQ), except for the Reliance and Effort items. These scales were dropped from the MFQ.

ased on Zelinski et al. (1980)

Table 4

Summary of Mean Age Difference and Prediction of Text Recall Performance for Eight Metamemory in Adulthood Subscales

Subscales	Age Differences	Predicts Performance for
Strategy	n.s.	Young
Task	Young > Old (Samples 1, 3)	Young and Old
Capacity	Young > Old (Samples 1, 2, 3)	Young
Change	Young > Old (Samples 1, 2, 3)	Young
Activity	n.s.	--
Anxiety	n.s.	--
Achievement	n.s.	Old
Locus	Young > Old (Samples, 1, 2, 3)	Old
Total Scale	Young > Old (Samples 1, 2, 3)	Young and Old

Based on Dixon and Hultsch (1983a)

Applications of Structural Equation Models

in Gerontological Research

Christopher Hertzog

Georgia Institute of Technology

Chapter to appear in:
K. W. Schaie (Ed.) Annual Review of Gerontology and Geriatrics (Volume 7).

Final revision, 10/30/86

Address Correspondence To:

Christopher Hertzog
School of Psychology
Georgia Institute of Technology
Atlanta, GA 30332-0170

Running Head: SEM Applications

Note: Preparation of this manuscript was supported in part by a Research Career Development Award (1 K04 AG00335-02) from the National Institute on Aging. I thank Tim Salthouse for helpful comments on an earlier draft of this review.

Abstract

This chapter reviews recent empirical research using structural equation models (SEM) to study aging and age-related phenomena. The emphasis in the chapter is on a conceptual and substantive treatment of SEM research, minimizing the mathematical and technical nuances that make these methods difficult for gerontological researchers to comprehend. The review begins by discussing an example of an SEM measurement model (essentially, a confirmatory factor analysis) that illustrates basic features of SEM while discussing the substantive issue of individual differences in information processing speed. The chapter then reviews in greater detail other SEM applications in gerontology, especially, cross-sectional and longitudinal SEM models that have been used to investigate age differences and age changes in intellectual abilities. The utility of SEM models for examining factorial invariance and for measuring and predicting individual differences in age-related change is illustrated by review of empirical findings in the domains of intelligence and personality. SEM research indicates that psychometric ability factors have invariant relationships to psychometric tests (equivalent factor patterns) with advancing age, but that correlations between factors increase. Longitudinal SEM results indicate substantial stability of individual differences in psychometric intelligence and personality in adulthood.

Introduction

The past several years have been marked by an accelerating rate of increase in sophisticated new methods for conducting valid and informative empirical research on nonexperimental data (e.g., Blalock, 1985a,b; Nesselroade & Baltes, 1979). Some of the more important advances have been in the domain of structural equation models (SEM). Traditionally, SEM usually refers to complex regression models (e.g., path analysis) that analyze causal relations among unobserved (latent) variables. An important component of SEM, therefore, is that part of the model that maps the latent variables onto variables we actually measure empirically (the observed or manifest variables). This part of SEM is usually termed the measurement model. The SEM measurement model is, essentially, a confirmatory factor analysis in which the observed variables are specified to be a linear combination of latent variables (factors). The part of the SEM specifying regression relationships among latent variables is the structural regression model.

In this paper I describe SEM applications, often consisting only of confirmatory factor analyses without a structural regression model, that address research questions of critical importance to gerontologists. Most of these applications are in the domain of psychometric intelligence and cognition, but they illustrate SEM techniques that can be used in other domains as well. This review avoids equations, derivations, or proofs, and does not discuss extensively the philosophy and methodological rationale for SEM in developmental research (see

Campbell & Mutran, 1982; Hertzog, 1985b, in press; Horn & McArdle, 1980; Nesselroade & Baltes, 1984; Rogosa, 1979; Schaie & Hertzog, 1982, 1985). Here the style is to emphasize concepts, not mathematics, and the goal is to discuss substantive implications as well as methodological advances.

Individual Differences in Information Processing Speed

One of the best documented findings in the gerontological literature is the age-related slowing in the speed of information processing (e.g., Salthouse, 1985). As humans grow older, the speed of elementary and complex cognitive processes slows. This slowing is independent of age changes in the peripheral nervous system and in psychomotor movement time, and has been argued to be a function of a primary aging process in the central nervous system itself (Birren, 1965). Most of the gerontological research has focused on mean age differences on tasks thought to assess different domains of information processing speed. There is little question that, on average, older persons demonstrate slower processing rates of information, even though the process of information analysis is usually found to be qualitatively similar in young and old persons (e.g., Petros, Zehr, & Chabot 1983).

The studies of the age-related slowing phenomenon have typically not examined issues of individual differences in the rate of slowing. This omission is curious, for even though primary aging, by definition, affects all individuals (Birren, 1965), the rate of aging may differ across individuals. Thus, even if one adopts the harsh view of age-related change as consisting solely of decrements caused by primary aging, accurate measurement of individual differences in rates of aging is important for scientific knowledge and applied problems (e.g., assessing pilot fitness after age 55). Assessment of the individuals must take into account the variation in aging rates, not just the average performance of individuals at a given chronological age.

To illustrate the features and advantages of an SEM measurement model, we will consider a study designed to measure individual differences in information processing speed (Hertzog, Raskind, & Cannon, 1986). Data were collected on reaction time (RT) measures of elementary verbal and nonverbal processes. The study used three RT tasks that had been used in the gerontological literature to study age differences in how quickly individuals accessed meaning of familiar nouns stored in semantic memory. The three tasks were: (1) Category Matching: subjects match a category noun label with an instance that is or is not a member of the category (e.g., FRUIT-APPLE vs. FRUIT-CHAIR); (2) Semantic Matching: subjects match two nouns that are or are not members of the same category (e.g., APPLE-PEAR vs. APPLE-CHAIR); and (3) Synonym Matching: subjects match two nouns that do or do not have the same meaning (e.g., THIEF-BURGLAR vs. THIEF-DANCER). Both Category Matching and Semantic Matching had two levels of nouns -- those of high or low typicality as instances of the category (e.g., APPLE is a high typicality fruit, whereas KIWI is a low typicality fruit). For all these variables, the score for an individual was his/her median RT for all correct judgements on Same (matching) trials. Data were collected on 55 persons (30 old, 25 young).

The purpose was to address several questions about these three RT tasks.

correlations among factors can be considerably higher than the correlations among any pair of observed variables. Disattenuation for measurement error is especially important in comparisons of different age groups. In SEM analysis of multiple groups, the estimates of group differences in variance and covariances among latent variables are not influenced by group differences in reliability, and differences in correlations among variables are not influenced by relative differences in reliability.

A third reason for conducting SEM is that the approach allows one to objectively evaluate the adequacy of a theoretical model, in terms of its fit to a set of variables (as predicted by theory). In addition to general indices of model fit to an entire data matrix, it is possible to formulate and test specific hypotheses regarding patterns of correlations (e.g., age group equivalence of factor correlations) and other SEM model parameters.

Figure 1 shows a factor model for the three semantic tasks, plus two other tasks (Simple RT and a Two-choice RT). In the simple RT, individuals pressed a button whenever a symbol appeared on the screen. In the two-choice task, subjects either pressed a button with the left hand if a left arrow appeared or pressed a button with the right hand if a right arrow appeared. Three latent variables were modeled (Simple RT, Two-Choice RT, and Semantic RT). These latent variables are depicted by the large circles. Their covariances are graphed as curved arrows ( ). Note that each latent variable (circle) has an arrow from it to the rectangles (labelled CATHI, CATLO, etc.). These rectangles represent each of the observed variables (RT tasks). The SMA factor has arrows pointing to the high and low typicality variables for Category Matching (CATHI, CATLO), the high and low typicality variables for Semantic Matching (SEMHI, SEMLO), and Synonym matching (SYN). These arrows represent regression coefficients for equations describing each observed variable as a weighted function of the latent variables. In standard factor analysis terms, they are factor pattern weights, or factor loadings. In our case, each observed variable is determined by only one latent variable, although this need not be the case. Note that each observed variable also has another circle pointing toward it. These circles represent unique components specific to the variable itself. The variances of these unique components include variance due to unreliable measurement error, as well as reliable variance not in common with other indicators of the latent variables. In the case of the Category Matching and Semantic Matching tasks, the model assumes that the high and low typicality variables will have a component specific to the type of matching task itself, and shared between the high and low typicality conditions. This assumption leads to specifying a residual covariance ( ) between these pairs of residuals. Thus, this model illustrates the three ways that observed variables can correlate with each other in SEM measurement models: they can correlate because (1) they are measures of the same latent variables, (2) they are measures of different latent variables that are correlated with each other, or (3) they have residual components (with respect to the latent variables actually in the model) that correlate with each other. What SEM models do is to attempt to estimate the values of the parameters in the SEM equations -- i.e., the regression coefficients, covariances, and variances in the model -- by using the sample variances and covariances among the observed variables.

Before evaluating the results of the model, a comment on the values of the regression coefficients (factor loadings) is in order. Note that several have a fixed value of 1.0 by hypothesis. Others (denoted by) are not fixed to a value, but instead are freely estimated by the computer program. Note also that the loadings for SEMHI and SEMLO have the same subscript (indicating that these are constrained equal to each other). These factor pattern coefficients can be quite confusing to someone familiar with standard factor analysis. Since the model is designed for explaining the variances and covariances among the observed variables, a single 1.0 loading is needed as an arbitrary constant that defines the units of measurement on the latent variable. This makes it possible to measure <u>the variances of the latent variables</u> themselves. The fact that more than 1 factor loading is fixed to 1.0 for each latent variable implies the additional assumption that these measures have equal relationships to the latent variables (as do SEMHI and SEMLO to SMA).

We estimated the model specified in Figure 1 by fitting it to the sample covariance matrix using the LISREL VI program (Joreskog & Sorbom, 1984). What did the analysis tell us about the research questions identified above? First, there are reliable individual differences in the semantic memory access speed (SMA) factor. The analysis provided us with three different ways of testing this conclusion. First, the overall fit of the model, as judged by its (likelihood ratio) X^2 test, was good. So too, was the LISREL adjusted goodness-of-fit index, which assesses goodness of fit in a manner less sensitive to sample size than the

X^2 test. So, as a whole, the model fit the entire data matrix well. The model also fit the sample covariances among the semantic RT tasks well. This was judged by examining the residual correlations among the items, removing from them the parts predicted by the model. These residual correlations were small, indicating that the covariances among the semantic RT tasks were well fit by the model. Finally, the estimated factor loadings for the SMA factor were large. After standardization, all the factor loadings for the SMA factor are very high, exceeding .9. The corresponding (standardized) unique variances were all less than .2. The analysis therefore indicated that the SMA factor is well defined -- and we can conclude that these RT tasks all are indeed measures of the same latent variable.

The third measurement issue relates to the equivalence of the RT measures between the young and the old age groups. Do they measure the SMA factor equivalently in the two age groups? We addressed this issue by estimating the model in the two age groups, while testing whether the regression coefficients (factor loadings) of Figure 1 could be constrained equal over the two groups. Forcing the groups to have equal factor loadings on the SMA factor did not result in a statistically significant increase in X^2. We concluded that the SMA factor is equivalently defined in the young and in the old adults.

Is there evidence of individual differences in the rate of slowing, as measured by RT tasks? Longitudinal data is required to answer this question definitively. But the cross-sectional design can give us some indirect evidence. Individual differences in rates of age-related change imply (1) increasing latent variable variances, and (2) <u>increasing latent variable covariances on variables that share common causes of change</u> (Hertzog, 1985a). Hertzog et al. (1986) found the estimated factor variance for SMA to be higher in the old than in the young. Furthermore, the correlation among two-choice RT and SMA was higher in the old group (.50) than in the young group (.31). Conversely, the correlation of simple RT and SMA was lower in the old group (.35 versus .45). This pattern of results is consistent with the hypothesis that the nonverbal and semantic two-choice RT tasks share a common cause of change from young adulthood to old age. Other explanations are also consistent with the results, of course, and 1) large sample replication as well as 2) analysis of longitudinal change data would be needed to provide more definitive evidence favoring the hypothesis of correlated individual differences in change.

<u>SEM Models for Psychometric Intelligence in Adulthood</u>

One of the real success stories in the brief history of SEM applications in gerontological research involves research on whether the aging process produces qualitative shifts in factor structure of psychometric intelligence. Intelligence has been one of the most widely studied construct domains in gerontology, and considerable effort was expended addressing the issue of whether aging altered the factor structure of intellectual abilities. The question is exptremely important, for it is in some senses a necessary first step to meaningful analysis of age changes in intelligence. Age changes in factor structure of intelligence would call into question the meaning of quantitative changes in levels of intelligence, as well as the meaning of individual differences in patterns of age-related change in intelligence (e.g., Baltes & Nesselroade, 1973). As discussed in some detail by Baltes and Nesselroade (1973; see also Schaie & Hertzog, 1985), a lack of factorial invariance across developmental levels could be produced by changes in the measurement properties of the psychometric tests (a lack of <u>measurement equivalence</u>) or a change in the fundamental properties of the constructs themselves (a lack of <u>construct equivalence</u>). With respect to psychometric intelligence, a lack of measurement equivalence might indicate a shift in the relative importance of performance-related processes (e.g., perceptual analysis of form, selective attention) necessary for adequate performance on the tests but clearly distinct from the construct(s) the tests were originally intended to measure. On the other hand, a lack of construct equivalence across age levels might indicate that there is a developmental shift in the organization of cognitive processes and their application to solving psychometric test items.

Prior to the use of SEM techniques for studying factorial invariance, the results from studies of factorial invariance presented a rather confusing picture (Reinert, 1970). The dominant hypothesis regarding adult age changes in intellectual factor structure has been one of <u>dedifferentiation</u>, in which intellectual ability factors become more highly interrelated with advancing age. Dedifferentiation (or reintegration) stands in apposition to the hypothesis of differentiation of intellectual factor structure in development from early childhood to adolescence. The differentiation hypothesis holds that abilities are first manifested as one or a very few abilities (such as general intelligence), but with increasing development, multiple intellectual abilities emerge.

Dedifferentiation implies a tendency to return to a factor structure like that of early childhood -- one or a very few ability factors. Extreme versions of dedifferentiation suggest actual factor collapse -- reductions in the number of factors needed to account for ability factors. Milder forms of differentiation would include (1) shifts in the factor pattern weights, (2) increases in communalities, and, in particular, (3) substantial increases in the correlations among ability factors. Certainly, part of the confusion that existed regarding the phenomenon of dedifferentiation centered around variable criteria for dedifferentiation (Olsson & Bergman, 1977). Perhaps the weakest form of dedifferentiation would consist of invariant factor pattern weights accompanied by increased factor correlations. Such a result would suggest tht the factors maintained invariant relationships to intelligence tests, but that there were some shifts in the relationships among the abilities themselves.

Tests of dedifferentiation hypotheses have often been examined by means of comparative factor analysis of cross-sectionally defined age groups, since longitudinal data appropriate to the issue have been relatively rare. Adequate tests of factorial invariance in multiple age groups requires use of simultaneous confirmatory factor analysis of covariance matrices from multiple age groups (Cunningham, 1978; Hertzog, 1985b; Schaie & Hertzog, 1985). The most crucial evidence for equivalence of factors is the test of equivalence in unstandardized factor pattern weights across the multiple groups. The critical problems with the early literature were that (1) the primacy of the test of factor pattern weights was not clearly understood, and (2) exploratory factor analysis often led to inconsistent results as a function of the fundamental indeterminacy of exploratory factor analysis solutions (Cunningham, 1978; Reinert, 1970).

The literature using confirmatory factor analysis has converged upon a common set of findings: (1) the number of factors present in different adult age groups seems remarkably consistent over the adult life span; (2) there are no major changes in which tests load on which ability factors (in Horn, McArdle, & Mason's, 1984, term, configural invariance); (3) the unstandardized factor pattern weights appear to be numerically equivalent in different age groups (metric invariance -- Horn et al. 1984); but (4) ability factors correlate more highly in old than in young or middle-aged populations, and (5) communalities appear to be higher in older groups. The first evidence for this conclusion came from studies conducted by Cunningham (1980, 1981), and an impressive number of other studies have produced similar findings (e.g., Hertzog & Schaie, 1986a; Hultsch, Hertzog, & Dixon, 1984; Stricker & Rock, 1985). Cunningham's studies are based upon comparisons of young subjects (15-32 years of age) with a sample of over 300 adults, ages 53 to 91. In both studies this adult sample is divided into young-old (53-68) and old-old (69-91) age groups. Adults were administered a battery of tests from Guilford and the ETS Reference Kit measuring multiple abilties. Cunningham (1980) compared three primary abilities (Verbal Comprehension, Sensitivity to Problems, and Semantic Redefinition). Cunningham (1981) did age-comparative factor analysis on 14 ability tests defining five primary abilities (Verbal Comprehension, Number Facility, Perceptual Speed, Symbolic Cognition, and Flexibility of Closure). After complex and somewhat unorthodox analyses, Cunningham (1980, 1981) found equal numbers of factors and invariance in unstandardized factor loadings across the three age groups. However, he found convincing evidence that factor covariances are greater in the adult groups than in the young sample. Similar data are reported by Stricker and Rock (1985), who applied SEM to analyze data from three age groups (20-29, 30-39, and 40-49) on the Graduate Record Examination. They drew random samples of 1,000 individuals from each age group, and performed separate SEM models on each age group. They did not explicitly test factor invariance by constraining parameters equal (probably, a decision based upon the pragmatic problem of too many variables, and hence, too many parameters to estimate given three age groups). The degree of equivalence in the separately estimated factor loadings on the GRE Verbal, Quantitative, and Analytical factors is startling. The standardized loadings generally differed by less than .05 across the three age groups. Factor correlations increased across the age-groups (without concomitant increases in standard deviations of the observed variables). Given the sample sizes, inferential statistics would not be needed to detect these differences as significant! Nevertheless, the differences are relatively substantial, given that all subjects sampled were less than 50 years of age. For example, the correlation between the GRE Verbal and GRE Quantitative factors increased from .53 in the 20-29 age group to .63 in the 40-49 age group. Data by the other studies cited above conform to this pattern -- equivalence in factor loadings, but age differences in factor covariances and communalities.

Any cross-sectional age group differences in factor structure are confounded with cohort differences (in analogy to the age/cohort problem in mean levels). Cunningham and Birren (1980) reported no time-lag differences in factor structure (factor loadings or factor covariance matrices) between student samples of widely different cohorts (but see below for some concerns regarding their SEM approach). Hayslip and Brookshire (1985) recently reported a time-lag comparison of two adjacent cohorts of older adults, in which they found invariant loadings on two factors they labelled as fluid and crystallized intelligence. It seems at this juncture that there are neither age differences nor cohort differences in factor loadings. There appears to be age-related changes in factor covariances that are much greater than the (possibly nonexistent) cohort differences in factor covariances.

Baltes, Cornelius, Spiro, Nesselroade, and Willis (1980) reported one SEM study that may be viewed as favorable toward a stronger form of dedifferentiation in adult intelligence factor structure. They conducted a factor analysis of a battery of 17 ability tests selected to mark second-order factors of fluid and crystallized intelligence. They did not do age comparative analysis; the sample consisted of 109 elderly adults (ages 60-89). Baltes et al. (1980) found substantial correlations among the ability tests in the elderly sample. A seven-factor model for the tests based upon an a priori primary ability factor structure fit the data well but produced high factor correlations, especially among the Induction, Figural Relations, and Experiental Evaluation primaries. The high relationship of Experiental Evaluation to Induction was surprising because the former had been conceptualized as a marker for the second-order factor of

Crystallized Intelligence, while the latter had been selected to measure Fluid Intelligence (e.g., Horn, 1978). Horn and Cattell have consistently argued that Fluid and Crystallized Abilities are positivley (but modestly) correlated. Baltes et al. (1980) subsequently fit models with fewer factors, and ultimately argued for a model that included a general factor (with largest loadings associated with the tests of reasoning ability), and three additional factors (Memory Span, Verbal Comprehension, and Perceptual Speed). They specifically argued that a Horn/Cattell fluid/crystallized model could not fit the data for their elderly sample (as it has for data in younger populations). Accordingly, they suggested that their results supported a dedifferentiation of factor structure; specifically, a collapse of the fluid and crystallized distinction into a general factor, and a collapse of the primary abilities determinined by fluid intelligence into a single factor highly related to general intelligence.

Baltes et al. (1980) results actually are ambiguous regarding the merits of the strong dedifferentiation hypothesis. Without age comparative analysis, it is difficult to know the degree to which the battery they administered would also produce high intercorrelations among abilities in different age samples. Their comparisons to younger age groups are based upon interpretations of research reports by Horn and Cattell (e.g., Horn, 1978), but these solutions were based upon exploratory factor analysis allowing small loadings of all variables on all factors. This approach tends to estimate lower factor correlations than SEMs with fixed zero loadings. Gustaffson (1984) recently reported data showing substantial correlations among primary abilities in young adults -- correlations much higher than one would expect given exploratory factor analysis results on fluid and crystallized intelligence. Moreover, Gustaffson (1984) argued that his general intelligence factor is isomorphic with fluid intelligence (a position consonant with Baltes et al. 1980). For now, it seems safest to conclude that the results of Baltes et al. (1980) support at least the mild dedifferentiation hypothesis -- i.e., increasing factor correlations.

Cornelius, Willis, Nesselroade, and Baltes (1983) reported an interesting analysis of additional data from the same sample, tested two years after the data analyzed by Baltes et al. (1980) were collected. In this data collection, additional measures of attention were administered. SEM was used to perform an extension analysis. Four factors (Reasoning, Crystallized Knowledge, Memory Span, and Perceptual Speed) were specified as oblique ability factors; the attention tests were then extended into this factor space by allowing free factor loadings for all of them on the four factors. A Continuous Paired Associates Recall measure loaded strongly on the Reasoning factor; the others loaded predominantly on Perceptual Speed. No attempt was made to extend the variables into the general factor model reported by Baltes et al. (1980); nor was an attempt made to identify separate Attention factor(s) and estimate their correlation(s) with Perceptual Speed independent of the extension analysis. Their results support the hypothesis that these attentional measures are more highly related to Perceptual Speed than other ability factors. Cornelius et al. (1983) concluded, on the basis of their own results and other studies, that there is little support for the hypothesis that attentional deficits mediate age changes in fluid abilities such as

Reasoning.

Longitudinal SEM Models of Age Changes in Intelligence

One of the most important classes of application for SEM is in modeling change in longitudinal data. Prior to the advent of SEM techniques for longitudinal factor analysis, conducting latent variable analysis in longitudinal data was a heroic task (see Bentler, 1973). SEM approaches enable modeling of latent variables in longitudinal data sets without great difficulty, providing that there are sufficient measures of the latent variable available. SEM models for longitudinal data also deal in a straightforward fashion with specification problems unique to longitudinal data (e.g., autocorrelated residuals in the measurement model). Details on longitudinal SEM models and their technical properties may be found in multiple references (e.g., Dwyer, 1983; Hertzog, 1985b, in press; Horn & McArdle, 1980; Joreskog, 1979; Joreskog & Sorbom, 1977, 1980; McArdle, 1986; Rogosa, 1979, 1980; Schaie & Hertzog, 1985). The discussion here will focus on illustrating a few advantages of the method in the context of reviewing published work in gerontology.

An adequate description of longitudinal SEM models requires first that we describe and discuss the properties of the longitudinal measurement model. Assume that we have longitudinal data across a particular age range collected in multiple birth cohorts. At each measurement point, one or more latent variables has been measured with multiple observed variables. The longitudinal measurement model requires (1) that a model for each occasion be specified, and (2) that the model specify the relationships between latent and observed variables across longitudinal occasions. Figure 2 shows a simple longitudinal factor model, as developed by Hertzog and Schaie (1986a) for Schaie's longitudinal data on intelligence. At each occasion, a single latent variable (general intelligence, or _g_) is measured by five observed variables (subtests of the Thurstone Primary Mental Abilities [PMA] test). The _g_ latent variable relates to itself over time as a function of the covariances between _g_ across occasions. Thus, age changes are reflected in (1) changes in _g_ factor variances with increasing age, and (2) _g_ factor covariances over time (e.g., _g_ at Age 1 with _g_ at Age 2). We shall consider the interpretation of these parameters below. The model of Figure 2 does _not_ provide a means by which residuals for the five subtests (the in Figure 2) relate, independent of _g_. As such the model is badly misspecified, for there will usually be reliable components of observed variables that are not part of the common factors (latent variables). If these components are not specified and estimated, then the estimates of factor loadings and factor covariances for _g_ will be biased. Fortunately, SEM permits modeling covariances among the residuals. In the Hertzog and Schaie (1986a) analysis, the differences in fit between models with and without the residuals was enormous -- and plausible parameter estimates could only be obtained with the correlated residuals. The type of model graphed in Figure 2 can be easily generalized to more than one latent variable at each occasion (see examples in Dwyer, 1983, and Horn & McArdle, 1980).

The longitudinal SEM measurement model provides a different basis for evaluating the dedifferentiation hypothesis than did the simultaneous factor analysis in multiple (cross-sectional) age groups. In the longitudinal analysis, dedifferentiation would be reflected in (1) differences in the number of factors at each occasion, (2) changes in the within-occasion factor covariances across occasions (e.g., Verbal Comprehension correlates more highly with Induction at Age 50 then at Age 30), (3) shifts in the factor loadings across occasions, and/or (4) increases in communalities across occasions. Longitudinal investigations of the dedifferentiation hypothesis using SEM are consistent with multiple groups comparisons: there is little evidence for change in the number of dimensions, or in the factor loadings across occasions, ruling out the stronger forms of the dedifferentiation hypothesis.

Cunningham and Birren (1980) used longitudinal data collected by Owens (see Cunningham & Owens, 1983) to test the dedifferentiation hypothesis. Eight subtests from the Army Alpha were measured on 96 males as college students, and again at ages 50 and 60. The Army Alpha tests were used to define three ability factors. As mentioned above, Cunningham and Birren (1980) found that no time-lag differences in factor structure between Owens' original young subjects and a recent group of university students. On the other hand, they found longitudinal differences in factor structure between young adulthood and age 50. Cunningham and Birren's results are inadequate for the purpose of localizing the changes to factor loadings, factor covariances, or both, owing to the nature of their analysis. They report convergence problems with their solution in many different parts of the analysis, which may be a function of the instability of their three factor solution. Part of their problems may have been a function of the way the confirmatory analysis was treated. In a preliminary analysis of the pooled correlation matrix, Cunningham and Birren (1980) found an estimated factor correlation of .82 between two factors. Deeming this unacceptable, they fixed the correlation to .70 and re-estimated the model. This practice was adopted in subsequent comparisons of equivalence over longitudinal occasions, which were even more unusual in that the longitudinal occasions were modeled as if they were independent data from different age groups. They also reported fixing to zero factor correlations estimated to be negative. This was done in order to maintain some correspondence to expectations that ability factor correlations ought to be positive. This pattern of outcomes and fixes suggests that Cunningham and Birren (1980) imposed questionable restrictions on the common factor space in order to salvage plausible parameter estimates for a misspecified model. In fairness, these problems were probably brought about by the fact that the eight Army Alpha subtests apparently did not adequately define primary ability factors conforming to simple structure. Nevertheless, the procedures employed appear to cast doubt on the validity of the hypothesis tests for equivalence of factor pattern weights and factor covariance matrices. Moreover, a direct likelihood ratio X^2 test of equivalence was actually not computed. They evaluated the hypothesis of equivalence in factor loadings by assessing of the overall fit of models with constrained equal factor loadings instead of computing the difference in X^2 between models imposing and relaxing the constraints. Taking these problems together, it is difficult to accept at face value Cunningham and Birren's (1980) suggestion that there are changes in factor loadings with advancing age. At best, one can argue that their results suggest some kind of age-related shifts in factor structure occurred in the longitudinal data.

Evidence that factor loadings remain invariant in longitudinal data have been reported by Lachman (1983) and by Hertzog and Schaie (1986a). Lachman (1983) analyzed data from the Penn State ADEPT project, including intelligence data used in the studies by Baltes et al. (1980) and Cornelius et al. (1983). Lachman's research focus was on relationships of perceived intellectual competence (control) to intelligence in the elderly. As a result, she used two markers for each of four factors (Fluid Intelligence, Crystallized Intelligence, Perceptual Speed, and Memory Span). Like Cunningham and Birren (1980), Lachman did not directly test the hypothesis of invariance in factor loadings; however, her model constraining factor loadings to be equal fit the data well. Hertzog and Schaie (1986a) did test directly the hypothesis of longitudinal invariance in g factor loadings, defined in the model depicted in Figure 2. They found the hypothesis of invariance to fit well over fourteen years of longitudinal age change for young, middle-aged, and old groups.

The latter two studies cannot be considered definitive evidence for invariant relationships between intelligence tests and factors in longitudinal data. In the case of Lachman's (1983) study, the older subjects were retested after only a two-year interval, which may not be sufficient time to observe qualititative shifts in factor loadings. Hertzog and Schaie's (1986a) analysis is more convincing, in that changes were not observed over fourteen-year intervals, but their analysis estimated g factor loadings, not primary ability loadings, and there is still a possibility that changes would be observed at the primary ability level in appropriate data sets. Moreover, neither of these studies can rule out the possibility that qualitative change in factor structure occurs differentially in groups likely to drop out of longitudinal samples. In principle this argument is difficult to address, but it is highly plausible. Individuals in phases of terminal decline (Riegel & Riegel, 1972) may be those most likely to show qualitative shifts in factor structure at the same time they experience losses in level of functioning. Nevertheless, a definitive test of the hypothesis of structural invariance in subpopulations who do remain in longitudinal studies has not yet been made. The evidence available does support the invariance hypothesis (and therefore, only the weak version of the dedifferentiation hypothesis).

Schaie, Willis, Hertzog, & Schulenberg (1986) examined whether factor loadings would change as a function of training skills needed to perform on psychometric ability tests. Subjects were either 1) trained on inductive reasoning, 2) spatial rotation ability, or 3) treated as a no-training control. One criticism of training studies has been that the very act of training the skills needed for test performance may change the measurement properties of the tests. Schaie et al. (1986) used a pretest-posttest design to test the invariance of factor structure for older adults before and after they were trained. Factor loadings were invariant in the no-training control group, but there were some subtle shifts in factor loadings in the training group. These changes were specific to the ability tested. For both groups, configural invariance still held after training -- the tests loaded on the same factors. However, loadings of one test on each factor were different after training -- rejecting the hypothesis of complete metric invariance. These changes did not appear sufficient to warrant a conclusion that the factor was no longer measured accurately after training. Nevertheless, the results suggest that care must be taken in analyzing levels of training gain; if possible, analysis at the level of the latent variables (factors) would be preferable to analysis of single indicators.

The capability of SEM to provide direct tests of invariance of factor loadings is only one of its merits. Another crucially important feature of these models is that they afford examination of the stability of individual differences as individuals age. Stability has multiple definitions (e.g., Baltes, Reese, & Nesselroade, 1977, Bengtson, Reedy, & Gordon, 1985; Kagan, 1980; Schaie & Hertzog, 1985). Stability of individual differences is distinct from stability in mean levels over time; it reflects the degree to which individual differences are consistent across time, and is often operationally defined in terms of the correlation of a variable with itself in longitudinal data. The magnitude of such a correlation (or its regression equivalent, the stability coefficient; Kessler & Greenberg, 1981) is inversely related to the magnitude of what Baltes and his colleagues have labeled "interindividual differences in intraindividual change" (Baltes, Reese, & Nesselroade, 1977; Baltes & Nesselroade, 1973). The more the heterogeneity in patterns of change across individuals, the lower the stability coefficient. The greater the tendency for individual change patterns to run parallel to the average (mean) change pattern, the greater the stability coefficient. A high degree of stability is central to Costa and McCrae's (1980;

Costa, 1986) argument of continuity in personality factors with aging (and more generally, to arguments of psychosocial continuity [see also Bengtson et al, 1985]). As discussed in more detail by Schaie and Hertzog (1985), the Joreskog/Sorbom type longitudinal SEM approach allows one to explicitly study the stability in terms of the variances and covariances of the latent variables over time.

The available evidence strongly suggests that there is a high degree of stability in individual differences in intelligence during adulthood. For example, Schaie et al. (1986) found virtually perfect stability of individual differences in a one-month retest of older individuals who were members of their no-training control group (see above). Stability over a one month period is hardly surprising. But Lachman (1983) reported perfect stability of fluid and crystallized intelligence over a two-year period in elderly participants in the ADEPT studies. Evidence for longer-term stability has been found in the Seattle Longitudinal Study (SLS) by Hertzog and Schaie (1986a). They found stability of individual differences (high _g_ factor covariances) and increases in the overall magnitude of individual differences (increased _g_ factor variances) in middle-aged and old participants in the SLS. When standardized, the correlations among _g_ at different times of measurement were generally at the .9 level or higher. The high factor covariances indicated that adults were preserving to a remarkable degree their relative orderings about the _g_ factor mean over the fourteen year interval in all age groups.

It is possible that the stability observed by Hertzog and Schaie (1986a) holds only for g, but not for the primary abilities measured in the Thurstone PMA battery. Hertzog and Schaie could not test this hypothesis at the level of primary ability factors, given that the PMA has but one measure of each primary ability. Stability in individual differences in the primary abilities was reflected in significant residual (specific) covariances for the PMA subtests across longitudinal occasions. Additional evidence for stability of individual differences at the primary ability level was obtained with a model which added five test-specific factors to the occasion-specific factors for the groups (see Joreskog and Sorbom, 1977 for explication of similar models in single group designs). Figure 3 shows the hypothesized factor pattern matrix for this model. The test-specific factors were forced to be orthogonal to g. The test-specific model allowed Hertzog and Schaie (1986b) to separate the variance in each measure due to test-specific covariance over time by removing it from the residual component. The proportion of variance due to test-specific covariance over time was substantial. For example, 45% of the variance in the PMA subtest Verbal Meaning was unique in the model with residual covariances, but only 12% of its variance was unique when the test-specific Verbal Meaning factor is added. Moreover, the residual variances in the model with test-specific factors were just noticeably larger than one would predict from the published reliabilities of the tests. For instance, the PMA manual reports an estimated reliabiity of .92 for Verbal Meaning. Thus it appears that there is substantial stability in individual differences at the primary ability level as well.

Stability of individual differences should not be confused with changes in the PMA means over time. There are age-related changes in mean performance levels in the SLS, but individual differences are relatively consistent around the means. This point is underscored when a simultaneous analysis of means and covariance structures was conducted by Hertzog and Schaie (1986b). Although there was high stability of individual differences in all three age groups, the means show differential patterns of change with increasing age. In the old, g exhibited large declines over the fourteen-year period (mean ages 58 through 72). On the other hand, the data for the middle aged group (mean ages 42 through 56) could be modeled as stable in both level and covariance structure. Thus the results seem to indicate a pattern of relative stability of performance levels during the decades of the 40's and 50's, but a shift to a pattern of performance decline following age 60.

Thus the available evidence strongly suggests that individual differences in intelligence remain highly stable in adulthood, at least for those individuals sufficiently advantaged to participate in longitudinal studies such as the SLS. Further investigation of the stability of individual differences in primary abilities is needed, as are studies of the conditions under which differential change (either positive or negative) is likely to be observed. One circumstance in which differential change might be expected is subsequent to ability training. Individuals might differ in the degree of training benefit, perhaps as a function of prior history of age-related change. Schaie et al. (1986) tested this hypothesis by comparing stability of individual differences in their induction training group, spatial training group, and no-training control group. Stability of individual differences was high for both training groups and not appreciably lower than in the control group. There was some suggestion of greater individual differences in gain for the Spatial training group. It appears that the type of training program used benefits most if not all individuals, and that the variance in change is small relative to the total variance in ability. The advantage of testing the differential stability hypothesis using SEM is that it is estimated for latent variables, thus avoiding the nasty problem of regression to the mean in variables containing measurement error (see Nesselroade, Stigler, & Baltes, 1980).

Structural Regression Applications: Selected Examples

One of the major risks in SEM use is that individuals will assume that one is necessarily doing "causal" analysis. This assumption is often erroneous, for causal analysis is not merely (or even primarily) a function of using latent variable regression models (see Hertzog, in press; Mulaik, 1986). The stance taken here is that many, if not most SEM applications in our field are descriptive, and not explanatory, research. In fact, one can argue that SEM is an optimal method for fully informative and valid descriptive work (e.g., Hertzog, in press). Nevertheless, the following discussion of empirical SEM work will be noticeably devoid of causal terminology.

There are to date relatively few cases of full applications of SEM in gerontological research -- that is, use of both the measurement model and the structural model to specify and estimate a system of regression equations for latent variables. It is much more common to observe single indicator path analytic studies. An excellent example of this type of work is the study by Caspi and Elder (1986) on antecedents of life satisfaction for older women differing in social class levels. They used data from the Berkeley Guidance Study to predict life satisfaction in old age (mean age = 70) from psychological and social variables measured 40 years earlier. The covariance matrices differed by social class (middle-class versus working class). Interviewer rating of emotional health at age 30 (operational definition not fully specified) predicted subsequent life satisfaction for middle class women, but not working class women. Conversely, degree of social involvement (operationalized as the number of groups with which an individual has active involvement) predicted life satisfaction for working class women, but not middle class women. Caspi and Elder's (1986) study illustrates some of the benefits of path analytic approaches. The structural regression coefficients are attenuated by measurement error, but this is unavoidable in archival analysis where multiple indicators are not available for the constructs of interest.

The effects of measurement error can be profound. This point is illustrated in a study by Lair, Hertzog, and Schulenberg (1985), who used LISREL to examine the stability of individual differences in personality. The data were from the Duke Adaptation Study, in which over 300 adults were measured with Cattell's 16PF (Form C) scale over a six year period (four waves of measurement separated by two years). Siegler (1983) summarizes results from the Duke Adaptation Study, and reports test-retest correlations in the .5 range for the different subscales of the 16PF. Costa (1986), in reviewing the literature on stability of personality, argued that these data show almost perfect stability of individual differences because of the likely degree of attenuation due to measurement error. His argument was based in part on the fact that Form C of the 16PF is less reliable than the combined Forms A and B more commonly used. While it is likely that the attenuation is substantial, what is needed is a point estimate of the disattenuated stability of individual differences. A common practice is to correct for attenuation by using published estimates of reliability, or alternatively, to invoke assumptions such as used by Heise (1969) to arrive at an estimate of stability in longitudinal data. Neither approach is fully satisfactory; in particular, use of corrections for attenuation can overestimate stability if the reliability estimates are inaccurate for the subpopulation under study. It is preferable to estimate the disattenuated stability directly using SEM techniques.

Lair et al. (1985) did so for the two second-order factors of neuroticism and extraversion. Three indicators of each factor were selected (Outgoing, Happy-Go-Lucky, and Venturesome for Extraversion; Stable, Controlled, and Tense for Neuroticism). An SEM analysis was used to estimate stabilities using a first-order autoregressive model (e.g., Joreskog & Sorbom, 1977). Autoregressive models can be used to study predictors of individual differences in change at the level of the latent variables. The autoregressive coefficients reflect the degree of stability in individual differences. Additional variables that also predict a latent variable, controlling for autoregression are in essence predicting individual differences in change between the two time points (Kessler & Greenberg, 1981; but see Rogosa & Willett, 1985). In the Lair et al. analysis, no additional latent variables, other than Neuroticism and Extraversion were measured, so the focus was merely on a descriptive analysis of individual differences in stability. Interest, then, centered on the magnitude of the autoregressive coefficients and residual variances. Figure 4 shows the model results for one-half of the sample. Although the original model estimated metric (unstandardized) regression coefficients, standardized coefficients are shown in Figure 4. As can be seen, the disattenuated stability coefficients (the regression of N2 on N1, etc.) are uniformly high, and in some cases close to 1. The differences between the zero-order (simple) regression coefficients and the LISREL estimates of stability are profound. Figure 4 shows why this is so; the factor loadings for both measures are small, and measurement error has been absorbed into the residual variances. The model did estimate correlated residuals (not shown in Figure 4), so it is the case that there is reliable but specific variance in the residual terms. Nevertheless, it is clear that the results support Costa's (1986) contention that these data indicate substantial stability of individual differences in the personality factors of neuroticism and extraversion.

The Lair et al. (1985) analysis is in some senses the minimum structural regression model that might be contemplated. No predictors nor outcomes of the two personality variables were included in the analysis, which focused exclusively on the stability of individual differences. The only conclusion that can be drawn about predictors of change in such models occur only in the situation when individual differences are <u>perfectly</u> <u>stable</u> over time (i.e., there are no individual differences in change, at the level of the latent variable, to be predicted). With stabilities in the range reported by Lair et al. (1985), there is relatively little variance residual to autoregression that could be accounted for by other variables, but nevertheless, the major problem is finding the appropriate variables.

The analysis by Lachman (1983), discussed above in terms of the longitudinal factor model, also included a structural regression analysis. Recall that Lachman (1983) found that the stability of individual differences in fluid and crystallized intelligence were essentially perfect (factor correlations of 1.0) for older persons measured two years apart. This result indicates that no analysis of predictors of change in these abilities would be particularly useful (although predictors of initial level of ability could be analyzed). However, Lachman (1983) did find less than perfect stability (a factor correlation of .70) for the memory span factor. Given that the LISREL estimate of stability is disattenuated for measurement error, one can conclude that about half of the variance in memory span is independent of autoregression, and hence, is change-related variance available for prediction by other variables.

Lachman (1983) investigated the degree to which perceptions of personal control and control specific to the domain of intelligence predicted change in memory span over the two-year period. To do so, she assessed the degree to which latent variables from these two domains (e.g., intellectual self-efficacy, defined as beliefs in one's own competence in situations requiring intelligent behavior), measured at the start of the study (Occasion 1) predicted memory span at Occasion

2. Lachman freely estimated the correlations of these variables with memory span at Occasion 1, thus requiring that any prediction of memory span at Occasion 2 be independent of the Occasion 1 relationships. Lachman (1983) found that a model specifying only autoregression of memory span on itself fit reasonably well, and that the fit could not be improved by adding lagged regression coefficients of general and intelligence-specific control beliefs. Figure 5 reproduces the essential results. A model focused on measuring predictors of change in Intellectual Self Efficacy showed (1) the stability of that factor to be relatively low, and (2) changes in Intellectual Self-Efficacy to be predicted by general internal locus of control at Occasion 1. High internals are more likely to maintain or increase their perceived intellectual self-efficacy over time.

Liang (1986) recently reported a model for the determinants of self-rated physical health in adults. Liang (1986) used data from the 1968 National Senior Citizens survey to predict self-ratings of physical health from four other latent variables (chronic illness, sick days, self-maintenance, and instrumental activities). The recursive SEM model specified multiple relations among these endogenous variables as well. All possible recursive paths were estimated (so the model is just-identified in structural equations), in the order of variables just listed. That is, Liang (1986) postulated that chronic illness would influence sick days, chronic illness and sick days would influence self-maintenance, and so on. Self-rated health was defined as a latent variable deterining ratings such as "how good is your health," and "how good is your health compared to others your age." Chronic illness was measured by self-nomination of three problems characterized as "lasting or continuing." Liang (1986) used the approach of having eight domains of chronic illness (e.g., circulatory, respiratory) treated as exogenous variables that determined the latent variable of chronic illness. By fixing the residual variance of chronic illness to zero, Liang (1986) implicitly defined the chronic illness variable to be a linear composite of the eight health conditions, with weights determined (identified) by the relationship of chronic illness to the other latent variables. Liang's results suggest a substantial relationship between self-report chronic disease conditions and self-rated health. The direct effect of chronic illness on rated health averaged -.36 across four subsamples. Liang (1986) computed effects decomposition for one of the samples, in which the relationship of a "cause" on an "effect" variable is partitioned into a direct effect (the actual regression weight of effect on cause) and indirect effects (the relationship mediated through other causes). The total effect (direct + indirect) of chronic illness on self-rated health was -.53. In addition, Liang (1986) found that instrumental activities (e.g., driving, taking a trip, gardening) had a substantial impact on self-rated health independent of chronic illness. The model and results are intriguing, and open several possibilities with respect to the relationship of subjective health perceptions to other variables, some of which are discussed by Liang (1986).

SEM Models for Measurement Properties of Scales

Schaie and Hertzog (1985) discuss in great detail the literature on SEM procedures for estimating reliability and equivalence of measurement properties across multiple populations. One recent application nicely illustrates two important concepts: (1) the distinction between scale reliability and stability of individual differences and (2) the use of alternate forms to reveal information about the measurement properties of scales. Hertzog and Nesselroade (in press) reanalyzed data originally collected by Nesselroade, Mitteness, and Thompson (1984). It consisted of self-ratings of elderly individuals of two mood state factors: Anxiety and Fatigue. The design involved a short-term retest, so that individuals were given the mood state questionnaires twice, with approximately one month intervening between administrations. The three measures of state Anxiety included Spielberger's State Anxiety scale and Forms A and B of Curran and Cattell's Eight State Questionnaire. The three measures of Fatigue were subsets of items from the Eight State Fatigue scale. Nesselroade et al. (1984) demonstrated that the Anxiety and Fatigue factors could be identified using confirmatory factor analysis, and that the stability of individual differences in Anxiety was substantial, but not perfect, over the one month period.

Hertzog and Nesselroade (in press) reanalyzed the Nesselroade et al (1984) data, focusing on estimating the measurement properties of the Forms A and B of the Eight State Questionnaire. The model, shown in Figure 6, closely resembles the longitudinal factor models described earlier. It specified that the three scales of state anxiety loaded on an Anxiety factor, and that there was a residual covariance for the Spielberger scale across the two measurement occasions. In a series of models, Hertzog and Nesselroade (in press) tested whether the Cattell Forms A and B could be considered parallel forms. They also tested whether the measurement properties of Forms A and B were equivalent across the two measurement occasions. The hypothesis of parallelism was tested by constraining factor loadings and residual variances (error variances of measurement) to be equal for Forms A and B. The test of equivalence over time was made by constraining these parameters equal across the first and second administrations of the questionnaires. The results showed clearly that (1) Forms A and B were parallel, and (2) that the measurement properties of Forms A and B were identical over the two occasions of measurement. The estimated reliability for Forms A and B was .89. Clearly, the Eight State Anxiety scales have excellent measurement properties in older populations.

The high reliabilities for the scales contrast with the moderate (but lower) stabilities of individual differences in the latent variable, Anxiety. As discussed above, the stability of individual differences is reflected in the covariance between the latent Anxiety factors over the two measurement points. Using the parameter estimates from the Hertzog and Nesselroade (in press) analysis, an estimate of .72 is obtained for the <u>disattenuated</u> correlation of Anxiety with itself over a one-month period. This correlation was certainly greater than zero, but less than 1.0, indicating individual differences in mood state change over the one-month interval. There is a marked contrast between the short-term stability of Anxiety and the long-term stability of intelligence and personality found in the studies already reviewed. Individuals who are anxious at Time 1 are likely to be anxious at Time 2, but only about 50% of the variance in self-reported anxiety at Time 2 can be predicted from anxiety levels at Time 1. The most important feature of the analysis, however, is that this stability of individual differences has been estimated in a way that disentangles it from reliability. Using SEM, one can reject the hypothesis that the less-than-perfect stability is a function of attenuation due to measurement error. Conversely, the analysis shows that the lability in mood states does not imply that the mood state measures are unreliable. Given that one would expect mood states to fluctuate, the lability of Anxiety, and the excellent measurement properties of Forms A and B, argue indirectly for the construct validity of the scales, and suggest that they measure something different from the personality trait of Anxiety, which has been shown to exhibit a high degree of stability of individual differences.

Concluding Comments

This paper has reviewed a number of recent research studies using SEM approaches to address important research questions in gerontology. In a sense, the SEM technology is in its adolescent phase. The last several years have witnessed major growth in techniques and modeling applications, although our understanding of the potential and pitfalls of SEM has not yet reached full maturity. On the other hand, use of these approaches in gerontological research seems more accurately characterized as being in its infancy. In a few isolated areas, mostly related to psychometric theory and practice, SEM measurement models have been used to improve our understanding of constructs and measures relevant to aging, and have led to some substantive advances in the literature. At this point, however, the major contributions of this class of technique -- particularly, applications of structural regression models for describing and explaining change -- have yet to be realized. Hopefully this chapter has succeeded in outlining some of the current (albeit modest) successes, and whetted our collective appetites for more in the future. A measure of progress would be if future Annual Reviews do not contain chapters on SEM applications per se -- because use of the technique was sufficiently widespread and well-understood that the results of SEM studies were treated as best covered in substantively oriented reviews of age-related phenomena.

REFERENCES

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

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

Baltes, P.B., Reese, H.W., and Nesselroade, J.R. (1977). Life-span developmental psychology: Introduction to research methods. Monterey, CA: Brooks-Cole.

Bengtson, V.L., Reedy, M.N., and Gordon, C. (1985). Aging and self-conceptions: Personality processes and social contexts. In J.E. Birren and K.W. Schaie (Eds.), Handbook of the Psychology of Aging (2nd Edition), (pp. 544-593). New York: Van Nostrand Reinhold.

Bentler, P.M. (1973). Assessment of developmental factor change at the individual and group level. In J.R. Nesselroade and H.W. Reese (Eds.) Life-span Developmental Psychology: Methodological Issues (pp. 145-174). New York: Academic Press.

Birren, J.E. (1965). Age changes in the speed of behavior: Its central nature and physiological correlates. In A.T. Welford and J.E. Birren (Eds.), Handbook of Aging, Behavior, and the Nervous System. (pp. 191-216). Springfield IL: Charles C. Thomas.

Blalock, H.M., Jr. (1982). Conceptualization and Measurement in the Social Sciences. Beverly Hills, CA: Sage.

Blalock, H.M., Jr. (Ed.). (1985a). Causal Models in Panel and Experimental Designs. Chicago IL: Aldine.

Blalock, H.M., Jr. (Ed.). (1985b). Causal Models in the Social Sciences (2nd Edition). Chicago IL: Aldine.

Botwinick, J. (1977). Intellectual abilities. In J.E. Birren and K.W. Schaie (Eds.), Handbook of the Psychology of Aging (pp. 580-605). New York: Van Nostrand Reinhold.

Campbell, R.T. and Mutran, E. (1982). Analyzing panel data in studies of aging: Applications of the LISREL model. Research in Aging, 4, 3-41.

Caspi, A. and Elder, G.H. (1986). Life satisfaction in old age: Linking social psychology and history. Psychology and Aging, 1, 18-26.

Cook, T,D. and Campbell, D.T. (1979). Quasi-experimentation: Design and Analysis Issues for Field Settings. Chicago: Rand McNally.

Cornelius, S.W. Willis, SL., Nesselroade, J.R. and Baltes, P.B. (1983). Convergence between attention variables and factors of psychometric intelligence in older adults. Intelligence, 7, 253-269.

Costa, P.T.Jr. (1986). Psychosocial continuity in adulthood: Personality, abilities, social support, and well-being. Paper presented at the 94th Annual Convention of the American Psychological Association, Washington, D.C.

Costa, P.T. Jr., and McCrae, R.R. (1980). Still stable after all these years: Personality as a key to some issues in adulthood and old age. In P.B. Baltes and O.G. Brim, Jr. (Eds.), Life-span Development and Behavior (Vol.3). (pp. 65-102). New York: Academic Press.

Cunningham, W.R. (1978). Principles for identifying structural differences: Some methodological issues related to comparative factor analysis. Journal of Gerontology, 33, 82-86.

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 difference in adulthood and old age. Multivariate Behavioral Research, 16, 3-22.

Cunningham, W.R. and Birren, J.E. (1980). Age changes in the factor structure of intellectual abilities in adulthood and old age. Educational Psychological Measurement, 40, 271-290.

Dwyer, J.H. (1983). Statistical Models for the Social and Behavioral Sciences. New York: Oxford University Press.

Gustafsson, Jan-Eric. (1984). A unifying model for the structure of intellectual abilities. Intelligence, 8, 179-203.

Hayslip, B., Jr., and Brookshire, R.G. (1985). Relationships among abilities in elderly adults. A time lag analysis. Journal of Gerontology, 40, 748-750.

Heise, D.R. (1969). Separating reliability and stability in test-retest correlation. American Sociological Review, 34, 93-101.

Hertzog, C. (1985a). An individual differences perspective: Implications for cognitive research in gerontology. Research on Aging, 7, 7-45.

Hertzog, C. (1985b). Applications of confirmatory factor analysis to the study of intelligence. In D.K. Delterman (Ed.), Current Topics in Human Intelligence. (pp.59-97). Norwood, N.J.: Ablex.

Hertzog, C. (in press). On the utility of structural regression models for developmental research. Chapter to appear in Baltes, P.B., Featherman, D., and Lerner, R.M. (Eds.) Life-span Development and Behavior. (Vol. 9). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hertzog, C. and Nesselroade, J.R. (in press). Beyond autoregressive models: Some implications of the trait-state distinction for the structural modeling of developmental change. Child Development.

Hertzog, C., Raskind, C.L., and Cannon, C.J. (1986). Age-related slowing in semantic information processing speed: An individual differences analysis. Journal of Gerontology, 41, 500-502.

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

Hertzog, C., and Schaie, K.W. (1986b). Stability and Change in Adult Intelligence: 2. Simultaneous Analysis of Longitudinal Means and Covariance Structures. Unpublished manuscript.

Horn, J.L. (1978). Human ability systems. In P.B. Baltes (Ed.), Life-span Development and Behavior: Vol.1 (pp. 211-256). New York: Academic Press.

Horn, J.L. and McArdle, J.J. (1980). Perspectives on mathematical/statistical model building (MASMOB) in research on aging. In L.W. Poon (Ed.), Aging in the 1980's: Psychological Issues (pp. 503-541). Washington, D.C.: American Psychological Association.

Horn. J.L., McArdle, J.J., and Mason, R. (1984). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. Southern Psychologist, 1, 179-188.

Hultsch, D.F., Hertzog, C., and Dixon, R.A. (1984). Text processing in adulthood : The role of intellectual abilities. Developmental Psychology, 20, 1193-1209.

Joreskog, K.G. (1979). Statistical estimation of structural models in longitudinal developmental investigations. In J.R. Nesselroade and P.B. Baltes (Eds.), Longitudinal Research in the Study of Behavior and Development (pp. 303-351). New York: Academic Press.

Joreskog, K.G. and Sorbom, D. (1977). Statistical models and methods for analyses of longitudinal data. In D.S. Aigner and A.S. Goldberger (Eds.), Latent Variables in Socio-economic Models (pp. 285-325). Amsterdam, North Holland.

Joreskog, K.G. and Sorbom, D. (1979). Advances in Factor Analysis and Structural Equation Models. Cambridge, MA: Abt Associates.

Joreskog, K.G. and Sorbom, D. (1980). Simultaneous Analysis of Longitudinal Data From Several Cohorts. Research Report 80-5. University of Uppsala, Dept. of Statistics.

Joreskog, K.G. and Sorbom, D. (1984). LISREL VI User's Guide. Mooresville, IN: Scientific Software.

Kagan, J. (1980). Perspectives on continuity. In O.G. Brim, Jr., and J. Kagan (Eds.), Constancy and Change in Human Development. Cambridge, MA: Harvard University Press.

Kessler, R.C. and Greenberg, D.F. (1981). Linear Panel Analysis. New York: Academic Press.

Lachman, M.E. (1983). Perceptions of intellectual aging: Antecedent or consequence of intellectual functioning? Developmental Psychology, 19, 482-498.

Lair, T.J., Hertzog, C., and Schulenberg, J.E. (1985). Stability of Personality in Adulthood: A Structural Equation Model. Paper presented at the 38th Annual Meetings of the Gerontological Society, New Orleans, LA.

Liang, J. (1986). Self-reported physical health among aged adults. Journal of Gerontology, 41, 248-260.

McArdle, J.J. (1986). Latent Growth Curves Within Developmental Structural Equation Models. Unpublished manuscript.

McArdle, J.J. & McDonald, R. P. (1984). Some Algebraic Properties of the Reticular Action Model for Moment Structures. British Journal of Mathematical and Statistical Psychology, 37, 234-251.

McDonald, R.P. (1978). A simple comprehensive model for the analysis of covariance structures. British Journal of Mathematical and Statistical Psychology, 31, 59-72.

McDonald, R.P. (1981). The dimensionality of tests and items. British Journal of Mathematical and Statistical Psychology, 34, 100-117.

Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. Psychological Bulletin, 89, 575-588.

Mulaik, S.A. (in press). Toward a Conception of Causality Applicable to

Experimentation and Causal Modelling. Child Development.

Nesselroade, J.R., and Baltes, P.B. (1979). Longitudinal Research in the Study of Behavior and Development. New York: Academic Press.

Nesselroade, J.R., and Baltes, P.B. (1984). From traditional factor analysis to structural-causal modeling in developmental research. In V. Sarris and A. Parducci (Eds.), Experimental Psychology in the Future (pp. 267-287). Hillsdale, NJ: Lawrence Erlbaum Associates.

Nesselroade, J.R., Mitteness, L.S., and Thompson, L.K. (1984). Short-term changes in anxiety, fatigue, and other psychological states in older adulthood. Research on Aging, 6, 3-23.

Nesselroade, J.R., Stigler, S.M., and Baltes, P.B. (1980). Regression toward the mean and the study of change. Psychological Bulletin, 88, 622-637.

Olsson, V. and Bergman, L.R. (1977). A longitudinal factor model for studying change in ability structure. Multivariate Behavioral Research, 12, 221-241.

Petros, T.V., Zehr, H.D., and Chabot, R.J. (1983). Adult age differences in accessing and retrieving information from long-term memory. Journal of Gerontology, 38, 589-592.

Riegel, K.F. and Riegel, R.M. (1972). Development, drop, and death. Developmental Psychology, 6, 306-319.

Reinert, G. (1970). Comparative factor analytic studies of intelligence throughout the human life span. In L.R. Goulet and P.B. Baltes (Eds.) Life-span Developmental Psychology: Research and Theory (pp. 467-484). New York: Academic Press.

Rogosa, D. (1979). Causal models in longitudinal research: Rationale, formulation, and estimation. In J.R. Nesselroade and P.B. Baltes (Eds.), Longitudinal Research in the Study of Behavior and Development. (pp. 263-302). New York: Academic Press.

Rogosa, D. (1980). A critique of cross-lagged correlation. Psychological Bulletin, 88, 245-258.

Rogosa, D. and Willett, J.B. (1985). Understanding Correlates of change by modeling individual differences in growth. Psychometrika, 50, 203-228.

Salthouse, T.A. (1985). A Theory of Cognitive Aging. Amsterdam: North Holland.

Schaie, K.W. and Hertzog, C. (1985). Measurement in the psychology of adulthood and aging. In J.E. Birren and K.W. Schair (Eds.), Handbook of the Psychology of Aging (2nd Edition). New York: Van Nostrand Reinhold.

Schaie, K.W., Willis, S.L., Hertzog, C., and Schulenberg, J.E. (1986). Effects of Cognitive Training upon Primary Mental Ability Structure. Unpublished manuscript.

Stricker, L.J. and Rock, D.A. (1985). The Structure of the GRE General Test for Older Examinees: A Confirmatory Factor Analysis. Unpublished Technical Report, Educational Testing Service, Princeton, NJ.

Figure Captions

1. Structural model for individual differences in three RT factors. SMA is a semantic memory access speed factor that determines covariances among five semantic RT task/task conditions: Category Matching (High and Low Typicality), Semantic Matching (High and Low Typicality), and Synonym Matching.

2. Longitudinal factor model for data from Schaie's Seattle Longitudinal Study. A general intelligence factor (g) determines covariances among five intelligence tests at each of three longitudinal occasions. Autocorrelated residuals among the five subtests across occasions are modeled, but not shown. Reprinted with permission from Hertzog and Schaie (1986a).

3. Factor pattern matrix for the longitudinal factor model including both occasion-specific general intelligence (g) and five test-specific factors for the primary abilities. Rows correspond to observed variables, columns to factors. For example, row 1 shows that V_1, Verbal Meaning at the first occasion, loads on g_1 (g at the first occasion) plus the Verbal Meaning factor (V). Reprinted with permission from Hertzog and Schaie (1986a).

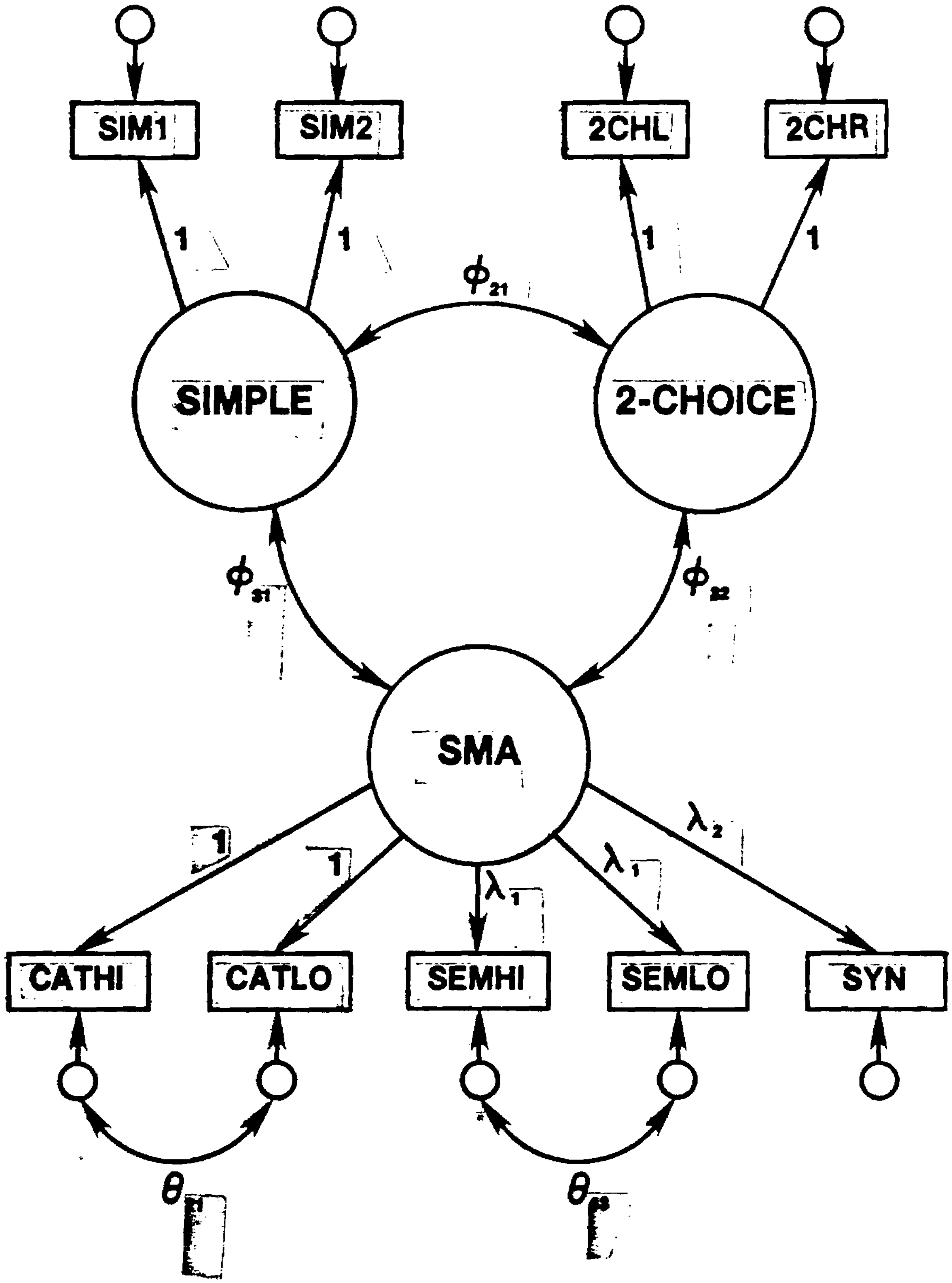
4. Model estimated by Lair et al. for two personality factors, Neuroticism (N)

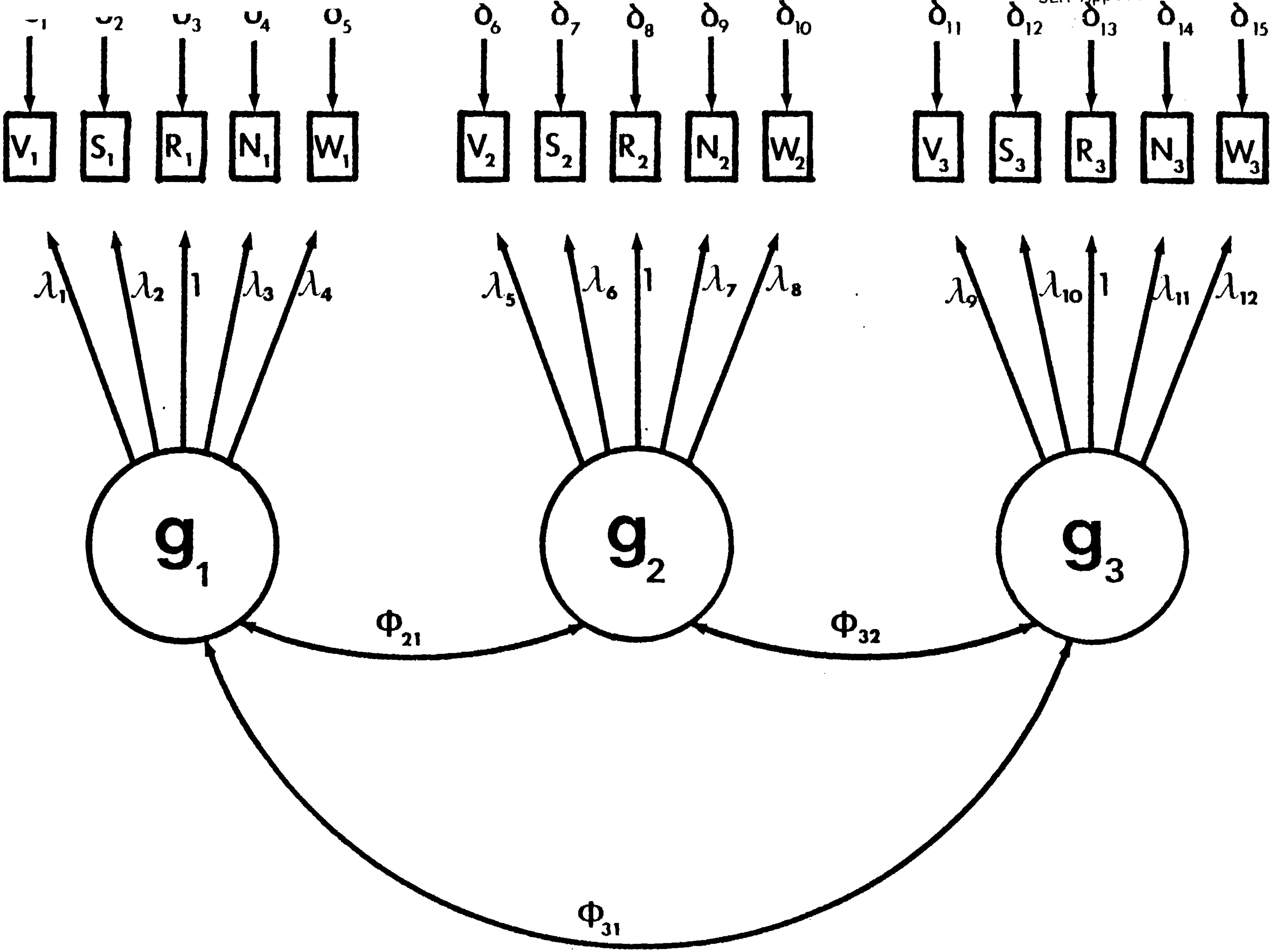
and Extraversion (E), measured at four longitudinal occasions. All regression coefficients have been rescaled to a standardized metric. Stability of individual differences in N and E is quite high over time, with all standardized stability coefficients exceeding .9.

5. Lachman's model of change in memory span in elderly adults. (INT = Internal Control, CHA = Chance Control, POW = Powerful Others Control, ISE = Intellectual Self-Efficacy, CIA = Concerns about Intellectual Aging, Ms = Memory Span). None of the personality factors significantly predicted change in memory span, with the only significant relationship being the autoregression (stability coefficient) of memory span on itself. Adapted with permission from Lachman (1983).

6. SEM for two mood state factors, Anxiety (ANX) and Fatigue (FAT) measured at two longitudinal occasions. For Anxiety, three measures were available, Spielberger's State Anxiety Scale (SPIEL) and two alternate forms of of the 8-State Anxiety scale (FORM A, FORM B). A series of models tested the measurement properties of these alternate forms and their relationship to SPIEL at the two occasions (see text). Reprinted with permission from Hertzog & Nesselroade (in press).

Figure 1

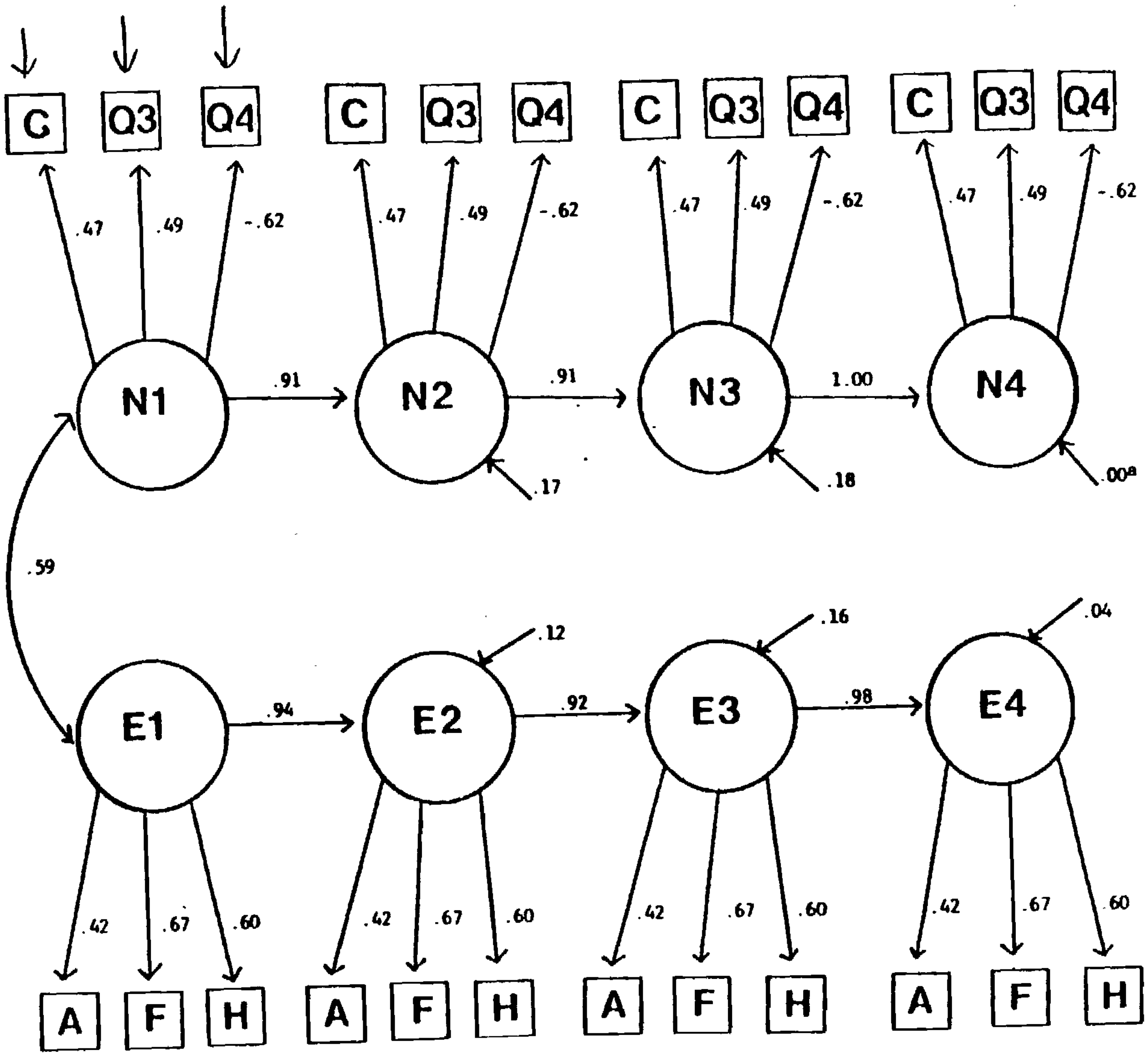

SIM1
SIM2
2CHL
2CHR
1
1
1
1
ϕ_{21}
SIMPLE
2-CHOICE
ϕ_{31}
ϕ_{32}
SMA
1
1
λ_1
λ_1
λ_2
CATHI
CATLO
SEMHI
SEMLO
SYN
θ_{21}
θ_{43}


δ6 δ7 δ8 δ9 δ10 δ11 δ12 δ13 δ14 δ15
V1 S1 R1 N1 W1 V2 S2 R2 N2 W2 V3 S3 R3 N3 W3
λ1 λ2 1 λ3 λ4 λ5 λ6 1 λ7 λ8 λ9 λ10 1 λ11 λ12
g1 g2 g3
Φ21 Φ32 Φ31

VARIABLES

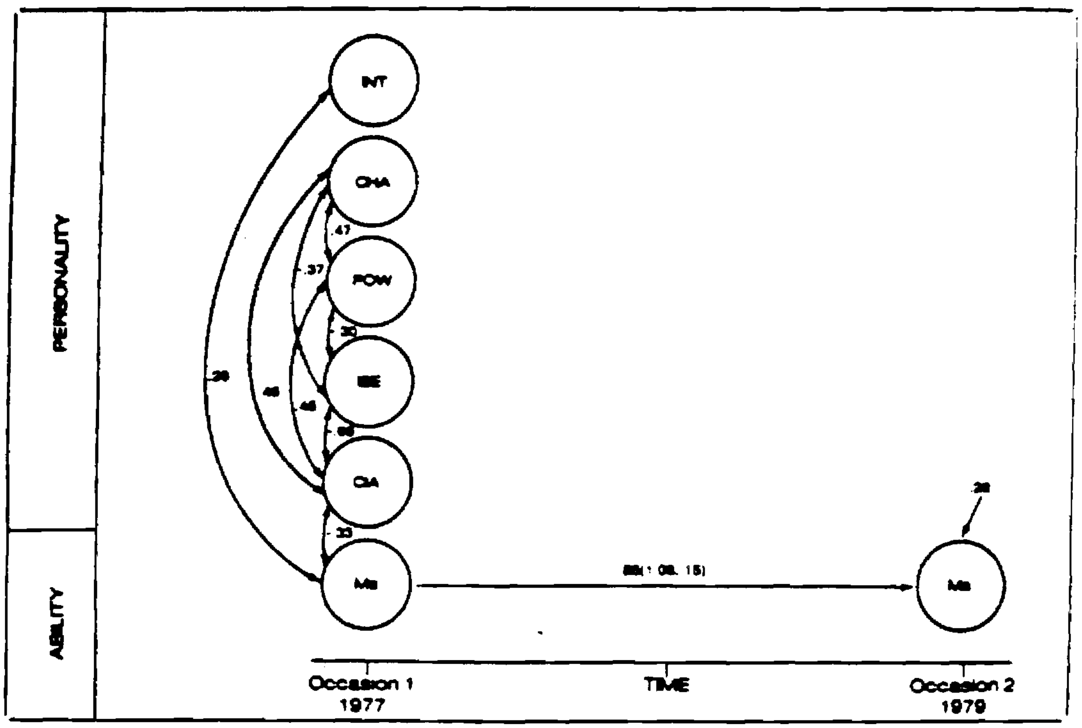
$$
\begin{array}{cc}
 & \begin{array}{cccccccc} g_1 & g_2 & g_3 & V & S & R & N & W \end{array} \\
\begin{array}{c} V_1 \\ S_1 \\ R_1 \\ N_1 \\ W_1 \\ \\ V_2 \\ S_2 \\ R_2 \\ N_2 \\ W_2 \\ \\ V_3 \\ S_3 \\ R_3 \\ N_3 \\ W_3 \end{array} &
\begin{bmatrix}
\lambda_1 & 0 & 0 & \lambda_5 & 0 & 0 & 0 & 0 \\
\lambda_2 & 0 & 0 & 0 & \lambda_7 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & \lambda_9 & 0 & 0 \\
\lambda_3 & 0 & 0 & 0 & 0 & 0 & \lambda_{11} & 0 \\
\lambda_4 & 0 & 0 & 0 & 0 & 0 & 0 & \lambda_{13} \\
 & & & & & & & \\
0 & \lambda_1 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & \lambda_2 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & \lambda_3 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & \lambda_4 & 0 & 0 & 0 & 0 & 0 & 1 \\
 & & & & & & & \\
0 & 0 & \lambda_1 & \lambda_6 & 0 & 0 & 0 & 0 \\
0 & 0 & \lambda_2 & 0 & \lambda_8 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & \lambda_{10} & 0 & 0 \\
0 & 0 & \lambda_3 & 0 & 0 & 0 & \lambda_{12} & 0 \\
0 & 0 & \lambda_4 & 0 & 0 & 0 & 0 & \lambda_{14}
\end{bmatrix}
\end{array}
$$

Add error varia
to Figure

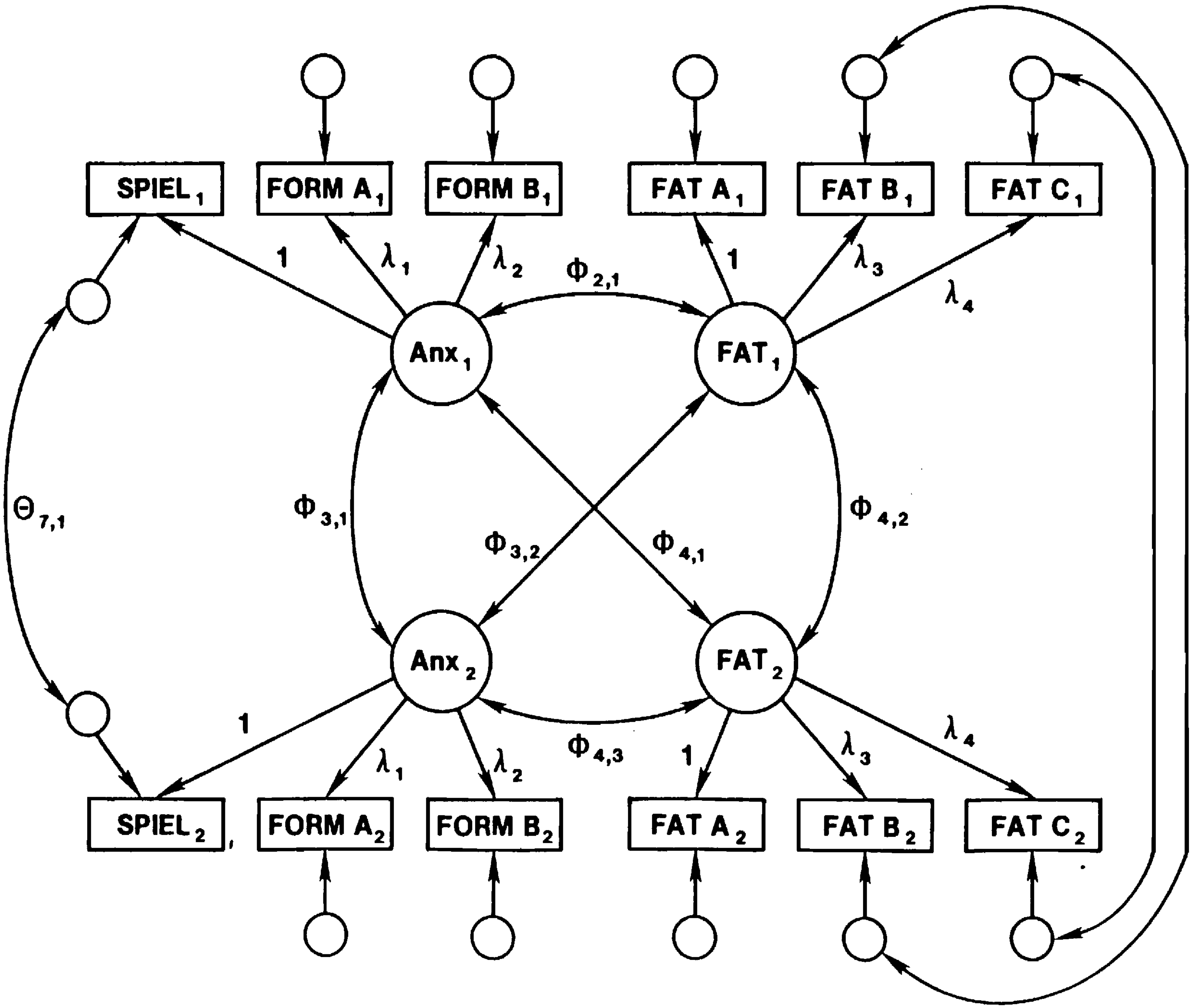


Accepted Longitudinal Model for the Exploratory Group

NOTE: All coeffecients are rescaled to a standardized metric.
[a]To correct for a Heywood case, this value was set to 0.



Causal model of memory span change. (INT = Internal Control, CHA = Chance Control, POW = Powerful Others Control, ISE = Intellectual Self-Efficacy, CIA = Concern About Intellectual Aging, Ms = Memory Span. Solid line indicates significant path consistent with predictions; parameter outside parentheses on straight path is the standardized regression coefficient; parameters inside parentheses are the unstandardized regression coefficient and its standard error, parameters on curved paths are correlation coefficients.)


SPIEL$_1$
FORM A$_1$
FORM B$_1$
FAT A$_1$
FAT B$_1$
FAT C$_1$
1
λ_1
λ_2
$\Phi_{2,1}$
1
λ_3
λ_4
Anx$_1$
FAT$_1$
$\Theta_{7,1}$
$\Phi_{3,1}$
$\Phi_{3,2}$
$\Phi_{4,1}$
$\Phi_{4,2}$
Anx$_2$
FAT$_2$
1
λ_1
λ_2
$\Phi_{4,3}$
1
λ_3
λ_4
SPIEL$_2$
FORM A$_2$
FORM B$_2$
FAT A$_2$
FAT B$_2$
FAT C$_2$

On the Utility of Structural Equation Models

for Developmental Research

Christopher Hertzog

Georgia Institute of Technology

Draft chapter prepared for:
P.B. Baltes, D. Featherman, & R. M. Lerner, (Eds.) Life-span Development and Behavior (Volume 10)
6/28/87

Address Correspondence To:

Christopher Hertzog
School of Psychology
Georgia Institute of Technology
Atlanta, GA 30332-0170

Running Head: Developmental Models

Author's Note: This chapter is dedicated to John Nesselroade, whose teaching, counsel, support, and friendship have been instrumental to my professional and personal growth. Preparation of this manuscript was supported in part by a Research Career Development Award (1 K04 AG00335-02) from the National Institute on Aging.

Abstract

Structural equation models (SEM) have become an increasingly popular technique for analysis of developmental research questions. However, a number of unfortunate misconceptions can be found in the literature regarding the nature, potential, and pitfalls of SEM. It is fallacious to assume that use of SEM techniques guarantees sound causal inference from correlational data; it is equally fallacious to argue that use of SEM for purposes other than testing causal models is an invalid misapplication of the method. In developmental research, important descriptive research questions can be shown to be linked to SEM models in two important ways: alternative SEM models may be used to provide direct statistical tests of important descriptive developmental hypotheses, and SEM model parameters can be interpreted with respect to fundamental issues in developmental analysis (e.g., estimating the degree to which differential developmental patterns alter distributions of individual differences). This paper develops the logic and procedures for implementing longitudinal SEM techniques to address descriptive developmental questions, with a brief illustration of the application of SEM to longitudinal factor analysis

Introduction

The past several years have been marked by a veritable explosion in multivariate methods suitable for analyzing developmental change (e.g., Nesselroade & Baltes, 1979). One of the more important advances have been in the area of structural equation models (SEM) for longitudinal data. As I use the term here, SEM includes "measurement models" specifying unobserved, latent variables (or factors) as determinants of observed or manifest variables. In other words, SEM subsumes confirmatory factor analysis as a special case of measurement equations alone. SEM provides a basis for empirical assessment of many of the more fundamental questions of interest in developmental research, particularly those questions involving correlates of individual differences in change. Although the intellectual roots of the approach (in sociology, econometrics, and psychology) were well understood prior to the decade of the 1970's (see, for example, Blalock's (1985a,b) edited volumes on the topic), much of the impetus for increased empirical application in the last 15 years was probably due to the development and distribution of computer algorithms for SEM. The contributions of Karl Joreskog, Dag Sorbom, and their colleagues on applying methods of modeling covariance structures to confirmatory factor analysis (e.g., Joreskog, 1969, 1971, 1974) and analysis of SEM with unobserved variables (e.g., Joreskog & Sorbom, 1977) culminated in the familiar LISREL model and computer program (Joreskog & Sorbom, 1984). Although their work is seminal, many important theoretical contributions and alternative computational approaches for covariance structures analysis have been made by others (e.g., Bentler, 1982, 1985; Bentler & Weeks, 1979; Browne, 1984; McArdle, 1980; McArdle & McDonald, 1984; McDonald, 1978, 1980).[1] Suffice it to say that the number of energetic and innovative scientists working in SEM theory and application gives us hope for continuing progress in this domain.

A number of papers have been written on the topic of developmental applications of SEM (e.g., Baltes, Reese, & Nesselroade, 1977; Campbell & Mutran, 1982; Hertzog, 1985b; Horn & McArdle, 1980; Labouvie, 1974; Nesselroade & Baltes, 1984; Rogosa, 1979; Schaie & Hertzog, 1985). Generally speaking, the most widely discussed application is the use of SEM to model change in longitudinal (panel) data in which a latent variable approach is possible given that multiple measures have been collected for each of the constructs of interest. Important, related treatments of longitudinal analysis have appeared in literature outside the developmental tradition (e.g., Dwyer, 1983; Kessler & Greenberg, 1981). As discussed by Schaie & Hertzog (1985), there are a number of important applications of SEM in developmental research, including analysis of 1) equivalence of empirical measures across different levels of development (or historical time), 2) factorial invariance, as assessed with SEM measurement models, and 3) structuring individual differences in developmental change at the level of the latent variables. A number of empirical applications of SEM have appeared in the developmental literature (e.g., Cunningham, 1981; Hertzog & Schaie, 1986a,b; Lachman, 1983; see the review by Hertzog, 1987).

As is often the case in an emergent paradigm, the early literature on SEM identified some prototypic models whose merits have since been critically evaluated -- often by the same scientists whose intial enthusiasm fueled the early growth. For example, Joreskog and Sorbom (1977) featured a general autoregressive SEM approach for panel data, and it formed the basis for Rogosa's (1979, 1980) methodological recommendations regarding causal analysis via cross-lagged regression analysis. Recently, however, Rogosa and colleagues (e.g., Rogosa & Willett, 1985a,b) have criticized autoregressive models as a method of structuring correlates of developmental change. In addition, others less enthusiastic initially about the approach have voiced caution about the uncritical and uninformed use of SEM (e.g., Cliff, 1983). Still others have analyzed weaknesses with the LISREL-type formulation, including distributional assumptions, use of full-information maximum likelihood estimation, effects of small sample sizes, etc. (e.g., Boomsma, 1982; Huba & Harlow, 1986). It seems evident that the field is in a period of expansion and, as it were, shaking out, in which original conceptions of the procedures and their utility are evolving as scientists gain further insight and experience regarding SEM's strengths and weaknesses. Furthermore, recent methodological advances in the area of structural modeling (e.g., Bentler, 1985; Browne, 1984; McArdle & McDonald, 1984) have opened new realms of possibilities regarding developmental applications of modeling techniques (e.g, McArdle & Epstein, 1987).

The advances that can be observed are indeed exciting, and presage what may eventually become a scientific revolution in dominant empirical approaches to research on life-span behavioral development. However, two types of inertia seem to be operating to hold back progress with respect to developmental applications of SEM. First, the amount of methodological expertise and technical knowledge required appears to be increasing geometrically, widening the gap between those able to apply SEM and those who cannot. This gap increases the chances of (1) abuse of the methods and (2) uncritical acceptance of questionable research by those who cannot distinguish valid from invalid applications of SEM. Moreover, the widening knowledge gap increases the likelihood that those who apply SEM will become increasingly estranged, scientifically speaking, from the community of developmental scholars who do not. Second, the complexity of the methods encourages perpetuation of incorrect notions about the differences between SEM and other alternatives. In particular, there is a great deal of confusion regarding appropriate methods for (1) exploratory verus confirmatory analysis and (2) the status of SEM for description versus explanation (i.e., causal models). These incorrect (and often implicit) ideas about the methods and their appropriate use leads to incorrect designs, invalid inferences, and unjustified criticisms of work that does and does not employ SEM. The perspective enunciated in this paper is, I believe, consistent with that of others who employ SEM in empirical research (see, in particular, McArdle & Epstein, 1987, for a similar view). It may prove novel to those who have learned to equate SEM with causal modeling.

This paper is intended to promote understanding of how SEM can be used (and misused) in developmental research. In order to do so, I shall first discuss the logic behind SEM. In particular, I shall discuss the importance of causal, or explanatory, models, and how SEM can be used for that purpose. I shall also discuss, however, how an incomplete perspective on SEM leads to (1) erroneous conceptions about applications of SEM for purposes of explanation as well as (2) misconceptions about the role of SEM in exploratory analysis. The use of SEM for exploratory analysis of longitudinal data is critical in developmental research, and I shall spend some time laying out the reasons why such exploratory analysis with SEM is both valid and useful. I then discuss briefly how longitudinal SEM models may be used for exploratory, descriptive research on psychological development.

The Logic of SEM in Developmental Research Applications

Description and Explanation in Developmental Research

Perhaps the basic reason for enthusiasm about SEM for developmental research is that it appears to provide a viable means of getting around the bane of developmental research: valid explanation of developmental change. From the first formal stirrings of the subdiscipline of life-span developmental psychology, its founders have argued for the importance of description, explanation, and modification of development, and have decried emphasis on description at the expense of explanation and modification (e.g., Baltes & Willis, 1977). In the case of adult intellectual development, cross-sectional data on age differences in intelligence are often interpreted in terms of ontogenetic decline under the presumption that biological aging is the cause of any age differences. It is by now well known that the chief argument against use of cross-sectional age differences to estimate ontogenetic change is that generational (cohort) differences are confounded with age changes in the cross-sectional design (e.g., Schaie, 1965; Baltes, 1968). However, a point that is often blurred is that there are two potential fallacies involved in the use of cross-sectional designs for assessing ontogenetic change: 1) the confounding of age changes with generational differences, which vititates the cross-sectional age differences as a valid description of age-correlated change, and 2) the assumption that age-correlated changes have biological causes -- i.e., are determined by biological mechanisms of aging (Baltes, Dittman-Kohli, & Dixon, 1984).

Improving on cross-sectional designs through more complicated sampling schemes can help to address the problem of confounded age, cohort, and period effects (e.g., Schaie, 1965, 1977). However, the validity of explanatory interpretations of data from longitudinal or cross-sectional sequences is at best marginally improved in the more sophisticated sampling designs (Baltes 1968; Labouvie, 1974, 1978; Labouvie & Nesselroade, 1985; Nesselroade & Labouvie, 1985). Schaie's sequential strategies and related methods may provide more valid description of age-correlated changes, but in and of themselves they do not isolate the explanations of such change (Labouvie, 1974; Nesselroade & Labouvie, 1985; Schaie & Baltes, 1975; Schaie & Hertzog, 1982). The causes of age-correlated change include but are not restricted to ontogenetic determinants. There are many age-graded (Baltes & Willis, 1977) determinants of change in our society that are not inherently ontogenetic (Featherman, 1985). An example in the United States is age-graded retirement policies, which have caused certain life transitions to occur regularly during the chronological age range of middle sixties (at least, for the birth cohorts reaching these ages in the latter half of the twentieth century). Sequential designs conducted during this epoch of American history might find changes in income dynamics, self concept, and social status between ages 60 and 70 that are replicated over multiple birth cohorts. Nevertheless, the validity of an ontogenetic interpretation of these changes would be questionable, being at best an indirect influence of biological aging on occupational performance and its subsequent impact on retirement policies (see Baltes et al. 1984, and Featherman, 1985, for further discussion of such issues).

Valid description of developmental phenomena is an important and often difficult enterprise, given the multiple design confounds that arise in both longitudinal and cross-sectional sequences (Baltes, Reese, & Nesselroade, 1977; Nesselroade & Labouvie, 1985; Schaie, 1977). At the present time, most life-span developmental research has been descriptive rather than explanatory. This is to be expected, of course, in an emergent discipline, given that valid description is, generally speaking, an important precursor of valid explanation. Thus, research paradigms that lead to valid descriptions of development are valuable, albeit not an endpoint in the research process.

It is important to distinguish description/explanation as a research goal from induction/hypothesis testing as a research method. Some seem to assume that descriptive research is inherently inductive and exploratory in nature (and hence, to be devalued), whereas explanatory research seeks verification of prior hypothesis about causal mechanisms (and hence, is to be applauded). Good explanatory research does generally involve hypothesis testing. We can be sure that we have isolated an explanation if we can use the explanatory theory to generate new, empirically testable hypotheses about the explanations of change. However, descriptive research need not be inductive searches for relationships among variables. It can be designed to test descriptive hypotheses (e.g., change is correlated with certain background conditions), or combine both approaches. Hypothesis testing is a property of good descriptive research as well. A longitudinal panel design which measures multiple indicators of related constructs (e.g., depression, self-concept, and perceived control) at arbitrarily spaced retest intervals without preconceptions about patterns of change over time is an <u>exploratory</u> descriptive design. It may not even be possible for the investigator to specify any hypotheses other than the usual statistical null hypotheses (e.g., no change in level of depression across longitudinal occasions) and their mundane alternatives. Many longitudinal studies fall into this category, at least at their inception (for examples, see the edited volume by Schaie, 1983). On the other hand, the investigator may design a study by selecting longitudinal sequences in such a way as to test a particular set of descriptive hypotheses. For example, the investigator might argue that age-correlated changes will be observed in levels of perceived control and depressive affect between ages 55 and 65, and that the changes in both constructs will be correlated with each other and with changes in perceived status in the workplace. Such descriptive hypotheses may or may not be based upon implicit explanatory hypotheses (e.g., perceived loss of status attainment in an employment setting determines changes in perceived control and depressive affect). Nevertheless, the hypothesis as stated above is still inherently descriptive in nature. The advantage of explicit testing of descriptive hypotheses is that the design and the statistical analysis can be tailored to the hypothesis, achieving gains in validity, statistical power, and relevance of the empirical observations for the substantive phenomena of interest.

SEM, Causal Modeling, and Developmental Explanation

The term "causal modeling" may eventually come to be understood as one of the most unfortunate labels ever created in social science research. On the one hand, it has led some to believe that it is somehow possible to achieve what they had been originally taught was impossible: the unambiguous assignment of cause and effect based upon statistical correlation. I shall not single out examples, but the recent developmental literature is rife with examples of the fruits of this misconception. SEM is not a magic window into causal relations among nonmanipulated variables. One can only use SEM to test the empirical predictions derived from theoretical causal models -- a subtle but crucially important distinction (e.g., James, Mulaik, and Brett, 1982; Mulaik, 1986; 1987). On the other hand, the term causal modeling also appears to have led to the more subtle but invidious counter-fallacy: namely, that use of SEM for descriptive purposes is somehow invalid science -- a misuse, as it were, of causal techniques. The implicit flow of the belief system appears to be something like the following:

"SEM is useful for testing causal models. Causal models require specific theoretical hypotheses. Specification of a SEM requires hypotheses about relationships among variables. Descriptive (or exploratory) research doesn't test causal hypotheses. Therefore, use of SEM in descriptive applications is a misuse of the technique."

Taken individually, all the statements prior to the conclusion are accurate. The primary logical fallacy inherent in the "therefore" is the assumption of equivalence of causal hypotheses and the hypotheses required for SEM specification. An (implicit) corollary of this fallacy is the assumption that a statistical test of a SEM is equivalent to a test of a causal model. Both assumed equivalences are in fact untrue: there is no necessary connection between a causal model and a SEM, nor is there a necessary relationship between statistical tests of a SEM and tests of causal hypotheses. SEM can be used in causal analysis, but only under restrictive conditions. SEM can also be used in descriptive analysis, and indeed, is particularly well-suited for certain kinds of descriptive research.

The Logic of Valid Causal Inference in SEM. The principal problem with causal inference from passive observation of systems of variables is well known — lack of experimental control for effects of variables in and outside of domain of variables studied. When multiple causal variables cannot be independently manipulated or held constant, the best available alternative is to use statistical control in order to measure cause/effect relations in a system of variables. SEM has evolved as a method for achieving this type of statistical control. It is an appropriate method for evaluating the plausibility of probabalistic models of causation (James et al., 1982; Mulaik, 1987; Steyer, 1985; Suppes, 1970) because it explicitly seeks 1) to analyze a system of variables under a set of specific assumptions about cause/effect relations in that system (i.e., a causal model) and 2) to reject models which are found to be inconsistent with empirical data. The assumptions are probabilistic assertions about cause/effect relations. For example, one might argue that, across the population of persons, there is some nonzero degree of causal effect of knowledge of memory functions for performance in laboratory memory tasks. We need not assume that this causal relation holds for all persons in all situations, given the complex nature of other causal influences. We do need to assume that the nature of the probabilistic cause/effect relationship, both with respect to the functional form of the relationship (e.g., linear vs. nonlinear) and with respect to the relationships to other causes of the effect variable (e.g., additive versus interactive, direct versus indirect causation) have been correctly specified in the causal model.

The inability to achieve experimental isolation of cause/effect relations has high costs. The general requirements for use of SEM in causal analysis are explained in much greater detail in several texts on the topic (e.g., Duncan, 1975; Dwyer, 1983; Heise, 1975; James, Mulaik, & Brett, 1982). Blalock (1985a,b) has recently provided excellent edited volumes summarizing some of the key papers in the area. Two key assumptions are difficult in practice to satisfy: <u>self-containment</u> and <u>equilibration of cause/effect relations</u>.

Self-containment is achieved when all relevant causes of a variable to be explained have in fact been included in the structural equation for that variable. The meaning of "relevant variables" is nicely delineated by James et al. (1982). Briefly, one need not include all causes of a given effect variable in a model. Instead, one needs to include all causes that are correlated with the causal variables that are the focus of the investigation. Provided that omitted causal variables are irrelevant to the action of the causal variables included in the model, then these residual causes do not perturb the causal model that is statistically estimated. Conversely, the omission of causal variables that are correlated with other causal variables in the system will thwart the efforts to achieve valid statistical isolation of magnitudes of causal influence. Equilibrium implies that the magnitude of cause/effect relations is stable across the interval of observation.

These assumptions are met by isolation of independent variables through experimental manipulation. In passively observed systems of variables, they must be satisfied by appropriate selection of both the observed variables and the density of observations in the sampling design. Any error in theory inevitably leads to error in the specification of the structural equations intended to reflect the causal system. Such specification error leads in turn to biased estimates of structural regression coefficients. Indeed, incorrect assumptions about direction of causal relations, direct versus indirect effects, etc., can lead to wholly erroneous conclusions about the causal system under study.

On the other hand, given that a causal model accurately reflects the assumptions and hypotheses of a causal theory, we can use the fit of an SEM to empirical data as a test of the causal theory. Causal analysis in passively observed systems of variables depends crucially on the logic of falsification. One seeks to falsify a causal theory about a system of variables by showing that it cannot account for the covariances among empirically measured variables (Popper, 1959). Inference is asymmetric, in that one can only conclude that a model is false. A model that is found to be consistent with empirical data has not been shown to be true, just merely consistent, for many alternative but false models may be consistent with an empirical data set. This feature of causal modeling is completely consistent with the philosophy of science and experimental tests of hypotheses, but it runs opposite to the scientist's natural desire to demonstrate the truth of an hypothesized cause/effect relationship. Thus one common mistake in causal modeling is to conclude that a significant structural regression coefficient in a system of equations demonstrates that a causal relationship exists between the two variables. In fact, all one has shown is that the hypothesis of direct influence is consistent with the data, given all other relationships specified to exist in the model. Alternative models may be found that predict no direct influence of the hypothesized causal variable but are still equally adequate in terms of their agreement with the sample data.

Once the properties of model falsification are understood, the hypothesis testing (via falsification) orientation of causal modeling may be seen as an advantage, not a liability. Indeed, one of the virtues of causal modeling is that it explicitly emphasizes the notion of risking a theory against empirical data. A model that has been falsified is far from worthless -- in fact, the falsification tells us something useful and important about the type of causal influences that are and are not operating in a given system of variables. A model that survives an empirical test has not been proven true, but it can be considered a useful approximation to reality that should be reevaluated and retested until its limits are realized. We learn far more by specifying, testing, and falsifying complex models of causal relations than by detecting atheoretical, statistically significant correlations among sets of variables (Meehl, 1978). Furthermore, the optimal use of causal modeling is when a study is deliberately designed to pit competing theories and models against one another -- seeking to falsify one set of models relative to a set of alternatives. As such, causal modeling represents nonexperimental science's logical equivalent to strong inference (Platt, 1964) in experimental research.

There are a number of research design issues that must be addressed before one can meaningfully use statistical methods to test a theoretically specified causal model. Blalock (1982) has emphasized the importance of the additional assumptions regarding research design and measurement issues needed for SEM under the rubric of an "auxiliary theory." We make a number of assumptions about the adequacy of our empirical measures and how they serve as indicators of a latent variable of interest. Many of these assumptions lie outside the measurement model as included, but they determine directly its adequacy. Given a theory of causal relations among constructs, one must insure that the measures used are in fact consistent with the causal process as specified. Let us say that I am interested in metamemory/memory behavior relationships (e.g., Dixon & Hultsch, 1983) and decide to examine a particular type of these relationships in a social setting, such as a cocktail party. If I argue that perception of competence in remembering faces and names influences social gregariousness by making a person more likely to converse with people who look familiar, then it makes little sense to measure perceived competence in this aspect of memory by giving people a more general metamemory questionnaire on incidents of forgetting, knowledge of memory strategies, etc. -- as is often done in the metamemory literature. To the extent that these measures combine multiple aspects of perceived competence that in theory vary across persons (e.g., I may be good at remembering names with faces, but terrible at remembering appointments) the measure may not be maximally valid with respect to the particular causal attribute operating in this class of situation. In Blalock's terms, use of the general metamemory measure makes the implicit auxiliary assumption that the dimension of metamemory operative in the social setting is in fact the dimension measured by the general metamemory scale. Even if the questionnaire contains specific questions about remembering names, the assumption that the summative scale score is the most appropriate metamemory measure implicitly assumes that the construct-valid components in items measuring perceptions of ability to remember names covary with the overall scale score. Evidence in favor of this hypothesis can be gathered outside of the SEM (e.g., demonstration of high item-total correlations for the questions involving name remembering) but the auxiliary assumption is not directly tested by the causal model. It should be noted that reliance on auxiliary measurement assumptions is not specific to causal modeling, but is a general requirement for any type of empirical research (Meehl, 1978; Messick, 1981).

The research design must establish a sequence and timing of empirical observation appropriate for studying the system of variables. In developmental applications, the research design will often involve collection of longitudinal data so that individual differences in change on both cause and effect variables may be measured. Critical issues for the longitudinal design include the age range covered and the spacing of measurement occasions (Dwyer, 1983; Heise, 1975; Kessler & Greenberg, 1981). The spacing must be wide enough to detect change, to enable assumptions of equilibrium in causal effects, and to avoid reactive effects of retesting that may alter the measurement properties of the variables, yet sufficiently frequent to isolate critical developmental phases of change and to capture functional forms of change.

These design and measurement decisions precede the statistical analysis, and in fact determine the extent to which the statistical analysis is truly informative with respect to the viability of the theoretical causal model. Given an appropriate design, one can then subject the causal model, as represented in the SEM, to a falsification test. One of the principal advantages of covariance structures programs such as LISREL (Joreskog & Sorbom, 1984) and EQS (Bentler, 1985) is that they use restricted estimation procedures (for example, full-information maximum likelihood estimation) that not only give consistent and unique parameter estimates but provide indices of fit of the model to the data. A poor overall statistical fit of the model, and/or implausible parameter estimates in particular parts of the model, would give cause to reject the theoretical model that generated the SEM provided that we assume the validity of the measures, the distributional assumptions of the SEM algorithm, and the appropriateness of the observational design for capturing the causal process under study.

In sum, the requirements for sound causal inference using SEM are indeed formidable -- and perhaps, rarely met in practice. The risk of misuse of the techniques is great and is compounded by naive beliefs that a program like LISREL can automatically convert a correlation matrix among variables into a set of statements about cause-effect relations. Small wonder that many psychologists have issued strong warnings about the method and its potential abuses (e.g., Cliff, 1983). As suggested above, such caveats are well taken, but we should not confuse bad causal analysis with exploratory use of SEM for descriptive purposes.

Disambiguating the exploratory/confirmatory distinction. I argue that use of SEM for descriptive research and exploratory model building can be a valid and informative research practice. A full appreciation of this argument requires acceptance of the premise that we need to cease using exploratory and confirmatory as universal labels describing the differences among various multivariate techniques, especially factor analysis. Some progress in this regard was made by Nesselroade and Baltes (1984) in their recent review of factor analysis and modeling techniques. Nesselroade and Baltes (1984) characterized traditional factor analysis and newer SEM approaches by distinguishing between research purpose (theoretical orientation) and analysis procedure. Both of these aspects of research method can be dichotomized into exploratory versus hypothesis testing (confirmatory) classes, as illustrated in Table 1. The cells in their 2 X 2 classification scheme were discussed in terms of convergence of purpose and method (e.g., use of exploratory factor analysis to perform exploratory research) on the one hand, and divergence of purpose and method (e.g., using exploratory factor analysis to test hypotheses about factor structure). The distinction between purpose and procedure is useful (see also Hertzog, 1985b). Certainly, much of the early literature in factor analysis of psychometric abilities -- particularly Thurstone's pioneering work on primary mental abilities (Thurstone, 1938; Thurstone & Thurstone, 1941; Thurstone, 1944) -- can be characterized as use of exploratory methods to do confirmatory analysis (perhaps, because confirmatory methods were not yet available; see Bechtoldt, 1974). With respect to selection of variables to define abilities and hyperplanes, Thurstone's work is certainly exemplary of hypothesis-testing by selection of marker variables to measure factors specified a priori. However, Nesselroade and Baltes' (1984) use of "confirmatory" and "exploratory" as labels for classification levels was in one sense unfortunate, for it directed them to focus on the divergence between exploratory analysis purposes and confirmatory techniques as exposure of an "Achilles heel" of confirmatory factor analysis (cell III of Table 1). Modified models, with the express purpose of improving statistical fit to sample data, may capitalize on chance. They stated:

> "Obviously, some modification of models is reasonable when one is elaborating a theoretical framework, but, in successive modification to fit a particular data set, there is surely a point reached where one is no longer doing hypothesis testing but, rather, exploratory analysis. What may have begun as a proper [emphasis added] hypothesis-testing study has become essentially an exploratory activity. The final model, then, should be viewed as a new hypothesis and ought to be subjected to a rigorous cross-validation against new data." (pp. 272-273)

Nesselroade and Baltes (1984) are absolutely correct on this score, and they are echoing here appropriate warnings regarding model modifications that can be found in almost every introductory treatment of SEM (e.g., Herting, 1985; Long, 1983a,b). In what sense, then, is their classification unfortunate? Only in that the scheme of Table 1 directs one to equate implicitly cell III, exploratory purposes/confirmatory techniques, with divergence (or inconsistency). A possible implication is that such a use of the technique is suboptimal if not actually improper.

Figure 1 presents a schematic encapsulating an alternative perspective on the difference between traditional multivariate techniques (e.g. factor analysis, canonical correlation) and the general class of covariance structures models. Figure 1 is a nested hierarchy, in which there are four levels (General Orientation, Specific Purpose, Analysis Technique, and Statistical Procedure). At the highest level, research is either explanatory or descriptive in orientation.[2] From this perspective, both descriptive and explanatory research may have either confirmatory or exploratory research purposes (or perhaps, some mixture of the two). The distinction between these is a function of whether the research is explicitly designed to test <u>substantive</u> research hypotheses. A substantive hypothesis is a statement of fact regarding constructs or relations among constructs, assuming the critical/realist perspective that constructs are things-in-the-world that can be measured (e.g., Cook & Campbell, 1979; Cronbach & Meehl, 1955; Messick, 1981). Confirmatory research tests substantive hypotheses formulated a priori; exploratory research does not.

In contrast to Nesselroade and Baltes (1984), <u>confirmatory</u> and <u>exploratory</u> are terms reserved <u>exclusively</u> for the level of research purposes. <u>Unrestricted</u> and <u>restricted</u> are the terms used for different analytic techniques, represented at the third level of the hierarchy. Traditional unrestricted factor analysis produces a nonunique solution that can be transformed (rotated) into an infinite number of alternative solutions (e.g., Mulaik, 1972). Restricted factor analysis achieves a solution by a model specification that reduces the number of unknown parameters in the linear equations so that each parameter is uniquely identified (Joreskog, 1969). The final solution in restricted analysis is determined, not indeterminate, and the focus becomes the degree to which the restricted model still provides a parsimonious fit to the sample data. The restricted/unrestricted distinction more aptly captures the basis for the factor analytic procedures themselves (e.g., Joreskog, 1969) and also serves to eliminate the semantically reflexive one-to-one identification of technique with purpose.

The fourth and lowest level of the hierarchy is based upon the distinction between <u>substantive hypothesis testing</u> and <u>statistical hypothesis testing</u>. A statistical hypothesis is a hypothesis about statistical parameters in some population. An example of a statistical hypothesis is a t-test for equality of two population means using data from independent samples. The fourth level discriminates multivariate research that tests specific hypotheses from research that relies on general fit assessments (e.g., an overall RMSR or alternative index of fit) or subjective model evaluations. It is not always the case that a statistical test is a meaningful test of a substantive hypothesis (e.g., Steyer, 1985). It is if and only if the substantive hypothesis drives the research design and leads to a correspondence between the alternative hypothesis (classically, H_1) of the statistical test and the substantive hypothesis. Rejection of an ANOVA F-test of the omnibus null hypothesis (all cell means equal) implies adoption of the alternative hypothesis (not all means equal). But this alternative hypothesis usually does not correspond to a specific substantive hypothesis regarding mean group differences. In theory, a planned comparison across levels of the factor <u>might</u> correspond more closely to a specific substantive hypothesis.

Clearly, there are many implications of the concepts implied in Figure 1, and I shall not attempt to delineate all the possibilities or limitations of this diagram as a heuristic for conceptualizing SEM research. There are many empty nodes in the hierarchy, in part because I am not interested in extending the conceptual scheme into a fully elaborated taxonomy of multivariate research methods.[3] Instead, I shall focus on the implications of the schematic in Figure 1 for understanding exploratory research from a modeling perspective. The distinction between statistical hypotheses and substantive hypotheses is critical in this regard.

In SEM, statistical hypotheses about model parameters are usually tested from sample data on the basis of distributional assumptions. This is generally done by computing the difference in likelihood ratio X^2 test statistics between two nested models (e.g., Joreskog, 1971, 1974). This difference can be treated as a likelihood ratio test of the validity of the restrictions imposed on the more restricted of the nested models. The procedure for calculating the statistical test is analagous to likelihood ratio tests in log-linear models for categorical data (Herting, 1985). This analogy is important, because usually such tests in log-linear analysis are essentially descriptive (and exploratory) in nature. The same is true in SEM! So the distinctions inherent in Figure 1 make it possible to argue that it is perfectly legitimate to conduct <u>exploratory research using restricted estimation techniques to perform statistical hypothesis tests</u>. This is the hierarchical chain highlighted by bold lines in Figure 1. And indeed, many of the published applications of LISREL fall in this category. For example, careful perusal of the recommended tests for group equivalence in simultaneous factor analysis (e.g., Alwin & Jackson, 1981; Joreskog 1971) indicates that they are statistical tests of group equivalence in entire factor analysis parameter matrices (e.g., the factor covariance matrix). The null hypothesis is that the groups have equal factor variances and covariances; the alternative hypothesis is that not all these parameters are equal across the two groups. Usually, these omnibus null hypotheses of group equivalence are examined <u>without</u> a priori hypotheses regarding the specific differences expected among the groups. These tests have the same status -- as statistical tests -- as do the overall tests in ANOVA or log-linear analysis. If the hypothesis is rejected, further study of the specific sources of the group differences is required.[4] They also have the same logical status vis a vis substantive hypotheses: they are exploratory, for no specific substantive hypothesis corresponds to the alternative hypothesis. They are descriptive because no plausible explanation for the differences is entertained as part of the study. The critical point is that such applications represent important and legitimate uses of SEM, but they are not confirmatory analyses!

How are we to explain, under this perspective, the Achilles heel identified by Nesselroade and Baltes (1984)? Such research (probably descriptive in nature) may be characterized as exploratory research being conducted with restricted methods, and without the use of specific hypothesis tests. Instead, a generic assessment of fit is used to assess model adequacy, and parameters are added until a reasonable level of fit is achieved. This is not a problem per se. In the absence of a strong foundation of theory and corroborating evidence, it is perfectly legitimate to conduct a specification search (Leamer, 1978) and this can be meaningfully done using restricted estimation procedures and model fit diagnostics (e.g., Herting & Costner, 1985; Saris, Pijper, & Zegwaard, 1980). Research that implicitly engages in a search for a new model specification becomes problematic when the investigator fails to recognize and appreciate the exploratory nature of the research and its inherent limitations. Any attempt to confer upon the research the status of a confirmatory analysis (via model disconfirmation) is inappropriate.

Moreover, interpretation of results and subsequent research activities must be based upon the exploratory nature of the analysis. Searching for a model that fits sample data may result in a final model that is specific to the sample at hand. Thus the model must cross-validate in an independent sample before any confidence can be placed upon the additional parameter estimates as reflections of population relationships. Even if the new parameters replicate, a confirmatory orientation to research would take such replication as a first and rather meager step. Subsequently, the theoretical meaning of the model in the context of the new parameters must be re-evaluated. This evaluation should generate new, empirically testable, hypotheses consistent with the interpretation given to the new parameters. Finally, a model based upon these new hypotheses must survive disconfirmation tests with an independent sample (and undoubtedly, a new set of observed variables) before it can be considered "confirmed."

Hypothesis Testing in SEM. Treatments of the topic of SEM usually distinguish between at least three different components of modeling: model specification, parameter estimation, and evaluation of model fit (e.g., Joreskog & Sorbom, 1979). Little can be said in this paper regarding the fundamentals of SEM. However, the treatment of descriptive hypothesis testing in SEM requires some discussion of statistical evaluation of model fit.

Hypothesis testing in SEM is critically dependent upon the use of restricted estimation techniques. Given a uniquely identified model, restricted approaches fit the model to the sample data, estimating the unknown parameters on the basis of the restrictions imposed upon other the model. In principle, the restricted model will not fit the data exactly, given 1) model misspecification, and 2) sampling fluctuation. Maximum likelihood estimation techniques rely upon assumptions regarding sampling distributions from covariance matrices (of variables that are distributed multivariate normal) to account for the expected lack of fit due to sampling fluctuation. The likelihood ratio X^2 test statistic may be used to determine whether the lack of fit of model to data is plausibly considered as statistical chance. The critical issues for evaluation of X^2 include statistical power and consequences of assumption violations, etc. (e.g., Bentler & Bonett, 1980; Huba & Harlow, 1986).

The logic of model comparisons, in terms of differential fit, depends upon the degree to which models are <u>nested</u>. One can always compare two models for relative fit, but such a comparison cannot be linked to a specific test of a statistical hypothesis unless the models are nested. The concept of nested models is fairly well discussed in the literature (e.g., Joreskog & Sorbom, 1979). Define two models, M_1 and M_2 with sets of specified parameters P_1 and P_2. Assume that M_2 contains more restrictions than M_1 (i.e., fewer parameters to be estimated, more degrees of freedom). Models M_1 and M_2 can be considered nested if 1) P_1 can be partitioned into two subsets, $\{P^*, P_3\}$, 2) P_2 can be partitioned into two subsets, $\{P^*, P_4\}$, 3) P^* is identical in both partitioned sets, and 4) the parameters in subset P_4 may be represented as additional restrictions on the parameters in subset P_3. In that case, we may treat M_1 and M_2 as having the same <u>basis</u> <u>specification</u> and differing <u>only</u> in the additional restrictions imposed on parameter subset P_3 to obtain the subset P_4. An example in restricted factor analysis would be two models specifying identical relationships of variables to factors, but differing in that one model allowed all factor correlations to be freely estimated (an oblique solution) and one model forced all factor correlations to be equal to zero (an orthogonal solution). Given identical basis specifications defining the factors in terms of observed variables, the model with all factor correlations fixed at zero is a more restricted model than the one allowing factor correlations to be free.

Each nested model generates a separate X^2 test statistic reflecting its fit to the sample data. The null hypothesis tested by X^2 is that the sample data matrix was drawn from a population matrix generated by the model as specified. Significant X^2 test statistics reject the hypothesis that the model as specified generated the sample matrix. In model comparisons, interest is not in the absolute levels of X^2 but in the relative difference in fit. Given nested models M_1 and M_2, as before, with tests of fit X^2_1 (with <u>m</u> <u>df</u>) and X^2_2 (with <u>n</u> <u>df</u>, $m < n$), respectively, then $X^2_1 - X^2_2$ is distributed asymptotically as X^2

with m - n df. Given the nesting, this difference in X_2 may be treated as a test of the null hypothesis that the additional restrictions imposed in P_4 are true (conditional on the truth of the basis specification). Joreskog (1974) describes a full range of applications of this type of hypothesis testing in SEM.

In the case of the two nested models -- one with orthogonal and one with oblique common factors -- the difference in X^2 may be taken as a test of the null hypothesis that all factor correlations equal zero.

Descriptive SEM in Developmental Research

Describing Development in Latent Variable Models

The preceding discussion suggests that exploratory research using restricted techniques to conduct meaningful statistical tests is a legitimate research practice. In the case of exploratory developmental research on longitudinal data, this approach is arguably an optimal exploratory method. Certainly, it is the case that meaningful descriptive research hypotheses are best structured by either complex exploratory methods (e.g., Meredith & Tisak, 1984, 1986) or restricted SEM techniques. This paper shall focus on the latter approach.

Why is SEM optimal for descriptive developmental research? There are two reasons. One is the relative inadequacy of traditional unrestricted multivariate techniques to structure longitudinal data in a way that relates meaningfully to descriptive research hypotheses of interest. The second is the ability of SEM to provide actual statistical tests of hypotheses of interest.

An empirical example will assist us to understand the limitations of simple unrestricted factor analysis in longitudinal research. Consider an analysis by Hertzog (1979) of data from Schaie's Seattle Longitudinal Study (SLS; see Schaie, 1983). Hertzog (1979) factored the correlations among five subtests from Thurstone and Thurstone's (1949) Primary Mental Abilities test. The five subtests are Verbal Meaning (a test of recognition vocabulary), Space (a test of ability to visualize spatial rotation in two dimensions), Reasoning (a test of induction), Number (a measure of addition skills), and Word Fluency (ability to generate words on the basis of nonsemantic rules -- in this case, words beginning with the letter "S"). The five tests were given to members of the SLS that participated at three times of measurement in the longitudinal study (1956, 1963, 1970). An exploratory factor analysis of these data recovers five test-specific factors, one for each ability, with tests from each occasion loading on that factor. Table 2 reports the salient factor weights and factor correlations from the oblique factor solution. Clearly, there is a high degree of consistency in individual differences on the subtests over time, such that the test-specific correlations are the largest in the longitudinal correlation matrix.

This analysis, although somewhat informative, is unsatisfactory as a reflection of change processes in the longitudinal data set. Baltes and Nesselroade (1970, 1973) identified the need to separate changes in factor structure (invariance in the relationships of variables to factors) from changes in factor scores (individual differences in changes in levels of the underlying factors). One must be able to differentiate these two types of change in order to have unambiguous inferences regarding the nature of development.

Why is the differentiation of the two types of change in latent variable models important? Changes in factor structure suggest the possibility of qualititative change in the nature of the factors, or at least, in the relationship of variables to factors. Substantively, qualitative changes in the underlying constructs are an issue of central importance in developmental psychology (Kagan, 1980; Lerner, 1986), and the hypothesis of qualitative change implies shifts in the factor structure of measures of the constructs under investigation (Baltes & Nesselroade, 1970, 1983; Labouvie, 1980a). If the same constructs are being measured at each age (or for different subpopulations), then one has <u>construct</u> <u>equivalence</u> (Schaie & Hertzog, 1982, 1985) or <u>conceptual</u> <u>equivalence</u> (Labouvie, 1980a).

Even if the same variables are valid measures of a given construct at different ages, they may not have equivalent measurement properties (i.e., equal reliabilities, equal metrics, equal validity; Baltes, Reese, & Nesselroade, 1977; Labouvie, 1980a,b; Schaie & Hertzog, 1985). Variables having equivalent measurement properties with respect to underlying constructs are defined as having <u>measurement</u> <u>equivalence</u> (e.g., Baltes & Nesselroade, 1973; Baltes et al., 1977; Labouvie, 1980a,b). A critical reason for careful examination of invariance in factor structure over time is that such invariance is prima facie evidence for both construct and measurement equivalence at different ages (Baltes & Nesselroade, 1970, 1973). On the other hand, changes in factor structure would suggest lack of construct equivalence, measurement equivalence, or both. One important implication of a lack of factorial invariance is that one could not have confidence in quantitative comparisons of age changes in levels on observed variables in the absence of measurement equivalence (Baltes & Nesselroade, 1973).

Multivariate longitudinal sequences are critical for investigation of the hypothesis of factorial invariance at different developmental levels, but only if the longitudinal sequences can be structured so as to directly reflect changes in factor structure with increasing age. The unrestricted factor solution of Table 2 is uninformative about the change in factor structure from Time 1 to Time 3, because the factors collapse over the different longitudinal occasions.

The calculation of test-specific factors also frustrates any ability to examine changes in factor scores across age levels. Assuming construct and measurement equivalence, then the factor scores become optimal operational definitions of the individual differences on the constructs of interest. Thus, providing that factorial invariance can be demonstrated, examination of the factor scores enables the developmental scientist to focus on critical developmental research questions (Baltes & Nesselroade, 1979): (1) are there average age changes in levels of multiple constructs; (2) are there differences in average age changes across different constructs (termed multidirectionality by Baltes and his colleagues [e.g., Baltes & Willis, 1977]) (3) are there individual differences in change (identified by Baltes et al. (1977) as <u>interindividual differences in intraindividual change</u>); and (4) what are the predictors of the individual differences in change?

The investigation of individual differences in developmental change is arguably one of the most critical, and yet most neglected, issue confronting life-span developmental psychology (e.g., Baltes et al., 1977; Baltes & Nesselroade, 1973; Hertzog, 1985a; Labouvie, 1980a). The logical converse of change is stability, and several authors have discussed multiple kinds of stability that conform to the questions about change identified above. The most common distinction drawn is between mean level stability (no change in average intraindividual change patterns) and stability of individual differences, indicating a lack of individual differences in change patterns (e.g., Baltes et al, 1977; Bengtson, Reedy, & Gordon, 1985; Costa & McCrae, 1980; Kagan, 1980; Mortimer, Finch, & Kumka, 1982; Schaie & Hertzog, 1985). Parallel profiles of developmental change across different individuals imply perfect stability of individual differences over time, as operationally defined by correlations of variables between different longitudinal occasions. This stability is logically and mathematically independent (in normally distributed variables) from mean level stability.

Baltes et al. (1977) provide an excellent illustration of the concept of perfect (covariance) stability, reproduced in Figure 2. Its top panel depicts individual differences in change; its bottom panel depicts data in which there is mean change as individuals grow older, but no individual differences in change (i.e., the lines are perfectly parallel). With observed variables, such parallelism will almost certainly not be observed because of measurement error. SEM techniques can determine whether such parallelism holds for the latent variables.

Obviously, the unrestricted factor solution reported in Table 2 cannot address the question of stability of individual differences over time either. Although the test-specific factor structure can be considered an indirect reflection of strong stability of individual differences, it is inadequate because (1) stability is estimated only at the level of the observed variables, not the factors, (2) the magnitude of the factor loadings is influenced by measurement error in the observed variables, thus confounding lack of stability with less-than-perfect reliability, and (3) the magnitude of the factor loadings is a function of stability of individual differences on the test-specific components as well as the components that are determined by the latent variables of interest. The effect of unreliability is to lower the loadings and reduce the apparent stability; the effect of stable individual differences in test-specific components will usually increase the factor loadings (relative to the stability of individual differences in the underlying constructs). Thus one cannot use the unrestricted factor solution to estimate the stability of individual differences in the latent variables, nor can one identify other variables that predict differential change patterns. We can only address the relative stability of each manifest variable, which is only in part determined by stability of individual differences in the latent variables the investigator truly wishes to study.

Using SEM for Longitudinal Factor Analysis

The interpretive problems with unrestricted factor analysis of longitudinal data were recognized some time ago (e.g., Tucker, 1964), and a number of technical improvements evolved, most based upon canonical analysis, for approaching the problem of defining a factor structure within each longitudinal occasion and examining its invariance over time (e.g., Bentler, 1973; Corballis, 1973; Corballis & Traub, 1970; Meredith & Tisak, 1984, 1986).

Restricted modeling techniques are also applicable to the problem. The critical advantage of SEM for descriptive longitudinal research is that it enables one to structure the data so that configuration of model parameters is directly relevant to all the major developmental questions identified above, including the issues of measurement equivalence. SEM does not merely lead to models relevant to the questions. Its second major benefit is that it provides a means of structuring statistical hypothesis tests that are directly linked to the major hypotheses of interest. That is, it is possible to test hypotheses that factor loadings are invariant over occasions, that factor variances increase over time, that individual differences in sets of latent variables are perfectly stable across age levels, and so on.

The use of SEM for longitudinal analysis illustrates the use of statistical hypothesis testing procedures to perform descriptive research. In the examples to be given below, taken from Hertzog and Schaie (1986, in press), only longitudinal factor analyses will be discussed. These analyses employ only the SEM measurement model, and do not utilize structural regression equations. The general principles apply equally, however, to models structuring regression relationships among the latent variables in a longitudinal data set.

The longitudinal factor analysis model discussed here is characterized by the a priori specification of <u>occasion-specific</u> factors -- that is, it specifies a separate factor structure at each longitudinal occasion (Horn & McArdle, 1980;

Joreskog 1979; Joreskog & Sorbom, 1977). Figure 3 illustrates a hypothetical model for two latent variables, each measured by three observed indicators, at three longitudinal occasions (age levels).[5] The factors are occasion-specific in that no regression relationships are allowed between the factors at one occasion and observed variables at the other occasion (e.g., the two factors at Time 2 have no regression coefficients leading to the observed variables at Times 1 and 3). The model allows the factor covariance matrix to be freely estimated, although the different elements of the matrix have different developmental interpretations.

The model has the potential for allowing for the test-specific relationships identified in Table 2, <u>but these test-specific relationships are modeled independently of the longitudinal factors</u>. These relationships are modeled as either 1) covariances among the residuals defined by the occasion-specific factors, or 2) <u>test-specific</u> factors (Jorekog, 1970) in addition to the occasion-specific factors. Generally speaking, the model shown in Figure 3 must be expanded to include one of these two options. Several investigators (e.g., Sorbom, 1975) have found that measurement model residuals in longitudinal data contain reliable components of variance which are specific to the observed variable. Failure to specify these specific relationships will bias estimates of the occasion-specific factor relationships (Hertzog & Schaie, 1986). Of course, the specification of subdiagonals of residual covariances may be considered as an empirically testable hypothesis, for it is nested with the model fixing all such residual covariances to zero.

The restricted longitudinal factor analysis of Schaie's longitudinal data differs fundamentally from the unrestricted factor analysis reported in Table 2. Hertzog and Schaie (1986, in press) used SEM to model an occasion-specific general intelligence factor (*g*) at each longitudinal occasion. The model tested by Hertzog and Schaie (1986a) is depicted in Figure 4. The *g* factor is a single occasion-specific factor. The model also specified residual covariances among each of the five PMA subtests to represent the test-specific relationships so prominent in the unrestricted factor analysis. Certainly these residual covariances are theoretically meaningful, for one would expect a component in each test specific to the primary ability measured (e.g., verbal comprehension ability) that will be independent of *g*. The residual covariances suggest that there is stability in individual differences in the primary ability components as well as in *g*.

Table 3 summarizes a series of hypothesis tests conducted on the data. The model with *g* and the residual covariances fits the data quite well, arguing for the validity of the general factor as a representation of the covariances among the observed measures. The overall X^2 is not significant, and the Bentler-Bonett normed fit index is well above .9, often taken as an informal criterion for model adequacy. Configural invariance appears to hold, in the sense that a single factor is plausible for all three occasions. This impression was confirmed by the small residual correlations (differences between predicted correlations, from the parameter estimates, and the actual sample correlations). Configural invariance at the level of *g* does not imply configural invariance in the primary abilities, of course, but these data cannot be used to address that question.

The critical test of equivalent factor pattern weights involves the comparison of the first and second models (specifying metric invariance). Table 3 shows that this test is not significant, and indeed, is less than the expected value of 8 (the number of df). Thus we can consider the hypothesis of equivalent factor loadings to be tenable, and can conclude that g is defined equivalently at each longitudinal occasion. A subsequent test of equal g factor variances was rejected. This suggests that individual differences are not of equal magnitude across the longitudinal occasions, one suggestion of different patterns of change. The estimated factor covariances were uniformly high, with standardized correlations all greater than .9. Thus the changes in variance were not associated with low autocovariance, which suggests that individual differences in change were not sufficiently large to cause major shifts in the rank order of individuals. One possible explanation, of course, is that the magnitudes of change were substantially correlated with initial level (with individuals at the bottom of the distribution more likely to decline than individuals at the top).

Hertzog and Schaie (1986) pooled the data used above with another longitudinal sample to form multiple age groups (young, middle-aged, and old) of individuals with longitudinal data. Simultaneous multiple group factor analysis was then performed, using SEM to test whether the solutions for the multiple age groups were equivalent. These analysis showed that the hypothesis of metric invariance in g factor loadings could not be rejected across the age groups (as well as across longitudinal occasions. The analyses also showed that all three age groups had comparable, high levels of covariance stability in g across the fourteen-year age interval.

Hertzog and Schaie (in press) extended the model to include the means of the PMA variables as well. A series of nested models were used to test descriptive hypotheses about the g factor means. The primary result was that, although all age groups had been shown to have comparable levels of stability of individual differences, the groups differed dramatically in g factor mean changes. The young adult group showed increment in g during early adulthood; the middle-aged group showed stability in g means, but the old group showed pronounced decline in g over the fourteen-year age interval. The analysis also used nested models to show that the model with g factor means alone was insufficient to account for the means of the PMA subtests. Instead, it was necessary to model mean changes in the test-specific components of the PMA that were orthogonal to g. This result was of course consistent with reports by Schaie (1983) and his colleagues that there are different patterns of mean age changes in intelligence for different PMA subtests. The important methodological point is that the approach of nested SEM models could be used to describe accurately the patterns of PMA test performance (at both the mean and covariance levels) on the basis of a latent variable model.

Concluding Comments

This paper has sought to show that SEM techniques provide a useful means for guiding descriptive research on developmental change. First, a theoretical position was articulated that argues strongly against superficial equation of SEM techniques with confirmatory and/or explanatory analysis. Although SEM techniques are certainly useful and appropriate in both instances, they are also highly useful in exploratory and descriptive research. Second, the basis for valid use of SEM in descriptive developmental research was briefly discussed and illustrated, emphasizing SEM applications to longitudinal data. As we have seen, SEM makes it possible to translate fundamental developmental questions into directly testable hypotheses about SEM parameters. The discussion here merely scratches the surface of potential applications of the methods. In particular, the use of SEM to specify full structural regression models makes it possible to perform descriptive research that tests hypotheses regarding lagged and simultaneous relationships (e.g., Dwyer, 1983; Joreskog & Sorbom, 1977; Rogosa, 1980; Schaie & Hertzog, 1985). Many of these applications involve the use of cross-lagged regression (or autoregression) models, but there are other important SEM approaches for developmental description that avoid the autoregressive approach (e.g., McArdle & Epstein, 1987).

It is important to emphasize in closing that the true power of SEM methods lies in the ability of the investigator to link statistical procedures to substantive questions. Given a clever design, the techniques outlined here can be used to make fine discriminations among meaningfully different theoretical models for behavioral development. Such applications would rightfully eschew any sort of generic, descriptive hypothesis testing (e.g., testing group equivalence of entire parameter matrices) in favor of statistical tests specifically tailored to reflect substantive hypotheses. With regard to theoretical model building, life-span developmental psychology may still need to crawl a little longer before walking. If so, then the type of approach described here may prove helpful in bringing us to our feet.

Footnotes

1. This paper uses the LISREL framework for a discussion of model specification and hypothesis testing, for two reasons. First, it is still the most widely available and familiar computer program for SEM, although Bentler's (1985) EQS may ultimately surpass it in use. Second, much of the work on longitudinal models has been done by Joreskog et al., and use of LISREL models and notation facilitates additional scholarship by the interested reader. The decision to make the LISREL model the basis for this paper should not be taken as an implied endorsement of the LISREL computer program over competing software.

2. The discussion of this hierarchical schematic is necessarily limited in length and ignores some important and subtle points. For example: One could include intervention as a general orientation, but it is sufficiently similar to explanation that its inclusion would make Figure 1 needless complicated; dichotomies such as restricted/unrestricted analysis techniques are actually more like fuzzy continua; research purposes are in practice some combination of confirmatory and exploratory goals, and even confirmatory research often unintentionally leads to novel scientific discovery unanticipated by the researcher.

3. For example, the empty node for unrestricted technique, statistical tests

reflects the difficulties inherent in formally testing specific statistical hypotheses with unrestricted multivariate methods. But many aspects of statistical tests are employed in such methods (e.g., X^2 test for the number of factors in Rao or Lawley/Maxwell maximum likelihood factor analysis; tests of multiple $R = 0$ in canonical correlation).

4. One can generate multiple comparisons for SEM parameters, including means, variances, correlations, and regression coefficients (within samples and between multiple groups) by specifying contrasts across parameter matrices and using the covariance matrix of the estimators to calculate appropriate standard errors. Post-hoc analysis of longitudinal changes in means, variances, etc. should logically be done in this fashion; simple comparison of confidence interval overlap would lack power and be technically incorrect, due to the dependence among the parameter estimates (e.g., Steiger, 1980). This is rarely done in practice.

5. In longitudinal sequences, occasions and age levels are not synonymous, which is an important design feature. The discussion of longitudinal factor models shall act as if they are. This enables unambiguous reference to "occasion-specific" features of the model, and makes it easier for the reader to refer to original sources (e.g., Joreskog & Sorbom, 1977) that discuss models where occasion is the primary definition of longitudinal variation, and where occasion of measurement and age are in fact synonymous (i.e., completely confounded, as in the single cohort longitudinal design).

REFERENCES

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

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

Baltes, P.B., Dittman-Kohli,F., & Dixon, R.A. (1984). New perspectives on the development of intelligence in adulthood: Toward a dual-process conception and a model of selective optimization with compensation. In P.B. Baltes & O.G. Brim, Jr. (Eds.), Life-span development and behavior (Vol. 6, pp. 34-76). New York: Academic Press.

Baltes, P.B., and Nesselroade, J.R. (1970). Multivariate longitudinal and cross-sectional sequences for analyzing ontogenetic and generational change: A methodological note. Developmental Psychology, 1, 162-168.

Baltes, P.B., and 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. and Nesselroade J.R. (1979). History and rationale of longitudinal research. In J.R. Nesselroade & P.B. Baltes (Eds.), Longitudinal research in the study of behavior and development. New York: Academic Press.

Baltes, P.B., Reese, H.W., and Nesselroade, J.R. (1977). Life-span developmental psychology: Introduction to research methods. Monterey, CA: Brooks/Cole.

Baltes, P.B., and Willis, S.L. (1977). Towards psychological theories of aging and development. In J.E. Birren & K.W. Schaie (Eds.), Handbook of the psychology of aging (pp. 128-147). New York: Van Nostrand Reinhold.

Bechtoldt, H.P. (1974). A confirmatory analysis of the factor stability hypothesis. Psychometrika, 39, 319-326.

Bengtson, V.L., Reedy, M.N., and Gordon, C. (1985). Aging and self-conceptions: Personality processes and social contexts. In J.E. Birren & K.W. Schaie (Eds.), Handbook of the psychology of aging (2nd ed., pp. 544-593). New York: Van Nostrand Reinhold.

Bentler, P.M. (1973). Assessment of developmental factor change at the individual and group level. In J.R. Nesselroade & H.W. Reese (Eds.), Life-span developmental psychology: Methodological issues (pp. 145-174). New York: Academic Press.

Bentler, P.M. (1982). Linear simultaneous equation systems with multiple levels and types of latent variables. In K.G. Joreskog & H. Wold (Eds.), Systems under indirect observation: Causality, structure, prediction (Part1) (pp. 101-129). Amsterdam: North Holland.

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

Bentler, P.M. (1985). Theory and implementation of EQS: A structural equations program. Los Angeles: BMDP Statistical Software.

Bentler, P.M., and Weeks, D.G. (1979). Interrelations among models for the analysis of moment structures. Multivariate Behavioral Research, 14, 169-185.

Blalock, H.M., Jr. (1982). Conceptualization and measurement in the social sciences. Beverly Hills, CA: Sage.

Blalock, H.M., Jr. (Ed.). (1985a). Causal models in panel and experimental designs. Chicago: Aldine.

Blalock, H.M., Jr. (Ed.). (1985b). Causal models in the social sciences (2nd ed.). Chicago: Aldine.

Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In K.G. Joreskog & H. Wold (Eds.), Systems under indirect observation: Causality, structure, prediction (Volume 1) (pp. 149-173). Amsterdam: North Holland.

Browne, M.W. (1984). Asymptotically distribution free methods in the analysis of covariance structures. British Journal of Mathematical and Statistical Psychology, 37, 62-83.

Campbell, R.T. and Mutran, E. (1982). Analyzing panel data in studies of aging: Applications of the LISREL model. Research in Aging, 4, 3-41.

Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. Multivariate Behavioral Research, 18, 115-126.

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

Corballis, M.C. (1973). A factor model for analyzing change. British Journal of Mathematical and Statistical Psychology, 26, 90-97.

Corballis, M.C., and Traub, R.E. (1970). Longitudinal factor analysis. Psychometika, 35, 79-98.

Costa, P.T. Jr., and McCrae, R.R. (1980). Still stable after all these years: Personality as a key to some issues in adulthood and old age. In P.B. Baltes & O.G. Brim, Jr. (Eds.), Life-span development and behavior (Vol. 3, pp. 65-102). New York: Academic Press.

Cronbach L.J., and Meehl, P.E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

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

Dixon, R.A. and Hultsch, D.F. (1983). Stucture and development of metamemory in adulthood. Journal of Gerontology, 38, 682-688.

Duncan, O.D. (1975). Introduction to structural equation models. New York: Academic Press.

Dwyer, J.H. (1983). Statistical models for the social and behavioral sciences. New York: Oxford University Press.

Featherman, D.L. (1985). Individual development and aging as a population process. In J.R. Nesselroade & A. Von Eye (Eds.), Individual development and social change: Explanatory analysis (pp. 213-241). New York: Academic Press.

Heise, D.R. (1975). Causal analysis. New York: John Wiley & Sons.

Herting, J.R. (1985). Multiple indicator models using LISREL. In H.M. Blalock, Jr. (Ed.), Causal models in the social sciences (2nd ed.) (pp. 263-319). Chicago: Aldine.

Herting, J.R., and Costner, H.L. (1985). Respecification in multiple indicator models. In H.M. Blalock, Jr. (Ed.), Causal models in the social sciences (2nd ed.) (pp. 321-393). Chicago: Aldine.

Hertzog, C. (1979). A structural equations analysis of adult intellectual development. Unpublished doctoral dissertation, University of Southern California, Los Angeles.

Hertzog, C. (1985a). An individual differences perspective: Implications for cognitive research in gerontology. Research on Aging, 7, 7-45.

Hertzog, C. (1985b). Applications of confirmatory factor analysis to the study of intelligence. In D.K. Delterman (Ed.), Current topics in human intelligence (pp.59-97). Norwood, N.J.: Ablex.

Hertzog, C. (1987). Applications of stuctural equation models in gerontological research. In K.W. Schaie (Ed.), Annual review of gerontology and geriatrics (Vol. 7).

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

Hertzog, C., and Schaie, K.W. (in press). Stability and Change in Adult intelligence: 2. Simultaneous analysis of longitudinal means and covariance structures. Psychology and Aging.

Horn, J.L. and McArdle, J.J. (1980). Perspectives on mathematical/statistical model building (MASMOB) in research on aging. In L.W. Poon (Ed.), Aging in the 1980's: Psychological issues (pp. 503-541). Washington, D.C.: American Psychological Association.

Huba, G.J. and Harlow L.L. (1986). Robust estimation for causal models: A comparison of methods in some developmental data sets. In R.M. Lerner & D.L. Featherman (Eds.), Life-span development and behavior (Vol. 6). New York: Academic Press.

James, L.R., Mulaik, S.A., and Brett, J.M. (1982). Causal analysis: Assumptions, models, and data. Beverly Hills, CA: Sage.

Joreskog, K.G. (1969). A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 34, 183-220.

Joreskog, K.G. (1970). Estimation and testing of simplex models. British Journal of Mathematical and Statistical Psychology, 23, 121-145.

Joreskog, K.G. (1971). Simultaneous factor analysis in several populations. Psychometrika, 36, 409-426.

Joreskog, K.G. (1974). Analyzing psychological data by structural analysis of covariance matrices. In D.H. Krautz, R.C. Atkinson, R.D. Luce, & P. Suppes (Eds.), Contemporary developments in mathematical psychology (Vol. 2, pp. 1-56). San Francisco, CA: W.H. Freeman.

Joreskog, 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.

Joreskog, K.G. and Sorbom, D. (1977). Statistical models and methods for analyses of longitudinal data. In D.S. Aigner & A.S. Goldberger (Eds.), Latent variables in socio-economic models (pp. 285-325). Amsterdam, North Holland.

Joreskog, K.G. and Sorbom, D. (1979). Advances in factor analysis and structural equation models. Cambridge, MA: ABT Associates.

Joreskog, K.G. and Sorbom, D. (1984). LISREL VI User's Guide. Mooresville, IN: Scientific Software.

Kagan, J. (1980). Perspectives on continuity. In O.G. Brim, Jr., & J. Kagan (Eds.), Constancy and change in human development. Cambridge, MA: Harvard University Press.

Kessler, R.C., and Greenberg, D.F. (1981). Linear panel analysis: Models of quantitative change. New York: Academic Press.

Labouvie, E.W. (1974). Developmental causal strucures for organism-environment interactions. Human Development, 17, 444-452.

Labouvie, E.W. (1978). Experimental sequential strategies for the exploration of ontogenetic and socio-historical changes. Human Development, 21, 161-169.

Labouvie, E.W. (1980a). Identity versus equivalence of psychological measures and constructs. In L.W. Poon (Ed.), Aging in the 1980's: Psychological issues (pp. 493-502). Washington, D.C.: American Psychological Association.

Labouvie, E.W. (1980b). Measurement of individual differences in intraindividual changes. Psychological Bulletin, 88, 54-59.

Labouvie, E.W., and Nesselroade, J.R. (1985). Age, period, and cohort analysis and the study of individual development and social change. In J.R. Nesselroade and A. Von Eye (Eds.), Individual development and social change: Explanatory analysis (pp. 189-212). New York: Academic Press.

Lachman, M.E. (1983). Perceptions of intellectual aging: Antecedent or consequences of intellectual functioning? Development Psychology, 19, 482-498.

Leamer, E.E. (1978). Specification searches. New York: John Wiley & Sons.

Lerner, R.M. (1986). Concepts and theories of human development. (2nd Ed.). New York: Random House.

Long, J.S. (1983a). Confirmatory factor analysis. Sage University paper series on Quantitative Applications in the Social Sciences, series no. 07-033. Beverly Hills and London: Sage.

Long, J.S. (1983b). Confirmatory factor analysis. Sage University paper series on Quantitative Applications in the Social Sciences, series no. 07-034. Beverly Hills and London: Sage.

McArdle, J.J. (1980). Causal modeling applied to psychonomic systems simulation. Behavior Research Methods and Instrumentation, 12, 193-209.

McArdle, J.J. and Epstein, D. (1985). Latent growth curves within developmental structural equation models. Child Development, 58, 110-133.

McArdle, J.J., and McDonald, R.P. (1984). Some algebraic properties of the reticular action model for moment structures. British Journal of Mathematical and Statistical Psychology, 37, 234-251.

McDonald, R.P. (1978). A simple comprehensive model for the analysis of covariance structures. British Journal of Mathematical and Statistical Psychology, 31, 59-72.

McDonald, R.P. (1980). A simple comprehensive model for the analysis of covariance structures: Some remarks on applications. British Journal of Mathematical and Statistical Psychology, 33, 161-183.

Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Meredith, W., and Tisak, J. (1982). Canonical analysis of longitudinal and repeated measures data with stationary weights. Psychometrika, 47, 47-67.

Meredith, W., and Tisak, J. (1986). Tuckerizing curves. Unpublished manuscript.

Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. Psychological Bulletin, 89, 575-588.

Mortimer, J.T., Finch, M.D., and Kumka, D. (1982). Persistence and change in development: The multidimensional self-concept. In P.B. Baltes & O.G. Brim, Jr. (Eds.), Life-span Development and Behavior (Vol. 4). New York: Academic Press.

Mulaik, S.A., (1972). The Foundations of factor analysis. New York: McGraw-Hill.

Mulaik, S.A. (1986). Toward a synthesis of deterministic and probabilistic formulations of causal relations by the functional relation concept. Philosophy of Science, 53, 313-332.

Mulaik, S.A. (1987). Toward a conception of causality applicable to experimentation and causal modelling. Child Development.

Nesselroade, J.R., and Baltes, P.B. (1979). Longitudinal research in the study of behavior and development. New York: Academic Press.

Nesselroade, J.R., and Baltes, P.B. (1984). From traditional factor analysis to structural-causal modeling in developmental research. In V. Sarris & A. Parducci (Eds.), Experimental psychology in the future (pp. 267-287). Hillsdale, NJ: Lawrence Erlbaum Associates.

Nesselroade, J.R., and 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.). New York: Van Nostrand Reinhold.

Platt, J.R. (1964). Strong inference. Science, 146, 347-353.

Popper, K.R. (1959). The logic of scientific discovery. London: Hutchinson.

Rogosa, D. (1979). Causal models in longitudinal research: Rationale, formulation, and estimation. In J.R. Nesselroade & P.B. Baltes (Eds.), Longitudinal research in the study of behavior and development (pp. 263-302). New York: Academic Press.

Rogosa, D. (1980). A critique of cross-lagged correlation. Psychological Bulletin, 88, 245-258.

Rogosa, D., and Willett, J.B. (1985a). Satisfying a simplex structure is simpler than it should be. Unpublished manuscript.

Rogosa, D., and Willett, J.B. (1985b). Understanding correlates of change by modeling individual differences in growth. Psychometrika, 50, 203-228.

Schaie, K.W. (1965). A general model for the study of development problems. Psychological Bulletin, 64, 92-107.

Schaie, K.W. (1977). Quasi-experimental research 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. (1983). Longitudinal studies of adult psychological development. New York: Guilford Press.

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 Press.

Schaie, K.W., and Baltes, P.B. (1975). On sequential strategies in developmental research and the Schaie-Baltes controversy: Description or explanation. Human Development, 18, 384-390.

Schaie, K.W. and Hertzog, C. (1982). Longitudinal methods. In B.B. Wolman (Ed.), Handbook of developmental psychology (pp. 91-115). Englewood Cliffs, NJ: Prentice-Hall.

Schaie, K.W. and 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.). New York: Van Nostrand Reinhold.

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

Steiger, J.H. (1980). Tests for comparing elements of a correlation matrix. Psychological Bulletin, 87, 245-251.

Steyer, R. (1985). Causal regressive dependencies: An introduction. In J.R. Nesselroade and A. Von Eye (Eds.), Individual development and social change: Explanatory analysis (pp. 95-124). New York: Academic Press.

Suppes, P. (1970). A probablisitic theory of causality. Amsterdam: North Holland.

Thurstone, L.L. (1938). Primary mental ablilties. (Psychometric Monographs, No. 1). Chicago: University of Chicago Press.

Thurstone, L.L., and Thurstone, T.G. (1941). Factorial studies of intelligence. (Psychometric Monographs, No. 2). Chicago: University of Chicago Press.

Thurstone, L.L. (1944). A factorial study of perception. (Psychometric Monographs, No. 4). Chicago: University of Chicago Press.

Thurstone, L.L. and Thurstone, T.G. (1949). Examiners Manual, SRA Primary Mental Abilities Test (Form 11-17). Science Research Associates, Chicago.

Tucker, L.R. (1964). The extension of factor analysis to three-dimensional matices. In N. Frederiksen & H. Gullikson (Eds.), Contributions to mathematical psychology. New York: Holt, Rinehart, and Winston.

TABLE 1

A Heuristic Scheme Classifying Research Purpose and Data Analysis Procedure Combinations

		Research Purpose (Theoretical Orientation)	
		Exploratory	*Hypothesis Testing*
Analysis procedure (method)	*Exploratory*	I	II
	Hypothesis testing	III	IV

Note: Adapted from Nesselroade & Baltes, 1984

TABLE 2

Unrestricted Factor Analysis in Sample 1: Five Factor Solution

Factor Loadings:

Occasions	Factors				
	V	S	R	N	W
1	.581 (5.648)	.846 (9.139)	.788 (4.845)	.935 (9.903)	.843 (10.577)
2	.728 (7.540)	.863 (9.216)	.893 (5.943)	.901 (9.522)	.895 (10.235)
3	1.120 (12.916)	.792 (8.742)	.643 (4.529)	.896 (9.300)	.802 (9.955)

Factor Correlations:

V	1				
S	.502	1			
R	.801	.580	1		
N	.430	.274	.493	1	
W	.493	.194	.466	.383	1

Note: All factor loadings of variables other than those shown were not significant. Factor loadings are scaled values (unscaled values in parenthesis).

From: Hertzog, 1979

TABLE 3

oodness of Fit Statistics for Alternative Longitudinal Models (Single Sample)

odel	χ^2	df	p	δ^a
. configural invariance	82.98	72	.17	.964
. metric invariance	87.20	80	.27	.962
. equal factor variances	112.90	82	.013	.951
. equal factor covariance matrices	121.78	84	.005	.947
odel Comparisons	$\Delta\chi^2$	Δdf	p	NS
-2 (tests metric invariance)	4.22	8	NS	.002
-3 (test equal variance in g)	25.70	2	<.001	.011
-4 (test equal g covariances)	8.88	2	<.05	.004

3entler-Bonett normed fit index

dapted by permission from Hertzog and Schaie (1986a)

Figure Captions

1. A schematic representation of the hierarchical relations among characteristics of multivariate statistical methods. Branches define different categories within level (e.g., explanation and description are types of General Orientation). Examples of methodological procedures typifying some of the most differentiated levels in the hierarchy are indexed by number. The schematic is a heuristic for understanding differentiation of concepts such as confirmatory analysis, restricted estimation techniques, and statistical hypothesis tests, and as such is not considered adequate for a full taxonomy of multivariate techniques (see text).

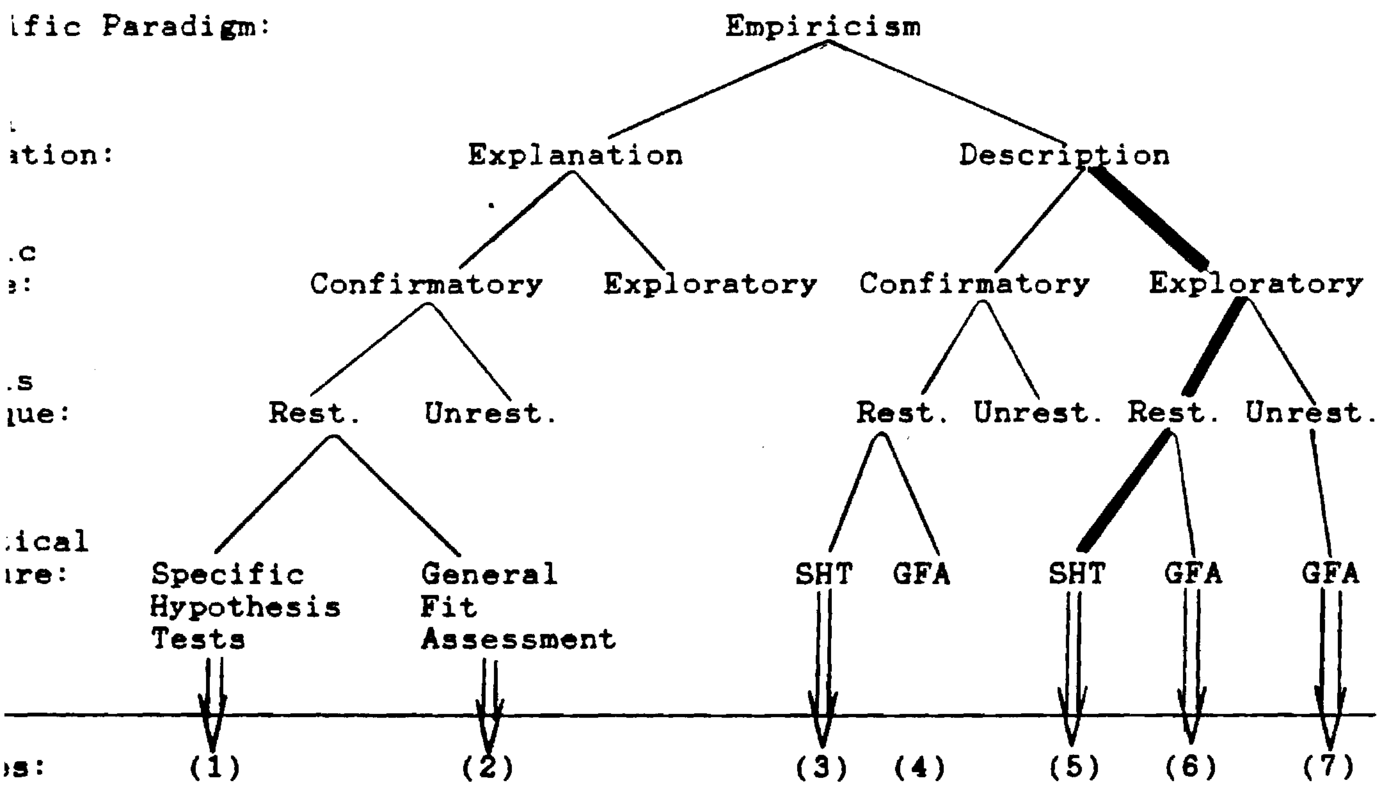
2. Examples of patterns of change typifying 1) mean stability versus average age change, and 2) covariance stability versus individual differences in change. In Example A, there is mean stability across all age levels, but substantial individual differences in change patterns between childhood and adulthood. In Example B, there is average age change between birth and childhood but perfect stability of individual differences. (Adapted from Baltes, Reese, & Nesselroade, 1977).

3. A SEM specification for a longitudinal factor model for six observed variables (y_1 through y_6) determined by two latent variables (or factors; __$_1$ and __$_2$) at the three longitudinal occasions (T_1 through T_3). There are therefore 6 occasion-specific factors (which are all correlated). Covariances

of residuals for specific observed variables (e.g., y_1's residual, __$_1$, with itself across T_1, T_2, and T_3) are also shown.

4. Occasion-specific model for general intelligence factor (g) tested by Hertzog and Schaie (1986). Residual covariances were specified but are not shown. Copyright, American Psychological Association. Reprinted with permission.

Figure 1



) Competing SEM models

) Single SEM model

) Cross-lagged regression, a priori hypotheses

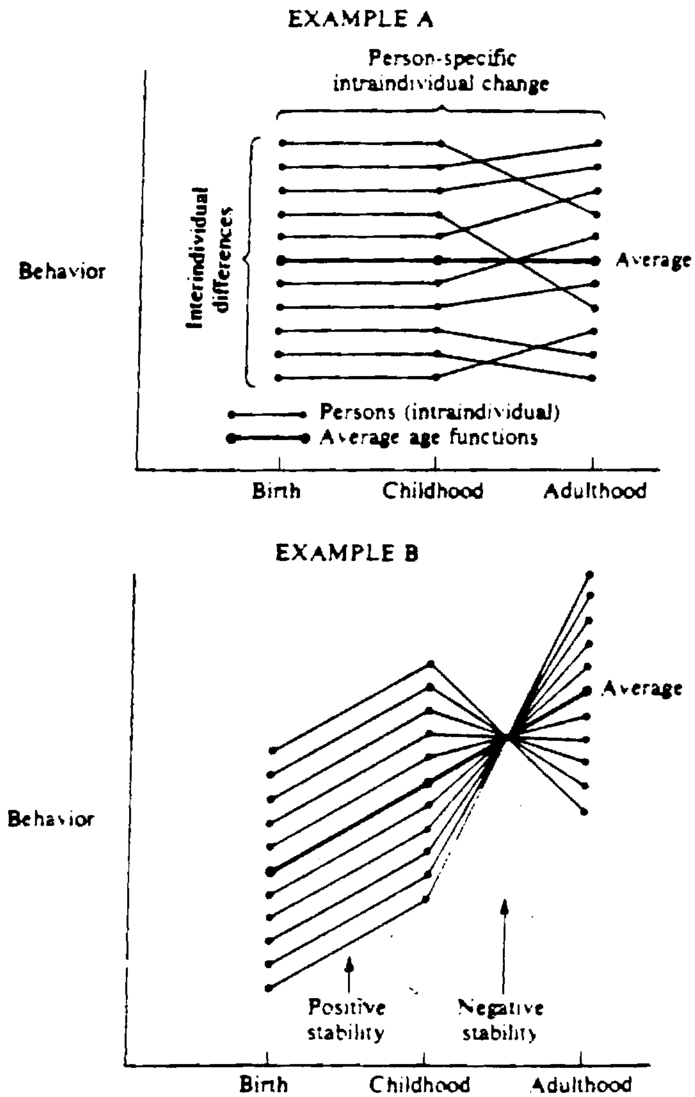
) Single factor analysis model

) models testing factorial invariance in multiple groups

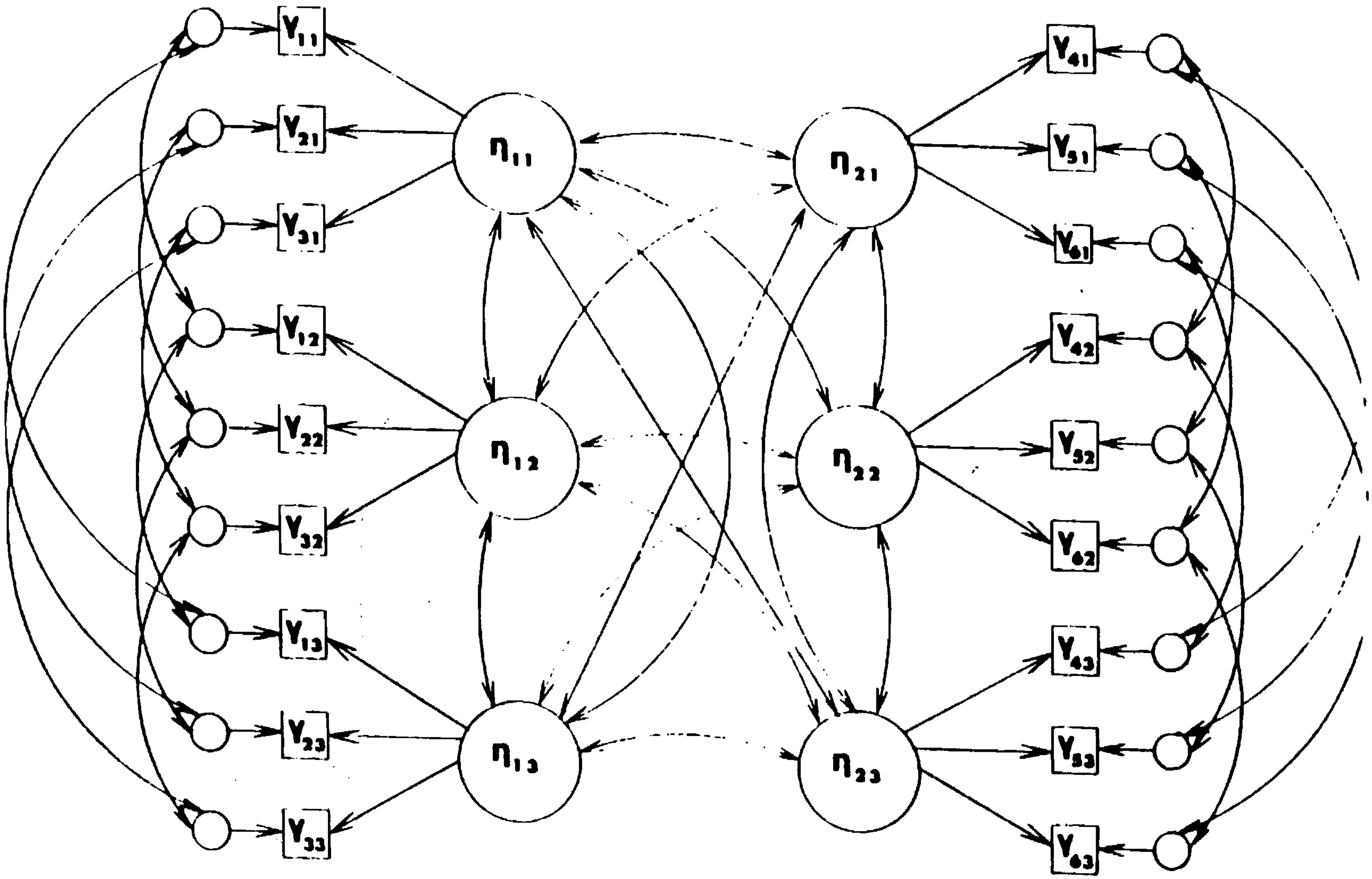
) specification searches using SEM

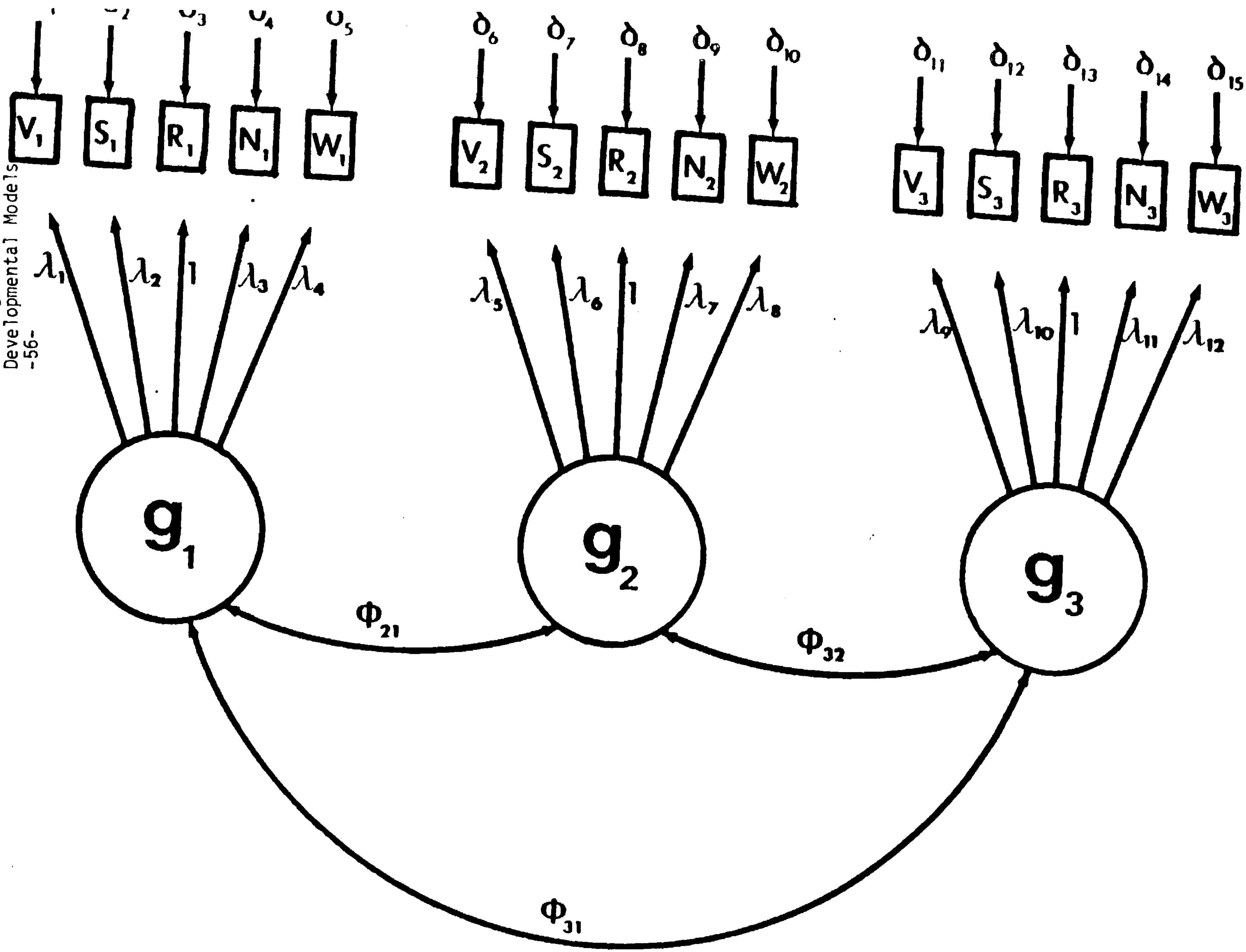
) traditional multivariate methods

Figure 2


EXAMPLE A
Person-specific
intraindividual change
Interindividual
differences
Behavior
Average
Persons (intraindividual)
Average age functions
Birth
Childhood
Adulthood
EXAMPLE B
Behavior
Average
Positive
stability
Negative
stability
Birth
Childhood
Adulthood

Figure


Y_{11}
Y_{21}
Y_{31}
Y_{12}
Y_{22}
Y_{32}
Y_{13}
Y_{23}
Y_{33}
η_{11}
η_{12}
η_{13}
η_{21}
η_{22}
η_{23}
Y_{41}
Y_{51}
Y_{61}
Y_{42}
Y_{52}
Y_{62}
Y_{43}
Y_{53}
Y_{63}


δ3
δ4
δ5
δ6
δ7
δ8
δ9
δ10
δ11
δ12
δ13
δ14
δ15
V1
S1
R1
N1
W1
V2
S2
R2
N2
W2
V3
S3
R3
N3
W3
λ1
λ2
1
λ3
λ4
λ5
λ6
1
λ7
λ8
λ9
λ10
1
λ11
λ12
g1
g2
g3
Φ21
Φ32
Φ31

The Influence of Cognitive Slowing

on Age Differences in Intelligence

Christopher Hertzog

Georgia Institute of Technology

Date of Submission: January 5, 1988

Address Correspondence to: Christopher Hertzog, School of Psychology, Georgia Institute of Technology, Atlanta, GA 30332-0170.

Running Head: Influence of Cognitive Slowing

Abstract

This study examines the relationship between the speed of basic perceptual and motor processes and psychometric intelligence. A large cross-sectional sample took a battery of psychometric tests measuring multiple primary abilities, using tests from the Educational Testing Services reference kit and the Thurstone and Thurstone (1949) Primary Mental Abilities (PMA) test used by Schaie (1983) in his longitudinal studies. The battery also included measures of basic perceptual speed and speed in working with the types of answer sheets used in the PMA. The cross-sectional data were consistent with other studies in the literature, showing large cross-sectional age differences for several abilities. Polynomial regression analyses, partialling for the speed measures, showed that (a) age trends were dramatically altered, attenuating (but not eliminating) the age differences, and (b) a substantial proportion of age-related variance is shared in common with speed. The PMA vocabulary test, Verbal Meaning, was more highly related to speed than other vocabulary tests. Moreover, there was a significant Age X Answer Sheet interaction, with a higher relationship between answer sheet speed and Verbal Meaning performance in older adults. These results suggest limited utility of Verbal Meaning as a test of verbal knowledge in older adults, due to a substantial speed component that is correlated with age. In general, the results suggest that theoretical interpretation of age differences in intelligence must attend to the role of cognitive speed, both at the level of the ability construct and at the level of psychometric test performance.

Introduction

A large body of evidence has accumulated to suggest that there are reliable age changes in multiple intellectual abilities (e.g., Botwinick, 1977; Cunningham, 1987; Horn & Donaldson, 1980; Schaie, 1983; Schaie & Hertzog, 1986). Although much of the available data is based on cross-sectional age differences, Schaie and Hertzog's (1983) cohort-sequential analyses of longitudinal and cross-sectional sequences confirm the existence of average age decline in intelligence that is independent of generational differences in intelligence.

Given the significant age changes, the question remains as to the cognitive processes that are implicated in the change. One important correlate of age changes in intelligence is age changes in the speed of information processing. Salthouse (1985a,b), building upon the seminal work of Birren and others (e.g., Birren, 1964, 1965, Birren, Woods, & Williams, 1980; Cunningham, 1980; Cunningham & Birren, 1980), has argued for a processing rate theory of aging, in which age-related slowing of information processing speed is hypothesized to be a primary cause of age-related decline in intelligence. Certainly there is evidence of (a) age-related changes in both information processing speed and intelligence (Salthouse, 1985a,b) and (b) evidence of correlations of information processing speed and intelligence in both old and young subjects (e.g., Cerella, DiCara, Williams, & Bowles, 1986; Salthouse, 1985a). Other evidence exists that may be interpreted as supporting the processing rate theory of speed/intelligence relationships. For example, Witt and Cunningham (1979) reported cross-lagged correlation analyses of Owen's longitudinal data that suggested a substantial lagged correlation of a "highly speeded" Relations factor (marked by the Army Alpha subtests Analogies and Following Directions) with Verbal and Numerical Intelligence factors. They suggested that slowed cognitive processes were driving age changes in these latter two factors. Cunningham (1985) has also reported analyses that suggest that age differences in test performance are primarily a function of number of items solved during a fixed time limit, rather than changes in the proportion of correct item responses.

Cornelius, Baltes, Willis, and Nesselroade (1983) found that measures of perceptual speed taken from the ETS Reference Kit (Ekstrom, French, Harman, & Price, 1976) correlated highly with fluid intelligence in a sample of elderly adults. Horn, Donaldson, and Engstrom (1981) found that partialling measures of clerical and perceptual speed (including the same ETS measures) from measures of fluid intelligence significantly reduced the magnitude of age differences in fluid intelligence for a sample of adults ranging roughly in age from 20 to 60. However, significant age differences in fluid intelligence were statistically independent of speed.

How does the processing rate theory account for such findings? In essence, it argues that the speed of basic information processing mechanisms (termed "cognitive mechanics" by Hunt, 1978) reflects the efficiency of central nervous system processing. In turn, the efficiency of information processing is a major determinant of intelligence (Carroll, 1980; Hunt, 1978, 1983). This perspective is consistent with findings of significant correlations of reaction time tasks measuring elementary cognitive processes with performance on tests that appear to be "power" rather than "speed" tests (e.g., Raven's Progressive Matrices, Nelson-Denny Vocabulary; Carroll, 1980; Jensen, 1985; Vernon, 1985). Jensen

(1982, 1985) has argued that these correlations are mediated by general intelligence, or g (cf. Horn, 1985). As pointed out by Salthouse (1982, 1985a), the relationship of information processing speed to intelligence, coupled with age-related cognitive slowing, should lead to age-related declines in intelligence.

There are, however, a number of issues that the processing rate theory must yet address (Salthouse, 1985a; Horn, 1985). Certain observations in the literature seem, at least on the surface, to be inconsistent with the theory. For example, Even though Horn et al. (1981) found that partialling for Perceptual Speed attenuated age differences in fluid intelligence, they also reported that the time taken to solve individual Induction and Spatial Visualization items did not significantly vary with age. More generally, there are both "strong" and "weak" versions of the processing rate theory, although the two types of theory are sometimes confused for one another in the literature. The strong version of the theory posits a general mechanism for slowing that causes reduced information processing speed <u>and</u> likewise causes intellectual decline with advancing age (Salthouse, 1985a). The weak version of the theory allows that there may be multiple, independent mechanisms that influence information processing speed. These mechanisms may all show, on average, age-related decline, but they act independently and have differential influence in the multivariate domain of intellectual abilities. The weak version of the theory is more consonant with the findings of multiple information processing speed factors, whether measured by psychometric (White & Cunningham, 1987) or experimental methods. Hertzog, Raskind, and Cannon (1986), using reaction time (RT) procedures, demonstrated that a semantic memory access speed factor can be measured independently of a factor measuring nonverbal choice RT in old and young age groups. Although both types of tasks show simimlar mean age differences in RT, they formed distinct factors with different patterns of correlation between old and young groups. Factors measuring different aspects of information processing speed may vary in magnitude of correlation with intelligence test performance and with age-related changes in intelligence. In some cases, evidence consistent with both the weak and strong forms of the theory (e.g., proportionality of mean age differences in RT across tasks) is taken as confirmation of the strong version of the theory (Cerella, 1985).

There are, however, additional points of view on speed-intelligence relationships in adulthood. A second relevant perspective on the problem emphasizes that certain ability constructs may be conceptualized a priori as being intimately related to real-time information processing speed. The ETS intelligence factor Perceptual Speed referred to above is an obvious example. However, other ability constructs are defined in ways that make them implicitly speed-related. For example, the ability Numerical Facility is defined by the ability to calculate answers to simple arithmetic problems (addition, subtraction, multiplication) and does not measure more advanced mathematical knowledge and skills. Implicitly, this ability is manifested in how rapidly one can add, not whether one can add, two column addition problems, since most adults studied are capable of addition (see Jensen, 1985). Clearly, Thurstone (1944; Thurstone & Thurstone, 1941) considered speed of successful item solution a defining characteristic of intelligence. Thus, apart from any hypothesized causal relationship between cognitive mechanics and intelligence, slowing of information processing speed may cause decline in ability constructs that by definition require rapid real-time processing. Both types of speed/ability relationship may be considered construct-relevant (Hertzog, 1985), in that the relationship is determined by slowing of cognitive processes that are conceptualized as components of the ability construct under study.

On the other hand, a third source of speed/intelligence relationships may be performance-specific. A psychometric test may be inadvertently designed in such a way as to maximize the importance of cognitive speed for high performance levels (Lorge, 1936). As is known from classic psychometrics, manipulation of item difficulty, arrangement of test formats, and selection of time limits relative to the number of test items affect in turn the degree to which the test is influenced by how rapidly problems are solved. In theory, then, it would be possible for a highly speeded test to manifest age changes in performance that are a function of slowed information processing speed and not a valid reflection of age changes in the ability construct the test was designed to measure. It is unlikely that slowed information processing speed can account for all age differences in intelligence test performance (Botwinick, 1977; Horn, 1985) but it may exaggerate the magnitude of the age effects observed.

The distinction between construct-relevant and performance-specific changes in information processing speed becomes important when evaluating Schaie's (1983) longitudinal data on intellectual change. A little appreciated fact regarding Schaie's sequential study is that it has used Form AM of Thurstone's

1948 PMA (designed to be appropriate for individuals ages 11 through 17). The practical implication of the use of Form AM is that the PMA tests Verbal Meaning (a test of recognition vocabulary) and Reasoning (a test of induction using letter series) have lower item difficulty levels than other forms of the PMA (including the 1962 revision). Indeed, all five Thurstone PMA subtests, which are administered to groups under time limits, are influenced by relatively substantial speed components in adults samples (Schaie, Rosenthal, & Perlman, 1953).

Schaie and Hertzog (1983) reported age-related decline in Verbal Meaning after age 60, whereas other longitudinal studies have found maintenance of verbal abilities until later in the life-span (Botwinick, 1977). Schaie and Hertzog (1983) hypothesized that the precocious age-related decline of PMA Verbal Meaning reflects a performance-specific influence of cognitive slowing on that subtest (independent of any relation of slowing in cognitive mechanics to the Verbal Comprehension ability). Consistent with this hypothesis, Schaie, Willis, Hertzog, and Schulenberg, 1987) recently reported new cross-sectional evidence that PMA Verbal Meaning loads as highly on the ETS Perceptual Speed factor as it does on a Verbal Comprehension factor marked by two ETS vocabulary tests.

The present report is part of a series of studies designed to assess the degree to which age differences in psychometric intelligence test performance are, when observed, reflections of slowing of the speed of intelligent thought, slowing in performance-related aspects of test performance, or both. A battery of psychometric tests was administered to groups of adults and university students; subsequently, subsamples of these individuals were tested in a set of reaction time tasks measuring basic information processing speed. This paper reports the results of analyses of the cross-sectional psychometric test data. The psychometric battery was designed to measure multiple intellectual abilities measured by Schaie and others, while simultaneously measuring perceptual and psychomotor speed. The speed measures included assessments of how rapidly individuals could mark computer answer sheets (as used in the Schaie longitudinal studies) when the correct answers were already provided in the test booklet.

Three general hypotheses were examined. First, large cross-sectional age differences were predicted for Perceptual Speed, Induction, and multiple spatial abilities, but minimal differences for Verbal Comprehension and Numerical Facility. The second hypothesis stated that age differences in psychometric abilities would be substantially reduced as a function of statistical adjustment for the perceptual and psychomotor speed variables. A central question was whether the ability measures would display any age-related variance after being partialled for the speed measures. The third hypothesis stipulated that older subjects would have a significantly greater relationship between the speed of working PMA answer sheets and psychometric test performance. For PMA Verbal Meaning, the specific predictions were (a) greater cross-sectional age differences on this test than on two ETS vocabulary tests, and (b) an age-related increase in the relationship between Verbal Meaning performance and PMA answer sheet speed.

Method

Subjects

There were 622 adult participants, ages 43-89, and 211 undergraduate students of the Pennsylvania State University. A substantial proportion of the adults were alumni of the Pennsylvania State University, supplemented by additional community-dwelling volunteers from the greater Harrisburg, PA, area. Adults were contacted by mail, newspaper, and television solicitations. A snowballing technique was also used to identify and recruit additional participants. Students were recruited through newspaper advertisments. All individuals were paid for their participation.

Measures

The psychometric battery administered to the sample is given in Table 1. The battery included four of the Thurstone PMA tests used by Schaie, but, more generally, it was designed to provide multiple measures of the following ability factors: Verbal Comprehension, Inductive Reasoning, Spatial Orientation, Spatial Visualization, Flexibility of Closure, Numerical Facility, Perceptual Speed, and PMA Answer Sheet speed. Test scoring followed the test manuals, using the number of correct responses, except for the Space, Number, and Object Rotation tests, where the scoring was correct minus incorrect responses. The ETS tests typically contain two parts, administered sequentially. Only the first part of the ETS tests were used in this study.[1] One exception was with the two ETS vocabulary measures: Vocabulary (V2) and Advanced Vocabulary (V4). For these tests, both parts were combined into a single test, without increasing the time limit. The Answer Sheet speed tests were scored by tallying the number of spaces marked during a fixed time limit.

Insert Table 1 about here

Procedure

Psychometric tests were administered in group sessions to small groups (usually, between 10 and 15 persons). Administration of the battery required two sessions of approximately 2.5 hours duration (including a 10 minute break at about the midpoint of the session). Subjects began by filling out a personal data questionnaire upon their arrival, followed by the psychometric tests, given in invariant order.

The statistical analysis is based on 210 students and 592 adults, ages 43-78, who had complete data on all psychometric tests.

Results

Age Differences in Intelligence

Age group MANOVA. To assess the magnitude of age differences in intelligence, we first ran a comparison of the 210 Penn State students with 342 adults, ages 43 through 78, that were college graduates. All other adults, including alumni with some college experience, were excluded from this analysis in order to maximize comparability of the adults to the student sample. The adults were divided into five seven-year age groups. A MANOVA was run on the 6

X 2 (Age X Sex) design for the 25 ability measures, yielding significant multivariate main effects for Age (approximate $F = 8.68$, $df = (125, 2544)$, $p <$ .001) and Sex ($F = 8.96$, $df = (25, 516)$, $p < .001$). No significant Age X Sex interaction was found. As predicted, significant cross-sectional age differences were found for all measures. The pattern of age differences varied by ability. Tukey's HSD tests on groups showed that age differences were substantial for all tests except measures of Verbal Comprehension and Numerical Facility. In most cases, adjacent seven-year age groups were significantly different, with monotonically decreasing performance from younger to older ages. In addition, students performed significantly better than all adult groups on all markers of Induction, Spatial Orientation, Spatial Visualization, and Flexibility of Closure. There were differences among the markers of Perceptual Speed. Identical Pictures and Number Comparisons showed large age differences, whereas Finding A's did not.

The pattern was markedly different for Verbal Comprehension and Numerical Facility. For all three markers of Verbal Comprehension, the students samples performed significantly poorer than adults. PMA Verbal Meaning showed significant late life age differences, with the oldest groups (mean ages 68 and 75) performing significantly lower than the other adult groups. However, the ETS Advanced Vocabulary and Vocabulary showed no such declines; indeed, the adult group with mean age 68 had the highest performance levels, although they differed significantly only from the students and the group with a mean age of 47. For Numerical Facility, the significant differences were primarily a function of poorer performance by the students. On the Addition subtest, none of the adult groups differed significantly. For Subtraction/Multiplication and PMA Number, significant late-life differences were found that were largest for Number.

Sex differences were associated with better performance by males on measures of Spatial Orientation, Spatial Visualization, and Flexibility of Closure. Females performed significantly better on the ETS Vocabulary test and the Finding A's test.

Polynomial Regression Analyses. A more precise and statistically powerful analysis of age differences in intelligence was conducted by means of polynomial regression analysis across age levels, using powers of Age. Data from the students was excluded; all 592 adults ages 43 to 78 were used in the analysis. The polynomial regression approach allowed a test of whether chronological age differences are linear over the 43 to 78 age range. Sex and Education were also used as independent variables. In the first analysis, Sex, the first through fourth (linear through quartic) terms of the Age, and their interaction (product) terms with Sex were used in hierarchical multivariate regression, using procedures for significance testing recommended by Cohen and Cohen (1983). These results paralleled the MANOVA on the high education sample, in that significant Age and Sex effects were found, but no Age X Sex interactions. An order two polynomial for Age (linear [Age_L], quadratic [Age_Q]) provided the best fit; the cubic and quartic Age terms did not approach statistical significance. At that point, Education was used in the analysis as an independent variable, and the interactions of Education and Sex with Age_L and Age_Q were evaluated. None of the interactions was significant. There were significant multivariate tests for Education (F = 9.14, df = (25, 548), $p < .001$), Sex (F = 10.96, df = (25, 548), $p < .001$), Age_L (F = 20.35, df = (25,548, $p < .001$), and Age_Q (F = 2.83, df = (25, 548, $p < .001$).

Table 2 reports the R^2 and regression coefficients for the individual subtests. Significant linear age differences were found on all subtests except ETS Vocabulary, Advanced Vocabulary, and Finding A's, with performance levels decreasing with advancing age. Significant linear and quadratic trends were found for PMA Verbal, suggesting substantial late-life age differences in performance on that test. There were large and significant linear differences on the Perceptual Speed and Answer Sheet speed measures. In both cases, the age difference between the 43 and 78 year-olds exceeded two standard deviations.

Effects for Education were in the expected direction — higher education associated with higher ability. Although, as expected, Education has significant relationships to many ability measures, the pattern of age trends appeared to be approximately parallel at different levels of education, and by inference, socioeconomic status (Gardner & Monge, 1977; Fozard & Nuttall, 1971). It is interesting to note that there were significant Education effects on the Answer Sheet measures. Clearly, the kind of visual search and attention switching required by these measures is correlated with education and general intellectual ability (suggesting that these tests are not merely measures of how fast one can move a pencil).

Insert Table 2 about here

Regression Analysis Using Speed Measures as Predictors

The polynomial analysis set the stage for the next step, which was designed to determine the degree of adjustment in cross-sectional age differences on ability measures brought about by using the Perceptual Speed and Answer Sheet speed measures as independent variables. First, composites for each primary ability factor were computed by summing z-scores for different measures of the eight abilities (as suggested by the arrangement of tests in Table 1) and then re-standardizing. The composite variable for Answer Sheet speed summed only the four PMA Answer Sheet variables, omitting the Crossing Digits test. The six ability composites were then used as dependendent variables in a multivariate polynomial regression. In the first step, the Perceptual Speed and Answer Sheet speed composites were entered into the regression equation. The second step entered the Age_L and Age_Q variables. Sex was entered in the third step. Finally, the last two steps entered product variables representing interactions of the Age trends with Answer Sheet and Perceptual Speed, respectively.

Table 3 reports the R^2 statistics for the polynomial regression analysis. Multivariate significance tests as of Step 3 (with all independent variables, save interaction terms, in the equation) showed significant effects for Age_L ($F = 20.14$, $df = (6, 581)$, $p < .001$), Age_Q, ($F = 3.23$, $df = (6, 581)$, $p < .01$), Sex ($F = 28.48$, $df = (6, 581)$, $p < .001$), Perceptual Speed ($F = 42.99$, $df = (6, 581)$, $p < .001$) and Answer Sheet ($F = 16.77$, $df = (6, 581)$, $p < .001$). The effects of both the speed variables were significant at $p < .01$ for all six abilities. Thus both Perceptual Speed and Answer Sheet speed have salient, <u>statistically</u> <u>independent</u> prediction of intellectual abilities in this sample of adults. Neither multivariate interaction test added at Steps 4 and 5 were significant, although there was a trend for an Age$_L$ X Answer Sheet speed for the Verbal Comprehension and Induction composites.

Insert Table 3 and Figures 1 through 6 about here

The significant multivariate effects for Age indicate that partialling for the speed measures did not eliminate age differences in intellectual abilities. Nevertheless, the statistical adjustment had a profound effect upon the function regressing ability on chronological age. As can be seen in Table 3, the increments in R^2 for Age, after the speed composites were already in the equation, did not exceed .03 for any composite. In contrast, the R^2 for the speed measures alone was quite substantial. One way to examine the effects of partialling for speed is to examine the pattern of the prediction equations derived from the regression analysis. Figures 1 through 6 plot the unadjusted and adjusted curvilinear regression functions relating Age to the composite ability scores. Adjustment for Perceptual Speed and Answer Sheet speed greatly reduced the negative slope relating Age to Spatial Relations, Spatial Visualization, Flexibility of Closure, and Induction. For example, the unadjusted age difference in Induction was just less than 2 standard deviations from age 43 to age 78. Adjusted for speed, the age difference spanned about half a standard deviation. Moreover, the significant curvilinearity in the adjusted functions for Induction, Spatial Relations, Spatial Visualization, and Flexibility of Closure reflected a tendency for performance levels to be relatively stable until the decade of the fifties.

On the other hand, the late-life decline suggested by linear Age effects on Numerical Facility and Verbal Comprehension were eliminated by partialling for the speed measures. The shape of both speed-adjusted functions changed to show positive Age gradients, levelling off in old age.

Table 4 reports the results of a commonality analysis (Pedhazur, 1982) partitioning the total R^2 for each composite into (a) variance predicted uniquely by the linear and quadratic terms for Age, partialling speed; (b) variance uniquely predicted by the speed composites, partialling Age; and (c) variance jointly predicted by both Age and speed (labeled "shared" R^2 in Table 4). There were substantial proportions of variance uniquely determined by the Perceptual Speed and Answer Sheet Speed composites. However, most of the predictive power of Age for the Spatial Relations, Induction, Visualization, and Flexibility of Closure variables was shared in common with speed. For example, Age accounted for about 20% of the variance in Spatial Relations, but only 3% of the variance in Spatial Relations is predicted by Age independent of speed. The last column in Table 4 gives the proportion of R^2 predicted by the linear and quadratic trends for Age that was statistically independent of speed. In the case of Spatial Relations, only about 14% of the age-related R^2 was independent of speed. Only Verbal Meaning and Numerical Facility failed to show significant amounts of shared prediction between Age and speed. These abilities were the ones that showed positive age-related trends.

Given that the PMA subtests used the PMA answer sheets, it was plausible that the degree of influence by the speed measures for the four abilities measured by the PMA is inflated by including the PMA subtests in the ability composites. The middle section of Table 4 reports the R^2 statistics for the four composites omitting the PMA tests. The results were essentially the same, except for Verbal Comprehension. The polynomial regression analysis showed significant effects for Perceptual Speed for all four composites. The Answer Sheet speed measure significantly predicted only three of the four new composite measures. The relationship of Answer Sheet speed to Verbal Comprehension, as well as the Age_L X Answer Sheet interaction, was eliminated when PMA Verbal Meaning was removed from the composite Verbal Comprehension score. It is interesting that the Answer Sheet composite still had significant prediction of the other composites, given that they were formed from ETS and Thurstone measures that did not require use of a computerized answer sheet.

Analysis of the PMA Subtests

Polynomial regression analysis. The previous finding of Answer Sheet speed relationships to composites omitting the PMA subtests suggested that the Answer Sheet Speed measures tap processes that are generally related to intellectual ability. Nevertheless, the question remained as to whether there were more substantial relationships of Answer Sheet usage to performance on the PMA subtests that required use of the answer sheets. A test of Schaie and Hertzog's (1983) hypothesis of age changes in the importance of speed for PMA Verbal Meaning performance demanded testing for an Age X Answer Sheet speed interactions for this variable alone. A new series of polynomial regression analyses were conducted to address these issues, using only the PMA subtests as dependent variables.

The composite Perceptual Speed and Answer Sheet variables were again used as independent variables. Although it would have been possible to partial the individual PMA subtests only for their corresponding Answer Sheet speed measure, this was not done, for two reasons. First, the Answer Sheet measures correlate substantially. Second, given that it appears that the Answer Sheet measures are markers for a latent variable (factor), use of the composite variable increased reliability of the independent variable, reducing loss of prediction due to measurement error.

The bottom part of Table 4 reports the commonality analysis associated with the additive model (through Step 3). As with the composite ability measures, adjustment for the speed variables eliminated negative age differences on PMA Verbal Meaning and PMA Number. Note also the large proportion of variance in PMA Verbal Meaning associated with speed. The adjustment for speed reduced but did not eliminate linear age differences on PMA Reasoning and PMA Space.

The hierarchical tests of the interaction of the Answer Sheet speed and Age (Age_L and Age_Q) showed a significant increment to R^2 for the PMA Verbal Meaning subtest ($\underline{F}$ = 6.09, $\underline{df}$ = (2, 555), $\underline{p}$ < .01). This effect was exclusively a function of a strong interaction with the Age_L component. No other interaction term was significant for any other PMA subtest.

Insert Table 4 about here

<u>Answer Sheet/PMA Subtest Correlations</u>. Another way to look at the

interaction effect for PMA Verbal Meaning is to calculate the correlation of PMA Verbal Meaning and the PMA Answer Sheet variable in different age groups. Table 5 reports the correlations of the Answer Sheet speed measures and the PMA subtests in four different age groups. Speed of marking the answer sheets correlated significantly with PMA test performance in all age groups. However, there was a marked tendency for increased correlations of individual differences in speed with individual differences in PMA Verbal Meaning performance. These correlations are not simply determined by group differences in the magnitudes of age differences. Partial correlations, removing Age from the Verbal Meaning and Verbal Meaning Answer Sheet scores, did not alter the picture of increasing correlations of the PMA variable and its Answer Sheet component.

Insert Table 5 about here

Discussion

The major focus of this research was an investigation of the relationship of Perceptual Speed and the speed of marking computerized psychometric answer sheets to individual differences in performance on speeded psychometric tests, including the PMA. The results showed that both the Perceptual Speed and Answer Sheet abilities strongly predicted individual differences in intelligence across the adult life span. The predictive power of the speed measures independent of other variables (Age, Sex, Education) was substantial, and in general, did not interact with these variables. Thus, independent of age, speed of basic perceptual and psychomotor processes is a highly salient correlate of individual differences in psychometric intelligence. This relationship holds for all tests used in this study, indicating that the correlation is not specific to whether or not the psychometric test uses computerized answer sheets.

There were also substantial age differences in psychometric intelligence, as predicted. The cross-sectional age differences in intellectual abilities found in this study are similar to those reported by several investigators, including Schaie and Horn (e.g., Botwinick, 1977; Cunningham, 1987; Horn & Donaldson, 1980; Schaie, 1983). There were substantial age differences on all measures of Induction, Spatial Relations, Spatial Visualization, Flexibility of Closure, and Perceptual Speed, with younger individuals scoring better than older individuals. However, unlike the speed relationship with psychometric intelligence, the age differences in intellectual abilities were substantially modified by partialling the polynomial regression of ability on Age for the two speed factors. Specifically, adjusting the polynomial age curves of composite primary ability measures for Perceptual Speed and Answer Sheet Speed substantially reduced <u>but did not eliminate</u> age differences on measures of Induction, Spatial Relations, Spatial Visualization, and Flexibility of Closure. Moreover, for all these abilities the statistical adjustment resulted in an estimated performance plateau during middle age, delaying the point at which linear decline in the ability became evident. The commonality analysis reported above showed that a small proportion of variance in psychometric intelligence predicted by age is in fact statisticlly independent of the two speed factors.

The findings reported here are consistent with earlier work by Horn (1980; Horn et al., 1981). Horn and his colleagues found that cross-sectional age differences in measures of fluid intelligence (defined by a linear composite of measures such as Letter Sets, Figural Relations, and Paper Folding) were significantly attenuated by partialling for measures of Perceptual Speed. Horn (1980) interpreted the adjustment of fluid intelligence curves by perceptual speed as an indication of the importance of processes such as attentional shift in producing age declines in fluid intelligence. His findings are replicated in this study at the level of the relevant primary abilities: Induction, Spatial Relations, Flexibility of Closure and Spatial Visualization. Adopting Horn's taxonomy of second-order factors (Horn, 1985), this suggests that age gradients for <u>both</u> fluid intelligence (Gf) and visualization (Gv) are attenuated by statistical adjustment for visual information processing speed. It is therefore tempting to speculate that a substantial proportion of early decline in these abilities is speed-related (Cunningham, 1980).

How is the speed relationship with intelligence to be interpreted? This issue is a point of dispute between some authors (Horn, 1985). The first issue is the nature of the speed construct being measured. The perceptual speed and the answer sheet variables correlated highly with each other, supporting the existence of a second-order speediness (Gs) factor (Horn, 1985). Factor analytic work just completed on these data (Hertzog, 1988) suggests that the Answer Sheet speed factor correlates highly (but not perfectly) with Perceptual Speed, and that these factors have somewhat different patterns of correlations with the other primary abilities. Moreover, the Answer Sheet composite variable predicted variance in intellectual abilities independent of Perceptual Speed. This latter finding is important because it indicates that there are probably <u>multiple</u> information processing speed factors with correlated but independent relationships to psychometric intelligence (White & Cunningham, 1987). Mediation of speed/intelligence relationships through a general speed factor would imply that the partial regression of primary abilities on the Answer Sheet variable, removing Perceptual Speed, ought to be zero.

What is the speed dimension tapped by the PMA answer sheet tests? It is not simple psychomotor speed, for Hertzog (1988) found a strong Answer Sheet factor independent of the Crossing Digits test (in which subjects cross off digits in a row as rapidly as possible). The nature of the markings on the answer sheet required use of visual scanning, focused attention, and shift of attention back and forth from test booklet and answer sheet, as well as the psychomotor component of marking the answer sheet itself. It appears likely that, although the Answer Sheet speed includes visuoperceptual processes common to the Perceptual Speed measures, that they emphasize to a greater degree the ability to rapidly shift and focus attention as part of a visual search process. It is interesting to note that education significantly relates to Answer Sheet speed measures, so it is questionable as to whether this measure should be interpreted as "low level" or "noncognitive" in nature. This significant relationship probably reflects the fact that educational attainment is determined by intellectual abilities correlated with Gs, although one could argue that education leads to advantages in strategy formation for this type of task.

This point brings back into focus the theoretical issue: is the speed/intelligence relationship reflective of construct-relevant or performance-specific influences of speed? In all probability, both types of influences operate to produce the empirical correlations. The results reported here and by Hertzog (1988) argue against the strong version of the processing rate theory: namely, that primary aging leads to a general cognitive slowing, manifested in a single speed factor, which is in turn the principal determinant of age-related decline in intellectual abilities. Cognitive slowing may be general, in the sense that it is ubiquitous, but it appears to be multidimensional in nature (Hertzog et al., 1986; Salthouse, 1985a,b). Nevertheless, the strong degree of shared prediction of primary abilities by the two speed factors and chronological age is consistent with the position that cognitive slowing is associated with declines on multiple intellectual abilities.

The clearest evidence for a performance-specific relationship of speed and psychometric intelligence involves the PMA Verbal Meaning subtest. Age curves for the Verbal Comprehension factor differ when PMA Verbal Meaning is or is not included in the composite variable. When Verbal Meaning is included, greater age differences in late life are observed, suggesting that the Verbal Meaning test is more susceptible to detecting age differences. Adjustment of the composite Verbal Comprehension measure for the two speed factors eliminates the late-life decline. More important, there was a strong Age_{linear} by Answer Sheet speed interaction in prediction of PMA Verbal Meaning scores. This interaction implies that speed of marking the answer sheets becomes an increasingly salient predictor of Verbal Meaning performance as individuals age. This pattern of results provides strong support for the hypothesis that Verbal Meaning is more influenced by performance-specific speed components than are the ETS Vocabulary tests used in this battery (Schaie & Hertzog, 1983). It stands in marked contrast to the Perceptual Speed/primary ability relationships, which appear to have approximately equivalent relationships across adulthood. The explanation proposed here is that performance on the Verbal Meaning test is influenced by speed in a way that reduces the construct validity of the test as a measure of vocabulary knowledge in older persons. This pattern is inconsistent with the notion that all speed/intelligence relations reflect the construct-relevant effects of general cognitive slowing (Salthouse, 1985a).

Certainly, use of answer sheets with the PMA is contraindicated by the present findings, and the current results support the efforts of Schaie to distribute a new version of the PMA that does not require answer sheet usage (Schaie, 1986). However, it is unclear at this time whether removal of the answer sheet itself will solve the problem. Given the strong relationship of the Answer Sheet speed factor to measures not employing answer sheets, it is likely that the present analysis is tapping into a more general speed-related phenomenon. It is therefore possible that the interaction with Verbal Meaning is a function of that measure's low to intermediate item difficulty, combined with a liberal time limit, rather than being a function of the use of the computer-type answer sheet per se.

What of the strong relationships of visual information processing speed factors to the other primary abilities? Although these relationships may indicate that speed of information processing is an important manifestation of multiple kinds of intelligence, there are other possible interpretations. It may be the case that psychometric tests in the Thurstone/ETS tradition implicitly place a heavy emphasis upon performance speed. However, the finding of no age X speed interactions suggest that there is a strong speed/intelligence relationship for adults of widely varying ages. This suggests that the finding by Cornelius et al (1983) of high correlations of Gf and Perceptual Speed in the elderly is probably not an indication that psychometric tests are invalid measures of intelligence in old adults. Instead, there appear to be substantial speed/intelligence relationships on these tests across the adult life span. This interpretation is consistent with multiple factor analytic studies that have found invariant relationships between psychometric tests and ability factors across different adult ages (Hertzog, 1987). Although ability factor correlations may increase with advancing age, there is no factor analytic evidence that the construct validity of the tests is adversely affected by the aging process (Cunningham, 1981; Hertzog & Schaie, 1986). In fact, a plausible explanation of increasing ability factor correlations is shared variance in change determined by individual differences in the rate of cognitive slowing (Cunningham, 1981; Hertzog, 1985).

Although it is difficult to directly compare the two studies, it does appear as if the speed-adjusted curves for Gf reported by Horn (1980; Horn et al., 1981) show greater decline in middle age than the speed-adjusted regression equations for Induction reported here. Horn's item pool may contain more items of high difficulty level than do the ETS and Thurstone tests used here. If correct, this explanation has several important implications, particularly with respect to the Thurstone PMA. First, it may be the case that a substantial proportion of the age changes observed by Schaie in his longitudinal studies is not loss of thinking capacity _per se_, but rather, slowing in rate of intelligent thought. This may be an unintended consequence of use of the adolescent form of the PMA, or alternatively, an indirect reflection of Thurstone's original definition of primary abilities as involving the ability to think quickly (Thurstone & Thurstone, 1941). This does not imply that the tests are invalid, or that the age differences they detect are epiphenomena (although this may well be the case for Verbal Meaning). The PMA tests may be valid and salient predictors of age changes in intelligence under real-time processing constraints (e.g., piloting an aircraft). They may, however, be subject to misinterpretation if the goal is to study loss of thinking capacity, especially prior to old age. For example, the 1948 PMA Reasoning test (Form AM) may conceivably _underestimate_ age changes in ability to solve complex induction problems because this level of problem difficulty is underrepresented in its item pool.

The latter point appears to represent a potential point of possible convergence between alternative positions on the speed/intelligence relationship offered by Cunningham (1980) and Salthouse (1985a), on the one hand, and Horn (1985) on the other. Horn (1985) downplays the importance of speed as an explanation of age changes in intelligence, arguing that changes in fluid intelligence are independent of speed. Horn (1985) is probably correct in arguing, in effect, that one cannot and should not use results such as those reported here to attribute most age changes in intelligence to cognitive slowing. In problems of higher difficulty, the speed with which individuals solve problems may have little relationship to the probability of a correct response (e.g., Horn, 1985). One probable reason is that impulsive responding, as a cognitive style, is counterproductive when careful problem analysis is required to maximize probability of a correct solution (Baron, 1985). Thus, the hypothesis that the age changes that are observed in the PMA and the ETS tests are, at least prior to age 60, highly correlated with slowing of intellectual thought, does not <u>necessarily</u> imply that age changes in intelligence are primarily a function of cognitive slowing. It may instead imply that (a) as a function of the way in which certain abilities are defined as involving real-time speed, or (b) the way the speeded tests such as PMA Verbal Meaning were constructed, cognitive slowing plays a major role in the age differences and age changes that have been observed.

In either case, these results raise profound questions about the limits on generalization from the large body of data on age differences and age changes in psychometric intelligence, including the seminal work of Schaie (1983) and his colleagues. What is needed at this point is a careful analysis of the actual processing mechanisms implicated in psychometric test performance and their relationship to age differences in psychometric intelligence.

References

Baron, J. (1985). Rationality and intelligence. New York: Cambridge University Press.

Birren, J. E. (1964). The Psychology of Aging. Englewood Cliffs, NJ: Prentice-Hall.

Birren, J. E. (1965). Age changes in the speed of behavior: Its nature and physiological correlates. In A. T. Welford & J. E. Birren (Eds.), Handbook of aging, behavior, and the nervous system (pp. 191-216). Springfield, Il: Charles C. Thomas.

Birren, J. E., Woods, A. M., & Williams, M. V. (1980). Behavioral slowing with age: Causes, organization, & consequences. In L. W. Poon (Ed.), Aging in the 1980's: Psychological issues. (pp. 293-308). Washington, D.C.: Amercian Psychological Association.

Botwinick, J. (1977). Intellectual abilities. In J. E. Birren and K. W. Schaie (Eds.), Handbook of the Psychology of Aging (pp. 580-605). New York: Van Nostrand Reinhold.

Carroll, J. B. (1980). Individual differences relations in psychometric and experimental cognitive tasks. Chapel Hill: UNC L. L. Thurstone Psychometric Lab (Report No. 163) (NTIS Document AD-A 086 057 ERIC Document ED 191 891).

Cerella, J. (1985). Information processing rates in the elderly. Psychological Bulletin, 98, 67-83.

Cerella, J. C., DiCara, R., Williams, D. & Bowles, N. (1986). Relations between information processing and intelligence in elderly adults. Intelligence, 10, 75-91.

Cohen, J. and Cohen, P.(1983). Applied multiple regression/correlation analysis for the behavioral sciences (2nd Edition). Hillsdale, NJ: Erlbaum.

Cornelius, S. W., Willis, S. L., Nesselroade, J. R., & Baltes, P. B. (1983). Convergence between attention variables and factors of psychometric intelligence in older adults. Intelligence, 7, 253-269.

Cunningham, W. R. (1980). Speed, age, and qualitative differences in cognitive functioning. In L.W. Poon (Ed.), Aging in the 1980's: Psychological Issues (pp. 327-331). Washington, D.C.: American Psychological Association.

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

Cunningham, W. R. (1987). Intellectual abilities and age. In K. W. Schaie (Ed.), Annual review of gerontology and geriatrics (Vol. 7, pp 117-134). New York: Springer.

Cunningham, W. R. & Birren, J. E. (1980). Age changes in the factor structure of intellectual abilities in adulthood and old age. Educational and Psychological Measurement, 40, 271-290.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). Manual for kit of factor-referenced cognitive tests. Princeton, NJ.

Fozard, J. L., & Nuttal, R. L. (1971). General aptitude test battery scores for men differing in age and socioeconomic status. Journal of Applied Psychology, 55, 372-379.

Gardner, E. F. & Monge, R. H. (1977). Adult age differences in cognitive abilities and educational background. Experimental Aging Research, 3, 337-383.

Hertzog, C. (1985). An individual differences perspective: Implications for cognitive research in gerontology. Research on Aging, 7, 7-45.

Hertzog, C. (1987). Applications of structural equation models in gerontological research. In K.W. Schaie (Ed.), Annual Review of Gerontology and Geriatrics (Vol. 7, pp. 265-293). New York: Springer.

Hertzog, C. (1988). Adult age comparisons in primary ability factor structure. Unpublished manuscript.

Hertzog, C., Raskind, C. L., & Cannon, C. J. (1986). Age-related slowing in semantic information processing speed: An individual differences analysis. Journal of Gerontology, 41, 500-502.

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

Horn, J.L. (1980). Concepts of intellect in relation to learning and adult development. Intelligence, 4, 285-317.

Horn, J. L. (1985). Remodeling old models of intelligence. In B. B. Wolman (Ed.), Handbook of intelligence: Theories, measurements, and applications (pp. 267-300). New York: John Wiley & Sons.

Horn, J. L. & Donaldson, G. (1980). Cognitive development in adulthood. In O. G. Brim and J. Kagan (Eds.), Constancy and change in human development (pp. 445-529). Cambridge, MA: Harvard University Press.

Horn, J. L., Donaldson, G. & Engstrom, R. (1981). Application, memory, and fluid intelligence decline in adulthood. Research on Aging, 3 33-84.

Hunt, E. (1978). The mechanics of verbal ability. Psychological Review, 85, 109-130.

Hunt, E. (1983). On the nature of intelligence. Science, 219, 141-146.

Jensen, A.R. (1982). Reaction time and psychometric g. In H.J. Eysenck (Ed.), A model for intelligence (pp 93-132). Berlin: Springer.

Jensen, A.R. (1985). The nature of the black-white difference on various tests: Spearman's hypothesis. The Behavioral and Brain Sciences, 8, 193-264.

Lorge, I. (1936). The Influence of the test upon the nature of mental decline as a function of age. Journal of Educational Psychology, 27, 100-110.

Pedhazur, E. J. (1982). Multiple regression in behavioral research: explanation and prediction (2nd d.). New York: Holt, Rinehart & Winston.

Salthouse, T. A. (1982). Adult cognition: An experimental psychology of human aging. New York: Springer.

Salthouse, T. A. (1985a). A theory of cognitive aging. Amsterdam: North-Holland.

Salthouse, T. A. (1985b). Speed of behavior and its implication for cognition. In J. E. Birren & L. W. Schaie (Eds.), Handbook on the psychology of aging (pp). New York: Van Nostrand Reinhold Company.

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 Press.

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

Schaie, K. W. & Hertzog, C. (1983). Fourteen-year cohort-sequential analyses of adult intellectual development. Developmental Psychology, 19, 531-543.

Schaie, K. W. & Hertzog, C. (1986). Toward a comprehensive model of adult intellectual development: Contributions of the Seattle Longitudinal Study. In R. J. Sternberg (Ed.), Advances in Human Intelligence (Vol. 3, pp. 79-118). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schaie, K. W., Rosenthal, F. & Perlman, R. M. (1953). Differential mental deterioration of factorially "pure" functions in later maturity. Journal of Gerontology, 8, 191-196.

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

Thurstone, L. L. (1944). A factorial study of perception. (Psychometric Monographs, Whole No. 4). Chicago: University of Chicago Press.

Thurstone, L. L. & Thurstone, T. G. (1941). Factorial studies of intelligence. (Psychometric Monographs, No. 2.) Chicago: University of Chicago Press.

Vernon, P. A. (1985). Speed of information processing and intelligence. Norwood, NJ: Ablex.

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

Witt, S. J. & Cunningham, W. R. (1979). Cognitive speed and subsequent intellectual development: A longitudinal investigation. Journal of Gerontology, 34, 540-546.

Footnotes

Author's Note: This study was supported by a grant from the National Institute on Aging (R01-AG06123). The author was also supported by a Research Career Development Award from NIA (K04-AG00335). The study was made possible by the support and assistance of Kathryn Towns of The Pennsylvania State University's Capitol Campus, who provided space for the project field office and assisted in staffing the project. Thanks to Angelica Brennan for coordinating data collection at the field site, to Constance J. Cannon, Cheryl L. Raskind, and Charles Uhlman for assistance in preparing the data for statistical analysis, and to Tim Salthouse and Andy Smith for critical comments on a first draft of this manuscript. Consultation of K. Warner Schaie during the design phase of the project is also gratefully acknowledged.

Address reprint requests to the author at School of Psychology, Georgia Institute of Technology, Atlanta, GA, 30332-0170.

1. Although use of but one of the two parts precluded calculation of alternate forms reliability estimates, subsequent factor analysis of the battery by Hertzog (1988) has shown the measures to have high communalities (and, thus, by inference, high reliabilities).

Table 1

Psychometric Battery in Harrisburg Study

Primary Ability	Measure	Source
Induction	PMA Letter Series	Thurstone & Thurstone, 1949
Induction	PMA Letter Sets	Thurstone & Thurstone, 1962
Induction	PMA Number Series	Thurstone & Thurstone, 1962
Spatial Relations	PMA Space	Thurstone & Thurstone, 1949
Spatial Relations	Card Rotation	Ekstrom et al., 1976
Spatial Relations	Cube Comparison	Ekstrom et al., 1976
Spatial Relations	Object Rotation	Schaie/Thurstone Test (STAMAT)
Spatial Visualization	Form Board	Ekstrom et al., 1976
Spatial Visualization	Paper Folding	Ekstrom et al., 1976
Verbal Comprehension	PMA Verbal Meaning	Thurstone & Thurstone, 1949
Verbal Comprehension	Advanced Vocab. (V3)	Ekstrom et al., 1976
Verbal Comprehension	Advanced Vocab. (V4)	Ekstrom et al., 1976
Perceptual Speed	Number Comparison	Ekstrom et al., 1976
Perceptual Speed	Picture Identity	Ekstrom et al., 1976
Perceptual Speed	Finding A's	Ekstrom et al., 1976
Flexibility of Closure	Hidden Patterns	Ekstrom et al., 1976

Flexibility of Closure	Hidden Figures	Ekstrom et al., 1976
Number Facility	PMA Number	Thurstone & Thurstone, 1949
Number Facility	Addition	Ekstrom et al., 1976
Number Facility	Subtraction & Multiplication	Ekstrom et al., 1976
Perceptual/Motor	Crossing Digits Test	Hertzog and staff
Perceptual/Motor	PMA Answer Sheet, VM	Adapted from Thurstone PMA
Perceptual/Motor	PMA Answer Sheet, S	Adapted from Thurstone PMA
Perceptual/Motor	PMA Answer Sheet, R	Adapted from Thurstone PMA
Perceptual/Motor	PMA Answer Sheet, N	Adapted from Thurstone PMA

Table 2

Univariable Polynomial Regressions of

Intelligence Tests on Education, Sex, and Age

Dependent Variables		Independent Variables	
	R^2	Education b (se)	Sex b (se)
PMA Reasoning	.317	0.672(0.081)***	-1.202(0.454)**
Letter Sets	.269	0.290(0.041)***	0.087(0.231)
Number Series	.272	0.351(0.043)***	0.972(0.240)***
PMA Space	.205	0.521(0.149)***	4.732(0.834)***
Card Rotations	.247	0.641(0.225)**	6.888(1.255)***
Cube Comparison	.159	0.049(0.051)	1.844(0.282)***
Object Rotation	.224	0.547(0.184)**	3.698(1.027)***
Form Board	.140	0.382(0.269).	9.676(1.503)***
Paper Folding	.190	0.074(0.030)*	0.863(0.167)***
PMA Verbal	.257	1.362(0.134)***	-0.451(0.748)
Vocabulary (V2)	.190	1.007(0.088)***	-1.280(0.489)**
Advanced Vocabulary (V4)	.217	1.174(0.097)***	-0.689(0.542)
Number Comparison	.178	0.135(0.046)**	-0.677(0.255)**
Identical Pictures	.296	0.454(0.092)***	0.043(0.514)
Finding A's	.068	0.467(0.115)***	-2.954(0.642)***
Hidden Patterns	.271	2.242(0.388)***	8.261(2.166)***
Hidden Figures	.189	0.260(0.052)***	1.652(0.292)***
PMA Number	.106	0.726(0.163)***	3.635(0.910)***

Addition	.063	0.485(0.107)***	0.954(0.599)
Subtraction / Multiplication	.127	0.842(0.142)***	1.598(0.794)*
Crossing Digits	.070	0.831(0.312)**	2.157(1.740)
Answer Sheet Reasoning	.145	0.337(0.623)***	0.050(0.349)
Answer Sheet Verbal	.293	0.505(0.098)***	0.684(0.545)
Answer Sheet Space	.251	0.598(0.096)***	0.723(0.533)
Answer Sheet Number	.294	0.714(0.114)***	0.080(0.638)

Dependent Variables	Independent Variables	
	AGE 1 b (se)	AGE 2 b (se)
PMA Reasoning	-0.317(0.025)***	-0.007(0.003)**
Letter Sets	-0.184(0.013)***	-0.004(0.001)**
Number Series	-0.132(0.013)***	-0.003(0.001)*
PMA Space	-0.448(0.046)***	-0.010(0.005)
Card Rotations	-0.830(0.069)***	-0.022(0.007)**
Cube Comparison	-0.127(0.015)***	-0.004(0.002)**
Object Rotation	-0.646(0.056)***	-0.026(0.006)***
Form Board	-0.583(0.082)***	-0.018(0.009)*
Paper Folding	-0.091(0.009)***	-0.002(0.001)
PMA Verbal	-0.318(0.041)***	-0.022(0.004)***
Vocabulary (V2)	0.004(0.027)	-0.003(0.003)
Advanced Vocabulary (V4)	0.065(0.030)*	-0.004(0.003)
Number Comparison	-0.134(0.014)***	-0.001(0.002)

Identical Pictures	-0.392(0.028)***	-0.004(0.003)
Finding A's	-0.060(0.035)	0.005(0.004)
Hidden Patterns	-1.429(0.118)***	-0.036(0.013)**
Hidden Figures	-0.124(0.016)***	-0.002(0.002)
PMA Number	-0.195(0.050)***	-0.015(0.005)
Addition	-0.093(0.033)**	-0.004(0.004)
Subtraction / Multiplication	-0.242(0.043)***	-0.007(0.005)
Crossing Digits	-0.528(0.095)***	-0.006(0.010)
Answer Sheet Reasoning	-0.144(0.019)***	-0.003(0.002)
Answer Sheet Verbal	-0.410(0.030)***	-0.005(0.003)
Answer Sheet Space	-0.334(0.029)***	-0.005(0.003)
Answer Sheet Number	-0.463(0.035)***	-0.116(0.004)**

* $p < .05$ AGE 1 - linear

** $p < .01$ AGE 2 - quadratic

*** $p < .001$

Table 3

R^2 and Change in R^2(ΔR^2) for Hierarchical Regressions of

Composite Ability Measures on Speed, Age, and Sex

Dependent Variable	Step 1	ΔStep 2		ΔStep 3	
	R^2	R^2	R^2	R^2	R^2
Spatial Relations	.372	.399	.027**	.472	.073**
Numerical Facility	.317	.339	.022**	.373	.044**
Induction	.455	.480	.025**	.493	.013**
Verbal Comprehension	.229	.256	.027**	.263	.007**
Spatial Visualization	.165	.194	.029**	.277	.083**
Flexibility of Closure	.389	.399	.010**	.466	.067**

Dependent Variable	Step 4		Step 5	
	R^2	ΔR^2	R^2	ΔR^2
Spatial Relations	.472	---	.473	.001
Numerical Facility	.375	.002	.375	---
Induction	.499	.006*	.500	.001
Verbal Comprehension	.273	.010*	.273	---
Spatial Visualization	.284	.007	.284	---
Flexibility of Closure	.469	.003	.472	.003

Note: Independent Variables at Each Step Were:

Step 1 - Perceptual Speed, Perceptual Speed;

Step 2 - $Age_{(linear)}$, $Age_{(quadratic)}$;

Step 3 - Sex;

Step 4 - $Age_{(linear)}$ x Answer Sheet, $Age_{(quadratic)}$ x Answer Sheet;

Step 5 - $Age_{(linear)}$ x Perceptual Speed,

$Age_{(quadratic)}$ x Perceptual Speed.

All R^2 were statistically significant.

Significance levels are shown only for ΔR^2.

* $p < .05$

** $p < .01$

Table 4

Commonality Analysis Partitioning R^2

into Unique and Shared Components

Composite Variables

Variable	Total	Age	Speed	Shared	Prop. Unique (age)
Spatial Relations	.399	.027	.199	.172	.138
Numerical Facility	.339	.022	.308	.009	.716
Induction	.480	.026	.244	.211	.108
Verbal Comprehension	.256	.027	.237	-.009	1.449
Spatial Visualization	.194	.029	.066	.099	.225
Flexibility of Closure	.399	.009	.234	.156	.056

Composite Variables (without PMA tests)

Variable	Total	Age	Speed	Shared	Prop. Unique (age)
Spatial Relations	.405	.032	.196	.177	.155
Numerical Facility	.344	.017	.314	.012	.554
Induction	.407	.021	.205	.181	.105
Verbal Comprehension	.149	.034	.148	-.033	31.399

PMA Tests Only

Variable	Total	Age	Speed	Shared	Prop. Unique (age)
PMA Space	.274	.011	.151	.113	.087

PMA Number	.251	.028	.224	-.002	1.057
PMA Reasoning	.481	.027	.244	.210	.113
PMA Verbal Meaning	.447	.020	.331	.095	.179

Table 5

Correlatons of Answer Sheet Variables with

Corresponding PMA Subtests in Four Separate Age Groups

	Students	Middle-Aged	Young-Old	Old-Old
N	210	148	242	172
Age Range	18-26	42-54	55-66	67-79
Verbal Meaning	.22	.37	.48	.62
Reasoning	.18	.19	.40	.33
Space	.35	.42	.40	.33
Number	.20	.42	.34	.43

Figure Captions

Figure 1. Plot of curvilinear regression for Induction (expressed in z-scores) on age, with an unadjusted curve (solid line) and a speed-adjusted curve (broken line).

Figure 2. Plot of curvilinear regression for Spatial Relations (expressed in z-scores) on age, with an unadjusted curve (solid line) and a speed-adjusted curve (broken line).

Figure 3. Plot of curvilinear regression for Spatial Visualization (expressed in z-scores) on age, with an unadjusted curve (solid line) and a speed-adjusted curve (broken line).

Figure 4. Plot of curvilinear regression for Flexibility of Closure (expressed in z-scores) on age, with an unadjusted curve (solid line) and a speed-adjusted curve (broken line).

Figure 5. Plot of curvilinear regression for Numerical Facility (expressed in z-scores) on age, with an unadjusted curve (solid line) and a speed-adjusted curve (broken line).

Figure 6. Plot of curvilinear regression for Verbal Comprehension (expressed in z-scores) on age, with an unadjusted curve (solid line) and a speed-adjusted curve (broken line).

FIGURE 1

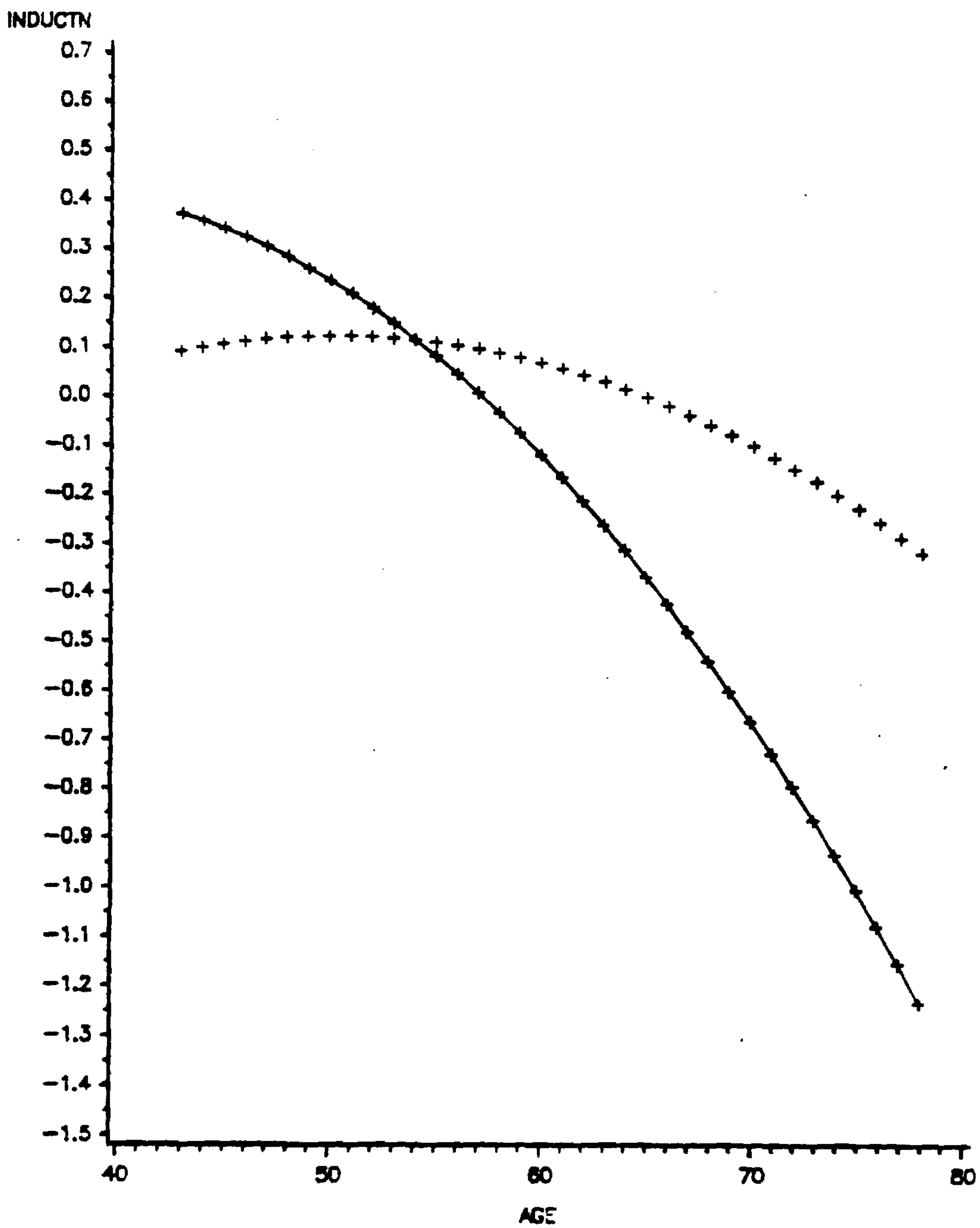

INDUCTN
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.6
−0.7
−0.8
−0.9
−1.0
−1.1
−1.2
−1.3
−1.4
−1.5
40
50
60
70
80
AGE

FIGURE 2

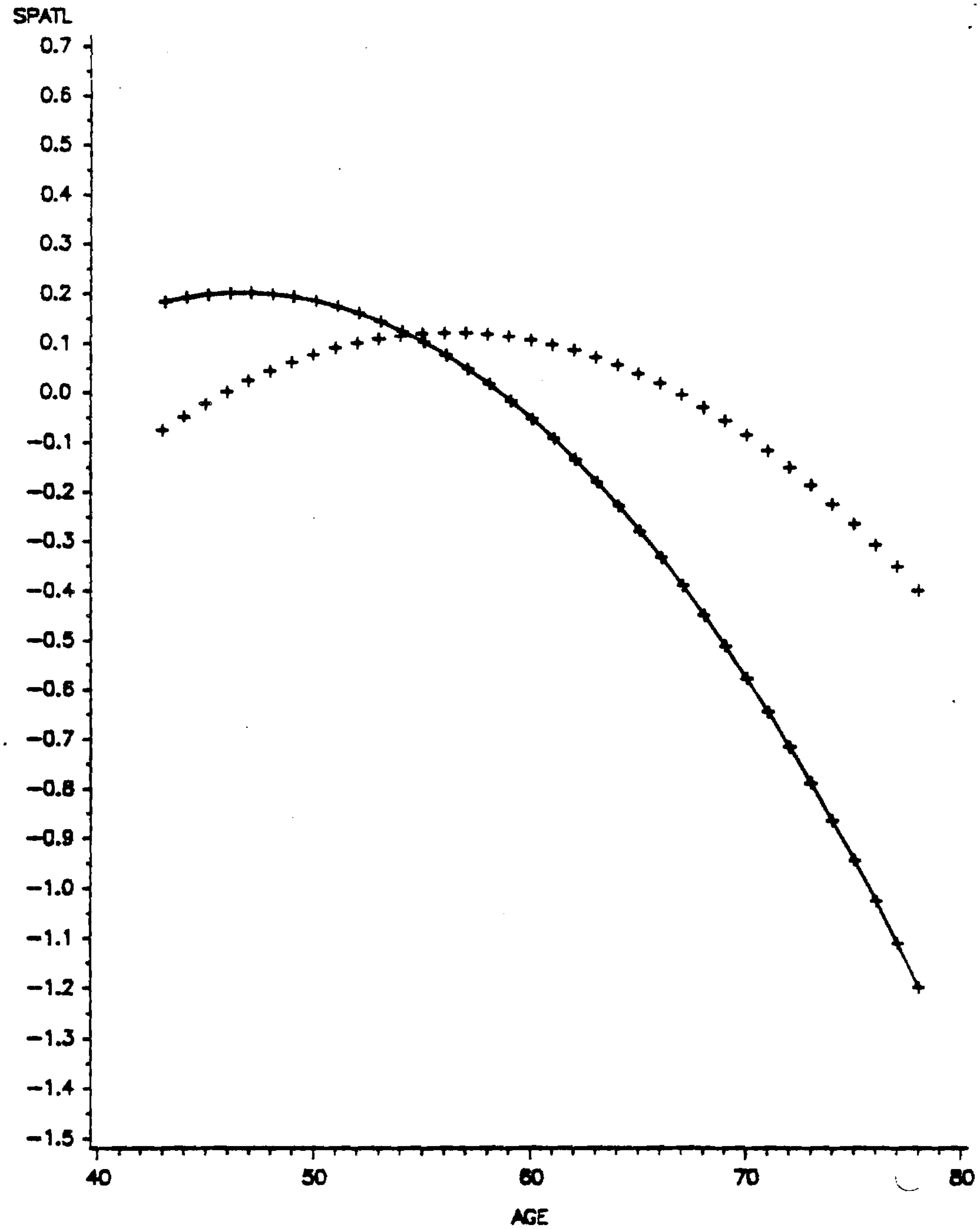

SPATL
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.6
−0.7
−0.8
−0.9
−1.0
−1.1
−1.2
−1.3
−1.4
−1.5
40
50
60
70
80
AGE

FIGURE 3

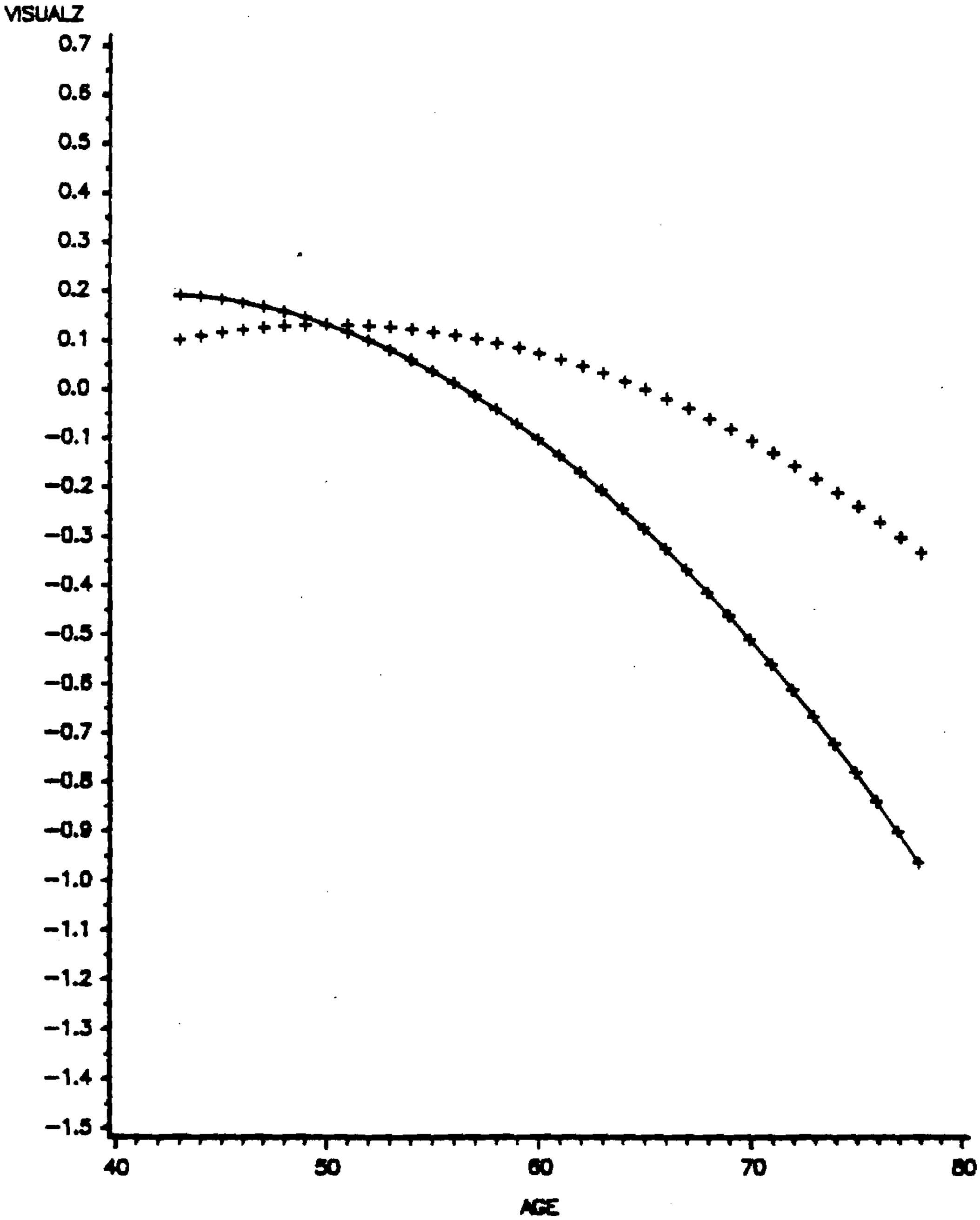

VISUALZ
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.6
−0.7
−0.8
−0.9
−1.0
−1.1
−1.2
−1.3
−1.4
−1.5
40
50
60
70
80
AGE

FIGURE 4

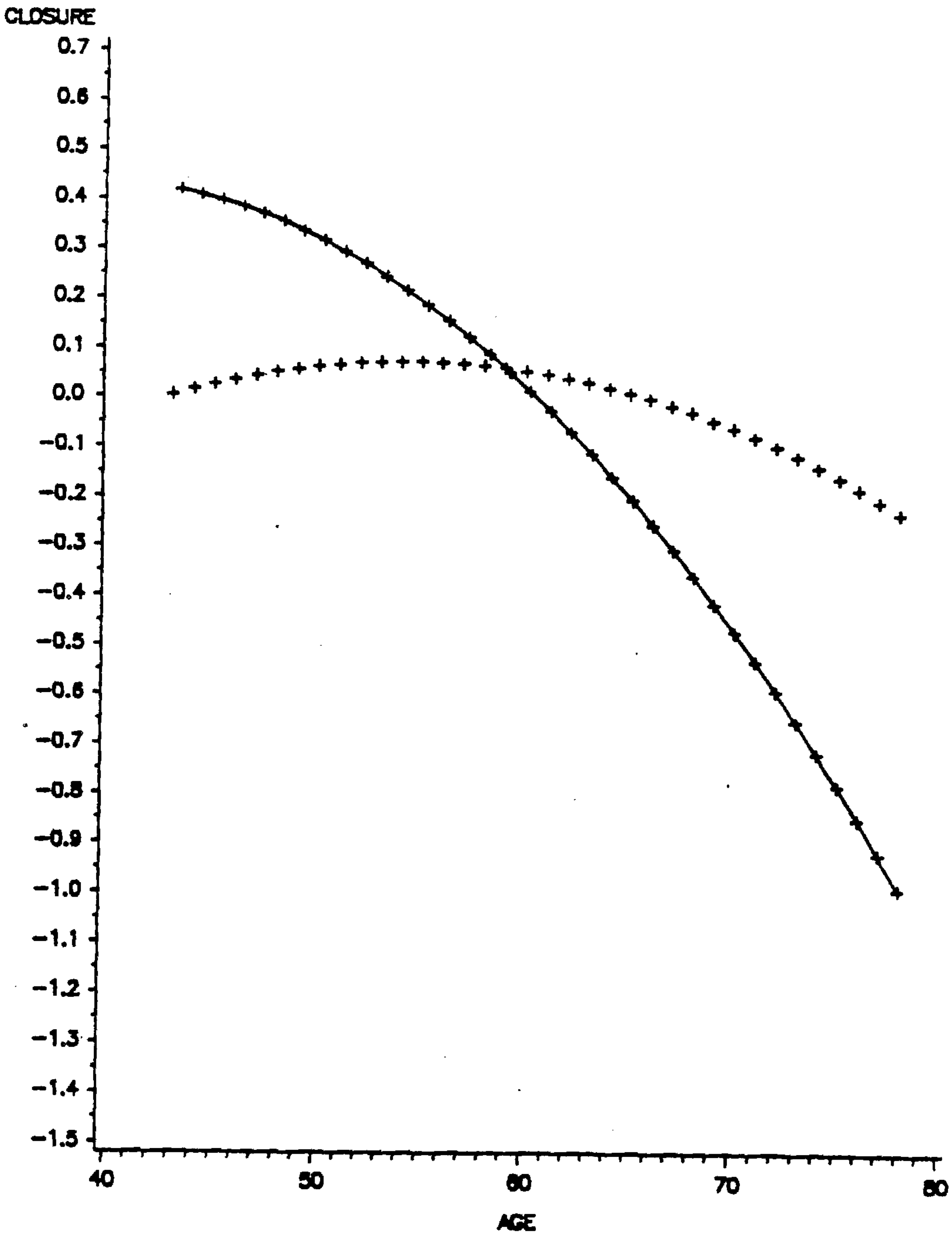

CLOSURE
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4
−0.5
−0.6
−0.7
−0.8
−0.9
−1.0
−1.1
−1.2
−1.3
−1.4
−1.5
40
50
60
70
80
AGE

FIGURE 5

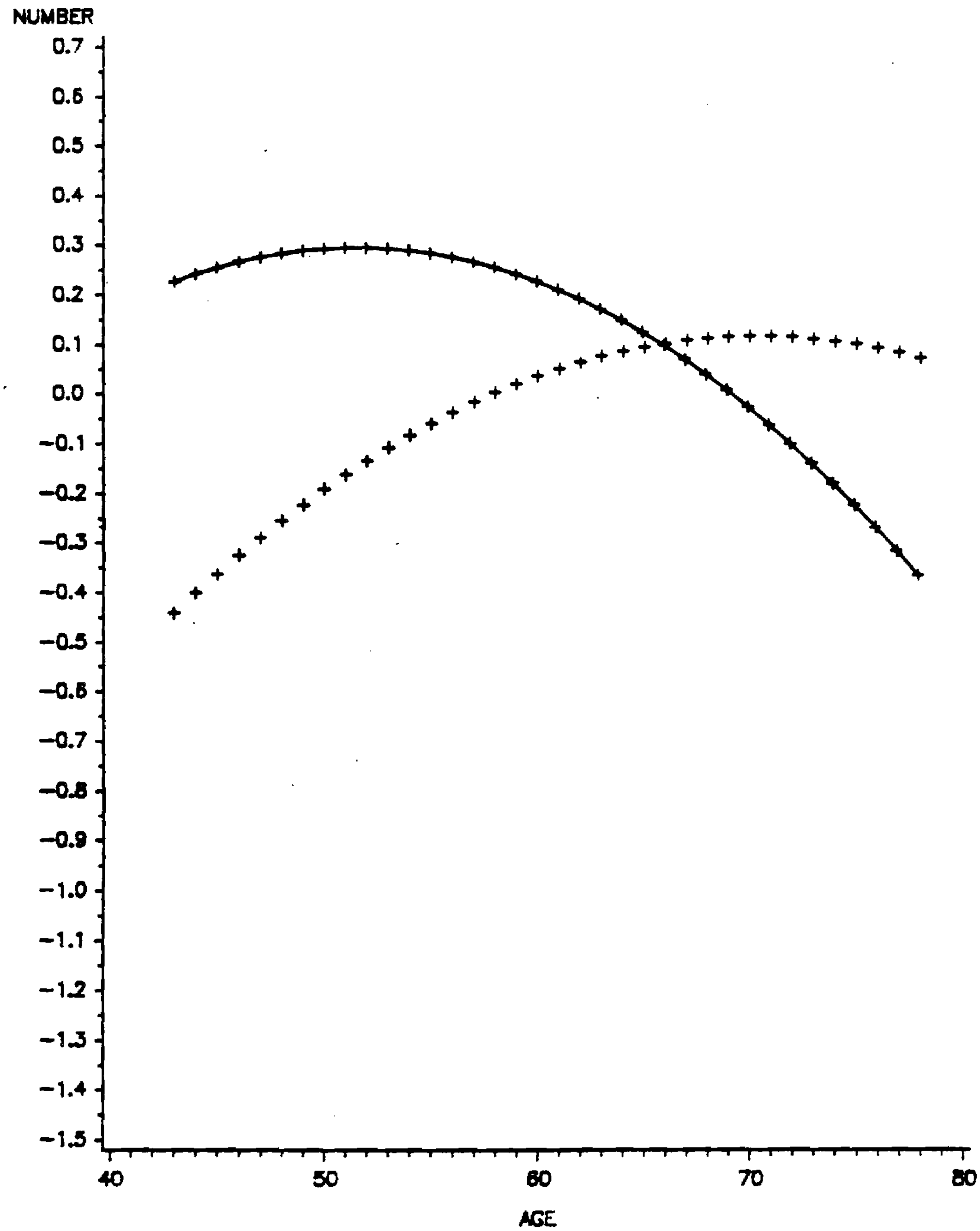
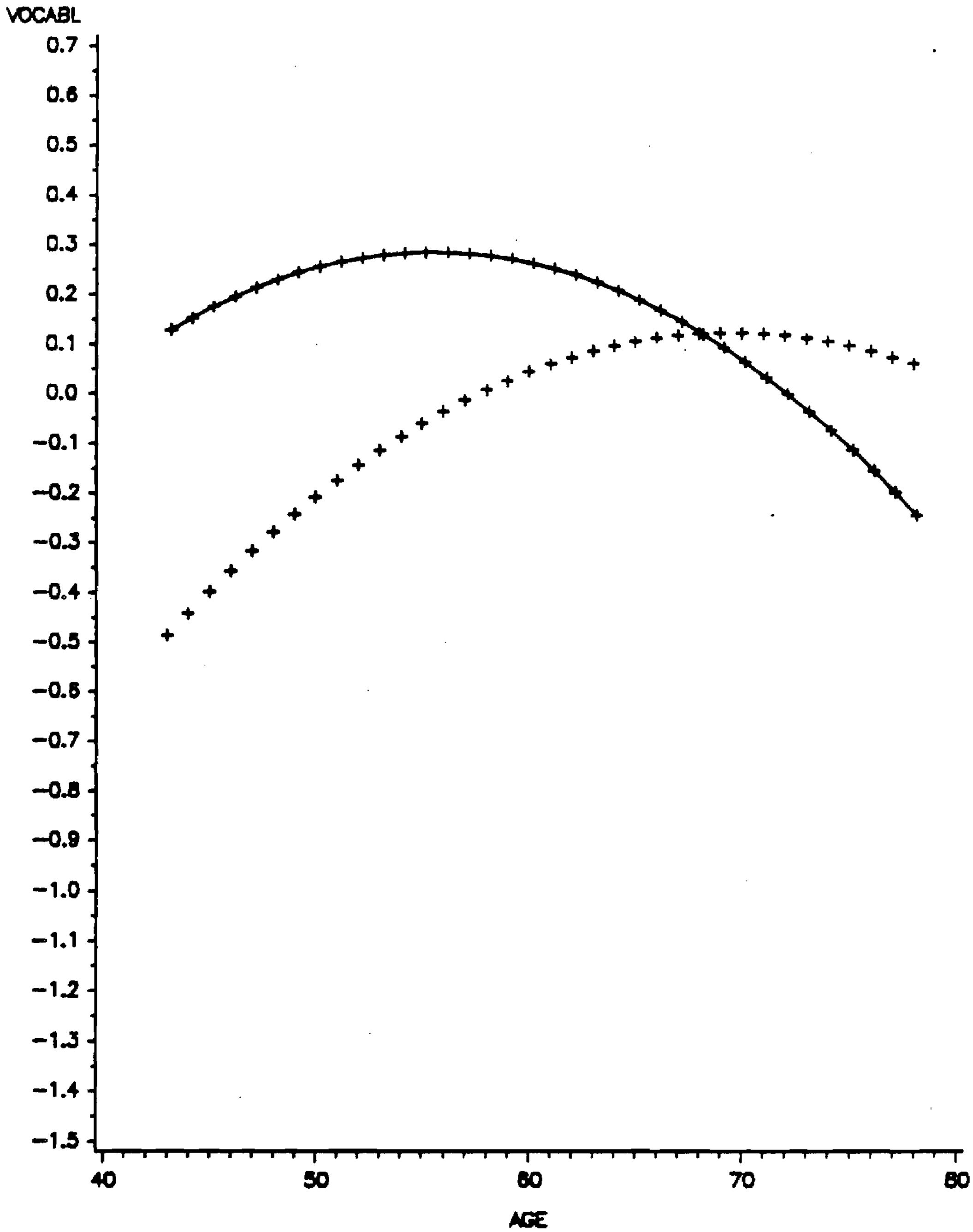

NUMBER
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
-0.9
-1.0
-1.1
-1.2
-1.3
-1.4
-1.5
40
50
60
70
80
AGE

FIGURE 6


VOCABL
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
-0.9
-1.0
-1.1
-1.2
-1.3
-1.4
-1.5
40
50
60
70
80
AGE

Memory Self-Knowledge and Self-Efficacy in the Aged

David F. Hultsch

Department of Psychology
University of Victoria
Victoria, B.C. V8W 2Y2

Christopher Hertzog

School of Psychology
Georgia Institute of Technology
Atlanta, GA 30332

Roger A. Dixon

Department of Psychology
University of Victoria
Victoria, B.C. V8W 2Y2

and

Heather Davidson

Department of Psychology
University of Victoria
Victoria, B.C. V8W 2Y2

Chapter to Appear in:

M.L. Howe & C.J. Brainerd, (Eds.). Cognitive development in adulthood: Progress in cognitive development research. New York: Springer-Verlag

Authors' Notes

The research conducted in our laboratories, and the preparation of this chapter, was supported by Research Grant 492-84-002 from the Social Sciences and Humanities Research Council of Canada to David Hultsch, and Research Grant AG06162 from the National Institute on Aging to Christopher Hertzog. Dr. Hertzog's work is also supported by a Research Career Development Award from the National Institute on Aging. Ms. Davidson is supported by a predoctoral fellowship from the Natural Sciences and Engineering Research Council of Canada. The cooperation of Robert K. Nielsen, M.D., the other physicians, and the members of the Annville Family Practice, Annville, Pennsylvania is deeply appreciated.

Following trends in cognitive psychology, gerontologists have been increasingly interested in how social and personality processes may contribute to cognitive functioning. In the domain of memory, this interest has led to suggestions that age-related changes in basic memory processes may be only one contributing factor in the typically observed decline in performance in later life. In particular, individuals' performance may be shaped not only by their actual skills, but also by their knowledge of the cognitive demand characteristics of the situation, and their perceptions of the likely outcomes of their behaviors in such a situation. Such self-knowledge and self-perceptions have been labeled metamemory. As originally proposed by Flavell and his colleagues (Flavell, 1971; Flavell & Wellman, 1977), emphasis was placed on knowledge about memory. In particular, they suggested that memory performance may be affected by (a) knowledge of the memory demand characteristics of particular tasks or situations; and (b) knowledge of potentially employable strategies relevant to a given task or situation. More recently, the concept has been expanded to include the individual's sense of self-efficacy with respect to memory, either generally, or in relation to a given task or situation. Several writers have suggested that perceived self-efficacy may be a particularly important determinant of memory-related behavior in older adults (Hultsch, Dixon & Hertzog, 1985; Lachman, Steinberg, & Trotter, 1987; West & Berry, 1987).

Two broad methodologies have been used to examine adults' memory self-knowledge and perceived memory self-efficacy. The most prevalent has relied on self-report questionnaires. In recent years, a number of these instruments have been developed for use with adults including the Short

Inventory of Memory Experiences (SIME, Herrmann & Neisser, 1978), the Memory Functioning Questionnaire (MFQ, Gilewski, Zelinski, Schaie, & Thompson, 1983), and the Metamemory in Adulthood instrument (MIA, Dixon & Hultsch, 1983b, 1984) (see Dixon, in press; Gilewski & Zelinski, 1986 for reviews of available questionnaires). In general, these questionnaires have assessed a variety of knowledge and self-efficacy dimensions in relation to a variety of "everyday" memory-demanding situations. In addition, the Dixon and Hultsch (1983b) instrument also examines several affective and motivational dimensions which may be associated with such memory-demanding situations.

The second methodology has indexed metamemory through the application of a number of experimental paradigms (see Cavanaugh, in press for a review). One widely used approach requires subjects to monitor their memory before, during, or following performance of a specific memory-demanding task. The focus may be on assessments of the memorability of particular items as well as the task as a whole. Experimentally based measures of metamemory have generally indexed memory self-knowledge and perceived self-efficacy in relation to "standard" laboratory tasks.

In the present chapter, we will focus on research that has used the questionnaire approach to index metamemory. Initial work with these questionnaires has produced promising results. There is evidence for their reliability and factorial validity (Dixon, in press; Gilewski & Zelinski, 1986). In addition, previous work has provided some indication of the presence of age-related differences on some dimensions of metamemory, as well as evidence of a number of linkages between individuals' self-knowledge and self-efficacy about memory and their actual performance on memory tasks (Chaffin & Herrmann, 1983; Dixon & Hultsch, 1983a; Dixon, Hertzog, & Hultsch, 1986; Zelinski, Gilewski, & Thompson, 1980). However, it is equally clear that despite these preliminary positive results, several fundamental issues remain unresolved. One set of issues relates to the definition and measurement of the metamemory construct itself. A second set of issues revolves around inconsistencies in the pattern of age-related differences on various dimensions of metamemory. Finally, a third set of issues is associated with the question of whether measures of metamemory are (or should be) veridical indicators of actual memory ability. In the following sections, we will examine these three sets of issues in turn.

The Definition and Measurement of Metamemory

Dimensions of Metamemory

Metamemory involves an essential central distinction between remembering and thinking about remembering. However, as Wellman (1983) has noted, the construct rapidly becomes fuzzy once we move away from this central distinction. It seems clear that there may be several dimensions of metamemory, but the question of how many dimensions are required to adequately define the domain remains unresolved. Four broad dimensions that may be relevant are suggested in Table 1. The first dimension

---

Insert Table 1 about here

---

reflects factual knowledge about memory tasks and memory processes. Examples of this dimension would include knowing that short lists of items are easier to remember than long ones, and that organizing the elements of a list of items is likely to improve recall. The remaining three dimensions reflect self-knowledge or perceptions about memory rather than factual knowledge. Memory monitoring involves self-knowledge about how one typically uses one's memory as well as the current state of one's memory; for example, reports of strategy use, feeling-of-knowing judgements (e.g., "I know that I know that"), and assessments of the accuracy of one's responses (e.g., "I got that right"). Memory-self efficacy refers one's sense of mastery within the memory domain. Examples of this dimension would include beliefs about memory capacity, short- and long-term changes in memory functioning, and the degree to which memory functioning is amenable to self-control. Finally, memory-related affect encompasses a variety of states that may be related to, or generated by, memory demanding situations including anxiety, depression, fatigue, and so on.

The dimensions outlined in Table 1 fit Wellman's (1983) broad definition of metamemory in that they all reflect cognitions about memory. However, they have received varying degrees of research attention. In particular, the bulk of work has focused on the first two dimensions. Indeed, developmental differences in memory knowledge were largely the basis for the original definition of metamemory proposed by Flavell (1971). Memory monitoring indices have been widely used in many experimental paradigms designed to examine metamemory. A focus on memory self-efficacy has been emphasized particularly by researchers interested in memory and aging. Attention to affective states generated by memory-demanding situations has received the least attention.

This diversity is reflected in the numerous questionnaires that have been developed to measure the metamemory. For example, some questionnaires such as the Memory Complaints Questionnaire make no distinction among different dimensions of metamemory, yielding only a total score (e.g., Zarit, 1982; Zarit, Cole, & Guider, 1981; Zarit, Gallagher, & Kramer, 1981). Inspection of the items from this and other "single score" measures suggests that multiple dimensions have been combined. Other questionnaires appear to examine a single dimension or facet of a dimension in some depth. For example, several questionnaires such as the SIME (Herrmann & Neisser, 1978) assess individuals' perceptions of the difficulty they have remembering within particular content domains (e.g., names, errands, conversations). Such questionnaires may be thought of as in-depth assessments of perceived memory capacity which can be considered to be one facet of memory self-efficacy. Finally, other questionnaires appear to explicitly tap several of the dimensions outlined in Table 1. For example, the MFQ (Gilewski et al., 1983) contains subscales that index aspects of memory monitoring and memory-self efficacy. Our own measure, the MIA (Dixon & Hultsch, 1983b, 1984), incorporates elements from all four of the dimensions noted above. Since the MFQ and MIA have been widely used in aging work, we will briefly describe these two questionnaires.

The 64 items of the MFQ are distributed into eight a priori scales as shown in Table 2. The instrument is a shortened version of a 92-item

---

Insert Table 2 about here

---

instrument originally developed by Zelinski et al. (1980). In the original sample, Cronbach's alpha for the various subscales ranged from .82 to .93, and three-year test-retest reliabilities ranged from .22 to .64 (Zelinski et al., 1980). Factor analysis of the a priori scales has revealed three common factors including Frequency of Forgetting, Seriousness, and Mnemonics/Retrospective Functioning. This factor structure replicated in young (29-39 years) and old-old (71+ years) age groups, but not a young-old (55-70 years) age group. In this latter group, the three factors were Frequency of Forgetting, General Rating, and Mnemonics.

The 120-item MIA is composed of eight factor analytically defined dimensions seven of which are summarized in Table 3. (In our recent work we have dropped the original Activity subscale for conceptual and measurement reasons). As summarized in Table 4, Cronbach's alpha for the various

---

Insert Tables 3 & 4 about here

---

subscales ranged from .71 to .92 across multiple samples. The demonstrated reliability of the MIA (and the MFQ) is reassuring. Several writers (e.g., Gilewski & Zelinski, 1986; Dixon & Hertzog, in press) have noted that many of the metamemory instruments now available have unknown or low levels of reliability. Further, there are important consequences of low reliability for determining convergent and discriminant validity, and for estimating metamemory/memory performance relationships (e.g., Dixon & Hertzog, in press; Rushton, Brainerd, & Pressley, 1983). We will return to this issue later.

Prior work with multiple samples has also indicated the subscales of the MIA are factorially valid (Dixon & Hultsch 1983b). More recently, Hertzog, Dixon, Schulenberg, and Hultsch (1987) tested the hypothesis that the eight subscales of the MIA tap higher order metamemory factors reflective of memory knowledge, memory self-efficacy, and memory-related affect dimensions. Data from the six separate studies involving a total of 750 subjects were combined to yield two half-samples for cross-validation purposes. Each of the samples were partitioned into young, middle-aged, and old groups in order to examine the consistency of factor structure at different ages. A multiple groups confirmatory factor analysis was conducted on the data using the first half sample to develop a model and the second half sample to validate it. Although the models did not fully cross-validate, both analyses indicated that there are at least two higher-order factors in the MIA. The first, labeled Memory Self-Efficacy, involves beliefs about competence associated with memory-demanding situations. The second, tentatively labeled Memory Knowledge, combined knowledge about memory and affect concerning memory. Factor loadings for Memory Knowledge were invariant across the three age groups; Strategy, Task, Achievement, and Anxiety consistently loaded on this factor. In contrast, there were significant age differences in the weights associated with the Memory Self-Efficacy factor. It was generally defined by Capacity, Change, Anxiety, and Locus, but loadings for the Change and Locus subscales were substantially higher in the old than in the young.

In sum, metamemory seems to be most productively considered as a multidimensional construct. At minimum, it is possible to differentiate memory knowledge from memory self-efficacy dimensions. Multiple facets may exist within these broad dimensions. In addition, there may be age differences in the structure of these dimensions. In particular, the composition of memory self-efficacy dimensions may be different for older as compared to younger adults. Specifically, perceptions of change in memory and perceptions of reduced control over memory are more salient for the elderly.

Convergent and Discriminant Validity

To this point, psychologists interested in constructing metamemory questionnaires have focused on developing and validating their own instruments, with relatively little attention to the similarity of their instrument to others. As discussed by several reviewers (Dixon, in press; Gilewski & Zelinski, 1986; Herrmann, 1982), further work investigating the validity of metamemory questionnaires for adult populations is critically needed. In particular, traditional issues associated with construct validity of psychological measures (e.g., Cronbach & Meehl, 1955; Messick, 1981) have yet to be addressed with respect to metamemory questionnaires (Dixon, in press; Dixon & Hertzog, in press). One issue is the degree of convergent validity between different metamemory questionnaires. As outlined above, different questionnaires emphasize different domains of metamemory. For example, the MIA measures affect about memory not explicitly assessed by other questionnaires. There is, however, reason to wonder whether memory self-efficacy, as measured by the MIA Capacity scale, is the same construct as measured by the MFQ or the SIME. The principal scales from the MFQ and the SIME are relatively similar to one another. Both query the respondent about frequency of forgetting problems in specific domains of memory, such as forgetting names. An overall frequency of forgetting score is then calculated by summing frequency of forgetting ratings across the domain of forgetting instances. Although there are some differences between the SIME and the MFQ in terms of the selection of forgetting instances and the specificity of question wording (see Gilewski

& Zelinski, 1986), it is reasonable to expect a high degree of convergent validity between the two scales.

If the MFQ and SIME also measure memory self-efficacy with their frequency of forgetting scales, then they should display convergent validity with some MIA scales (particularly Capacity). There is, however, a greater opportunity for divergence between the MIA and the other two scales, given differences in the way self-efficacy questions are phrased in the MIA (see Table 3). There is at this time some limited evidence that the MIA and the SIME are significantly correlated. Cavanaugh and Poon (1985) found evidence of a substantial correlation between the MIA Capacity scale and the total score on the SIME, but with a very small sample of older adults. There has been, prior to the work described in this chapter, no information relating the MFQ to the MIA or the SIME.

A second issue regarding the validity of metamemory questionnaires is their discriminant validity from other, theoretically related constructs. Can memory self-perceptions be differentiated from well-established constructs such as locus of control, self-esteem, and personality? Surprisingly, little effort has been made to this point to demonstrate that cognitive psychologists have not merely re-discovered such well-known constructs and given them a different label! This is an obvious concern for the MIA, where it seems quite plausible that measures of perceived locus of control or anxiety regarding memory might measure nothing more than general locus of control and trait anxiety. In addition, it is plausible that such measures are highly influenced by the emotional state of the respondent at the time of the self-rating. Zarit (1982) has suggested that older adults' complaints about their memory may reflect their degree of depression as much, if not more, than their actual memory capacity. Indeed, some studies have found higher correlations between memory complaints and depression than between memory complaints and actual memory performance (e.g., Kahn, Zarit, Hilbert, & Niederehe, 1975; West, Boatwright, & Schleser, 1984). Although such findings do not necessarily indicate that perceived memory self-efficacy is determined by depression, they do indicate that close examination of the issue is required.

In order to address the validity of the MIA and MFQ, we have recently conducted a major study which used confirmatory factor analysis to explicitly evaluate the convergent validity of the two questionnaires and their discriminant validity from related constructs. Several methodologists have noted that confirmatory factor analysis is an ideal tool for assessing validity (e.g., Bentler, 1978; Hertzog, 1985; Joreskog, 1974). Like its counterpart, structural equation models, confirmatory factor analysis estimates relations among latent variables (factors) which are disattenuated for measurement error. This means, for example, that it is meaningful to test whether a factor correlation is 1.0 in a population (because 1.0 is a true upper bound of the latent variable correlation; see Joreskog, 1974). A meaningful operational definition of convergent validity, then, is that two latent variables have a disattenuated correlation of 1.0. Similarly, confirmatory factor analysis is helpful in assessing discriminant validity. With this approach, low correlations among factors imply a low degree of shared variance in the latent construct (and not any artifact of poor reliability). Thus, showing low correlations among factors implies discriminant validity of the constructs reflected in those factors.

The validation study included two samples drawn from rather different populations. One sample was drawn from a medium-size western Canadian city (Victoria, British Columbia). The second sample was drawn from a semi-rural area in the eastern United States (Annville, Pennsylvania). The Victoria sample consisted of 378 individuals (100 university students, 278 adults ages 55-78 years), whereas the Annville sample included 447 adults (age range 20-78 years). Additional details regarding the samples may be found in Hultsch, Hertzog, and Dixon (1987).

Convergent validity of the MIA and MFQ. Do the MIA and the MFQ measure the same dimensions of metamemory? Clearly, the MFQ does not measure the same aspects of memory-related affect, but there appears to be substantial overlap in other domains. The critical dimension, given our previous discussion, is memory self-efficacy. The Hertzog, Dixon, Schulenberg, and Hultsch (1987) analysis suggested that the MIA scales of Capacity, Change, Anxiety, and Locus all loaded on a dimension we interpreted as Memory Self-Efficacy. A content analysis of the MFQ suggested that its Global Rating Scale, Memory Problems, and Remote Memory scales, do indeed measure interrelated aspects of Memory Self-Efficacy. The MFQ also measures in great depth self-reported memory problems for reading materials. This also seemed to be an aspect of Memory Self-Efficacy, although it seemed plausible that its specificity would cause a less-than-perfect correlation of it with a more general Memory Self-Efficacy factor.

Our content analysis of the MFQ is consistent with an unpublished study of the factor structure of the MFQ reported by Gilewski et al. (1983). Their confirmatory factor analysis of the MFQ found one strong factor which seems to be Memory Self-Efficacy (including high loading on the Memory

Problems and Memory Rating indicators). Two weaker factors were also reported, that in essence appeared to be dominated by single measures: Strategy Use and Perceived Seriousness. The Gilewski et al. (1983) analysis thus appears to identify a strong Memory Self-Efficacy factor in the MFQ that in theory should converge with the Memory Self-Efficacy factor in the MIA found by Hertzog, Dixon, Schulenberg, and Hultsch (1987).

Aside from Memory Self-Efficacy, there are other overlapping scales between the two questionnaires. Both contain a self-reported strategy use scale. Both contain a scale asking individuals to assess perceived change in memory capacity (MIA Change, MFQ Retrospective). Although these may be primary markers of a Memory Self-Efficacy factor, it is plausible that indicators of perceived change would form a factor independent of Memory Self-Efficacy.

Hertzog, Hultsch, and Dixon (1987) conducted an extensive series of confirmatory factor analyses designed to evaluate the degree of convergence between the MIA and MFQ scales. The analysis was conducted in three phases. First, an exploratory model was developed using the data from the Annville sample. Second, this model was replicated (cross-validated) in the Victoria adult sample. Finally, a multiple groups factor analysis was run to determine age-related invariance in the joint factor structure of the two scales. We shall only summarize the salient results of the analyses here.

The exploratory analysis of the Annville sample forced immediate reassessment of some of our hypotheses about the factor structure of the metamemory scales. First, the zero-order correlations among the MIA Task, Strategy, Locus, and Achievement scales were lower than in the original

Dixon and Hultsch (1983b) samples. These correlations, generally below .3, indicated that we would not be able to replicate the second factor (labelled Memory Knowledge) found by Hertzog, Dixon, Schulenberg, and Hultsch (1987). In fact, the low correlation of the MIA Task with either the MIA or the MFQ Strategy scales was a surprising finding, for it indicated that simply stripping the more affect-related MIA scales such as Achievement off the factor would not suffice in defining a memory knowledge factor. On the other hand, the high correlation of the two strategy use scales suggested that a convergent Strategy factor could be modeled. Second, the Seriousness scale of the MFQ had virtually zero correlations with other scales and was hence eliminated from the analysis.

After a series of model building exercises, we arrived at a basic specification that appeared to account for the structure of the two questionnaires. The model specified two Memory Self-Efficacy factors, one for each questionnaire, a MFQ Reading Self-Efficacy factor (marked by problems in remembering materials from novels and problems remembering newspapers and magazines), a strategy use factor, a memory-related affect factor (marked chiefly by MIA Achievement), a Change factor, (marked by MIA Change, MFQ Retrospective, and Locus), and MIA Task (treated as a single indicator).

The model fit well in both samples. The LISREL Adjusted Goodness-of-Fit index was .945 for the Annville sample and .943 for the Victoria sample. This index has a maximum of 1.0 when a model fits a set of data perfectly, and fits greater than .9 are usually considered excellent for this index.

Table 5 reports the factor loadings for this model for both the Annville sample and the Victoria sample. We estimated standardized solutions in both samples which can act to accentuate sample differences in the model's parameter estimates. Moreover, the two samples differed widely in age range, so differences might be expected both because of sample differences in variances and age differences in the factor

----

Insert Table 5 about here

----

structure of metamemory. The similarity, then, of the standardized factor loadings reported in Table 5 is impressive. Indeed, when _t_-tests of the differences between sample estimates were computed, none of the estimates differed significantly from each other. The largest difference, that for the loading of Anxiety on the Memory-Related Affect factor, yielded a _t_-test of 1.79, which is not significant at the 5% level of confidence.

We did conduct a more direct test of age differences in factor loading (see Hertzog, Hultsch, & Dixon 1987 for more details). The two samples were split into two age groups. The old sample, approximated the age range of the Victoria sample, and then a middle-aged group was constructed from the 20-55 year olds.[1] In the Victoria sample, the student group defined a young age group to supplement the old group already analyzed. We then ran a series of simultaneous multiple group factor analyses (Joreskog, 1971) designed to test group differences in factor loading. None of the groups differed significantly in factor loading. The result was somewhat

---

[1]Most of these subjects were age 33 and older.

surprising, for it disagreed with the findings of Hertzog, Dixon, Schulenberg, and Hultsch (1987), who found significant age group differences in the loading of the MIA scales on a Memory Self-Efficacy factor. The explanation of the discrepant results is happily straightforward. There were significant age group differences in the correlation of the Change factor with the Memory Self-Efficacy factor. As might be expected, perceived change was more highly correlated with self-efficacy in the old groups. In the Hertzog, Dixon, Schulenberg, and Hultsch (1987) analysis of the MIA, it was not possible to separately estimate the Change factor. Thus, the different relationship of change to self-efficacy was absorbed, as it were, into the loading of MIA Change and MIA Locus on the Memory Self-Efficacy factor. In sum, it appears that the factor solutions were replicable across multiple groups, and that the different age groups had equivalent factor loadings, but differed in factor correlations.

Table 6 reports the factor correlations for the four different age groups.[2] Table 6 includes crucial information regarding the

----

Insert Table 6 about here

----

question of convergent validity: the correlation between the MIA and MFQ Memory Self-Efficacy factors. We can now see the benificial effect of using confirmatory factor analysis to assess convergent validity. As discussed above, if it were the case that the two scales measure the same

[2]The models computed factor covariance matrices, but these have been standardized for ease of interpretation and discussion.

Memory Self-Efficacy construct, then the correlation between the two Memory Self-Efficacy factors ought to be 1. The first row of correlations in Table 6 shows that these correlations were uniformly large, albeit larger in the Annville than in the Victoria groups. Correlations of this magnitude justify the conclusion that the two scales are, indeed, measuring the same construct. It should be noted, however, that formal tests of this hypothesis, achieved by constraining the factor correlation to equal 1 (see Joreskog, 1974), are statistically significant for the Victoria samples. Practically speaking, the divergence from 1 is relatively trivial, but some patterns in the estimated factor correlations indicate differences between the MIA and MFQ Memory Self-Efficacy factors in their relations to other factors. For example, the Reading Self-Efficacy factor (taken from the MFQ) correlates more highly with the MFQ Memory Self-Efficacy factor than with the MIA Memory Self-Efficacy factor. It seems likely that the minor differences are a function of method variance associated with the different types of questions and Likert response formats; indeed, Hertzog, Hultsch, and Dixon (1987) report results from a model consistent with this hypothesis. Thus, with respect to the major metamemory factor, Memory Self-Efficacy, the MIA MFQ converge to measure the same construct. The viability of the Strategy Use factor also indicates convergence of the MIA and MFQ in measuring self-reported use of memory strategies (especially, external aids).

The correlations reported in Table 6 also support the conclusion that there are multiple dimensions of metamemory. In all four samples there is a small, negative relationship between Memory Self-Efficacy and Strategy Use, with individuals low in self-efficacy more likely to report more strategy use. All four samples also show virtually zero correlations between Memory Self-Efficacy, measured either by the MIA or the MFQ, and the MIA Task scale. This finding buttresses the contention that knowledge about how memory functions is independent of memory self-efficacy beliefs.

Discriminant validity. We are now conducting a series of multivariate analyses exploring the discriminant validity of the metamemory factors from related constructs. At this point we can draw some preliminary conclusions from patterns of simple correlations of the metamemory scales with scales measuring personality, locus of control, and affective states.

Table 7 reports correlations of the MIA scales with measures of personality taken from the Jackson Personality Inventory and the Jackson Personality Research Form for the Annville sample. The highest correlations involve the MIA Anxiety scale with the personality scales Anxiety, Self-Esteem, and Conformity. This finding supports the hypothesis that part of the variance in MIA Anxiety is associated with the personality dimension of Neuroticism (and of course, more specifically, trait Anxiety). With respect to the first three scales listed in Table 7, Locus, Capacity, and Change, the correlations are relatively low. There therefore appears to be little cause for concern that Memory Self-Efficacy, as measured by these scales, is a re-expression of basic personality dimensions.

Insert Table 7 about here

It is also interesting to note the low correlations of MIA Locus with the three scales taken from the Levenson Internal/External Locus of Control scale (Internal, Powerful Others, and Chance). Levenson (1981) has argued that internal control orientations (as assessed by her Internal scale) are distinct from two aspects of an external control orientation (as measured by Powerful Others and Chance). One would expect a high correlation of MIA Locus with Levenson Internal if, in fact, the MIA Locus scale was just a re-expression of a general internal control orientation. The small correlations reported in Table 7 provide no basis, then, for arguing that Locus is just another measure of generalized locus of control. This result fits well with work by Lachman (1983, 1986), who has shown that control beliefs are both general and domain-specific. Lachman also demonstrated that control beliefs about intellectual functioning predicted performance on intelligence tests, whereas the Levenson scales did not. It does appear from Table 7 that there are modest correlations of Levenson Internal Locus of Control with all three indicators of Memory Self-Efficacy, and that Memory Self-Efficacy correlates more with internal than external locus of control.

The remaining correlations in Table 7 are from two measures of psychological well-being and depression. The Veit/Ware scales refer to revisions by Veit and Ware (1983) of a questionnaire measuring aspects of well-being and distress that was originally administered as part of a government study of health and well-being in the United States. Similar measures of well-being and distress were taken from the Center for Epidemiological Studies Depression scale (CES-D; see Radloff, 1977). The Veit/Ware and CES-D scales are scored in opposite directions of the relationships. Examination of these correlations shows that the MIA scales generally correlate at low levels with the measures of perceived well-being and depression. The major exception, again, is the MIA Anxiety scale.

There is little cause for concern, however, that perceived memory self-efficacy is indistinguishable from depression.

As indicated above, a second set of issues involves the degree to which responses on the MIA are influenced by concurrent affective state -- particularly when responding to questions about control, anxiety, and achievement motivation. Table 8 reports the correlations of the MIA with two sets of mood state measures: (a) scales from Cattell's Eight State Questionnaire, and (b) mood adjective rating scales, with adjectives used from the Profile of Mood States and Lebo and Nesselroade (1978). These correlations were consistent with those already reported for personality, control, and depression. MIA Anxiety showed salient correlations with mood states of anxiety and depression. None of the other MIA scales showed much relationship to the mood rating variables.

Insert Table 8 about here

In sum, our preliminary analyses of the discriminant validity data suggest that (a) MIA Anxiety has substantial relationships to trait anxiety and related affective states; (b) none of the other MIA scales shows strong relationships to these variables; (c) the Memory Self-Efficacy factor identified in the MIA is not merely generalized locus of control, although there appears to be a modest relationship between internal locus of control and Memory Self-Efficacy; and (d) the small correlations seem to rule out the hypothesis that Memory Self-Efficacy ratings are highly related to depression or concurrent depressive affect.

Age and Sex Differences in Metamemory

Researchers interested in metamemory and aging have, of course, examined the basic question of the existence of age and other group (e.g., gender) differences in metamemory. The answer to this basic question has not been particularly straightforward. Specifically, the consistency and robustness of such group differences is somewhat unclear. This lack of clarity may be due, in part, to some of the measurement issues noted above. Studies have varied widely in the definition of the construct and the particular instrument used to operationalize it. However, inconsistencies have also appeared in work administering the same questionnaire to different samples. Nevertheless, examination of multiple data sets has begun to paint a reasonably consistent picture that permits us to unravel some of the confusion present in the literature.

Age differences

As noted above, the literature presents a plethora of conclusions regarding adult age differences in metamemory. For example, although several studies have failed to find age differences in reported use of memory strategies (e.g., Dixon & Hultsch, 1983b; Gilewski et al., 1983; Perlmutter, 1978), others have reported that older adults use fewer strategies than younger adults (Weinstein, Duffy, Underwood, MacDonald, & Gott, 1981). There has also been disagreement about whether older adults report more memory failures in everyday activities than younger adults. Whereas some studies (e.g., Gilewski et al., 1983: Perlmutter, 1978) report negative age differences, others (e.g., Sunderland, Harris, & Baddeley, 1983) have found that younger adults actually reported more such incidents that older adults. Similarly, a mixed pattern of results appears for indicators of perceived memory abilities or capacities. Although Dixon and Hultsch (1983b), Gilewski et al. (1983), and Zelinski et al. (1980) found older adults had a poorer perception of their memory for various content domains than younger adults, Chaffin and Herrmann (1983) found a mixed pattern of results (including positive, equivalent, and negative age differences) across domains, and Bennett-Levy and Powell (1980) reported a positive age effect on their measure.

Given the diversity of measurement approaches used in the work summarized above, it is extremely difficult to sort out age effects from measurement effects. However, data on multiple samples are available for at least two multidimensional metamemory instruments: the MIA and the MFQ. In the case of the MIA, data are available for seven samples varying in size and nationality (Cavanaugh & Poon, 1985; Dixon & Hultsch, 1983b; Gutman, 1987; Hultsch et al., 1987). The nature of the samples and the pattern of age differences emerging on the various subscales of the MIA in the different samples are shown in Table 9. Although there are

----

Insert Table 9 about here

----

inconsistencies across the samples, the pattern of results suggests that there may be reliable age differences on the Capacity, Change, and Locus subscales. Significant age differences on these indicators are observed in most of the samples. Further, in most instances, the age effects associated with these subscales are accounting for substantial portions of variance (Capacity: range = 3-10%; Change: range = 13-37%; Locus: range = 3-19%). Across multiple samples, then, there is evidence to suggest that, compared to younger adults, older adults see themselves as having less memory capacity, report that their memory has declined over the years, and believe that there is little that they can do to enhance their memory or prevent its deterioration.

It should be noted that the age differences observed appear to be most pronounced when a contrast is drawn between a young university student sample and middle-aged and older community residents. For example, Hultsch et al. (1987) tested two large samples which differed in their demographic characteristics. One sample was designed to represent the entire adult age range, and consisted of younger, middle-aged, and older adults, none of whom was enrolled full-time in university at the time of testing. Hierarchical regression analyses of the data from this sample indicated significant linear effects related to age on the Capacity and Change subscales, and a marginally significant trend on the Locus subscale. The other sample was designed to be comparable to the "traditional" cross-sectional sample typically used in cognitive aging research. In this case, the younger adults were university students and the middle-aged and older adults were healthy, community-dwelling volunteers. In this sample, the significant age effects, which included the Capacity, Change, and Locus subscales, were generally due to differences between the youngest group and the remaining groups. Differences among the various older groups were generally not significant. Similarly, a hierichical regression analysis conducted on the data from the middle-aged and older age groups from this sample did not show any significant linear trends. This suggests that age differences at the mean level within the middle to older age ranges may be relatively fragile. Thus, some of the discrepant findings in the literature may be partially due to the nature of the subjects sampled from these portions of the life span.

In the case of the MFQ, results from four separate samples are available. The sample characteristics and pattern of age effects across them are shown in Table 10. The pattern of results suggests consistent age

Insert Table 10 about here

differences are present on the Retrospective Functioning subscale. There are less consistent indications of differences on the Global Rating and memory problems associated with reading. In general, it appears that the MIA and MFQ are differentially sensitive to detecting mean age differences. Such differences are more likely to be found with the MIA than with the MFQ. The strongest evidence for this conclusion comes from the Hultsch et al. (1987) study which administered both questionnaires to two large samples. As indicated in Tables 9 and 10, age differences were found on both measures in the traditional sample contrasting young university students and older community residents. However, even in this case, the magnitude of the effects were generally smaller in the case of the MFQ than in the case of the MIA. In the other sample that sampled community residents (non-students) from the entire age range, several significant age differences emerged on the MIA, but only trends were observed in the case of the MFQ.

The differential sensitivity of various questionnaires to age differences may be related to the phrasing of the questions. For example, Hultsch et al. (1987) found no age differences on subscales consisting of questions that ask people to report the extent to which they experience episodes of forgetting in particular domains (e.g., MFQ Frequency of Forgetting). In contrast, age differences were observed on subscales consisting of questions that ask people to rate their memory relative to some unspecified anchor (e.g., MIA Capacity). Age differences are particularly apparent on subscales consisting of questions that ask people to rate their memory relative to the anchor of their own past performance (e.g., MIA Change; MFQ Retrospective Functioning). One possible explanation of this pattern is that, although older adults perceive that their memory has declined from previously higher levels of functioning, they do not view this loss as a "problem" either because their current level of functioning conforms to what they expect, or because the incidents of forgetting do not seriously interfere with achieving everyday goals. Sunderland, Watts, Baddeley, & Harris, (1986) have presented data that are consistent with this latter notion.

Sex differences

The question of whether there are gender differences in memory knowledge and perceptions is unresolved. In some instances, samples have been composed of individuals of predominately one gender (e.g., Dixon & Hultsch, 1983b). In other instances, although both genders have been represented in the sample, differences between them have not been explicitly examined. In our recent analysis (Hultsch et al., 1987), we found some evidence for gender differences on the MIA and MFQ that are consistent across samples, although they do not account for large amounts of variance. Specifically, women appear to report more strategy use and greater anxiety associated with memory-demanding situations than men. Significant differences were observed on the MIA Strategy subscale in both samples, and on the MFQ Mnemonics subscale for in one sample. Differences were also observed on the MIA Anxiety subscale in both samples. In all instances, however, the effects accounted for 3% or less of the variance.

Metamemory/Memory Relationships

One of the thorniest questions facing researchers interested in metamemory concerns the relationship between thinking about remembering and actually remembering. The most straightforward view assumes there should be close linkages between these two activities. For example, work which has conceptualized metamemory largely as factual knowledge about memory tasks and strategies quickly leads to the hypothesis that actual performance in memory-demanding situations may be wholly or partially dependent on such metamemorial knowledge (Flavell & Wellman, 1977). Experimental work, particularly with children, has provided some support for the notion that performance differences among different age groups may be related to differences in knowledge of task demands and strategy use (Cavanaugh & Perlmutter, 1982).

It is becoming increasingly clear, however, that the relationship between metamemory and actual memory performance is not straightforward. For example, as Herrmann (1982, 1984) and others have pointed out the evidence supporting the predictive validity of the various metamemory questionnaires is relatively limited. The general pattern of results suggests that correlations in the .20 to .30 range are typical. Such findings have led some writers to question the validity of self-report measures of metamemory, and to reject their use as substitutes for performance measures in clinical settings (e.g., Sunderland et al., 1986).

As we noted earlier, some questionnaires have been developed without attention to the usual steps associated with instrument development. Thus, it is possible that some of the difficulty may be related to measurement problems. For example, zero-order correlations of metamemory scales with memory performance may be attenuated by measurement errors. It is well known that the maximum population correlation between two variables is the product of their reliabilities (Nunnally, 1978). Given low reliability in the metamemory scale, we could get low correlations because the scale is a poor measure, not because metamemory and memory are unrelated. As Rushton et al. (1983) point out, scales that aggregate multiple items should have better reliability than other metamemory measures, and thus correlate higher with performance measures. The MIA scales fulfill this criterion.

Of course, aggregate scales are still, to a lesser degree, unreliable. Hence they are also subject to attenuation due to measurement error. One of the features of our work is the use of structural equation models to estimate correlations among latent variables such as the metamemory dimensions discussed above. The principal advantage of such approaches is that they <u>completely</u> disattenuate correlations between variables for measurement error (Schaie & Hertzog, 1985). In other words, when one estimates a latent variable for metamemory and a latent variable for memory performance, the maximum possible correlation between these latent variables is not bounded by reliability: the latent variables' correlation ranges between -1 and 1. This disattenuation is accomplished by using multiple measures of each latent variable in the structural equation model

(Joreskog, 1974; Long, 1983). The structural equation approach also has the advantage of maximizing the validity of the latent variable (e.g., Embretson, 1983).

We have also argued that metamemory is a multidimensional construct (e.g., Dixon & Hertzog, in press), and studies which have used unidimensional measures have almost unanimously failed to find significant relationships with performance (e.g., Kahn et al., 1975). Similarly, it is clear that memory performance is itself a multidimensional construct. As a result of the multidimensional nature of the domains, then, the obtained pattern of metamemory-memory correlations will be a function of which indicators were selected from which domain.

Evidence for the importance of domain specificity is convincing, although the exact patterns remain to be clarified. Dixon and Hultsch (1983a), for example, found that several metamemory subscales predicted memory for text performance in several samples. In addition, there was evidence to suggest age-related differences in the pattern of the correlations. It appeared that two indicators of memory knowledge (Task and Strategy) were the best overall predictors of performance for all adults in all samples. Certain age differences were evident however. Younger adults' performance was predicted by what they knew about retrieval strategies and physical reminders (Strategy), what they believed about their capacity to perform on given tasks (Capacity), and what they knew about memory tasks and processes in general (Task). In contrast, older adults' performance was predicted by what they knew about memory tasks and precesses in general (Task), their level of motivation to achieve in memory-demanding situations (Achievement), and their belief in the degree of control they exercise over their memory functioning (Locus). These results suggest the possibility that the performance of older adults is more related to their beliefs about their memory self-efficacy than is the case for younger adults. However, this may not be the case for all performance measures. In another analysis, Dixon et al. (1986) examined the relationship of the MIA to performance on a battery of psychometric measures. In this instance, moderate correlations (range: .25 to .53) were observed mostly between the Strategy and Task subscales and several memory and verbal comprehension tests. The subscales reflecting memory self-efficacy showed little relationship to these traditional cognitive performance measures.

These results suggest that certain metamemory-memory performance relationships may be more likely to appear with performance tasks that are high in ecological validity. Such a finding is not surprising considering the fact that most metamemory questionnaires solicit self-evaluations of memory capacity in relation to everyday memory situations. Direct support for this notion comes from studies that have reported significant correlations of various components of metamemory with story recall but not with word recall (e.g., Sunderland et al., 1983; Zelinski et al., 1980). Similarly, Berry, West, and Scogin (1983) found that self-reports about memory predicted performance on a set of everyday memory tasks better than performance on a set of traditional laboratory tasks. Such results suggest the need for more careful consideration of the domains of metamemory and memory performance being examined. Indeed, it may be suggested that metamemory may be most relevant for memory-related behaviors typically not considered at all. For example, the decision to enter a memory-demanding situation in the first place may be determined in part by the individual's perceptions of their self-efficacy in such situations.

Despite some consistencies, it is clear that the relationship between individuals' self-knowledge and self-efficacy about their memory and memory performance is complex. There is sufficient evidence to reject a straightforward interpretation of individuals' self-reports about their memory as veridical reports of their experience. For example, Sunderland et al. (1983), examining the relationship between reports of memory problems and performance for patients suffering from closed head injuries, showed that patients' self-reports showed only weak correlations with their actual performance, whereas relatives reports of the patients' problems showed stronger relationships with the same measures. Similarly, some other studies using community-dwelling adults have found instances of negative correlations between subjective ratings of memory and memory performance (e.g., Dixon et al., 1986; Cavanaugh & Poon, 1985). Thus, individuals with poor memories may be poor at recalling instances of memory failure resulting in an overestimate of memory abilities. The absence of evidence for veridical self-reports of memory suggests that metamemory questionnaires are of limited usefulness for clinical purposes

However, it is becoming increasingly clear that cognitive processes do not operate in isolation from personality and social processes. If one accepts the possibility of interfaces among such processes, we should not necessarily expect self-reports about memory to be veridical indicators of actual memory ability. Assessments of components such as memory self-efficacy will vary considerably across individuals, and perhaps within individuals across even relatively brief intervals of time. The question is whether these individual differences in accuracy are systematic, and whether they relate to other behaviors in memory-demanding situations. If older persons' perceptions of their memory prove to be one link in a process relating the social and cognitive domains, then the metamemory construct is of interest even if it is not a substitute for performance measures.

Summary and Conclusions

Several multidimensional metamemory questionnaires (including the MIA and MFQ) with demonstrated reliability and factorial validity have been developed. Our recent work, summarized above, has shown that the MIA and MFQ converge to measure memory self-efficacy and strategy use. In addition, the MIA measures aspects of memory-related affect and knowledge about memory-demanding situations. We have also demonstrated the discriminant validity of the MIA with respect to various personality traits and states. Recent research also suggests that there are consistent and reliable age differences in metamemory, mostly related to a sense of self-efficacy. Older adults perceive less capacity, greater change, and less control than younger adults. These differences appear to be relatively substantial when the comparison is between young university students and older adults. Differences within the middle to older ages ranges are less robust. In addition, the MIA appears to be somewhat more sensitive to age-related differences than the MFQ. The issue of metamemory/memory performance relations remains somewhat unclear. Generally, predictive relationships are relatively modest. Latent variable analysis may help clarify the ambiguities associated with measurement error problems. We are currently pursuing this strategy with our data set. In addition, a focus on additional measures of memory-related behaviors such as the decision to enter memory-demanding situations may be required. In sum, we have made considerable progress in the domains of measurement and descriptive research related to metamemory. Attention must now be turned toward the development of an explanatory process model which will permit understanding of how the cognitive and personality/social domains interact to produce the behaviors we have labeled metamemory.

REFERENCES

Bennett-Levy, J., & Powell, G.E. (1980). The subjective memory questionnaire (SMQ). An investigation into the self-reporting of 'real life' memory skills. British Journal of Social and Clinical Psychology, 19, 177-188.

Bentler, P.M. (1978). The interdependence of theory, methodology, and empirical data: Causal modeling as an approach to construct validation. In D.B. Kandel (Ed.), Longitudinal research on drug abuse: Empirical and methodological issues (pp. 267-302). Washington: Hemisphere.

Berry, J., West, R.L., & Scogin, F. (1983, November). Predicting everyday and laboratory memory skill. Paper presented at the meeting of the Gerontological Society of America, San Francisco, CA.

Cavanaugh, J.C. (in press). The importance of awareness in memory aging. In L.W. Poon, D.C. Rubin, & B.A. Wilson (Eds.), Everyday cognition in adulthood and old age. Cambridge: Cambridge University Press.

Cavanaugh, J.C., & Perlmutter, M. (1982). Metamemory: A critical examination. Child Development, 53, 11-28.

Cavanaugh, J.C., & Poon, L.W. (1985, August). Patterns of individual differences in secondary and tertiary memory performance. Paper presented at meeting of the American Psychological Association, Los Angeles.

Chaffin, R., & Herrmann, D.J. (1983). Self reports of memory ability by old and young adults. Human Learning, 2, 17-28.

Cronbach, L.J., Meehl, P.E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

Dixon, R.A. (in press). Questionnaire research on metamemory and aging: Issues of structure and function. In L.W. Poon, D.C. Rubin, & B.A. Wilson (Eds.), Everyday cognition in adulthood and old age. Cambridge: Cambridge University Press.

Dixon, R.A., & Hertzog, C. (in press). A functional approach to memory and metamemory development in adulthood. In F.E. Weinert & M. Perlmutter (Eds.), Memory development across the life-span: Universal changes and individual differences. Hillsdale, N.J.: Lawrence Erlbaum.

Dixon, R.A., Hertzog, C., & Hultsch, D.F. (1986). The multiple relationships among Metamemory in Adulthood (MIA) scales and cognitive abilities in adulthood. Human Learning, 5, 165-177.

Dixon, R.A. & Hultsch, D.F. (1983a). Metamemory and memory for text relationships in adulthood: A cross-validation study. Journal of Gerontology, 38, 689-694.

Dixon, R.A., & Hultsch, D.F. (1983b). Structure and development of metamemory in adulthood. Journal of Gerontology, 38, 682-688.

Dixon, R.A., & Hultsch, D.F. (1984). The Metamemory in Adulthood (MIA) instrument. Psychological Documents, 14, 3.

Embretson (Whitley), S. (1983). Construct validity: Construct representation versus nomothetic span. Psychological Bulletin, 93, 179-197.

Flavell, J.H. (1971). First discussant's comments: What is memory development the development of? Human Development, 14, 272-278.

Flavell, J.H., & Wellman, H.M. (1977). Metamemory. In R.V. Kail, Jr., & J.W. Hagen (Eds.), Perspectives on the development of memory and cognition (pp. 3-34), Hillsdale, N.J.: Lawrence Erlbaum.

Gilewski, M.J., & Zelinski, E. (1986). Questionnaire assessment of memory complaints. In L.W. Poon (Ed.), Handbook for clinical memory assessment of older adults (pp. 93-107). Washington, DC: American Psychological Association.

Gilewski, M.J., Zelinski, E.M., Schaie, K.W., & Thompson, L.W. (1983, August). Abbreviating the metamemory questionnaire: Factor structure and norms for adults. Paper presented at the meeting of the American Psychological Association, Anaheim.

Gutman, M.P. (1987). Memory perceptions and memory performance in older adults. Unpublished Master's thesis, Department of Psychology, University of Victoria, Victoria, BC.

Herrmann, D.J. (1982). Know thy memory: The use of questionnaires to assess and study memory. Psychological Bulletin, 92, 434-452.

Herrmann, D.J. (1984). Questionnaires about memory. In J.E. Harris & P.E. Morris (Eds.), Everyday memory, actions and absent-mindedness (pp. 133-152). Academic Press: London.

Herrmann, D.J., & Neisser, U. (1978). An inventory of everyday memory experiences. In M.M. Gruneberg, P.E. Morris, & R.N. Sykes (Eds.), Practical aspects of memory (pp. 35-51). London: Academic Press.

Hertzog, C. (1985). Applications of confirmatory factor analysis to the study of intelligence. In D.K. Detterman (Ed.), Current topics in human intelligence (pp. 59-97). Norwood, N.J.: Ablex.

Hertzog, C., Dixon, R.A., Schulenberg, J., & Hultsch, D.F. (1987). On the differentiation of memory beliefs from memory knowledge: The factor structure of Metamemory in Adulthood scale. Experimental Aging Research, 13, 101-107.

Hertzog, C., Hultsch, D.F., & Dixon, R.A. (1987). Evidence for the convergent validity of two self-report metamemory questionnaires. Unpublished manuscript, School of Psychology, Georgia Institute of Technology.

Hultsch, D.F., Dixon, R.A., & Hertzog, C. (1985). Memory perceptions and memory performance in adulthood and aging. Canadian Journal on Aging, 4, 179-187.

Hultsch, D.F., Hertzog, C., & Dixon, R.A. (1987). Age differences in metamemory: Resolving the inconsistencies. Canadian Journal of Psychology, 41, 193-208.

Joreskog, K.G. (1971). Simultaneous factor analysis in several populations. Psychometrika, 36, 409-426.

Joreskog, K.G. (1974). Analyzing psychological data by analysis of covariance matrices. In D.H. Krantz, R.C. Atkinson, R.D. Luce, & P. Suppes (Eds.), Contemporary developments in mathematical psychology (Vol. 2, pp. 1-56). San Francisco: W. H. Freeman.

Kahn, R.L., Zarit, S.H., Hilbert, N.M., & Neiderehe, G. (1975). Memory complaint and impairment in the aged: The effects of depression and altered brain function. Archives of General Psychiatry, 32, 1569-1573.

Lachman, M.E. (1983). Perceptions of intellectual aging: Antecedent or consequence of intellectual functioning. Developmental Psychology, 19, 482-498.

Lachman, M.E. (1986). Locus of control in aging research: A case for multidimensional and domain-specific assessment. Journal of Psychology and Aging, 1, 34-40.

Lachman, M.E., Steinberg, E.S., & Trotter, S.D. (1987). The effects of control beliefs and attributions on memory self-assessments and performance. Unpublished manuscript, Department of Psychology, Brandeis University.

Lebo, M.A., & Nesselroade, J.R. (1978). Intraindividual differences of mood change during pregnancy identified in five P-technique factor analyses. Journal of Research in Personality, 12, 205-224.

Levenson, H. (1974). Activism and powerful others: Distinctions within the concept of internal-external locus of control. Journal of Personality Assessment, 38, 377-383.

Long, J.S. (1983). Covariance structure models: An introduction to LISREL. Beverly Hills, CA: Sage.

Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. Psychological Bulletin, 89, 575-588.

Nunnally, J.C. (1978). Psychometric theory (2nd ed.). New York: McGraw-Hill.

Perlmutter, M. (1978). What is memory aging the aging of? Developmental Psychology, 14, 330-345.

Radloff, L. (1977). The CES-D scale: A self-report depression scale for research in the general population. Journal of Applied Psychological Measurement, 1, 385-401.

Rushton, J.P., Brainerd, C.J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. Psychological Bulletin, 94, 18-38.

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.

Sunderland, A., Harris, J.E., & Baddeley, A.D. (1983). Do laboratory tests predict everyday memory? A neuropsychological study. _Journal of Verbal Learning and Verbal Behavior_, _22_, 341-357.

Sunderland, A., Watts, K., Baddeley, A.D. & Harris, J.E. (1986). Subjective memory assessment and test performance in elderly adults. _Journal of Gerontology_, _41_, 376-384.

Veit, C., & Ware, J.E., Jr. (1983). The structure of psychological distress and well-being in general populations. _Journal of Consulting and Clinical Psychology_, _51_, 730-742.

Weinstein, C.E., Duffy, M., Underwood, V.L., MacDonald, J. & Gott, S.P. (1981). Memory strategies reported by older adults for experimental and everyday learning tasks. _Educational Gerontology_, _7_, 205-213.

Wellman, H.M. (1983). Metamemory revisited. In M.T.H. Chi (Ed.), _Trends in memory development research_ (pp. 31-51). Basel: Karger.

West, R.L., & Berry, J.M. (1987, August). Self-efficacy and memory performance: Measurement issues. In J.M. Berry & J.C. Cavanaugh (Chairs), _Cognitive and memory self-efficacy in adults_. Symposium conducted at meeting of the American Psychological Association, New York.

West, R.L., Boatwright, L.K., & Schleser, R. (1984). The link between memory performance, self-assessment, and affective status. _Experimental Aging Research_, _10_, 197-200.

Williams, S.H., Denney, N.W., & Schadler, M. (1983). Elderly adults. Perception of their own cognitive development during the adult years. International Journal of Human Development, 16, 147-158.

Zarit, S.H. (1982). Affective correlates of self-report about memory of older adults. International Journal of Behavioral Geriatrics, 1, 25-34.

Zarit, S.H., Cole, K.D., & Guider, R.L. (1981). Memory training strategies and subjective complaints of memory in the aged. The Gerontologist, 21, 158-164.

Zarit, S.H., Gallagher, D., & Kramer, N. (1981). Memory training in the community aged. Effects on depression, memory complaint, and memory performance. Educational Gerontology, 6, 11-27.

Zelinski, E.M., Gilewski, M.J., & Thompson, L.W. (1980). Do laboratory tests relate to self-assessments of memory ability in the young and old? In L.W. Poon, J.L. Fozard, L.S. Cermak, D. Arenberg, & L.W. Thompson (Eds.), New directions in memory and aging. Proceedings of the George A. Talland Memory Conférence (pp. 519-544). Hillsdale, NJ: Lawrence Erlbaum.

Table 1

Hypothetical Dimensions of Metamemory

Dimension	Content
Memory Knowledge	Factual knowledge about memory tasks, processes, strategies, etc.
Memory Monitoring	Self-knowledge about current memory use, contents, states, etc.
Memory Self-Efficacy	Beliefs about memory abilities, strengths, weaknesses, etc.
Memory-Related Affect	Affective states generated by or associated with memory-demanding situations

Table 2

A Priori Subscales of the Memory Functioning Questionnaire (MFQ)

Subscale	Sample Item
. General rating	1. How would you rate your memory in terms of the kinds of problems you have? (+ = no problems)
. Retrospective functioning	2. How is your memory compared to what it was ... (a) one year ago? (+ = much better)
. Frequency of forgetting	3. How often do these present a memory problem for you ...(a) names? (+ = never)
. Frequency of forgetting when reading novels	4. As you are reading a novel, how often do you have trouble remembering what you have read ...(a) in opening chapters, once you have finished the book? (+ = never)
. Frequency of forgetting when reading newspapers and magazines.	5. When you are reading a newspaper or magazine article, how often do you have trouble remembering what you have read ... (a) in the opening paragraphs, once you have finished the article? (+ = never)
. Remembering past events	6. How well do you remember things which occurred ...(a) last month? (+ = very good)
. Seriousness	7. When you actually forget in these situations, how serious of a problem do you consider the memory failure to be ...(a) names. (+ = not serious)
. Mnemonics	8. How often do you use these techniques to remind yourself about things ...(a) keep an appointment book. (+ = never)

3ased on Gilewski et al., 1983)

Table 3

Seven Dimensions of the Metamemory in Adulthood (MIA) Instrument

Dimension	Description	Sample Item
. Strategy	Knowledge and use of information about one's remembering abilities such that performance in given instances is potentially improved (+ = high use)	Do you write appointments on a calendar to help you remember them?
. Task	Knowledge of basic memory processes, especially as evidenced by how most people perform. (+ = high knowledge)	For most people, facts that are interesting are easier to remember than facts that are not.
. Capacity	Perception of memory capacities as evidenced by predictive report of performance on given tasks. (+ = high capacity)	I am good at remembering names.
. Change	Perception of memory abilities as generally stable or subject to long-term decline. (+ = stability)	The older I get the harder it is to remember things clearly.
. Anxiety	Feelings of stress related to memory performance. (+ = high anxiety)	I do not get flustered when I am put on the the spot to remember new things.
. Achievement	Perceived importance of having a good memory and performing well on memory tasks. (+ = high achievement)	It is important that I am very accurate when remembering names of people.
. Locus	Perceived personal control over remembering abilities. (+ = internality)	Even if I work on it my memory ability will go downhill.

3ased on Dixon & Hultsch, 1983b)

Table 4

Internal Consistency (Cranbach's Alpha) for
Seven MIA Subscales for Multiple Samples

bscale	Sample				
	Dixon & Hultsch (1983b) N=120	Dixon & Hultsch (1983b) N=108	Dixon & Hultsch (1983b) N=150	Hultsch et al. (1987) N=388	Hultsch et al. (1987) N=342
rategy	.86	.86	.85	.84	.82
sk	.83	.81	.83	.78	.78
pacity	.86	.82	.86	.85	.81
ange	.93	.90	.91	.91	.92
xiety	.83	.84	.83	.87	.86
hievement	.76	.78	.79	.78	.76
cus	.71	.78	.77	.78	.71

Table 5

Factor Loadings of Metamemory Scales

For Annville and Victoria

	Factors											
	MSE_{MIA}		MSE_{MFQ}		RD		STRAT		AFF		CHANGE	
Scale	A	V	A	V	A	V	A	V	A	V	A	V
STRAT	0	0	0	0	0	0	1.0*	1.0*	0	0	0	0
TASK	0	0	0	0	0	0	0	0	0	0	0	0
CAP	.844	.843	0	0	0	0	0	0	.235	.314	0	0
CHNGE	.514	.612	0	0	0	0	0	0	0	0	.500	.505
ANX	-.633	-.702	0	0	0	0	0	0	.332	.165	0	0
ACH	0	0	0	0	0	0	0	0	.847	.944	0	0
LOC	.250	.279	0	0	0	0	0	0	.343	.300	.336	.335
G.RATING	0	0	.448	.544	0	0	0	0	0	0	0	0
RETRO	0	0	.151	.220	0	0	0	0	0	0	.334	.345
FORGET	0	0	.853	.850	0	0	0	0	0	0	0	0
READ1	0	0	0	0	.856	.882	0	0	0	0	0	0
READ2	0	0	0	0	.849	.822	0	0	0	0	0	0
PAST	0	0	.650	.741	0	0	0	0	0	0	0	0
MFQSTRAT	0	0	0	0	0	0	-.719	-.695	0	0	0	0

Note: *Denotes fixed parameter. All 0 entries were fixed by hypothesis.

Abbreviations: Factors: MSE_{MIA} - Memory Self-EFficacy, MIA scale; MSE_{MFQ} - Memory Self-Efficacy, MFQ Scale; RD - Memory Self-Efficacy, Reading; STRAT - Memory Strategy Use: AFF - Memory-Related Affect.
Scales: STRAT - MIA Strategy Use: CAP - MIA Capacity; CHA - MIA Change; ANX - MIA Anxiety; ACH - MIA Achievement; LOC - MIA Locus of Control; G.RATING - MFQ Global Memory Rating; RETRO - MFQ Retrospective; FORGET - MFQ Problems; READ1 - MFQ Problems Remembering Novels; READ2 - MFQ Problems Remembering Magazines; PAST - MFQ Remote Memory; MFQSTRAT - MFQ Strategy Use.

Table 6

Metamemory Factor Correlations
in Four Groups

FACTORS	GROUPS			
	VICTORIA (OLD)	ANNVILLE (OLD)	ANNVILLE (MIDDLE-AGED)	VICTORIA (YOUNG)
MSE_{MIA}, MSE_{MFQ}	.879	.963	1.004*	.836
MSE_{MIA}, RD	.667	.730	.736	.567
MSE_{MFQ}, RD	.742	.749	.806	.576
MSE_{MIA}, STRAT	-.194	-.368	-.258	-.177
MSE_{MFQ}, STRAT	-.117	-.205	-.227	-.152
RD, STRAT	-.113	-.074	-.124	-.139
MSE_{MIA}, AFF	-.096	-.210	-.131	-.166
MSE_{MFQ}, AFF	.087	-.148	-.005	.105
RD, AFF	-.057	-.157	.005	-.029
STRAT, AFF	.320	.293	.267	.376
MSE_{MIA}, TASK	.089	.095	-.004	.103
MSE_{MFQ}, TASK	.049	.195	.057	.188
RD, TASK	.006	.225	.144	.238
STRAT, TASK	.082	.087	.274	.124
AFF, TASK	.320	.178	.180	.269
MSE_{MIA}, CHANGE	.367	.637	.336	.127
MSE_{MFQ}, CHANGE	.324	.488	.172	.105

Abbreviations: Factors: MSE_{MIA} - Memory Self-Efficacy, MIA scale; MSE_{MFQ} - Memory Self-Efficacy, MFQ Scale; RD - Memory Self-Efficacy, reading; STRAT - Memory Strategy Use; AFF - Memory-Related Affect.

* The maximum likelihood estimate for the factor covariance did, when rescaled, convert to a correlation greater than 1.0. This can happen in latent variable models of this kind, especially when factor loadings are constrained equal over groups.

Table 7
Correlations of MIA Subscales
with Personality and Locus
of Control Measures

	MIA Subscales						
Measure	LOC	CAP	CHA	ANX	ACH	STRAT	TASK
JPI Anxiety	-.050	-.082	-.073	.438	.216	.177	-.045
JPI Affect	.073	.045	-.076	.238	.244	.193	.045
JPI Energy	.173	.248	.219	-.281	.011	-.087	.029
JPI Self-esteem	.082	.315	.255	-.423	-.035	-.094	.017
JPI Conformity	-.048	-.169	-.145	.388	.226	.234	.060
PRF Endurance	.194	.209	.182	-.218	.109	-.110	-.003
Internal	.160	.170	.156	-.140	.091	-.012	.170
Others	-.037	-.054	-.076	.253	.175	.143	.086
Chance	-.014	-.041	-.048	.169	.063	.028	-.063
VW DEP	.050	.124	.092	-.361	-.187	-.054	.002
CES-D DEP	-.004	-.122	-.107	.289	.175	-.013	-.057
CES-D WB	-.040	-.166	-.133	.234	.133	-.018	-.035
VW WB	.067	.138	.145	-.309	-.144	.003	.011

Abbreviations: JPI -Jackson Personality Inventory; PRF - Jackson Personality Research Form; VW Dep - Veit/Ware Depression; CES-D DEP - CES-D Depression scale; CES-D WB - CES-D Well-Being scale; VW WB - Veit/Ware Well-Being scale; LOC-MIA Locus of Control; CAP-MIA Capacity; CHA-MIA Change; ANX-MIA Anxiety; ACH-MIA Achievement; STRAT-MIA Strategy; TASK-MIA Task .

Table 8
Correlations of MIA Subscales
With Mood State Variables

Mood State Variable		LOC	CAP	CHA	ANX	ACH	STRAT	TASK
		MIA Subscales						
Cattell	Anxiety	-.075	-.128	-.112	.358	.100	.065	-.073
Cattell	Fatigue	-.116	-.194	-.211	.270	.039	.079	-.019
Cattell	Depression	-.129	-.260	-.265	.416	.105	.035	-.093
Cattell	Well-Being	.119	.142	.090	-.187	-.019	-.029	-.010
MA	Anxiety	-.090	-.200	-.152	.361	.101	.048	-.038
MA	Fatigue	-.080	-.130	-.105	.203	.032	.053	-.022
MA	Depression	-.059	-.144	-.088	.192	.116	-.021	-.042
MA	Well-Being	.124	.023	.094	-.240	-.023	-.022	.023
MA	Vigor	.170	.233	.192	-.252	.030	-.080	-.005

Abbreviations: MA - Mood Adjective Rating Scale; LOC-MIA Locus of Control; CAP-MIA Capacity; CHA-MIA Change; ANX-MIA Anxiety; ACH-MIA Achievement; STRAT-MIA Strategy; TASK-MIA Task

Table 9
Summary of Significant Age-Related Differences
for Subscales of the MIA Over Multiple Samples

Sample Characteristics/ Sources	MIA Subscales						
	STRAT	TASK	CAP	CHA	ANX	ACH	LOC
1. N=120; 2 age gps. (18-37, 50-81) Dixon & Hultsch (1983b)		-	-	-			-
2. N=108; 3 age gps. (21-30, 40-58, 60-84) Dixon & Hultsch (1983b)			-	-			-
3. N=150; 3 age gps. (21-39; 40-58; 60-74) Dixon & Hultsch (1983b)		-	-				-
4. N=46; 2 page gps. (M=19.0, M=76.9) Cavanaugh & Poon (1985)			-	-			
5. N=360; 4 age gps. (20-26, 55-61, 62-68, 69-78) Hultsch et al. (1987)	+		-	-			-
6. N=415; continuous age sample 20-78 Hultsch et al. (1987)			-	-			-(t)
7. N=376; 4 age gps. (54-61, 62-68, 60-75 76-93) Gutman (1987)				-		+	

+ = older adults score significantly higher ($p < .01$) than younger adults;
- = older adults score significantly lower ($p < .01$) than younger adults;
(t) = trend toward significance at $p < .05$

Abbreviations: STRAT - MIA Strategy; TASK - MIA Task; CAP - MIA Capacity; CHA - MIA Change; ANX - MIA Anxiety; ACH - MIA Achievement; LOC - MIA Locus of Control

Table 10

Summary of Significant Age-Related Differences

for Subscales of the MFQ Over Multiple Samples

ımple Characteristics/ ɔurce	G.RATING	RETRO	FORGET	READ1	READ2	PAST	SERIOUS	MFQSTRAT
. N=639; 3 age gps. l6-54, 55-70, 71-89) ilewski et al. (1983)	–	–	–	–	–	–	–	
. N=264; 3 age gps. l6-54, 55-70, 71-89) ilewski et al. (1983)	–	–	–	–	–	–		
. N=360; 4 age gps.)-26, 55-61, 62-68, 69-78) ıltsch et al. (1987)	–	–		–	–			
. N=415; continuous ge sample 20-78 ıltsch et al. (1987)		-(t)			-(t)			+(t)

= older adults score significantly higher ($p < .01$) than younger adults;
= older adults score significantly lower ($p < .01$) than younger adults;
t) = trend toward significance at $p < .05$.

bbreviations: G.RATING - MFQ Global Memory Rating; RETRO - MFQ Retrospective; FORGET - MFQ roblems; READ1 - Problems Remembering Novels; READ2 - Problems Remembering Magazines; PAST MFQ Remote Memory; SERIOUSNESS - MFQ Seriousness; MFQSTRAT - MFQ Strategy Use

Stability and Change in Adult Intellectual Development: 2. Analysis of Means and Covariance Structures

Christopher Hertzog

Georgia Institute of Technology

and

K. Warner Schaie

The Pennsylvania State University

Running Head: Stability of Adult Intelligence

Address Correspondence To: Christopher Hertzog

School of Psychology

Georgia Institute of Technology

Atlanta, GA 30332-0170

Abstract

This study conducted an analysis of data on psychometric intelligence from the Seattle Longitudinal Study. We simultaneously estimated longitudinal factors, their covariance structure, and their mean levels. Data on five Thurstone Primary Mental Abilities (PMA) subtests were available for 412 adults, ages 22-70 at first test, who were tested three times at seven-year intervals. The model specified a general intelligence factor at each longitudinal occasion (age) for three different age groups (young, middle-aged, old), as well as a residual specific to each of the five PMA subtests. A previous longitudinal factor analysis had shown high stability of individual differences (covariance stability) in general intelligence for all three age groups over the fourteen-year longitudinal interval. We extended that model to estimate general factor means. Whereas all three age groups showed high levels of covariance stability, they differed sharply in their mean profiles. The young group showed increasing levels of general intelligence, the middle-aged group had stable levels of intelligence, and the old group showed salient, approximately linear, decline. There were also substantial age group differences in mean levels of intelligence. The analysis revealed that not all mean age changes in PMA subtests could be accounted for solely on the basis of age changes in mean levels of general intelligence; instead, different profiles of age changes in test-specific residuals were required to adequately model the data. The patterns of stability in middle-age, followed by mean decline and high covariance stability in old age, suggest a normative developmental transition from a stability pattern to a decline pattern of general intelligence, with the inflection point occuring somewhere around age 60.

Introduction

An important issue in the study of adult intellectual development concerns whether levels of intelligence remain stable with advancing age. There is general agreement that the average level of performance on certain psychometric measures of intelligence declines with age, although there is great debate as to the ubiquity of decline, the proper interpretation of declines in psychometric performance, when it occurs, and the practical importance of the magnitudes of age-related decline (e.g., Baltes, Dittman-Kohli, & Dixon, 1984; Botwinick, 1977; Dixon, Kramer, & Baltes, 1985; Horn, 1985; Horn & Donaldson, 1976, 1980; Schaie, 1983). At the center of the disagreements in the literature regarding aging and intelligence has been Schaie's longitudinal studies of aging and primary mental abilities (see Schaie, 1983). The debate between Horn, Schaie, and others (e.g., Baltes & Schaie, 1976; Horn & Donaldson, 1976) covered a large number of issues associated with Schaie's sequential design, psychometric tests, and alternate theories and interpetations of aging and intelligence. Subsequent work by Schaie and Hertzog (1983) re-examined the issues with additional, new data from Schaie's sequential samples. Their cohort-sequential analyses identified clear cohort differences in certain psychometric tests, and identified statistically significant changes in multiple psychometrically defined abilities. For all five subtests of Thurstone's Primary Mental Abilities, declines in performance (whether measured by longitudinal or cross-sectional sequences), were negligible until the decade of the 50's. Declines that were observed in the fifties and early sixties were small in magnitude, but became increasingly large after mean age 60. A somewhat surprising result, given earlier cross-sequential results from Schaie's data, was that the longitudinal sequences suggested decline after mean age 60 in all PMA subtests, although the decline began later for the PMA subtest Verbal Meaning (a test of recognition vocabulary). Schaie and Hertzog (1983) argued that these results required some minor modification of previous positions regarding the age of onset of intellectual decline, but that they supported the major conclusions of 1) age-confounded cohort differences in cross-sectional studies; 2) relative stability of mean performance levels into the decade of the fifties, with substantial declines only after age 60; and 3) some differences across subtests in the onset and magnitude of age-related performance declines (see also Dixon et al. 1985).

Although most of the gerontological literature has focused on the issue of stability of mean levels of intelligence with aging, mean stability is but one type of stability that can be assessed in longitudinal data. Another important type of stability is stability of individual differences (e.g., Baltes, Reese, & Nesselroade, 1977; Kagan, 1980; Schaie & Hertzog, 1985). This stability reflects the degree to which individuals differ in their developmental patterns of change (Baltes et al, 1977; Nesselroade & Labouvie, 1985; Schaie & Hertzog, 1985). Whereas stability of means is reflected in equivalent mean values at different developmental times, stability of individual differences is reflected in the covariance of a variable with itself over two points in time (see Baltes et al. 1977). In this paper, we shall refer to stability of individual differences as covariance stability (see Hertzog & Nesselroade, in press).

In the first paper in this series, Hertzog & Schaie (1986) demonstrated that there is substantial covariance stability in intelligence across the adult life-span. Hertzog and Schaie (1986) used a longitudinal factor analysis of data from the Seattle Longitudinal Study to show that 1) a general intelligence factor, *g*, could be identified for three age groups (labelled young, middle-aged, and old), 2) this *g* factor was defined equivalently by the PMA subtests in each age group, and showed invariant factor loadings across longitudinal occasions; 3) the covariance stability of *g* was high in all age groups, with longitudinal correlations of *g* with itself at or above .9 between successive longitudinal occasions, even in the older group; and 4) substantial covariance stability in the five primary ability subtests, independent of *g*, as reflected in the proportion of variance in the PMA subtests determined by "test-specific" factors.

Hertzog and Schaie's (1986) results support the hypothesis that age changes in *g* are relatively consistent for same-aged individuals; although there are individual differences in change patterns, these differences produce shifts in relative ordering of individuals that are small in magnitude relative to the overall population variance in *g*. It is interesting that covariance stability was high in age ranges where Schaie & Hertzog (1983) detected decline in the individual PMA subtests -- namely, after age 60. This finding suggests only modest individual differences in the magnitudes of late-life decline in *g*.

This paper reports a series of additional analyses designed to examine explicitly the mean level stability of *g*, while simultaneously estimating stability of individual differences in *g*. The results of these analyses demonstrate the independence of these the two types of stability in the domain of psychometric intelligence. The analyses also were used to examine the question of inflection point for shifts from stability to decline in general intelligence.

The simultaneous examination of mean and covariance stability in longitudinal data is made possible by use of structural equation models to analyze means of latent variables (e.g., McArdle & McDonald, 1984; Sorbom, 1982). The longitudinal factor analyses, omitting the means, reported by Hertzog & Schaie (1986) constitute an important precursor to simultaneous analysis of mean and covariance structures. Hertzog & Schaie (1986) found metric invariance in the *g* factor loadings between groups and across longitudinal occasions of measurement. Metric invariance is defined as equivalence in the unstandardized regression weights of variables on factors (see Horn, McArdle, & Mason, 1984). As discussed by several developmental methodologists (e.g., Baltes & Nesselroade, 1973; Labouvie, 1980a,b; Schaie & Hertzog, 1985), an assumption of metric invariance is essential for allowing unambiguous interpretation of quantitative differences in mean levels of factor scores. The demonstration of metric invariance in *g* insures that *g* is measured in equivalent units of measurement, so that differences in *g* factor means are uncontaminated reflections of mean level differences in the latent variable (see Labouvie, 1980a,b; Schaie & Hertzog, 1985, for further discussion of this issue).

Given evidence of metric invariance, the simultaneous analysis of means and covariance structures requires introduction of the means into the structural equations of the longitudinal factor model already used by Hertzog & Schaie

(1986). The critical questions of interest were 1) what is the magnitude of mean age changes in g at the different age levels studied; 2) do age differences and age changes in g fully account for the mean changes in PMA subtests, or must different developmental trends of PMA means be modeled to account fully for the information in the means; and 3) is there evidence for independence of stability of g means from covariance stability of g?

Method

Subjects

The subjects in this study were participants in the Seattle Longitudinal Study (SLS) conducted by Schaie and associates (Schaie, 1983). The population consisted of members of a health maintenance organization (HMO) in the greater Seattle, Washington area. In order to minimize the probability of selection differences over time, the population was defined as all members of the organization as of 1956, the initial year of the longitudinal study. All participants were unpaid volunteers who answered questionnaires and took part in a single session psychometric test session. The participants, adults spanning the age range of 20 through 74 at first test, represented a range of socioeconomic and ethnic groups (although the population defined by the HMO membership in 1956 was predominantly Caucasian and somewhat more affluent than the general Seattle population. Further details on the population and sampling procedures may be found in Schaie (1983).

Sequential Sampling Design

The longitudinal samples studied here are a subset of the sequential samples collected in the SLS. The sampling plan of the SLS is discussed more fully in Schaie (1983), and the present sample is defined explictly in Hertzog and Schaie (1986). Briefly, we restrict our analysis here to two fourteen-year longitudinal samples (first tested in either 1956 or 1963). The data from the two longitudinal sequences were partitioned into a hybrid sequential data matrix described in Table 1. The partitioned data matrix forms three age groups for simultaneous analysis.

Variables

As part of a larger psychometric battery, all subjects were administered the 1948 version of the SRA Primary Mental Abilities Test, Form AM 11-17 (Thurstone & Thurstone, 1949). The 1948 PMA includes five subtests, all of which are timed and have significant speed components in adult samples (see Schaie & Hertzog, 1983): (a) Verbal Meaning — a test of recognition vocabulary; (b) Space -- a test of spatial relations requiring mental rotation of figures in a two-dimensional plane; (c) Reasoning — a test of inductive reasoning requiring recognition and extrapolation of patterns of letter sequences; (d) Number -- a test of the ability to solve simple two-column addition problems quickly and accurately; and (e) Word Fluency -- a test of the ability to retrieve words from semantic memory according to an arbitrary syntactic rule (words beginning with the letter "s"). Scoring followed the PMA manual: Verbal Meaning and Reasoning were scored in terms of the number of correct items, Space and Number were scored by subtracting incorrect items (comission errors) from the total number of correct items, and Word Fluency was scored by tallying the number of unique, admissable words generated during the allotted time.

Models and Statistical Procedures

The longitudinal factor model employed is an application of a generic longitudinal model described in some detail by Joreskog and Sorbom (1977; see also Hertzog, 1985; Horn & McArdle, 1980; Schaie & Hertzog, 1985). A detailed description of the model may be found in Hertzog and Schaie (1986). The model specified an occasion-specific g factor at each longitudinal occasion. The factor covariance matrix modeled the variances and covariances of g at the different occasions of measurement, and the residuals in the PMA subtests were modeled as having test-specific covariances (e.g., the residuals for Verbal Meaning were allowed to covary across longitudinal occasions). The specification of longitudinal models including factor means is relatively complex (Joreskog & Sorbom, 1980, 1984; McArdle, in press). Appendix A provides the formal notation and model specifications. The critical features are 1) a vector of location constants, analagous to grand means, 2) representation of latent variable means as regressions on a fixed constant, and modeled in the LISREL GAMMA parameter matrix, 3) the assumption that the means of all residuals are 0 in the population. The vector of location constants identifies an intercept for each observed variable (PMA subtest). In longitudinal analysis in multiple groups, these location parameters are constrained equal both across longitudinal occasions and between the multiple age groups. Given data containing neither group differences nor longitudinal changes in means, this location parameter vector would perfectly account for the mean structure. Thus, the model with factor means will be meaningful only if there are either group differences or longitudinal changes in mean that the model may attempt to structure as a function of the factor means. Identification of the location parmaeters and the factor means is achieved by fixing the mean of *g* to 0 for one age group at one longitudinal occasion. In the models reported, we fixed the *g* mean for the middle-aged group at the first occasion (mean age 42) at zero. This procedure then enables the remaining factor means to be estimated as deviations from this reference point (see Joreskog & Sorbom, 1984; Sorbom, 1982 for additional details). The fact that factor means are modeled as regression of factors (i.e., *g*) on a constant requires the assumption that the means of the residuals are 0. This is an unlikely assumption, given that we expect age trends in mean levels to vary across PMA subtests (independent of their relationship to *g*). It is, however, possible to estimate residual component means by moving these parameters into the "latent variable" vector in LISREL, as illustrated in Appendix A. This specification results in a model like the one used in Bentler's EQS program (Bentler, 1985). in which factors and residuals are modeled in the same parameter matrix.

All models were estimated in either LISREL V or VI (Joreskog & Sorbom, 1984) using maximum likelihood estimation. In structural modeling, model fit can be assessed by likelihood ratio X^2, as well as relative fit indices provided by the program. These indices are of less value in models with means, however, so we report a decomposition of overall model fit into (a) fit of the covariance structure model and (b) fit of the mean structure model (see Bentler & Bonett, 1980; Sobel & Bohrnstedt, 1985). The relative fit index for the means may be interpreted as an index of the proportion of information in the mean structure, adjusted for location parameters, accounted for by the model.

The procedures used here are unabashedly exploratory in nature. The goal is to use modeling techniques to explore descriptive developmental hypotheses about the longitudinal mean and covariance structures of the PMA subtests. This use of a generic longitudinal factor model is an appropriate application of structural equation techniques, which are ideal for exploratory multivariate modeling of longitudinal data (Hertzog, in press; McArdle, in press). This study cannot and should not be considered to represent a confirmatory analysis, in the philosophical sense of the term.

Results

The first model we estimated fixed the **g** factor means at 0 in all three age groups, but allowed all location parameters to be freely estimated. This model fits the 15 means of each age group with 15 freely estimated location parameters. There is a one-to-one correspondence between location parameters and sample means, and as such the location parameters are just-identified. This model is therefore saturated with respect to the means, using Bentler and Bonett's (1980) definition. The fit of the model, denoted M_S, is reported in Table 2. As expected, this model fit the same as the model ignoring means reported by Hertzog and Schaie (1986), and yielded an identical longitudinal factor solution. A second preliminary model, following recommendations of Bentler and Bonett (1980), was a null model in the means. This model specified 5 location parameters, one for each PMA subtest, and constrained these parameters to fit the means of all three longitudinal occasions for all three age groups. Thus the 45 population means were fit with 5 location parameters. This null model, M_N, would have a fit equal to the saturated model, M_S, if there were no group differences or longitudinal changes in PMA subtest means to structure as part of the analysis. There was, however, a substantial, statistically significant difference between the two models, as seen in the first model comparison reported in the lower part of Table 2. Clearly there was longitudinal and age group variation in the PMA means, and the task of the analysis was to structure this variation in terms of the longitudinal factor model.

The first substantive model of interest specified **g** factor means in all three age groups. In order to identify the factor means, the occasion 1 factor mean of **g** in the middle-aged group was fixed at zero; all other factor means were freely estimated. Interpretation of the fit of these substantive models must be made on the basis of relative differences from the null and saturated models, so that one can evaluate fit to the means ignoring (assuming) the basis specification and fit of the longitudinal factor model (Bentler & Bonett, 1980; Sobel & Bohrnstedt, 1985). In essence, the difference between the null and saturated models marks a range of possible fits of models structuring means in the longitudinal analysis. The critical question is how close a model with structured means comes to the fit of the model that is saturated in the means (or conversely, how far it has come from the poor fit of the null model.

As shown in Table 2, this first substantive model, $\dot{M}_1$, improved meaningfully on the fit of the null model, although there was still a significant difference between M_1 and M_5. The relative fit of the new model is best indexed by the Sobel/Bohrnstedt relative fit index, denoted as DELTA in Table 2. The fit of .49 indicates that about half of the variation in the means had successfully been structured by M_1.

One interesting outcome of model M_1 was that the g factor means for the middle-aged adults were not significantly different from zero, relative to their standard errors. In models of this type, these estimated factor means are scaled as deviations from the fixed zero mean (age 42 for the middle-aged population). Therefore, the finding of essentially zero g means at ages 49 and 56 for the middle-aged group indicated no statistically significant change in mean level of g over this age range. A second model, M_2, incorporated this feature by fixing the g means to 0 for all three ages of the middle-aged group. This model did not fit more poorly than M_1.

The fact that M_2 fit significantly worse than M_5 implied that the assumption of no mean variation in the residuals for the PMA factors had to be abandoned. That is, it was not possible to model age group differences and age changes in PMA means solely as a function of age differences and age changes in g factor means. Apparently, the primary abilities measured by the PMA have variations in the means that are saliently different from the behavior of the g factor means.

A logical possibility is that there are group differences in subtest-specific means, but no differential age changes in the primary ability means. Our previous work (Hertzog & Schaie, 1986) modeling both g and PMA test-specific factors provided a convenient means of testing this hypothesis. We specified a model that included both occasion-specific (*g*) factors and test-specific factors, one for each PMA subtest. We then allowed the test-specific factor means to be estimated, achieving identification of the test-specific factor means by fixing all five test-specific factor means for the middle-aged group to 0. Given the rejection of models M_1 and M_2, and the decision to fit an alternative model specifying residual factor means, we did not wish to assume mean stability in *g*, as suggested from the M_2-M_1 comparison. Indeed, it was possible that the stable *g* factor means in the middle-aged group in the previous models were an artifact of model misspecification. We therefore used model M_1 as the basis for the test-specific factor mean model, allowing the *g* factor means at ages 49 and 56 to be freely estimated in the middle-aged group.

Table 2 reports the fit of this third substantive model, M_3. The model fit significantly better than M_1, indicating group differences in residual means. However, the model still did not approximate the fit of M_5, requiring rejection of Model M_3. Nevertheless, there were statistically significant age group differences in test-specific factor means. It was also still the case that the *g* factor means did not differ significantly between ages 42 and 56 for the middle-aged group. We concluded that there were age group differences in PMA subtest means, but that there are also differential age changes for the PMA subtest means, independent of *g*. We also concluded that it was still plausible to maintain the assumption of no age changes in *g* in the middle-aged group.

We next proceeded by fitting a series of models allowing residual means.

This approach proceeded in exploratory fashion. Large mean residuals (differences between sample means for the PMA subtests and PMA means predicted from the model parameters) and salient LISREL modification indices were used to indicate a need for structuring additional mean parameters. Unlike M_3, these models specified a separate PMA residual "factor" at each longitudinal occasion, permitting both g and the PMA residuals from g to display age-related change. After a series of model modifications, we arrived at a model that did not differ significantly from the saturated model. This model allowed residual means for Word Fluency, Number, Verbal Meaning, and Space. This modified model, denoted M_4 in Table 2, achieved a relative fit index of .97 to the means, indicating excellent fit. Of course, this fit was achieved by adjusting to the sample moments, and can therefore be treated only as a descriptive index of the success of the model modification process.

One of the major reasons for fitting additional models to the means was to insure that the estimated age changes and age differences in g means were not inappropriately biased by the incorrect assumption of no residual means. Hertzog and Carter (1982) previously demonstrated that group differences in intelligence factor means were affected by the specification error of zero residual means. Table 3 reports the g factor means for the four substantive models, M_1 through M_4. Irrespective of the model, the relative pattern of g factor means in the three age groups remained the same. g increased from mean age 30 to mean age 37 in the young group, and then remained relatively stable through age 44. g exhibited mean stability from mean age 42 through mean age 56 in the middle-aged group. Finally, g showed substantial decline from mean age 58 through mean age 72 in the old group. The mean decline in g in the old group was roughly linear over the fourteen-year period. The comparable pattern of g mean behavior is particularly important in Model M_4, where it was most likely that the apparent age changes in g estimated in Models M_1 through M_3 would change as a function of specifying longitudinal changes in the PMA residuals as well. The fact that conclusions regarding the behavior of g means were not altered by specifying longitudinal variation in PMA residual means indicated that the mean patterns were unlikely to be an artifact of model specification.

Approximate 99% confidence intervals around the factor means can be calculated by subtracting and adding 2.5 standard errors to the estimated g factor means. Inspection of Table 3 clearly showed that these 99% confidence intervals did not include 0 for any of the freely estimated means in the old and young groups. As these means are deviation contrasts from the middle-aged g means, we concluded there were reliable age group differences in means. The significant differences included comparisons between the different groups at roughly comparable ages. That is, the young group at age 44 (Occasion 3) differed significantly from the middle-aged group at age 42 (Occasion 1), as did the middle-aged group at mean age 56 (Occasion 3) from the old group at mean age 58 (Occasion 1). Although the hybrid sequential design does not completely unconfound age changes and cohort differences, it seems likely that these differences reflect cohort differences in the mean levels of g.

Table 4 reports the residual means estimated in the final model, M_4. These means must be interpreted with care. They represent mean patterns in the PMA subtests orthogonal to the trends mediated through *g*. The first feature of note involves the residual means for Word Fluency and Number in the middle-aged group. Although the *g* means show no age-related changes in the middle-aged, the residuals for Word Fluency and Number do. There are small but statistically significant declines in Word Fluency and Number between mean ages 42 and 56. The second noteworthy feature of the residual means in Table 4 is that it is apparent that the large age group (cohort) differences in *g* overestimate age group differences in Number and Verbal Meaning. This is shown by the large negative means in the young group for these two PMA subtests, as well as the large positive means for Number for the old group. Finally, there appears to be modest levels of decline in Space for the old group (between mean ages 58 and 65) that is greater than the decline in Space predicted by *g*.

We do not report here the other parameter estimates from the longitudinal factor solution (e.g., factor covariances, factor loadings) because they differed trivially from the solution ignoring means reported by Hertzog and Schaie (1986). However, one question remained regarding the factor covariance matrix for *g*. As reported in Hertzog and Schaie (1986), there was an age-related increase ing factor variance in the old group, as well as the greater variance in the old group relative to the middle-aged and young groups. One possible explanation of these differences is that they are methodological artifact. The old group was formed by pooling over a larger age span in order to achieve acceptable sample size for structural analysis (refer back to Table 1). In the present context, it was possible that the developmental changes in *g* factor means would differ if the youngest age group (mean age 53 at first occasion; age range 50 to 56) were omitted from the analysis. To address this question, we redefined the old group to include only the individuals 57 and older at first test, and reran the longitudinal model with this subsample. Briefly, this analysis showed 1) similar age declines in g means, but of greater magnitude; 2) higher variability in g in the old group, but 3) more homogeneity of g variance across the three longitudinal occasions. Thus it appears that the increasing variability in g over time, found in the full sample reflected differences in developmental patterns from ages 50 to 65, as opposed to heterogeneity of developmental trajectories for same-aged individuals in the latter part of the adult life span. The analysis thus provides further support to the argument of an inflection point around age 60, in which age decrements in PMA performance begin to accelerate. The increased variability in g in the older group is not, however, merely a methodological artifact of age group definition.

Discussion

The results from this analysis amplify and accentuate several issues regarding age changes in psychometric intelligence. First, the results extend Schaie's (1983) work on age patterns in multiple primary intellectual abilities to the level of general intelligence, as measured by the g factor defined from the PMA subtests. We found a pattern of age changes in g factor means highly consistent with previous univariate results (e.g., Schaie & Hertzog, 1983). There were small increases in g in early adulthood (through mean age 32), stability in g means through middle age (until mean age 56), and substantial decline in late life. We explicitly tested the hypothesis that there was no decline in g in the middle-aged group at two different junctures, and could not reject the hypothesis. Moreover, the age changes that were estimated as part of this hypothesis test were so small as to be trivial in importance. On the other hand, we did find evidence of some decline in the middle-aged group on the PMA subtests Word Fluency and Number, independent of g.

The results also suggest substantial cohort differences in g means. The age groups differ not only in terms of mean age at initial test but also in birth cohort membership. The fact that the middle-aged group at mean age 56 performs significantly better on g than does the old group at mean age 58 surely indicates salient cohort differences in these data, as already detailed by Schaie (1983).

The unique contribution of this study, in terms of estimating age changes in PMA means, stems from the fact that the mean differences are estimated at the level of the g factor. Because these estimates are based upon the simultaneously estimated factor pattern weights, they represent optimal estimates of g factor means that are not contaminated by mean patterns specific to the primary abilities themselves. Moreover, the analysis permitted the evaluation of mean trends in the primary abilities after they have been residualized with respect to g (see below).

An additional contribution of the present analysis is that it permitted independent evaluation of mean stability and covariance stability in g. These results demonstrate concretely the independence of these two types of stability. In all three age groups, individual differences in g were highly stable over the fourteen-year period. Yet each age group showed dramatically different age trends in g. In the young group, g increased to a stable plateau. In the middle-aged group, g means remained stable, but in the old group, substantial g decline was observed.

The change in mean patterns across the age groups, coupled with the high degree of covariance stability across the life span, has important implications for several prominent hypotheses about adult intellectual development. It is often the case, especially recently, that g is identified with basic intelligence (e.g., Jensen, 1982). Given 1) the widely-accepted notion that there is multidirectionality in age trends in ability, such that some but not all abilities show age-related declines (e.g., Baltes et al. 1984; Botwinick, 1977; Horn & Donaldson, 1980, Salthouse, 1982) and the 2) accepted argument that it is measures of fluid intelligence (Horn, 1985; Horn & Donaldson, 1980), or alternatively, Wechsler-type performance tests (Botwinick, 1977; Salthouse, 1982) that manifest early decline, one would expect that g, as measured here, would be the prime candidate for evidencing decline from age 25 to age 55. To the contrary, it appears to be the case that g manifests both mean stability and covariance stability in middle age in the Seattle Longitudinal Sample.

How can this discrepancy be explained? Certainly not on the basis of challenging the validity of the g factor estimated in these data. The g factor loadings estimated here are highly consistent with those found by Thurstone and Thurstone (1941) for these tests, and show a pattern of loadings consistent with a plethora of studies from the psychometric literature. The best indicator of g in the PMA, judged from our factor loadings, is Reasoning. This subtest, a measure of induction, was also identified by Horn and Donaldson (1976) as the best indicator of the Horn-Cattell second-order fluid intelligence factor in the PMA. Yet the Reasoning means in all age groups were well fit by the models specifying no age-related changes in *g* in the middle-aged group. Although we have estimated the single higher-order *g* factor here, as opposed to fluid intelligence, Gustaffson (1984) recently reported hierarchical factor results from multiple intelligence tests that suggest that the *g* factor is isomorphic with fluid intelligence.

Thus it would seem that the hypothesis of early decline in *g* is not supported by these data. The best model for the development of *g* in middle-age is a model of stability in both means and individual differences. One could argue that the generalizability of these results is limited due to the longitudinal design; that is, that there is early decline in individuals who drop out of longitudinal studies. However, the finding of mean stability of *g*, even in a select subpopulation, argues against the ubiquity of early age declines in *g*. Moreover, there is evidence in these data of decline in two PMA subtests, Word Fluency and Number, in the middle-aged group. We suggest that, barring the sort of nonnormative events that lead to early mortality, individuals appear to maintain stable performance levels of *g* into the decade of the fifties.

However, after that age period, the developmental pattern of *g* changes dramatically. After mean age 58, we found substantial, statistically significant decrements in mean levels of *g*. This decline was observed in an age group in which the covariance stability of g remained quite high. These results, then, offer little support to the hope that age-related decline in g is somehow nonnormative, or restricted to a small subpopulation of older individuals. We did find increased variance in g in the middle-aged and older groups, suggesting some small differences in developmental trajectories between those in their fifties and those in their sixties. However, the increases in g variance in the older group -- crucial to the argument of different developmental trajectories in old age -- were removed when the criterion for inclusion in the old group was restricted to individuals 57 and older at first test. Individual differences in g were significantly larger in the old group, even after this adjustment.

The fact that it was necessary to fit residual mean factors, varying in age patterns, provides support for the arguments of Baltes and colleagues that intelligence is both multidimensional and multidirectional in its development. But we find little support for the argument that old age is characterized by substantial interindividual differences in intraindividual change in intelligence (e.g., Baltes & Baltes, 1980; Baltes et al., 1984). To the contrary, these findings of differential age group patterns in g means, coupled with high degree of covariance stability in all age groups, suggest a relatively normative developmental transition in g. That is, it appears that most individuals make a transition from a stability to a decline pattern of g development at some point between age 55 and age 70, with individual differences in the age of onset of this transition.

It is important to note that these inferences are based upon population parameters, and that there are some individuals who do not show salient decline even into old age (Schaie, 1983). Nevertheless, the results suggest that the heterogeneity of developmental trends in g during old age is small, relative to the population variance.

The high degree of covariance stability is a descriptive phenomenon, and should not be assumed to demonstrate the validity of biological causes of age changes in g. Furthermore, stability does not imply immutability, and Schaie and Willis (1986) have demonstrated significant training gains in Reasoning in individuals with prior histories of Reasoning declines (all of whom were, in fact, part of the samples used in the present analysis).

In a sense, these results contradict aspects of the arguments made by both sides of the earlier debate regarding the nature of intellectual decline manifested in the Seattle Longitudinal Study (Baltes & Schaie, 1976; Horn & Donaldson, 1976). The results appear, however, consistent with the updated perspectives of both Horn (1985) and Baltes and his colleagues (e.g., Baltes et al., 1984). The key involves an assessment of the kinds of abilities measured in timed psychometric tests such as the Thurstone PMA, and hence, the nature of the g factor extracted from it. Evidence from a number of studies have shown that Thurstone-type tests of primary abilities (including those measured by the ETS Reference Kit (Ekstrom, French, Harman, & Price, 1976) are highly correlated with speed of basic perceptual processes in adult samples (Cornelius, Willis, Nesselroade, & Baltes, 1983; Horn, Donaldson, & Engstrom, 1981). Schaie originally selected the adolescent form of the PMA for administration to adult samples, and this form has limited item difficulty and substantial speed components in adult samples (e.g., Schaie, Rosenthal, & Perlman, 1953). The g factor estimated in this study was marked as highly by PMA Verbal Meaning as by PMA Reasoning. We have recently shown a strong relationship of PMA Verbal Meaning to a Perceptual Speed factor independent of its relationship to other vocabulary tests (e.g., ETS Advanced Vocabulary; Schaie, Willis, Hertzog, & Schulenberg, in press). Thus it appears that the PMA was constructed so as to maximize variance determined by what might be termed the mechanics of intelligence (e.g., Hunt, 1978): i.e., the speed of basic cognitive processes needed for rapid decisions of low to moderate difficulty. Given that age-related slowing in information processing speed is a highly normative developmental phenomenon (e.g., Birren, 1974; Salthouse, 1985), we can construct the following argument. The PMA manifests little age change in g prior to age 55 because g, as operationally defined by the PMA, emphasizes speeded solution of problems of limited difficulty. However, after the decade of the fifties the age-related slowing in information processing speed becomes a salient limiting factor in PMA performance, and g begins to decline dramatically. Individual differences in decline are minimized because 1) the PMA items are not optimally sensitive to the type of cognitive processes likely to maximize psychometric test performance in superior old adults (e.g., strategies for solving difficult problems, cognitive styles, metacognitive processes -- Baron, 1985; Dixon, 1985; Sternberg, 1985) and 2) the ability domain covered by the tests is highly limited, excluding the types of abilities most likely to show increment and differential growth in adulthood, such as social cognition, domain-specific procedural knowledge, expertise, and post-formal reasoning (Berg & Sternberg, 1985; Dittman-Kohli & Baltes, 1986; Dixon et al., 1985; Labouvie-Vief, 1985; Rybash, Hoyer, & Roodin, 1986). Although important gains can be made by studying these other domains of cognition, we maintain that a continuing study of cognitive mechanics, as they relate to performance on intelligence tests, remains a continuing priority for gerontology. A formal test of the cognitive mechanics interpretation of psychometric test performance in adulthood requires investigation of the nature of the information processing skills tapped by Thurstone-type tests, research now ongoing in several laboratories.

REFERENCES

Baltes, P.B., & Baltes, M. M. (1980). Plasticity and variability in psychological aging: Methodological and theoretical issues. In G. Gurski (Ed.), Determining the effects of aging on the central nervous system. Berlin, Schering.

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., Dittman-Kohli, F., & Dixon, R.A. (1984). New perspectives on the development of intelligence in adulthood: Toward a dual-process conception and a model of selective optimization with compensation. In P.B. Baltes & O.G. Brim, Jr. (Eds.), Life-span development and behavior (Vol. 6, pp. 34-76). New York: Academic Press.

Baltes, P.B., Reese, H.W., & Nesselroade, J.R. (1977). Life-span developmental psychology: Introduction to research methods. Monterey, CA: Brooks-Cole.

Baltes, P.B., & Schaie, K.W. (1976). On the plasticity of intelligence in old age: Where Horn and Donaldson fail. American Psychologist, 31, 720-725.

Baron, J. (1985). Rationality and intelligence. New York: Cambridge University Press.

Bentler, P.M. (1985). Theory and implementation of EQS: A structural equations program. Los Angeles: BMDP Statistical Software.

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

Berg, C.A., & Sternberg, R.J. (1985). A triarchic theory of intellectual development during childhood. Developmental Review, 5, 334-370.

Birren, J.E. (1974). Translations in gerontology -- from lab to life: Psychophysiology and the speed of response. American Psychologist, 29, 808-815.

Botwinick, J. (1977). Intellectual abilities. In J.E. Birren and K.W. Schaie (Eds.), Handbook of the Psychology of Aging (pp. 580-605). New York: Van Nostrand Reinhold.

Cornelius, S.W. Willis, S.L., Nesselroade, J.R. & Baltes, P.B. (1983). Convergence between attention variables and factors of psychometric intelligence in older adults. Intelligence, 7, 253-269.

Dittman-Kohli, F., & Baltes, P.B. (1986). Towards a neo-functionalist conception of adult intellectual development: Wisdom as the prototypical case of intellectual growth. In C. Alexander & E. Langer (Eds.), Beyond formal operations: Alternative endpoints to human development. New York: Oxford University Press.

Dixon, R.A. (in press). Questionnaire research on metamemory and aging: Issues of structure and function. In L.W. Poon, D.C. Rubin, & B.A. Wilson

(Eds.), Everyday cognition in adulthood and old age. New York: Cambridge University Press.

Dixon, R.A., Kramer, D.A., & Baltes, P.B. (1985). Intelligence: Its life-span development. In B. B. Wolman (Ed.), Handbook of intelligence: Theories, measurements, and applications(pp. 469-518). New York: John Wiley & Sons.

Gustafsson, Jan-Eric. (1984). A unifying model for the structure of intellectual abilities. Intelligence, 8, 179-203.

Hertzog, C. (1985). Applications of confirmatory factor analysis to the study of intelligence. In D.K. Detterman (Ed.), Current Topics in Human Intelligence. (pp.59-97). Norwood, N.J.: Ablex.

Hertzog, C. (in press). On the utility of structural regression models for developmental research. In Baltes, P.B., Featherman, D., & Lerner, R.M. (Eds.) Life-span Development and Behavior. (Vol. 10). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hertzog, C., & Carter, L. (1982). Sex differences in the structure of intelligence: A confirmatory factor analysis. Intelligence, 6, 287-303.

Hertzog, C. & Nesselroade, J.R. (in press). Beyond autoregressive models: Some implications of the trait-state distinction for the structural modeling of developmental change. Child Development.

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

Horn, J.L. (1985). Remodeling old models of intelligence. In B. B. Wolman (Ed.), Handbook of intelligence: Theory, measurements, and applications (pp. 267-300). New York: John Wiley & Sons.

Horn, J.L., & Donaldson, G. (1976). On the myth of intellectual decline in adulthood. American Psychologist, 31, 701-719.

Horn, J.L., & Donaldson, G. (1980). Cognitive development in adulthood. In O.G. Brim, Jr., & J. Kagan (Eds.), Constancy and change in human development (pp. 445-529). Cambridge, MA: Harvard University Press.

Horn, J.L., Donaldson, G., & Engstrom, R. (1981). Apprehension, memory, and fluid intelligence decline in adulthood. Research on Aging, 3, 33-84.

Horn, J.L. & McArdle, J.J. (1980). Perspectives on mathematical/statistical model building (MASMOB) in research on aging. In L.W. Poon (Ed.), Aging in the 1980's: Psychological Issues (pp. 503-541). Washington, D.C.: American Psychological Association.

Horn. J.L., McArdle, J.J., & Mason, R. (1984). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. Southern Psychologist, 1, 179-188.

Hunt, E. (1978). The mechanics of verbal ability. Psychological Review, 85, 109-130.

Jensen, A.R. (1982). Reaction time and psychometric g. In H. J. Eysenck (Ed.), A model for intelligence. New York: Springer-Verlag.

Joreskog, K.G. & Sorbom, D. (1977). Statistical models and methods for analyses of longitudinal data. In D.S. Aigner & A.S. Goldberger (Eds.), Latent Variables in Socio-economic Models (pp. 285-325). Amsterdam, North Holland.

Joreskog, K.G. & Sorbom, D. (1980). Simultaneous Analysis of Longitudinal Data From Several Cohorts. Research Report 80-5. University of Uppsala, Dept. of Statistics.

Joreskog, K.G. & Sorbom, D. (1984). LISREL VI User's Guide. Mooresville, IN: Scientific Software.

Kagan, J. (1980). Perspectives on continuity. In O.G. Brim, Jr., & J. Kagan (Eds.), Constancy and Change in Human Development. Cambridge, MA: Harvard University Press.

Labouvie. E.W. (1980a). Identity versus equivalence of psychological measures and constructs. In L.W. Poon (Ed.), Aging in the 1980's: Psychological issues (pp. 493-502). Washington, D.C.: American Psychological Association.

Labouvie, E.W. (1980b). Measurement of individual differences in intraindividual changes. Psychological Bulletin, 88, 54-59.

Labouvie-Vief, G. (1985). Intelligence and cognition. In J.E. Birren & K.W. Schaie (Eds.), Handbook of the psychology of aging (2nd Ed: pp. 500-530). New York: Van Nostrand Rheinhold.

McArdle, J.J. (in press). Latent growth curves within developmental structural equation models. Child Development.

McArdle, J.J. & McDonald, R. P. (1984). Some Algebraic Properties of the Reticular Action Model for Moment Structures. British Journal of

Mathematical and Statistical Psychology, 37, 234-251.

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. 35-60). New York: Van Nostrand Reinhold.

Rybash, J.M., Hoyer, W.J., & Roodin, P.A. (1986). Adult cognition and aging: Developmental changes in processing, knowing, and thinking. New York: Pergamon Press.

Salthouse, T.A. (1982). Adult cognition: An experimental psychology of human aging. New York: Springer.

Salthouse, T.A. (1985). A Theory of Cognitive Aging. Amsterdam: North Holland.

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 Press.

Schaie, K.W., & Hertzog, C. (1983). Fourteen-year cohort-sequential analyses of adult intellectual development. Developmental Psychology, 19, 531-543.

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 Edition). New York: Van Nostrand Reinhold.

Schaie, K.W., Rosenthal, F., & Perlman, R. M. (1953). Differential mental deterioration on factorially "pure" functions in later maturity. Journal of Gerontology, 8, 191-196.

Schaie, K.W., Willis, S.L., Hertzog, C., & Schulenberg, J.E. (in press). Effects of cognitive training upon primary mental ability structure. Psychology and Aging.

Sorbom, D. (1982) Structural equation models with structured means. In K.G. Joreskog & H. Wold (Eds.), Systems under indirect observation: Causality, structure, prediction (pp. 183-195). Amsterdam: North Holland.

Sobel, M.E., & Bohrnstedt, G.W. (1985). Use of null models in evaluating the fit of covariance structure models. In N.B. Tuma (Ed.), Sociological methodology. San Francisco: Jossey-Bass.

Sternberg, R. J. (1985) Beyond IQ: A triarchic theory of intelligence. New York: Cambridge University Press.

Thurstone, L.L., & Thurstone, T.G. (1941). Factorial studies of intelligence. (Psychometric Monographs, No. 2). Chicago: University of Chicago Press.

Thurstone, L.L. & Thurstone, T.G. (1949). Examiners Manual, SRA Primary Mental Abilities Test (Form 11-17). Science Research Associates, Chicago.

Authors' Note

This article reports data collected as part of the Seattle Longitudinal Study, which has been supported over an extended period of time by grants from the National Institutes of Health, the National Institute of Child and Human Development, and the National Institute on Aging. Currently, the study is supported by Grant R01-AG04770 from the National Institute on Aging. The first author's effort on this project was also supported by a Research Career Development Award from the National Institute on Aging (K04-AG00335).

The cooperation and support from members and staff of the Group Health Cooperative of Puget Sound is gratefully acknowledged.

Correspondence concerning this article should be addressed to Christopher Hertzog, School of Psychology, Georgia Institute of Technology, Atlanta, GA 30332-0170.

Table 1

Reparameterized Sequential Sample for
Multiple Group Analysis

	Sample	Cohort (mean birth year)	Age (means)	N
Group 1				
			30, 37, 44	109
	1	1931	25, 32, 39	21
	1	1924	32, 39, 46	26
	2	1938	25, 32, 39	22
	2	1931	32, 39, 46	40
Group 2				
			42, 49, 56	160
	1	1917	39, 46, 53	27
	1	1910	46, 53, 60	32
	2	1924	39, 46, 53	51
	2	1917	46, 53, 60	50
Group 3				
			58, 65, 72	143
	1	1903	53, 60, 67	28
	1	1896	60, 67, 74	15
	1	1889	67, 74, 81	13
	2	1910	53, 60, 67	48
	2	1903	60, 67, 74	18
	2	1896	67, 74, 81	21

TABLE 2

Goodness-of-Fit Comparisons for Longitudinal Factor Model with Means

Model	χ^2	df	F[a]	p
1. M_s (saturated)	287.68	248	.352	.048
2. M_n (null in means)	642.02	288	.785	.000
3. M_1 (*g* factor means)	467.59	280	.572	.000
4. M_2 (*g* factor means; all 0 in Middle-Aged)	470.08	282	.575	.000
5. M_3 (*g* and test-specific factor means)	338.76	270	.414	.003
6. M_4 (*g* and residual means for V,S,N,W)	299.05	254	.366	.027

Model Comparison

	M_n		M_s						
	$\Delta\chi^2$	Δdf	$\Delta\chi^2$	Δdf	δ[b]	Comparison	$\Delta\chi^2$	Δdf	$\Delta\delta$
1. M_s	–	–	–	–	–	–	–	–	–
2. M_n	–	–	–	–	–	2-1	354.34	40	
3. M_1	174.43	8	179.91	32	.492	–	–	–	–
4. M_2	171.94	6	182.40	34	.485	3-4	2.49	4	.007
5. M_3	303.26	18	51.08	22	.857	4-5	128.83	10	.365
6. M_4	342.97	34	11.37	6	.968	3-6	168.54	28	.483

Abbreviations: V - Verbal Meaning; S - space; R - Reasoning; N - Number;
W - Word Fluency

[a] LISREL fitting function at minimum

[b] Relative fit index for fit to the mean structure (see text)

[c] Comparison of subscripted models (e.g., 2-1 compares M_n to M_s)

TABLE 3

g. Factor Means for Alternative Longitudinal Models

	Mean Age	Model [a] M_1	M_2	M_3	M_4
Young					
g_1 [b]	30	1.61(0.60) [c]	1.62(0.59)	8.54(3.26)	2.82(0.65)
g_2	37	2.76(0.57)	2.78(0.57)	10.11(3.49)	3.99(0.65)
g_3	44	2.70(0.56)	2.71(0.55)	9.87(3.30)	3.50(0.62)
Middle-Aged					
g_1	42	0*(-)	0*(-)	0*(-)	0*(-)
g_2	49	0.10(0.17)	0*(-)	0.14(0.16)	0*(-)
g_3	56	-0.20(0.18)	0*(-)	-0.20(0.17)	0*(-)
Old					
g_1	58	-3.96(0.61)	-3.97(0.60)	-10.96(4.48)	-4.20(0.64)
g_2	65	-4.61(0.61)	-4.62(0.61)	-12.41(4.64)	-4.78(0.64)
g_3	72	-6.55(0.65)	-6.57(0.64)	-13.28(4.24)	-6.22(0.66)

Note: Asterisks denote fixed factor means.

[a] Model designation corresponds to Table 2 and text

[b] Model Subscript denotes longitudinal occasion

[c] Standard errors in parentheses

TABLE 4

Residual Means in Final Model (M_4)

Variable	Occasion	Age Group: Young	Middle-Aged	Old
Verbal Meaning	1	-5.10(1.01)[a]	0*	0.26(0.98)
Verbal Meaning	2	-4.75(1.07)	0*	1.09(1.05)
Verbal Meaning	3	-3.65(1.03)	0*	-0.49(1.08)
Space	1	0.58(1.15)	0*	-1.19(1.01)
Space	2	0.98(1.22)	0*	-2.68(1.01)
Space	3	1.76(1.20)	0*	-2.56(1.03)
Reasoning	1	0*	0*	0*
Reasoning	2	0*	0*	0*
Reasoning	3	0*	0*	0*
Number	1	-5.56(1.32)	0*	3.71(1.23)
Number	2	-5.58(1.40)	0.28(0.44)	5.12(1.28)
Number	3	-6.03(1.31)	-1.62(0.43)	3.38(1.27)
Word Fluency	1	-1.45(1.48)	0*	4.98(1.45)
Word Fluency	2	-3.56(1.59)	-1.43(0.68)	2.77(1.46)
Word Fluency	3	-1.18(1.60)	-2.08(0.69)	2.36(1.49)

Note: Asterisks denote fixed 0 parameters.

[a] Standard errors in parentheses

EFFECTS OF COGNITIVE TRAINING UPON PRIMARY MENTAL ABILITY STRUCTURE

K. Warner Schaie

Sherry L. Willis

Christopher Hertzog

and

John E. Schulenberg

The Pennsylvania State University

Running Head: Cognitive Training and Ability Structure

Abstract

This paper reports results of the first empirical test, as far as we know, of the assumption of structural invariance of latent constructs from pre- to posttest in cognitive training research on the elderly. 401 participants of the Seattle Longitudinal Study aged over 62 years received a 5-hour test battery at pre- and post-test that included 16 ability tests, marking the 5 primary abilities of Spatial Orientation, Inductive Reasoning, Numerical Ability, Verbal Ability, and Perceptual Speed. 229 of our subjects received 5 hours of individual training on either Spatial Orientation or Inductive Reasoning. Restricted factor analysis using the LISREL algorithm was used to test the hypothesis of measurement equivalence across test occasions, separately for the controls and for each of the training groups. The regression of observable marker variables on their latent ability factors was found to be virtually undisturbed by test-retest effects. 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 ability.

Effects of Cognitive Training Upon Primary Mental Ability Structure

K. Warner Schaie, Sherry L. Willis, Christopher Hertzog,
and John E. Schulenberg
The Pennsylvania State University

Introduction

During the past few years there has been a growing interest in research that investigates the question whether the cognitive performance of older adults can be improved by means of training interventions. Such training may be directed towards remediating clearly identified deficit (cf. Schaie & Willis, in press; Willis, 1985), or towards improving the performance of elderly persons with unknown prior ability status (Baltes & Willis, 1982; Sterns & Sanders, 1980). In all these studies the primary concern is not to prove that it is possible to improve subjects' performance on some measure by "teaching the test," but rather to show that there has been training gain on a more general latent construct or ability factor. To attain this objective it is generally necessary to design a transfer-of-training type study that assesses variables that are hypothesized to benefit from the proposed training regime as well as others that should not improve if training is ability-specific. Assessment of the convergent and divergent validity of the training paradigm for the latent ability constructs, moreover, requires multiple markers (observed variables) for all of the abilities to be included in the study.

A critical assumption that underlies the evaluation of the effects of cognitive training research in the elderly at the ability level is the supposition that the projection of the observed marker variables upon the latent ability factors of interest remain equivalent from pre- to posttest. If this assumption is true, then training gains can be interpreted unambiguously. That is, training can be interpreted as increasing levels of performance without altering the nature of the performance. If the assumption is false, then it is possible that the intervention many have produced change in ability structure, and estimates of level changes could be biased. For example, differential near transfer effects of training could result in changing the factor loadings for one or more of the ability markers. More seriously even, changes in structure could obscure training gains at the factor level that would have been observed had the structure remained invariant. The hypothesis that factorial invariance has been maintained across the training intervention can best be tested by applying methods of restricted factor analysis, such as the LISREL algorithm (Joreskog & Sorbom, 1984). This paper reports the first empirical test, as far as we know, of the assumption of measurement equivalence from pre- to posttest in a cognitive training study.

Method

Subjects

Our sample consisted of 401 participants (177 men and 224 women) over the age of 62 years from the Seattle metropolitan area, who had been participants in the Seattle Longitudinal Study (SLS) since 1975 or earlier (Schaie, 1983). All subjects are, or had been members of the Group Health Cooperative of Puget Sound, a health maintenance organization. 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 subjects were community dwelling, and most were Caucasian.

Design and Procedure

Training Paradigm. All subjects received a 5-hour ability test battery at pre- and posttest. 229 of our subjects received five hours of individual cognitive training on either Spatial Ability (N = 1 8) or on Inductive Reasoning (N = 111). The remaining 172 subjects were testing only controls.

Classification of participants. Subjects' test performances on the Thurstone PMA Reasoning and Spatial Orientation measures were classified as having remained stable or having declined over the prior 9 to 14 year interval (1970/75-1984). Subjects entered the study at different points in time (from 1956 though 1975); performance in 1970/75 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. There were 170 subjects (42.4% of sample) who were classified as having remained stable on the training target abilities, while 231 subjects (57.6%) had decline on one or both of the abilities.

Assignment of subjects. Subjects were assigned to either Reasoning or Space training programs, based on their performance status. Subjects who had declined on Reasoning, but not on Space, 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 2 1/2 hour sessions conducted in small groups. Training involved 5 one-hour 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. Subjects were paid $100 for participation in the study.

Measures

The pre-posttest battery involved psychometric measures representing five primary mental abilities, including the Thurstone Primary Mental Ability measures (Thurstone, 1948) administered at previous SLS assessments. Each ability was represented by three to four marker measures (see Table 1). All tests are slightly speeded.

Insert Table 1 about here

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 5 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 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 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).

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, based on 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, 180 degrees). Cognitive strategies to facilitate mental rotation which were focused upon in training included: 1) Development of concrete terms for various angles; 2) Practice with manual rotation of figures prior to mental rotation; 3) Practice with rotation of drawings of concrete, familiar objects prior to introduction of abstract figures; 4) Subject-generated names for abstract figures; and 5) 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 (in press).

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 (Joreskog & Sorbom, 1984) to perform simultaneous, multiple group, confirmatory factor analysis (see Joreskog, 1971, and Schaie & Hertzog, 1985, for a further discussion 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

$$\underline{y} = \Lambda \eta + \varepsilon \quad (1),$$

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

$$\Sigma = \Lambda C \Lambda' + \Theta \quad (2),$$

where Λ is as before, C is the covariance matrix of the η, and Θ is the covariance matrix of the ε s. Equation (2) should be recognized as a restricted factor analysis model that can be generalized to a multiple group model (Joreskog & Sorbom, 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 the equations in (2) 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 Ø) 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^2 test statistic. Differences in Chi^2 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 upon the model parameters) that is nested within a less restrictive model would be accepted over the less restrictive model if the difference in Chi^2 between the two models is not significant. Conversely, if the difference in Chi^2 is significant, then the less restrictive model would be accepted.

In multiple groups analysis it is necessary to estimate factor models using covariance metric and sample covariance matrices rather than to standardize the solution and analyze separately standardized correlation matrices. Standardization could obscure invariant relationships because of group differences in observed variances (Joreskog, 1971). Our approach is to estimate 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. Since standardized statistics are easier to interpret, we generally report parameter estimates that have been rescaled to a standardized metric, using a SAS PROC Matrix program (Hertzog & Cannon, 1985). This program extends formulae supplied by Joreskog (1971; see also Alwin and Jackson, 1980) to handle both multiple groups and longitudinal factors (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 models so that the reader may evaluate the statistical significance of individual parameter estimates against 1) a null hypothesis that each parameter is equal to zero or 2) that group differences in unconstrained parameters are statistically reliable. In general, parameters that exceed their standard errors by a ratio of 2:1 are reliably different from zero at a 5% (per comparison) alpha level.

This study also employs LISREL VI to perform structural regression analysis of the pretest-posttest data. All structural regression models were performed by using the β matrix in LISREL. The structural model was therefore

$$\eta = \beta\eta + \zeta$$

where β is the matrix of regression coefficients and LISREL estimates the coefficients in β as well as the covariance matrix of regression residuals, Ψ.

Results

Factor Analysis of Pre-test Data

The analysis was begun by attempting 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: 1) Induction, 2) Space, 3) Perceptual Speed, 4) Number, and 5) 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. The basic specification that was ultimately used in confirmatory model testing is listed in Table 2.

Insert Table 2 about here

Invariance Across Stable and Unstable Groups

Given that the training analysis classified groups by prior developmental history (i.e., stable levels of intelligence versus declining levels of intelligence; see Schaie & Willis, in press), it was necessary to evaluate the invariance of the ability factor structure across the stable and decline groups. Although this analysis was of interest in its own right, it was also required by our need to pool data over stable and decline groups to get large sample sizes for the more crucial tests of invariance between the training and control groups. The model shown in Table 2 was tested for invariance in factor pattern weights ( ), factor covariance matrices ( ), and residual covariance matrices ( ). The test of this model provided a reasonable fit (Chi^2 = 509.22, *df* = 260, GFI Stable = .822, GFI Unstable = .892). Relaxation of the constraints on the residual variances, however, yielded a significant improvement (Chi^2 = 463.17, *df* = 243, GFI Stable = .847, GFI Unstable = .899; difference in Chi^2 = 46.05, *df* = 243, $p < .001$). Further relaxation of constraints upon factor covariances and factor weights while maintaining equivalent factor patterns did not result in significant improvements in fit. The results of these tests showed that the stable and unstable groups could be considered to have equivalent factor pattern weights and factor covariance matrices, while the hypothesis of equal residual variances had to be rejected. This configuration suggests complete invariance of the solution in the common factor space for the stable and unstable groups. We therefore concluded that pooling the data over the stable and unstable developmental pattern cases was justified.

As noted in the description of the sample, there are also differences in the proportion of males and females represented in each training condition. In order to evaluate the possibility of confounded sex differences, we first analyzed for sex differences in factor structure. Paralleling the results with stable and unstable groups, no salient differences were found.

Invariance Across Training and Control Groups

Given the invariance of the factor solution for the pretest data across both prior developmental history and gender, the stage had been set for the main thrust of the question addressed by this paper -- testing for changes in factor structure as a function of training. The approach used was to specify a longitudinal factor analysis model for the pretest and posttest data for each of the three training condition groups: Controls, Inductive Reasoning training, and Spatial Orientation training. We began by estimating the model in the control group,and then estimating a structural regression model for the data in that group from preto posttest. Following the control group analyses, we proceeded to estimate the longitudinal measurement model and the structural regression models in the two training groups[1].

The basic longitudinal factor model specified the structure posited in Table 2 at both pretest and posttest. In addition, correlated residuals were specifed for each scale across pretest and posttest. For example, the model included a residual covariance of the residual for PMA Space at pretest and PMA Space at posttest. Previous work with longitudinal models of intelligence measures has shown these correlated residuals to be present (e.g., Hertzog & Schaie, in press; Sorbom, 1975). The fit of the basic model was adequate, although not perfect (Chi^2 = 552.37, df = 399, GFI = .838).

The basic measurement model was used as the basis for evaluating the hypothesis of longitudinal invariance in the factor pattern weights ( ). A model constraining the corresponding loadings equal between pretest and posttest showed some indication of strain on the model (Chi^2 = 574.84, df = 412, GFI = .833). The change in fit was just significant at the .05 but not the .01 level (change in Chi^2 = 2 .47, df = 13, $p < .05$). 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 $Chi^2 = 2.92$, $\underline{df} = 1$, $p > .10$), nor did the LISREL goodness of fit index 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 3 provides the standardized factor loadings, standardized unique variances and correlated errors, and factor intercorrelations for the accepted model.

Insert Table 3 about here

At this point we proceeded to test a structural regression model for the control group. The starting point was an isolated stability model (Hertzog, 1986), positing regression of the five posttest factors on their pretest counterparts (autoregressions; see Joreskog & Sorbom, 1977), covariances among the pretest factors, and residual variances for the posttest factors. This model may be considered an isolated stability model because it includes no relationships among posttest factors except as mediated through the autoregression coefficients (stable individual differences from pre- to posttest). The fit of the isolated stability model was compared to the measurement model's fit in order to evaluate the adequacy of the structural regression part of the model. The isolated stability pattern did not provide as good a fit as the unconstrained factor covariances of the measurement model (Chi^2 = 628.49, df = 442, GFI = .813; difference in Chi^2 = 53.65, df = 30, $p < .01$). We proceeded to evaluate the cross-lagged coefficients for significance, using the modification indices as a guide. Ultimately, two statistically significant coefficients were found (the regression of posttest Inductive Reasoning on pretest Numerical Ability, and the regression of posttest Verbal Ability on pretest Perceptual Speed. These changes improved the fit (Chi^2 = 614.33, df, = 440, GFI = .822), so that the difference of the structural regression model from the measurement model was no longer statistically significant (difference from measurement model, Chi^2 = 39.49, df = 28, $p > .05$). No other plausible model improvements were identified.

Figure 1 depicts the standardized regression coefficients of both the isolated stability and the final models. Clearly, the control groups's data are dominated by the stability of individual differences from pretest to posttest. In both models, the standardized autoregression coefficients are close to unity. Table 4 provides the estimated factor variances at pretest and posttest, and the unstandardized regression coefficients. As indicated by the factor variances, individual differences generally increased slightly in magnitude from pretest to posttest in the control group, but the rank order of individuals about the group mean are highly consistent over time and the estimated cross-lagged coefficients are quite small.

Insert Figure 1 and Table 4 about here

The control group analysis provides a benchmark against which to evaluate the changes in factor structure and individual differences in the training groups. The analysis in each of these groups paralleled the analysis in the control group -- testing first the longitudinal invariance of factor structure and then the structural regression model.

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 = 574.43, *df* = 399, 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 = 599.00, *df* = 412, GFI = .767; change in Chi^2 = 24.57, *df* = 13, $p < .05$). This statistically reliable difference was not surprising, given difference 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 = 16.15, *df* = 3, $p < .001$). It appeared that most of the lack of fit (i.e., 16 of 25 Chi^2 units with only 3 of 13 df) of the model specifying invariant pattern weights could thus be attributed to the Induction scales. In turn, the only significant difference in factor loadings among the Induction indicators involved the Word Series scale. Note that this was also the scale that showed some shift in the control group; however, the 1 df test of the difference was significant in the Induction group ($Chi^2 = 12.42$, $p < .001$), whereas it had not been in the control group (see above). Table 5 provides the standardized factor loadings, standardized unique variances, correlated errors, and factor intercorrelations for the accepted measurement model for the Induction training group (all factor loadings being constrained equal over time, with the exception of the loadings for the Word Series measure).

Insert Table 5 about here

All tests of the structural model proceeded with the same basic specification and longitudinal invariance in factor pattern weights, excepting Word Series. The level of fit of this measurement model ($Chi^2 = 586.85$, $df = 411$, $GFI = .770$) was subsequently used to evaluate the fit of the structural regression models. An isolated stability model for the Induction group did not fare badly, not being significantly different from the measurement model ($Chi^2 = 628.53$, $df = 441$, $GFI = .757$; change in $Chi^2 = 41.68$, $df = 30$, $p < .05$). We did test, by hypothesis, the model adding cross-lagged regressions predicting posttest Induction from the pretest ability factors. This model fit marginally better than the isolated stability model, with the regression of Inductive Reasoning on Perceptual Speed identified as the only salient cross-lagged coefficient. This is a model in which the autoregressive coefficients and the regression coefficient of Inductive Reasoning on Perceptual Speed were freely estimated. This model did not differ significantly from the measurement model, and thus was accepted as the preferable model (Chi^2 = 621.09, $\underline{df}$ = 440, GFI = .756; change in Chi^2 = 32.42, $\underline{df}$ = 29, $p > .35$). The standardized stability and cross-lagged coefficients for this final model, as well as of the isolated stability model are depicted in Figure 2. As is evident for both models, the stability coefficients are all close to unity.

Insert Figure 2 about here

The unstandardized structural regression coefficients and the estimated factor variances at pretest and posttest for the Induction group were given in Table 4. Comparison with the data from the control group also presented there shows that both groups are characterized by high stability of individual differences and increasing variance over time.

Space Training Group. The basic longitudinal factor model did not fit the Space training data as well as it fit the data for the other two groups (Chi^2 = 685.07, df = 399, GFI = .746). The test of invariant factor pattern weights over time resulted in a marginally salient reduction in fit (Chi^2 = 707.61, df = 412, GFI = .740; change in Chi^2 = 22.54, df = 13, $p < .05$). An assessment of the lack of fit attributable to the Space factor indicators (analogous to the search for lack of fit in the Inductive Reasoning factor for the Induction training group) revealed that the Object Rotation test was carrying most of the stress in the model with respect to invariant factor pattern weights. By freeing the factor loading for Object Rotation and constraining the loadings of the other markers for the Spatial Orientation factor (as well as the markers for the other factors) to be invariant over time, an acceptable measurement model was found (Chi^2 = 700.83, df = 411, GFI = .742; change in Chi^2 compared to the model = 15.76, df = 12, $p < .20$). Table 6 provides the standardized factor loadings, standardized unique variances, correlated errors, and factor intercorrelations for the accepted model.

Insert Table 6 about here

The isolated stability model provided a relatively poor fit as compared to the accepted measurement model (Chi^2 = 761.01, df = 441, GFI = .724; change in Chi^2 = 60.18, df = 30, $p < .001$). By hypothesis, the model that included cross-lagged regressions predicting posttest Spatial Orientation from the other factors at pretest was tested; however, this model did not significantly increase the fit, and none of the cross-lags were significant. We did identify, on the basis of the modification indices, two small but salient cross-lagged regression coefficients: (1) posttest Verbal Ability on pretest Perceptual Speed; and (2) posttest Perceptual Speed on pretest Verbal Ability. A model positing these two cross-lagged regression coefficients, as well as the autoregressive coefficients was tested (Chi^2 = 747.25, df = 439, GFI = .731). Although this model was a significant improvement over the isolated stability model (change in Chi^2 = 13.76, df = 2, p = < .001), it still provided a marginally, but significantly worse fit to the data than did the accepted measurement model (change in Chi^2 = 46.32, df = 28, $p < .02$). Nevertheless, based on the modification indices, no other plausible model could be identified and this final structural model was accepted. Figure 3 illustrates the standardized autoregressive and cross-lagged coefficients for both the final model and the isolated stability model.

Insert Figure 3 about here

The unstandardized structural regression coefficients and the estimated factor variances at pretest and posttest for the Space training group were given in Table 4. As was the case for the other two groups, individual differences were quite stable over time, and factor variances tended to increase over time.

Discussion and Conclusions

The hypothesis of structural invariance across cognitive training intervention in an elderly sample was investigated in this study by means of restricted factor analysis using the LISREL paradigm. A five-factor measurement model was first identified on the basis of the pretest data for all subjects. Next, the equivalence of factor structure was tested across subsets of subjects that had declined or remained stable, and across subsets aggregated by gender. Factorial invariance was tested with a model that constrained factor loadings and factor correlations across groups. In both instances, stability status and gender, factorial invariance across groups was found to be acceptable.

Having demonstrated structural invariance over time in the control group, an isolated stability model (i.e., one that permits only autocorrelated but no cross-lagged regression coefficients) was then tested across the pre- and posttest data. Stability coefficients for the ability factors were above .90, and ability measures at pretest predicted approximately 97 percent 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.

The same model was next tested separately for each of the groups that received the training intervention. 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 pre- to posttest were only slightly lower for the Inductive Reasoning and Spatial Orientation training groups. The perturbations in the projections of the observed variables upon 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 or any of the target-ability extraneous observable-latent relationships. For the Induction training group an improved fit could be obtained when the across occasion constraint upon 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. This finding suggests that differential effects of the training procedure upon the various marker variables, usually referred to as "near transfer," had consequences also for the observable/latent relationship for some but not for other markers of the target training ability at posttest. In both instances where significant change in optimal factor loading occurred at post-test in the training group (but not in the controls) the stimuli involved were the most "concrete" markers of the target ability. It is conceivable that training may have led to increasingly routinized response on these variables, somewhat reducing their contribution as a marker of the latent ability.

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. Employing a set of multiple indicators for a latent variable, such as were provided 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 (1) we have indeed trained on the latent variable, (2) we can unambiguously interpret individual differences and mean changes in the latent variable as a function of training, and (3) we can identify which indicators are reactive to training in terms of shifting measurement properties.

Results of this study suggest that the regression of observed marker variables on their latent ability factors is virtually undisturbed by test-retest effects over brief test intervals (two to four weeks in our case) when no ability-specific intervention occurs between test occasions. When ability-specific cognitive training intervenes, no structural changes are observed for those abilities that were not subject to intervention (far transfer). However, slight shifts did occur in the optimal regression weights for the different markers for the training target ability. This finding suggests that training studies that wish to assess training effects at the latent ability level should include procdures such as those reported here. Factor regression weights used to estimate factor scores at pre- and posttest can then be separately estimated to assure equivalence of ability factors across occasions.

In our study, regression weights computed for the best fitting posttest model were only trivially different from those estimated on the basis of the best-fitting pretest model across all training groups. It is noteworthy, however, that both retest and training result in increased variability for the latent variables. In effect this means that practice and other interventions have counter-intuitively increased rather than reduced individual differences in cognitive performance. As the analysis of changes in level of performance has shown (cf. Schaie & Willis, in press), most subjects who declined or remained stable over the prior fourteen 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 <u>not</u> 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., & 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.

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.

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 Serivce.

Hertzog, C. (1986). 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 (Vol. 8). Hillsdale, NJ: Erlbaum.

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

Hertzog, C., & Schaie, K. W. (In press). Stability and change in adult intelligence: I. Analysis of longituinal covariance structures. Psychology and Aging, 1.

Joreskog, K. G. (1971). Simultaneous factor analysis in several populations. Psychometrika, 36, 409-426.

Joreskog, K. G., & Sorbom, 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 Publishers.

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

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 Psychological 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. (In press). Can decline in adult intellectual functioning be reversed? Developmental Psychology, 22.

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., & Schaie, K. W. (1983). Alphanumeric rotation test. Unpublished manuscript, The Pennsylvania State University, University Park, PA.

Author Note

The research reported in this article was supported by grant # AG03544 from the National Institute of Aging. We gratefully acknowledge the enthusiastic cooperation of members and staff of the Group Health Cooperative of Puget Sound. Dr. Hertzog is now at the School of Psychology, Georgia Institute of Technology.

Reprint requests should be sent to K. Warner Schaie, Department of Individual and Family Studies, S-110 Human Development Building, The Pennsylvania State University, University Park, PA 16802.

Footnotes

[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 1

Intellectual Abilities Measurement Battery

Primary Ability	Test	Source
Inductive Reasoning	PMA Reasoning	Thurstone, 1948
	ADEPT Letter Series (Form A)	Blieszner, Willis, & Baltes, 1981
	Word Series	Schaie, 1985
	Number Series	Ekstrom, French, Harman, & Derman, 1976
Spatial Orientation	PMA Space	Thurstone, 1948
	Object Rotation	Schaie, 1985
	Alphanumeric Rotation	Willis & Schaie, 1983
Perceptual Speed	Finding A's	Ekstrom et al., 1976
	Number Comparison	Ekstrom et al., 1976
	Identical Pictures	Ekstrom et al., 1976
Numerical Ability	PMA Number	Thurstone, 1948
	Addition	Ekstrom et al., 1976
	Subtraction & Multiplication	Ekstrom et al., 1976
Verbal Ability	PMA Verbal Meaning	Thurstone, 1948
	ETS Vocabulary II	Ekstrom et al., 1976
	ETS Vocabulary IV	Ekstrom et al., 1976

Table 2

Specification of the Factor Loading Pattern Matrix

for the Accepted Measurement Model

	Ability Factor				
Variable	Induction	Space	Perc. Speed	Number	Verbal
PMA Reasoning	1[a]				
ADEPT Letter Series	1				
Word Series	1				
Number Series	1				
PMA Space		1[a]			
Object Rotation		1			
Alphanumeric Rotation		1			
Finding A's			1		
Number Comparison			1[a]	1	
Identical Pictures			1		
PMA Number				1[a]	
Addition				1	
Subtraction & Multiplication				1	
PMA Verbal Meaning			1		1
Vocabulary II					1[a]
Vocabulary IV					1

(Table 2 Footnote)

Note. A "1" indicates that the given variable loads on that factor and a blank indicates that the given variable does not load on that factor. The Phi matrix is symmetrical and free, the Theta matrix is diagonal. [a]To identify the metric of each factor, these elements were fixed to 1.ØØ.

Table 3

Rescaled Solution for the Accepted Pretest-Posttest Measurement Model for the Control Group

Variable	Factor Loadings[a]					Unique Variance		Unique Auto-Correl.
	Induction	Space	Perc. 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

(Table 3 Continues)

Ability Factor

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

[a]Since the accepted measurement model included the factor loadings being set equal over time, this matrix was identical for both pretest and posttest.

Table 4

Estimated Factor Variances and Unstandardized Structural

Regression Coefficients for the Three Training Conditions

	Training Group					
	Controls		Induction		Space	
	Factor Variance[a]					
Factor	Pretest	Posttest	Pretest	Posttest	Pretest	Posttest
Induction	33.41	38.39	29.25	36.22	31.83	37.95
	(3.96)	(4.67)	(4.66)	(5.98)	(4.61)	(5.43)
Space	66.24	77.29	66.99	62.09	60.47	70.36
	(9.94)	(11.59)	(11.87)	(11.48)	(10.88)	(11.93)
P. Speed	15.91	17.00	6.38	6.77	6.92	9.11
	(3.46)	(3.73)	(2.39)	(2.53)	(2.06)	(2.68)
Number	85.58	93.95	97.68	107.75	68.88	72.27
	(10.67)	(11.73)	(14.93)	(16.50)	(11.48)	(12.16)
Verbal	27.55	27.82	12.58	14.01	30.38	25.42
	(3.59)	(3.48)	(2.08)	(2.41)	(4.79)	(4.01)

(Table 4 Continued)

Unstandardized Structural Regression Coefficients[b]

Factor	Regression of Posttest Upon Pretest		
	Controls	Induction	Space
Induction	.984 (.041)	.913 (.078)	1.072 (.036)
Space	1.046 (.040)	.920 (.048)	1.039 (.061)
P. Speed	.988 (.035)	1.046 (.044)	1.034 (.055)
Number	1.034 (.023)	1.045 (.023)	1.027 (.028)
Verbal	.997 (.025)	1.011 (.046)	.777 (.048)
Speed on Verbal	.070 (.026)	-	.079 (.027)
Induction on Number	.068 (.025)	-	-
Induction on P. Speed	-	.492 (.191)	-
Verbal on P. Speed	-	-	.244 (.103)

Note. Values in parentheses are standard errors.

[a]These values are from the accepted measurement model for the given training group.

[b]These values are from the accepted structural regression model for the given training group.

Table 5

Rescaled Solution for the Pretest-Posttest Measurement Model for the Induction Training Group

	Factor Loading[a]					Unique Variance		Unique
Variable	Induction	Space	Perc. Speed	Number	Verbal	Pretest	Posttest	Auto-Correl.
PMA Reasoning	.897					.190	.199	.242
ADEPT Letter Series	.897					.202	.190	.138
Word Series	1.031[b]					.146	.185	.208
Number Series	.773					.446	.362	.343
PMA Space		.825				.262	.373	.532
Object Rotation		.921				.137	.166	.264
Alphanumeric Rotation		.825				.234	.292	.415
Finding A's			.605			.643	.624	.726
Number Comparison			.512	.334		.407	.331	.534
Identical Pictures			.780			.368	.412	.558
PMA Number				.921		.145	.157	.411
Addition				.959		.085	.074	.411
Subtraction & Multiplication				.902		.194	.180	.812
PMA Verbal Meaning			.737		.273	.229	.189	.320
Vocabulary II					.880	.141	.288	.242
Vocabulary IV					.893	.240	.165	.783

(Table 5 Continues)

Ability Factor

	Pretest					Posttest				
Factor	Induction	Space	P. Speed	Number	Verbal	Induction	Space	P. Speed	Number	Verbal
Pretest										
Induction	-									
Space	.696	-								
P. Speed	.805	.768	-							
Number	.794	.574	.726	-						
Verbal	.560	.242	.405	.466	-					
Posttest										
Induction	.986	.715	.859	.768	.512	-				
Space	.654	.933	.770	.564	.157	.694	-			
P. Speed	.822	.747	1.007	.750	.417	.876	.790	-		
Number	.773	.550	.732	.995	.467	.760	.546	.772	-	
Verbal	.568	.295	.455	.476	.951	.536	.195	.459	.480	-

[a]Since the accepted measurement model included the factor loadings being set equal over time, this matrix was identical for both pretest and posttest with the exception of Word Series (see footnote b).

[b]At posttest, this value was .809.

Table 6

Rescaled Solution for the Pretest-Posttest Meaasurement Model for the Space Training Group

	Factor Loadings[a]					Unique Variance		Unique Auto-.
Variable	Induction	Space	Perc. Speed	Number	Verbal	Pretest	Posttest	Correl.
PMA Reasoning	.936					.138	.112	.126
ADEPT Letter Series	.873					.275	.203	.283
Word Series	.900					.200	.180	.514
Number Series	.716					.545	.427	.384
PMA Space		.829				.364	.263	.301
Object Rotation		1.013[b]				.117	.168	.442
Alphanumeric Rotation		.848				.321	.242	.338
Finding A's			.504			.782	.709	.726
Number Comparison			.514	.391		.323	.320	.617
Identical Pictures			.852			.313	.242	.668
PMA Number				.855		.247	.289	.564
Addition				.942		.129	.097	.327
Subtraction & Multiplication				.871		.256	.227	.506
PMA Verbal Meaning			.790		.145	.269	.196	.337
Vocabulary II					.911	.183	.155	.385
Vocabulary IV					.940	.085	.150	.980

(Table 6 Continues)

Ability Factor

	Pretest					Posttest				
Factor	Induction	Space	P. Speed	Number	Verbal	Induction	Space	P. Speed	Number	Verbal
Pretest										
Induction	–									
Space	.601	–								
P. Speed	.887	.680	–							
Number	.567	.298	.658	–						
Verbal	.583	.334	.490	.334	–					
Posttest										
Induction	.988	.585	.825	.592	.546	–				
Space	.633	.958	.715	.324	.362	.635	–			
P. Speed	.891	.636	.966	.630	.589	.865	.658	–		
Number	.574	.299	.665	1.001	.328	.607	.324	.650	–	
Verbal	.608	.341	.533	.361	.909	.573	.340	.609	.388	–

[a]Since the accepted measurement model included the factor loadings being set equal over time, this matrix was identical for both pretest and posttest with the exception of Object Rotation (see footnote b).

[b]At posttest, this value was .846

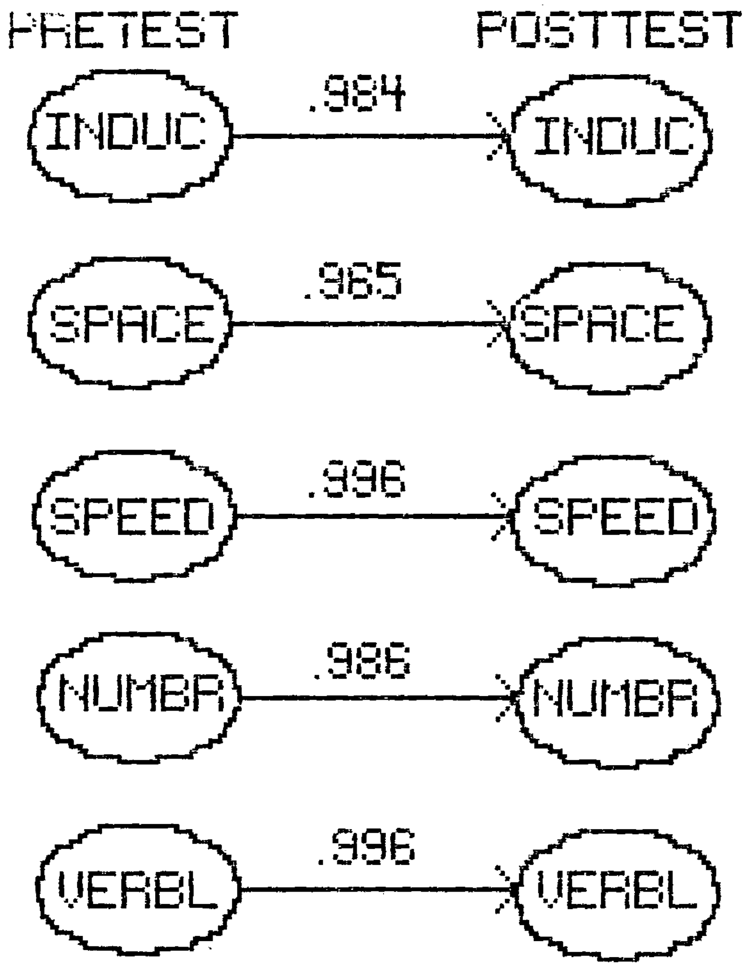
Figure Captions

Figure 1. Standardized stability and cross-lagged coefficients for the isolated stability and final models estimated for the control group.

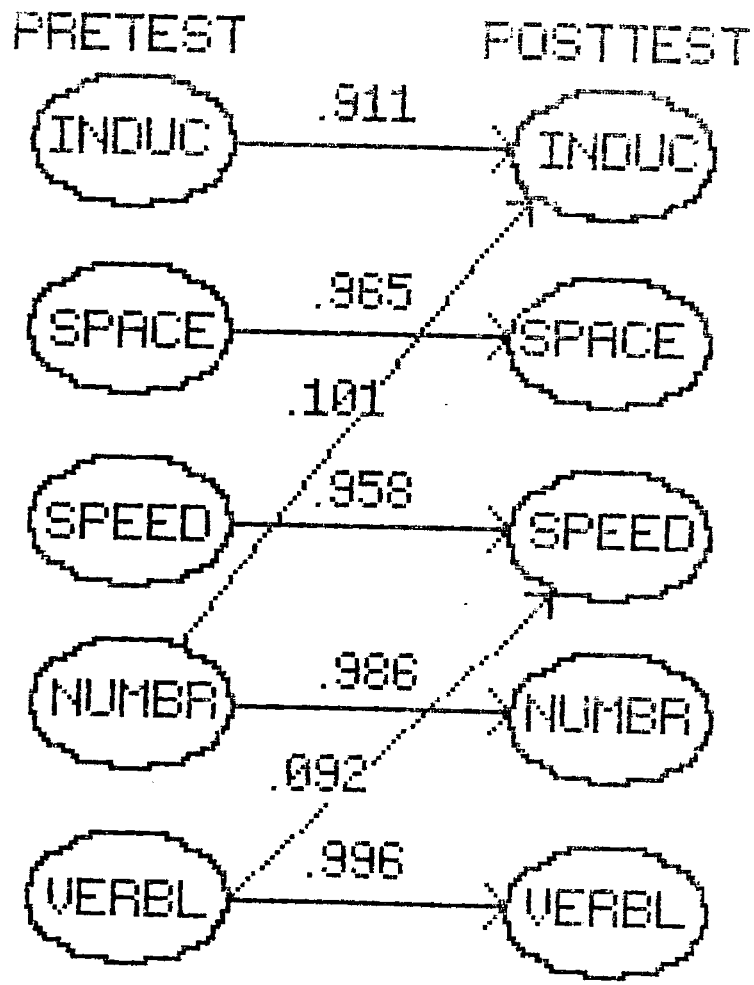
Figure 2. Standardized stability and cross-lagged coefficients for the isolated stability and final models estimated for the induction training group.

Figure 3. Standardized stability and cross-lagged coefficients for the isolated stability and final models estimated for the space training group.

Isolated Stability Model

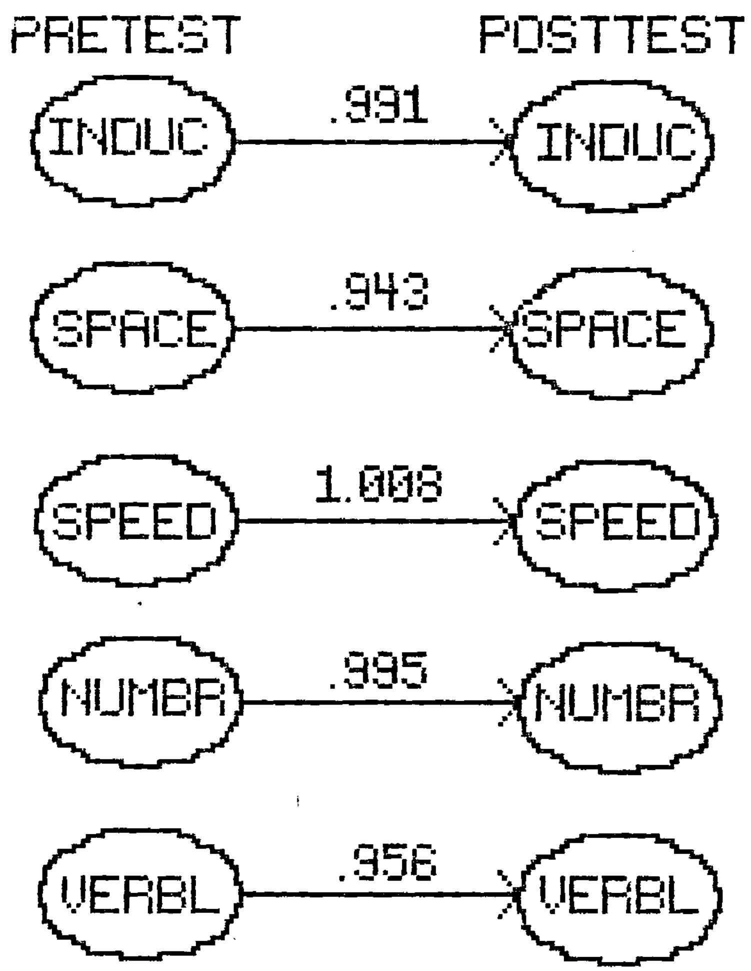


Final Model

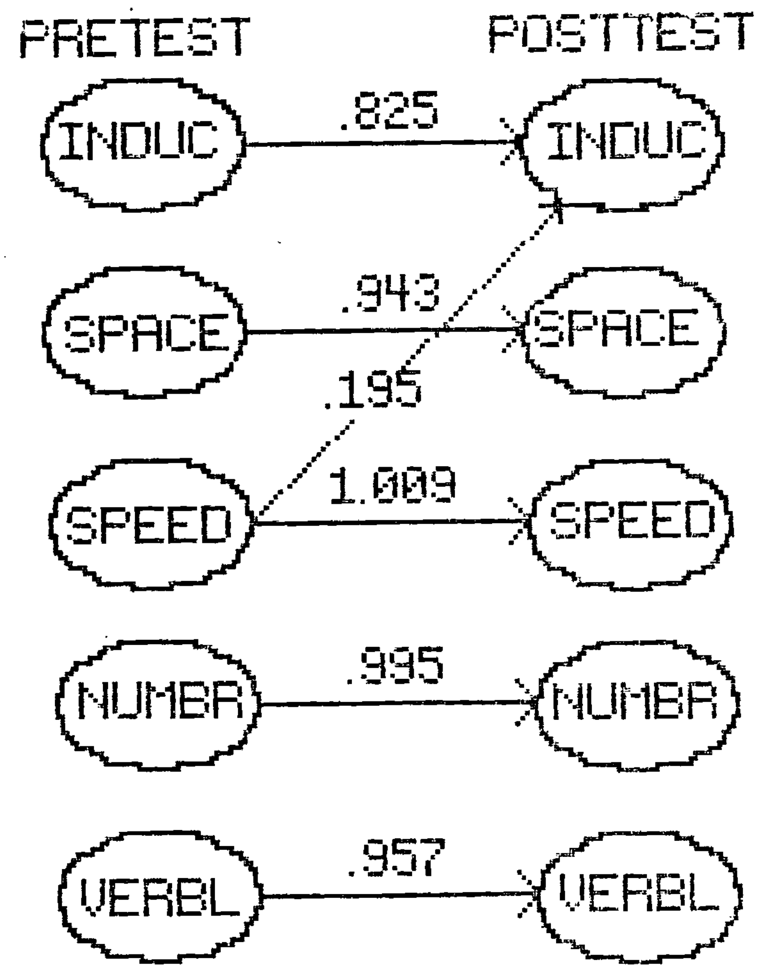


INDUCTION TRAINING GROUP

Isolated Stability Model



Final Model



SPACE TRAINING GROUP