Active

Project #: G-42-604 Cost share #: G-42-327 Rev #: 0
Center # : R6010-4A0 Center shr #: F6010-4A0 OCA file #:
Work type : RES
Contract#: 5 R01 AG06162-04 Mod #: INITIATION Document : GRANT
Prime #: Contract entity: GTRC

Subprojects ? : N
Main project #:

Project unit: PSYCH Unit code: 02.010.154
Project director(s):
HERTZOG C K PSYCH

Sponsor/division names: DHHS/PHS/NIH / NATL INSTITUTES OF HEALTH
Sponsor/division codes: 108 / 001

Award period: 871201 to 890229 (performance) 890228 (reports)

Sponsor amount	New this change	Total to date
Contract value	138,371.00	138,371.00
Funded	33,800.00	33,800.00
Cost sharing amount		7,000.00

Does subcontracting plan apply ?: N

Title: SHORT TERM CHANGE IN MEMORY/METAMEMORY IN THE ELDERLY

PROJECT ADMINISTRATION DATA

OCA contact: E. Faith Gleason 894-4820

Sponsor technical contact

ZAVEN KHACHATURIAN, PH.D., ASSOC DIR
NEUROSCIENCE & NEUROPSYCHOLOGY AGING
PROGRAM, NATIONAL INST. ON AGING
NATIONAL INSTITUTES OF HEALTH

Sponsor issuing office

MARIAN PARK
(301)496-1472
GRANTS MGT OFFICAL, NIA
BLDG.31, RM 5C-39
BETHESDA, MD 20892

Security class (U,C,S,TS) :
Defense priority rating : N/A
Equipment title vests with: Sponsor

ONR resident rep. is ACO (Y/N): N
supplemental sheet
GIT X

NO EQUIPMENT MAY BE PURCHASED DURING THE LAST 6 MONTHS OF THIS GRANT PERIOD.

* Administrative comments -
INITIATION. NIH IS OPERATING UNDER A CONTINUING RESOLUTION, INTERIM FUNDING FOR 3 MONTHS IS PROVIDED, PENDING ESTABLISHMENT OF FINAL FUNDING LEVELS.

GEORGIA INSTITUTE OF TECHNOLOGY
OFFICE OF CONTRACT ADMINISTRATION

NOTICE OF PROJECT CLOSEOUT

Closeout Notice Date 08/28/90

Project No. G-42-604 Center No. R6010-4A0

Project Director HERTZOG C K School/Lab PSYCHOLOGY

Sponsor DHHS/PHS/NIH/NATL INSTITUTES OF HEALTH

Contract/Grant No. 5 R01 AG06162-04 Contract Entity GTRC

Prime Contract No. _____

Title SHORT TERM CHANGE IN MEMORY/METAMEMORY IN THE ELDERLY

Effective Completion Date 890630 (Performance) 890930 (Reports)

Closeout Actions Required:	Y/N	Date Submitted
Final Invoice or Copy of Final Invoice	Y	_____
Final Report of Inventions and/or Subcontracts	Y	_____
Government Property Inventory & Related Certificate	N	_____
Classified Material Certificate	N	_____
Release and Assignment	N	_____
Other _____	N	_____

Comments _____

_____

Subproject Under Main Project No. _____

Continues Project No. G-42-625/R6010-3A0

Distribution Required:

Project Director	Y
Administrative Network Representative	Y
GTRI Accounting/Grants and Contracts	Y
Procurement/Supply Services	Y
Research Property Managment	Y
Research Security Services	N
Reports Coordinator (OCA)	Y
GTRC	Y
Project File	Y
Other _____	N
_____	N

NOTE: Final Patent Questionnaire sent to PDPI.

Final Report

Research grant from the National Institute on Aging, (R01-AG06162, "Short Term Change in Memory and Metamemory in the Elderly." Christopher Hertzog, Principal Investigator. 12/1/84 - 06/30/89.

I. Publications

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.

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

Hertzog, C. (1986). On pooling covariance matrices for multivariate analysis. Educational and Psychological Measurement, 46, 349-352.

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

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

Dixon, R. A., & Hertzog, C. (1988). A functional approach to memory and metamemory development in adulthood. In F. Weinert & M. Perlmutter (Eds.), Memory development: Universal changes and individual differences (pp. 293-330). Lawrence Erlbaum Associates.

Dixon, R. A., Hultsch, D. F., & Hertzog, C. (1988). The Metamemory in Adulthood (MIA) questionnaire. Psychopharmacology Bulletin, 24, 671-688.

Hultsch, D. F., Hertzog, C., Dixon, R. A., & Davidson, H. (1988). Memory self-knowledge and self-efficacy in the aged. In M. L. Howe & C. J. Brainerd (Eds.), Cognitive development in adulthood: Progress in cognitive development research (pp. 65-92). New York: Springer.

Dixon, R. A., Hultsch, D. F., & Hertzog, C. (1989). A manual of 25 three-tiered structurally equivalent texts for use in aging research. Technical Report No. 2, Collaborative Research Group on Cognitive Aging.

Hertzog, C. (1989). Using confirmatory factor analysis for scale development and validation. In M. P. Lawton & A. R. Herzog (Eds.), Special research methods for Gerontology (pp. 281-306). New York: Baywood Press.

Hertzog, C., Hultsch, D. F. & Dixon, R. A. (1989). Evidence for the convergent validity of two self-report metamemory questionnaires. Developmental Psychology, 25, 687-700.

Usala, P. D., & Hertzog, C. (1989). Measurement of affective states in adults: Evaluation of an adjective rating scale instrument. Research on Aging, 11, 403-426.

Hertzog, C., Dixon, R. A., & Hultsch, D. F. (1990). Relationships between metamemory, memory predictions, and memory task performance in adults. Psychology and Aging, 5, 215-227.

Hertzog, C., Van Alstine, J., Usala, P. D., Hultsch, D. F., & Dixon, R. A. (1990). Measurement properties of the Center for Epidemiological Studies Depression Scale (CES-D) in older populations. Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2, 64-72.

Hertzog, C., Dixon, R. A., & Hultsch, D. F. (in press). Metamemory in adulthood: Differentiating knowledge, belief, and behavior. In T. M. Hess (Ed.), Aging and Cognition: Knowledge Organization and Utilization. Amsterdam: North Holland.

Hertzog, C., Dixon, R. A., & Hultsch, D. F. (1990). Intraindividual change in text recall of the elderly. Unpublished Manuscript.

Usala, P. D., & Hertzog, C. (1990). Evidence for differential stability of state and trait anxiety in adults. Unpublished manuscript.

II. Papers Presented

Dixon, R. A., Hertzog, C., Schulenberg, J., & Hultsch, D. F. Adult age differences in the second-order factor structure of metamemory. Paper presented at the Eighth Biennial Meeting of the International Society for the Study of Behavioral Development, Tours, France, 1985.

Hertzog, C., Dixon, R. A., Schulenberg, J., & Hultsch, D. F. Adult age differences in the factor structure of metamemory. Paper presented at the 93rd Annual Convention of the American Psychological Association. Los Angeles, CA, 1985.

Dixon, R. A., Hultsch, D. F., & Hertzog, C. Twenty-five structurally equivalent texts for use in aging research. Paper presented at the 94ht Annual Convention of the American Psychological Association, Washington, D.C., 1986.

Hultsch, D. F., Hertzog, C., & Dixon, R. A. Memory self-knowledge and memory performance in the aged. Paper presented at the 9th Annual Convention of The American Psychological Association, Washington, D.C., 1986.

Dixon, R. A., Hultsch, D. F., Hertzog, C., & Comish, S. More on verbal ability and text structure effects on text recall in adulthood.

Paper presented at the First Annual Meeting of the Society for Cognitive Aging Research, Atlanta, GA, May 1987.

Hertzog, C., Hultsch, D. F., & Dixon, R. A. What do metamemory questionnaires measure? A construct validation study. Paper presented at the 95th Annual Convention of the American Psychological Association, New York, NY, 1987.

Usala, P., & Hertzog, C. Measurement properties of affective states in adult populations. Paper presented at the 49th Annual Meeting of the Gerontological Society, Washington, D.C., November 1987.

Van Alstine, J., & Hertzog, C. The validity of the mental health inventory as a measure of subjective well-being in the elderly. Paper presented at the 40th Annual Meeting of the Gerontological Society, Washington, D.C., November, 1987.

Hertzog, C., Dixon, R. A., & Hultsch, D. F. Relationships between metamemory and memory task performance in adults. Paper presented at the 41st Annual Meeting of the Gerontological Society of America, November 1988.

Hertzog, C., Mobley, M. I., Saylor, L. L., Hultsch, D. F., & Dixon, R. A. Stability of metamemory in adults: A longitudinal study. Paper presented at the 97th Annual Convention of the American Psychological Association, New Orleans, LA, August 1989.

Usala, P. D., & Hertzog, C. Evidence of differential stability of states and traits in adults. Paper presented at the 97th Annual Convention of the American Psychological Association, New Orleans, LA, August 1989.

III. Summary of Research Plan and Findings

General background

The general aim of the research grant was to conduct studies of the construct validity of two metamemory questionnaires in adult populations, while in the process relating multiple dimensions of metamemory to memory performance. Metamemory may be defined as knowledge, beliefs, and cognitions individuals have about memory. Metamemory is thought to play an important role in determining individuals' behavior in memory demanding situations, including laboratory memory studies. Gerontologists have studied metamemory in order to understand how knowledge of memory processes and strategies, beliefs about one's own ability to remember, and awareness of the status of information held in memory may contribute to effective use of memory by older adults.

At the time the grant proposal was written, there was a great deal of confusion in the literature regarding the nature of metamemory, whether there were age differences in

different kinds of metamemory, and the role metamemory played in the functioning of older persons. Much of the confusion seemed to involve disparate, and perhaps inadequate, methods for measuring multiple aspects of metamemory. The proposed research was designed to address the issue of the construct validity of several methods for measuring metamemory, while at the same time tackling important substantive questions regarding metamemory, memory, and aging. A number of specific research questions were enumerated in the original and competing continuation grant proposals, which can be most succinctly summarized in the context of reviewing major findings from the project.

Design and Samples

In order to achieve the specific aims, several data collections were undertaken:

A. 1985 Annville cross-sectional study. In 1985 447 adults, ages 20 to 78, were drawn from the Lebanon Valley, PA, recion, most of whom were members of a family health practice in Annville, PA. These individuals were administered a two-session battery that included (1) two metamemory questionnaires, (2) measures of mood states, (3) a depression screening measure, (4) measures of locus of control and self-efficacy, (5) a personality inventory, (6) a vocabulary test, and (7) a memory test, consisting of three trials of free recall of words and three trials of free recall of texts. About half the subjects received a version of the memory test obtaining performance predictions, a common method of measuring metamemory.

B. 1987 Annville sequential study. In 1987 we retested 240 of the original 1985 participants on the same set of measures, although there were two changes: (1) all participants made performance predictions, and (2) a new text was employed. We also added a new cross-sectional sample of 335 adults in the same 20-80 age range.

C. Cornwall P-technique study. In 1986 we began a small panel study, in which 7 elderly women from a retirement community participated weekly for a period of up to two years. The women filled out metamemory questionnaires, mood state scales, and the depression scale, and then performed on a computerized recognition memory task and a text memory task.

D. 1989 Atlanta study. In the summer of 1989 we tested 158 university students and 188 adults, ages 45-75, using metamemory questionnaires, measures of mood state, and free recall of words. The study featured a more elaborate performance prediction paradigm, replicating and extending the procedures used in the Annville studies.

Summary of Findings

One of the major questions was whether we could resolve the inconsistencies in the literature regarding age differences in metamemory. Our first publication (Hultsch, Hertzog, & Dixon, 1985) outlined our hypothesis that measures of metamemory that tapped beliefs about one's own memory ability, and changes in memory ability during adulthood, were most likely to show age differences. We argued that these measures tapped a construct of memory self-efficacy, and that increased age was associated with lower memory self-efficacy. Hertzog, Dixon, Schulenberg, & Hultsch, 1987) reanalyzed data collected by Dixon and Hultsch on one metamemory questionnaire: the Metamemory in Adulthood (MIA) instrument. Hertzog et al. (1987) used confirmatory factor analysis to show that there were higher order factors in the MIA, and that the strongest higher-order factor had loadings consistent with the argument that it was a memory self-efficacy factor.

Our first analysis of data from the Annville study was an examination of age differences in metamemory scales, using the MIA and the Memory Functioning Questionnaire (MFQ), another widely used metamemory questionnaire (Hultsch, Hertzog, & Dixon, 1987). The study also analyzed data from a sample of Victoria, British Columbia, residents, collected using the same design as the Annville study (data collection was funded by a Canadian research grant to D. F. Hultsch). Our principal finding was that age differences were only observed on measures thought to have the highest relationship to memory self-efficacy: MIA Capacity, MIA Change, and MIA Locus. The proportion of variance accounted for by age was modest, and differed across the two samples. Differences were largest when University of Victoria students were compared to a sample of Victoria adults (ages 55 - 78); however, we did observe significant cross-sectional age effects for these variables in the Annville sample as well. Polynomial regression analyses showed that these effects were linear across the 20 - 78 age range. However, we found no age differences within the adult Victoria sample across the 55 - 78 age range. Age differences on the MFQ were small and nonsignificant in the Annville sample, despite the fact that the main scale of the MFQ, the Frequency of Forgetting scale, appears to be highly similar to the MIA Capacity scale. The complex pattern of results confirmed our expectation that memory self-efficacy was the principal source of age differences in metamemory, but also showed that (1) age differences were most likely when young college students are compared to older adults, and (2) depend upon the type of questionnaire measure employed.

We then examined the convergent validity of the two metamemory questionnaires (Hertzog, Hultsch, & Dixon, 1989).

Our confirmatory factor analyses showed that the MIA and MFQ did have strong convergent validity in measuring memory self-efficacy, strategy use, and perceived change in memory self-efficacy. Memory self-efficacy factors from both scales correlated about .9 in old and young samples. We also found that the factor loadings were age-invariant, although there was a major age-related increase in the correlation between perceived change and memory self-efficacy. Moreover, we found that memory self-efficacy had a very low correlation with declarative knowledge about memory functioning (measured by the MIA Task scale) and low correlations with self-reported use of memory strategies in everyday life.

A set of analyses currently being written up for publication clarified the mystery of why the MIA and MFQ could show strong convergent validity and yet show divergent patterns of age differences. The MFQ measures of perceived problems with memory, emphasizing forgetting, correlate more highly with measures of depression, anxiety, and well-being (as well as the higher-order factor of psychological distress). In our cross-sectional samples, as with other community samples, older persons report lower levels of depression and negative affect than do younger adults (Hertzog, Van Alstine, Usala, Hultsch, & Dixon, 1990). So lower levels of self-reported negative affect in the old offset lower perceived self-efficacy in memory, eliminating age differences in self-reported memory problems. In order to conduct these analyses, we found it necessary to do extensive preliminary analyses of the item factor structure of the affective state, depression, and control measures, some of which we published because of their relevance to issues of measuring affect in elderly samples (Hertzog, Van Alstine, Usala, Hultsch, & Dixon, 1990; Usala & Hertzog, 1989). Hertzog (1989) also published a book chapter using these results to illustrate issues associated with comparative item factor analysis.

Hertzog, Dixon, & Hultsch (1990) reported results from the 1985 Annville sample relating metamemory questionnaires to memory task performance predictions. Predictions have traditionally been used as a measure of knowledge about memory functioning and awareness of the contents of one's own memory. Individuals higher in self-knowledge should, according to traditional approaches, have more accurate predictions of their memory performance. Our analyses showed that initial performance predictions are determined principally by memory self-efficacy, as measured by the MIA and MFQ, and that these predictions are actually not very accurate (correlations of the first prediction and subsequent recall performance are about .2 in the Hertzog et al. (1990) report and in the 1987 and 1989 studies). Our multiple trial paradigm then showed, consistent with other studies, that accuracy of predictions improved after task

experience, and we used structural equation models to show that this shift in correlations was best modeled as an effect of previous performance on the next prediction. We did find, however, a greater accuracy in prediction of initial text recall performance than for word recall performance, which might suggest that adults have more accurate beliefs about their text recall abilities. The results of this study supported the view of performance predictions as task-specific self-efficacy beliefs that are sensitive to age and determined initially by more general self-efficacy beliefs, as measured by metamemory questionnaires.

These memory self-efficacy beliefs are not necessarily accurate. There are major problems in establishing valid criteria for accuracy, of course, and it is by no means clear that the best benchmark for accuracy of memory self-efficacy beliefs are correlations of individual differences in beliefs with individual differences in laboratory memory tasks. Our data show that these correlations are typically relatively low in adult samples (usually between .2 and .3). Hertzog, Dixon, and Hultsch (in press) summarize these findings. In the context of our construct validity studies, one can argue that the questionnaires do validly measure memory self-efficacy beliefs and other beliefs about memory, but that these beliefs have relatively low predictive validity for performance on measures of free recall (words and texts). Task-specific performance predictions may have slightly better predictive validity, but these too are generally low unless and until individuals have direct experience with the task.

Our first analyses of the longitudinal data generated by the 1987 data collection were reported by Hertzog, Mobley, Saylor, Hultsch, and Dixon (1989) and by Usala & Hertzog (1990). We find that memory self-efficacy beliefs are highly stable over a two-year period, both in terms of mean levels and individual differences. Longitudinal correlations for the memory self-efficacy measures are as high or higher than the longitudinal correlations for free recall performance. However, not all self-report measures show this kind of stability in our sample. Usala and Hertzog (1990, unpublished) showed that the same subjects produce self-ratings of trait and state anxiety that have differential stability: trait anxiety self-ratings are as highly stable as are metamemory beliefs, but state anxiety measures show much lower stability of individual differences. It appears, then, that memory self-efficacy beliefs are consistent and enduring.

In a different vein, we are analyzing data from the Cornwall p-technique study. The first manuscript produced from this study examined intraindividual change and intraindividual variability in text recall performance

(Hertzog, Dixon, & Hultsch, 1990, unpublished). We found substantial amounts of variability in performance, which has important implications for assessment of older persons with text recall measures (as with the Wechsler Memory Scale's Logical Memory subtest).

In summary, the major finding of the present study is that memory self-efficacy beliefs can be reliably and validly measured by questionnaires, that these beliefs determine performance predictions, and that the beliefs are stable over long periods of time. Nevertheless, the beliefs are not necessarily accurate. Future research should be directed at understanding the positive and negative impact that memory beliefs have on memory-related behaviors, determining the extent to which beliefs can be (or should be) modified by intervention designs, and the degree to which suboptimal use of strategies in memory performance situations are determined by negative memory self-efficacy beliefs.

IV. Ongoing Analyses

Although this is a "final" report, analyses of the data produced by the project are ongoing. We have, despite hard work, only begun to mine the wealth of the existing data sets. For example, data from the 1989 study are being reported in a master's thesis by a Georgia Tech student, and will be combined with analyses of data from the 1987 study to report further results on the relations between memory self-efficacy and memory performance predictions. We are now conducting dynamic factor analyses of the mood variables from the Cornwall study, in order to relate intraindividual flux in mood to intraindividual variability in memory task performance. Future manuscripts will be provided to NIA as they are produced.

V. Appendix: Selected Publications

Experimental Aging Research, Volume 13, Number 2, 1987, ISSN 0734-0664

On the Differentiation of Memory Beliefs from Memory Knowledge: The Factor Structure of the Metamemory in Adulthood Scale[1]

CHRISTOPHER HERTZOG[2], ROGER A. DIXON[3], JOHN E. SCHULENBERG[4], AND DAVID F. HULTSCH[3]

This study was designed to test the hypothesis that there are multiple factors of metamemory present in the Metamemory in Adulthood (MIA) questionnaire. Data on seven MIA scales from six separate studies of memory/metamemory relationships (total $N = 750$) were combined to yield two half-samples for cross-validation purposes. The samples were partitioned into young, middle-aged, and old groups. A multiple group confirmatory factor analysis was then 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 involves beliefs about self-efficacy in using memory. The second factor combines knowledge about memory and affect concerning memory (e.g., achievement motivation). The analyses also indicated that the factor loadings for the second factor, tentatively labelled Knowledge, were invariant across the three age groups, but that there were age differences in the Self-Efficacy Beliefs factor loadings. The differences were localized to age-related increases in the loadings for the MIA Change and Locus scales. The two factor solution has potential for resolving conflicting results in the literature regarding age differences in both metamemory and metamemory/memory performance relationships.

The construct of metamemory—the knowledge and beliefs one has about memory functioning (one's own and others')—has been a focus of a growing body of research in cognitive gerontology [7; 12]. Although the nature of knowledge and beliefs regarding memory functioning are undoubtedly of interest in their own right, one important reason for the current emphasis on studying metamemory and aging is that metamemory may have an influence on memory-related behaviors of adults. Whether a person remembers an incident, a face, or a fact may be partly determined by what s/he knows or believes must be done to remember accurately. Moreover, a person's system of self-beliefs about memory, including whether s/he expects to be able to remember a given piece of information, may influence how s/he will behave in a given memory-demanding situation.

Psychologists interested in metamemory have conceptualized it primarily in terms of veridical knowledge—that is, the accuracy of one's knowledge about how one's own memory functions. Critical questions for studies deriving from this point of view have been whether and when an individual is capable of accurately monitoring and assessing his or her own memory processes. Developmental psychologists interested in metamemory have sought to identify the emergence of a child's awareness of memory functioning [3; 11]. The emphasis on accuracy of knowledge about memory has led to a set of operational definitions of metamemory that includes recall prediction accuracy, accuracy of memory monitoring, accuracy of information about memory strategies, and level of general knowledge regarding memory processes and functions (e.g., optimal mnemonic strategies) [6].

In addition to veridical knowledge about memory, other investigators, including ourselves, have examined a related but different aspect of metamemory: the subjective belief system of an individual regarding his or her own memory. We argue that, independent of both knowledge of how memory functions and concurrent awareness of memory functioning (as studied in memory monitoring paradigms, [6]), an individual also has a set of beliefs about his or her own memory capacities in some universe of environmental settings [5; 8; 9; 10]. These beliefs may be accurate or inaccurate, but irrespective of their validity, they may profoundly influence the cognitive behaviors and affective states of an individual. That is, beliefs about aspects of one's memory (e.g., skills, proclivities, reliability, weaknesses, developmental changes) may influence memory performance in the absence of veridical memory knowledge—or even, conceivably, when such knowledge is available but contrary to one's

[1]This project was supported by a research grant (7 R01 AG06162) from the National Institute on Aging. The first author was also supported by a Research Career Development Award (1 K04 AG00290) from the National Institute on Aging. Address correspondence to C. Hertzog at the School of Psychology, Georgia Institute of Technology, Atlanta, GA 30332-0170.

[2]From the School of Psychology, Georgia Institute of Technology, Atlanta, Georgia, U.S.A.

[3]From the Department of Psychology, University of Victoria, Victoria, British Columbia, CANADA.

[4]From the Department of Child Development and Family Studies, Purdue University.

TABLE 1

Descriptive Charactistics of Participants

Sample and Variable	Age Group			Total N
	Young	Middle-Age	Old	
EXPLORATORY SAMPLE				
Sample 1				
Sample Size	$n=60$		$n=60$	$N=120$
Mean Age	22.9		66.9	
Mean Years of Education	13.0		12.4	
Mean Vocabulary (ETS V-1)	14.1		16.2	
Sample 2				
Sample Size	$n=36$	$n=36$	$n=36$	$N=108$
Mean Age	32.3	48.5	68.5	
Mean Years of Education	13.7	13.6	13.3	
Mean Vocabulary (ETS Advanced)	10.0	11.3	10.3	
Sample 3				
Sample Size	$n=50$	$n=50$	$n=50$	$N=150$
Mean Age	32.0	49.5	68.9	
Mean Years of Education	13.6	12.8	10.9	
Mean Vocabulary (ETS Advanced)	8.2	10.2	8.1	
VALIDATION SAMPLE				
Sample 4				
Sample Size	$n=60$	$n=60$	$n=60$	$N=180$
Mean Age	23.10	44.17	66.83	
Mean Years of Education	13.82	12.45	12.72	
Mean Vocabulary (ETS V-1)	7.64	7.75	10.32	
Sample 5				
Sample Size	$n=24$	$n=24$	$n=24$	$N=72$
Mean Age	20.50	44.71	68.55	
Mean Years of Education	12.87	11.92	12.37	
Mean Vocabulary (ETS Advanced)	6.70	6.32	9.97	
Sample 6				
Sample Size	$n=60$		$n=60$	$N=120$
Mean Age	28.71		67.20	
Mean Years of Education	13.28		12.11	
Mean Vocabulary (ETS Advanced)	8.70		8.71	

beliefs. Whereas most research in metamemory and aging has addressed questions pertaining to the relationships between memory knowledge and memory performance, there is relatively little research pertaining to memory beliefs and their influences on memory performance [8].

Nevertheless, the distinction between these two aspects of metamemory may serve to explain some of the discrepant results in the literature on metamemory in adulthood. That is, it is at least evident that different operational definitions of the construct yield different patterns of adult age differences in metamemory [5; 8]. Age differences are often found in performance-based indicators of metamemory (e.g., spontaneous use of effective acquisition strategies) but not with other methods of assessing metamemory such as some tasks tapping memory monitoring or general knowledge, beliefs, or opinions about memory functions [4; 17; 18; 19].

Dixon and Hultsch [10] developed a metamemory questionnaire, the Metamemory in Adulthood (MIA) scale, for the purpose of measuring multiple, selected aspects of memory knowledge, memory beliefs, and memory-related affect in older populations. One motivation for development of the MIA was the need for psychometrically sound metamemory questionnaires [7; 12; 13]. The MIA has been shown to have good psychometric properties and to yield significant prediction of adult individual differences in text recall [9; 10].

A review of extant metamemory instruments suggested to us that there may be three higher-order metamemory factors operating to a greater or lesser degree in most metatmemory instruments: general memory knowledge, beliefs about memory self-efficacy, and memory-related affect [14]. To determine the theoretical merit and empirical utility of the delineation of metamemory into memory knowledge, memory beliefs, and

TABLE 2

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

Dimension	Description	Sample Item
1. Strategy	Knowledge of one's remembering abilities such that performance in given instances is potentially improved; reported use of mnemonics, strategies, and memory aids. (+ = high use)	Do you write appointments on a calendar to help 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	Rating of influence of anxiety and stress on performance. (+ = high anxiety)	I do not get flustered when I am put on the spot to remember new things.
7. Achievement	Perceived importance of having a good memory and 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)	It's up to me to keep my remembering abilties from deteriorating.

(Source: Dixon & Hultsch 1983 [10])

memory-related affect, two related steps are required. First, evidence concerning the validity of a differentiated representation of metamemory must be examined. Second, the differential prediction of memory performance by multiple dimensions of metamemory must be evaluated.

In the present study, we are concerned with the first of these two steps. Dixon and Hultsch [10] used item factor analysis to identify the eight MIA scales. In this paper we report the results of factor analyses designed to determine whether these MIA scales form higher-order metamemory factors that correspond to the knowledge, self-efficacy, and affect dimensions. In particular, our chief interest was to investigate whether a "subjective" memory beliefs factor could be differentiated from a memory knowledge factor. A secondary goal was to find out whether these factors had similar factor structures at different ages.

METHOD

Participants

This study is an analysis of data collected by Dixon and Hultsch as part of a series of studies on metamemory and memory performance. The recruitment procedures and population characteristics are described in greater detail in Dixon and Hultsch [10]. Briefly, participants were paid volunteers from a small city in central Pennsylvania. They were white, predominantly female, community dwelling adults who generally reported themselves to be in good to excellent health. The participants were drawn from six different studies. Their mean ages, educational status, and vocabulary scores are given in Table 1. In this analysis, the partipants were divided into two half-samples for cross-validation purposes – an exploratory half and a validation half (see Table 1).

Materials

We analyzed data from the Metamemory in Adulthood (MIA) scale, a self-report instrument using a modified Likert-type response format. In the MIA, individuals are asked to rate on a five-point scale statements either describing themselves and their memory or indicating knowledge of general memory processes. The MIA consists of eight subscales: (1) Strategy – use of memory strategies; (2) Task – knowledge of memory tasks and processes; (3) Capacity – assessment of one's own memory capacities; (4) Change – perceptions of change in one's own memory functioning; (5) Activity – activities supportive of memory; (6) Anxiety – state anxiety regarding memory performance; (7) Achievement – achievement motivation with respect to memory; and (8) Locus – perceptions of control in memory-demanding situations (see Table 2). Items for each of these factors were either adapted from previous metamemory instruments [19] or were developed by Dixon and Hultsch as part of the original instrument development study [10]. Additional information on the nature of the scales may be found there.

As our purpose was to analyze the factor structure of the MIA subscales, we restricted our interest to the covariance matrices among seven of the MIA subscales. Activity was dropped from the analysis because it is not a homogenous scale, as judged by low internal consistency reliability estimates [10]. Moreover, activity seems to measure behaviors that are determinants and outcomes of metamemory rather than metamemory *per se*.

Procedure

Given the relative lack of knowledge regarding metamemory factor structure, we opted to conduct exploratory analyses on the factor structure of the MIA. A point of departure was our hypothesis that there are three separate domains of Metamemory: Self-efficacy-beliefs, Knowledge, and Affect [14]. However, as the MIA was not developed for the purpose of measuring higher-order factors, it was an open question as to whether the MIA scales would factor in this way.

Therefore, factor analysis on the exploratory half-sample used a mixture of unrestricted and restricted factor analysis techniques. First, we conducted unrestricted maximum likelihood factor analyses that were used to assess the number of factors in each age group. Second, we used the LISREL VI program [16] to perform two sets of restricted factor analyses. The first set determined the optimal factor structure for the older adults, while the second set assessed the invariance of factor structure across the three age groups in a simulataneous factor analysis [15; 16]. Unstandardized (covariance metric) models were estimated, but final results were rescaled into a quasi-standardized metric recommended by Jöreskog [15]. This rescaling is not an option in LISREL VI, but was obtained by use of a SAS PROC MATRIX program written by the first author. Although some concern has been raised about this scaling in structural equation models [1], it is an appropriate scaling in multiple groups factor analysis when factor loadings are constrained equal [15]. We also calculated a relative goodness-of-fit index (GFI) that indicates how well a model fits a set of data, independent of the sample size. This index, analogous to the Bentler-Bonett normed fit index, was calculated from the LISREL fitting function of each model by $(F_o-F_k)/F_o$, where F_k is the fitting function for the model of interest and F_o is a null model of simultaneous no association in all age groups [2]. This index is a multiple groups generalization of the Bentler-Bonett normed fit index [2].

RESULTS

Exploratory Sample Analyses

The results of the preliminary factor analysis on the exploratory half-sample supported our hypothesis that the covariance structure of the MIA subscales could not be adequately fit by either one or two factors. Multiple factors were necessary to fit the covariance structure of all age groups. A plot of the eigenvalues suggested three factors in each age group, an impression reinforced by the χ^2 goodness of fit tests for the unrestricted solutions.

The simultaneous LISREL analysis of the three age groups from the exploratory sample specified a highly restricted factor model composed of the hypothesized three factors in the MIA: Self-Efficacy Beliefs, Knowledge, and Affect. Specifically, the Knowledge factor was defined by Strategy and Task, the Self-Efficacy Beliefs factor was formed by Capacity and Change, and the Affect factor was composed of Achievement, Locus, and Anxiety. This model was ultimately abandoned due to its inadequate fit to the data. In particular, the Affect factor was ill-defined. Subsequent models specified two factors, Knowledge and Self-Efficacy Beliefs, with an additional residual covariance between Achievement and Anxiety. In addition, the Locus, Achievement, and Anxiety scales appeared to relate to both the Knowledge and Self-Efficacy factors. A two factor model incorporating these modifications fit the data reasonably well ($\chi^2 = 62.58$, $df = 24$, $p < .001$, GFI = .924).

We then tested the equality of factor pattern matrices across the three age groups. There was a significant loss of fit in the model constraining the factor loadings for Knowledge and Self-Efficacy Beliefs to be equal over all three groups ($\chi^2 = 104.37$, $df = 40$; change in $\chi^2 = 41.79$, $df = 16$, $p < .001$). This outcome indicated that the hypothesis of equal factor loadings should be rejected. A third model showed this lack of fit to be specific to nonequivalent factor loadings on the Self-Efficacy Beliefs factor. A model specifying equal factor pattern weights on Knowledge, but not Self-Efficacy Beliefs, fit about as well as the model with no equality constraints ($\chi^2 = 73.34$, $df = 32$, GFI = .910; difference in $\chi^2 = 10.76$, $df = 8$, n.s.). A final model deleted two nonsignificant residual covariances in the old group that had been added in exploratory model building. Although this final model still had significant lack of fit ($\chi^2 = 75.72$, $df = 34$, $p < .001$), it was considered to be an adequate final model, in part becuase of its .907 GFI.

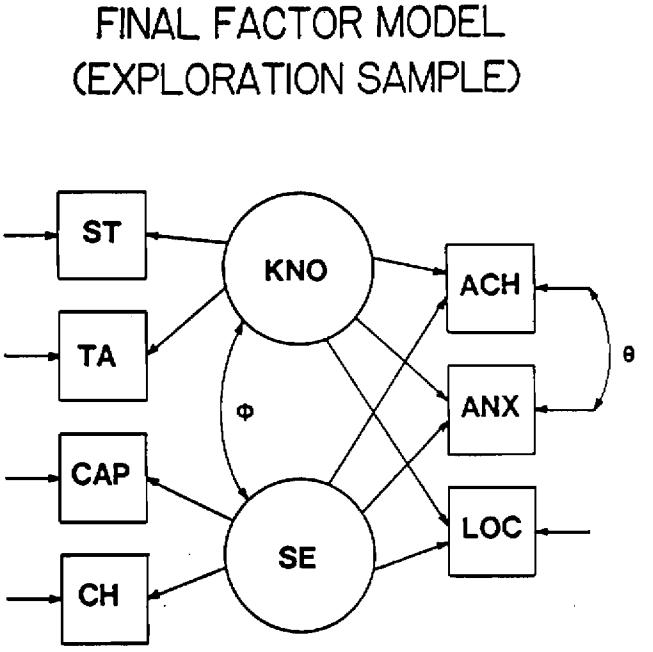


FIGURE 1. Final Model for metamemory factors in the exploratory half-sample. Straight arrows denote factor loadings, regressing variables (rectangles) on circles (factors); Is the factor covariance between Self-Efficacy Belief and Knowledge; is the residual covariance between Anxiety and Achievement.

Figure 1 shows the specification for the final exploratory model. Table 3 provides the rescaled results from each age group for the final model. Most of the age differences in factor loadings may be attributed to the loadings of Change and Locus on Self-Efficacy Beliefs, which were higher in the old group. The correlation of Knowledge and Self-Efficacy Beliefs was small and negative in the young and middle-aged groups, and larger and negative in the old group. This indicates that high levels of Knowledge were associated with lower levels of Self-Efficacy Beliefs.

Validation Sample Analysis

Although an identically specified model fit the validation half-sample relatively well ($\chi^2 = 73.73$, $df = 34$, $p < .001$, GFI = .886), shifts in relative magnitudes of certain loadings (especially Achievement on Self-Efficacy Beliefs and Knowledge) were found. Moreover, the residuals and LISREL modification in-

TABLE 3

Rescaled Two Factor Model (Exloratory Sample)

	Factor Loadings and Communalities		
	Knowledge	Self-Efficacy	H^{2*}
Old Group			
Strategy	.605	0	.290
Task	.499	0	.240
Capacity	0	.804	.520
Change	0	1.131	.981
Anxiety	.232	−.502	.423
Achievement	.608	.463	.324
Locus	.460	.914	.482
Middle-Aged Group			
Strategy	.605	0	.644
Task	.499	0	.320
Capacity	0	.804	.759
Change	0	.750	.529
Anxiety	.232	−.540	.460
Achievement	.608	.226	.427
Locus	.460	.416	.377
Young Group			
Strategy	.605	0	.314
Task	.499	0	.202
Capacity	0	.804	.748
Change	0	.384	.224
Anxiety	.232	−.454	.282
Achievement	.608	.438	.329
Locus	.460	.619	.497

Factor correlations: Old, −.581; Middle-Aged, −.174; Young, −.229.

Note: Rescaled estimates are as follows: factor loadings, on average, are standardized regression coefficients [15], communalities are proportions of common variance in each age group.

*Communalities

TABLE 4

Rescaled Two Factor Model (Validation Sample)

	Factor Loadings and Communalities		
	Knowledge	Self-Efficacy	H^{2*}
Old Group			
Strategy	.543	0	.264
Task	.343	.287	.276
Capacity	.308	.558	.410
Change	0	.739	.626
Anxiety	.548	−.723	.630
Achievement	.744	0	.591
Locus	.376	.667	.524
Middle-Aged Group			
Strategy	.543	0	.270
Task	.343	.164	.259
Capacity	.308	.558	.412
Change	0	.943	.644
Anxiety	.548	−.700	.680
Achievement	.744	0	.584
Locus	.376	.544	.368
Young Group			
Strategy	.543	0	.345
Task	.343	.273	.321
Capacity	.308	.558	.571
Change	0	.635	.418
Anxiety	.548	−.776	.730
Achievement	.744	0	.512
Locus	.376	.272	.394

Factor correlations: Old, .071; Middle-Aged, .107; Young, .373.

Note: Rescaled estimates are as follows: factor loadings, on average, are standardized regression coefficients [15], communalities are proportions of common variance in each age group.

*Communalities

dices indicated that the fit of the model could be improved. This outcome was perhaps unsurprising, given that the sample covariance matrices differed significantly between the exploratory and validation half-samples. An alternative model was developed in exploratory fashion. In the final version, Task loaded on Self-Efficacy Beliefs, Capacity loaded on Knowledge, and Achievement loaded only on Knowledge. The factor loadings for Knowledge were still constrained equal across the three age groups. This alternative model for the validation sample fit quite well ($\chi^2 = 38.76$, $df = 37$, n.s., GFI = .940).

Table 4 gives the rescaled solution for the validation half-sample. Although there are obvious similarities in the solutions, the differences in the factor loadings, as well as the different sign of the factor correlations, leads us to conclude that the model did not fully replicate across half-samples. In the validation analysis, nonsignificant *positive* correlations were found between Knowledge and Self-Efficacy Beliefs. Although this outcome buttresses the distinction between the two factors, it is inconsistent with the results from the exploratory half-sample.

DISCUSSION

Our contention that multiple factors could be differentiated from the seven MIA subscales was supported by this analysis. Most strongly supported was our hypothesis that a dimension of self-efficacy beliefs regarding memory functioning can be identified and differentiated from knowledge of how memory functions. In all three age groups we identified a Self-Efficacy Beliefs factor that had very small correlations with the Knowledge factor. Thus multiple factors do seem to be present in the MIA.

Although these findings are informative, they must be viewed as tentative given the differences between the exploratory and validation half-samples. In both analyses we found strong support for the distinction between Knowledge and Self-Efficacy factors in the MIA. However, there were differences in the factor correlation between these two half-samples that we cannot explain at this time. A possible explanation involves individual differences in the accuracy of Self-Efficacy Beliefs; i.e., the degree to which Knowledge and Beliefs correlate in subgroups of individuals. With these data, however, we cannot determine whether the six samples differ in variables that might predict degree of concordance between memory knowledge and memory beliefs.

The results of the analysis in both half-samples did not support our original hypothesis that there is a separate Affect dimension present in the MIA. This hypothesis seemed to be supported, to a small degree, by the residual covariance between Anxiety and Achievement in the exploratory sample analysis. However, this relationship was not replicated in the validation sample, where instead the Achievement and Anxiety measures had strong loadings on the factor labelled Knowledge.

This pattern in the validation half-sample suggests that our second factor may be an amalgam of Knowledge and Affect rather than a pure Knowledge factor. In both half-samples, Achievement and Anxiety had substantial loadings on this factor. However, a content analysis of Anxiety scale items suggests another alternative consistent with the interpretation of the factor as knowledge about memory. The Anxiety items appear to be clearly divisible into two sets: (1) questions regarding how anxious the respondent is in memory-demanding situations, and (2) knowledge about how high levels of anxiety inhibit memory functioning. A question for future research is whether the strong loadings of affect scales on the Knowledge factor are differentially a function of selected items arguably reflective of Knowledge regarding affect-cognition relationships. Nevertheless, we recognize that the interpretation of this factor as knowledge about memory may be contraindicated in subsequent work.

Tests of the equality of factor loadings for the two factors showed reliable age differences in the weights associated with the Self-Efficacy Beliefs dimension. Change and Locus relate more highly to this factor in the old than in the young. This qualitative difference in factor structure is intriguing and suggests some caution is in order when interpreting mean group differences on these scales. In particular, the substantial age differences in the loading of the Change scale suggest that there may be differences in the construct measured when young versus old persons are asked to judge whether their memory has improved (or declined) over a period of time (e.g., 10 years). Clearly, perceptions of change are more highly associated with anxiety about memory and perceptions of reduced control over memory in the elderly than in younger adults.

The distinction between Self-Efficacy Beliefs and Knowledge may help to explain some of the discrepant findings in the metamemory literature. Inconsistency across studies in the pattern of age differences in metatmemory measures may not reflect inconsistency in some global metamemory construct; instead, these differences may reflect differences in the degree to which different operational definitions of metamemory measure Knowledge, Self-Efficacy Beliefs, or some combination of the two. With references to the Dixon and Hultsch [10] data on age differences in the MIA scales, it appears that age differences in Self-Efficacy Beliefs may be more common (and more pronounced) than age differences in Knowledge. With the exception of Task, the MIA scales yielding significant age differences emphasize evaluation of self-efficacy beliefs (i.e., Capacity, Change, and Locus).

A possible explanation of these results goes as follows. Older persons are relatively familiar with how their own memory functions. Nevertheless, perhaps because they are (1) often less able to monitor specific on-line demands of memory tasks, (2) often unable to spontaneously produce and effectively utilize acquistion strategies in laboratory settings, and (3) sensitized to the potential negative implications of memory failures in real life, they may believe themselves to be less competent in memory than younger adults. This account argues, then, that older persons are on average similar to young persons in familiarity with how memory functions, but perceive themselves as being less competent in memory than younger adults.

The question arises as to whether these subjective perceptions of self-efficacy are, at the level of the individual person, accurate or inaccurate [5; 8; 20]. Given the modest correlations between metamemory and memory performance reported in the literature [6; 9], and the finding here that the correlation between Knowledge and Self-Efficacy Beliefs is relatively low in the older groups, we argue that these perceptions of low memory self-efficacy by older persons are not necessarily accurate. Negative self-efficacy beliefs may mediate phenomena such as older persons' frequent failure to employ mnemonic strategies even though they are aware that such strategies can improve performance.

Research exploring a model for the development of negative self-efficacy beliefs should be given high priority [5]. As a next step, however, replication of the differentiated factor structure of metamemory is necessary. Given the differences we encountered in cross-validation, it is evident that additional research is needed to demonstrate conclusively the differentiation of these two metamemory factors. We are currently analyzing data from a large scale validation study of the MIA and Gilewski and Zelinski's metamemory questionnaire [12] that should provide more definitive answers.

For now, there is sufficient empirical support for cautioning researchers on metamemory in adulthood to consider the existence of multiple dimensions of metamemory when interpreting or designing research. At least we should refrain from generalizing to a global metamemory construct on the basis of theoretically undifferentiated operational definitions. If further confirmation of the distinction between memory self-efficacy beliefs and memory knowledge is obtained, then studies simultaneously measuring these constructs and examining their differential relationships to both memory performance and memory-related behaviors will be of paramount importance.

REFERENCES

1. Acock, A.C., & Fuller, T.D. Standardized solutions using LISREL on multiple populations. *Sociological Methods and Research,* 1985, *13,* 551-557.

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

3. Brown, A.L. Knowing when, where, and how to remember: A problem of metacognition. In R. Glaser (Ed.), *Advances in Instructional Psychology.* Hillsdale, NJ: Lawrence Erlbaum Associates, 1978.

4. Bruce, P.R., Coyne, A.C., & Botwinick, J. Adult age differences in metamemory. *Journal of Gerontology,* 1982, *37,* 354-357.

5. Cavanaugh, J.C., Kramer, D.A., Sinnott, J.D., Camp, C.J., & Markley, R.P. On missing links and such: Interfaces between cognitive research and everyday problem solving. *Human Development,* 1985, *28,* 146-168.

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

7. Dixon, R.A. Questionnaire research in metamemory and aging: Issues of structure and function. In L.W. Poon, D.C. Rubin, & B.A. Wilson (Eds.), *Cognition in everyday life: Research approaches, aging effects, and enhancement methods,* in press.

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

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

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

11. Flavell, J.H., & Wellman, H.M. Metamemory. In R.V. Kail, Jr., & J.W. Hagen, (Eds.), *Perspectives on the development of memory and cognition*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1977.

12. Gilewski, M.J., & Zelinski, E.M. Assessment of memory complaints in the community-dwelling elderly. In L.W. Poon (Ed.), *Handbook of clinical memory assessment for older adults*. Washington, D.C.: American Psychological Association, 1987.

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

14. Hultsch, D.F., Dixon, R.A., & Hertzog, C. Memory perceptions and memory performance in adulthood and aging. *Canadian Journal of Gerontology*, 1985, *4*, 179-187.

15. Jöreskog, K.G. Confirmatory factor analysis in several populations. *Psychometrika*, 1971, *36*, 409-426.

16. Jöreskog, K.G., & Sorbom, D. *LISREL VI user's guide*. Mooresville, IN: Scientific Software, 1984.

17. Lachman, J.L., Lachman, R., & Thronesbery, C. Metamemory throughout the adult life span. *Developmental Psychology*, 1979, *15*, 543-551.

18. Murphy, M.D., Sanders, R.E., Gabriesheski, A.S., & Schmitt, F.A. Metamemory in the aged. *Journal of Gerontology*, 1981, *36*, 185-193.

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

20. West, R.L., Boatwright, L.K., & Schleser, R. The link between memory performance, self-assessment, and affective status. *Experimental Aging Research*, 1984, *10*, 197-200.

EDUCATIONAL and PSYCHOLOGICAL MEASUREMENT

A QUARTERLY JOURNAL DEVOTED TO THE DEVELOPMENT AND APPLICATION OF MEASURES OF INDIVIDUAL DIFFERENCES

(Continued on inside front cover)

VOLUME FORTY-SIX, NUMBER TWO, SUMMER 1986

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1986, 46

ON POOLING COVARIANCE MATRICES FOR MULTIVARIATE ANALYSIS[1]

CHRISTOPHER HERTZOG[2]
Department of Individual and Family Studies
The Pennsylvania State University

It is often useful to analyze a covariance matrix pooled over multiple groups. This note shows that the pooled covariance matrix cannot in general be calculated from the weighted average of the separate covariance matrices for the multiple groups. The correct equation for calculating the pooled covariance matrix adjusts for group differences in sample means.

THE covariance structures techniques developed by Jöreskog and others (e.g., Bentler and Weeks, 1979; Jöreskog, 1971; Jöreskog and Sörbom, 1979) have popularized simultaneous factor analysis in multiple populations. Such analyses require the use of covariance matrices rather than correlation matrices, given the expectation that, in general, only raw score (unstandardized) factor pattern matrices can be expected to replicate across populations (Meredith, 1964). Although these techniques require the analysis of separate sample covariance matrices for each population, there are occasions when the pooled covariance matrix taken across populations is of interest. For example, Jöreskog (1971) suggested using the pooled sample covariance matrix to test the viability of the common factor model prior to engaging in simultaneous analysis in the multiple groups. Another use of the pooled covariance matrix is when

[1] This work was supported by grants AG 04611 and AG 05165 from the National Institute on Aging.
[2] Address correspondence to the author at School of Psychology, Georgia Institute of Technology, Atlanta, GA 30332-0170.

simultaneous factor analysis has demonstrated that the factor pattern matrix may in fact be considered invariant across multiple populations. In such cases it may be argued that the populations were selected from a parent population for which the same common factor model holds (Meredith, 1964). Under this argument, it may be considered desirable to estimate the factor pattern parameters of the parent population in the pooled sample. Finally, multiple group analysis is often used for cross-validation of a factor model in multiple, random samples from a single population. If the factor model has been successfully cross-validated, it is appropriate to pool the sample data and re-estimate the model using all the data.

Usually the original raw data are available to the investigator, and the pooled covariance matrix can be calculated directly. However, when the analysis uses covariance matrices obtained indirectly from an archival source, then a pooled covariance matrix must be calculated from the separate covariance matrices of each group. This note shows that such pooling cannot in general be achieved by averaging the covariance matrices (even when such averaging takes into account unequal sample sizes). Instead, the pooled covariance matrix can only be obtained by taking into account group differences in the means of the variables.

The problem for pooling the covariance matrices is that each group's covariance matrix is based upon the cross-products corrected by the group mean, not the pooled (grand) mean. To explicate this point, let us assume that data have been collected on a set of variables $x = x_1, \ldots, x_p$ for $g = 1, \ldots, G$ separate groups. In the gth group, we define the pooled sample mean of the pth variable as

$$\bar{x}_p = \frac{1}{N} \sum N^{(g)} \bar{x}_p{}^{(g)}$$

where $\bar{x}_p{}^{(g)}$ is the sample mean in the gth group, $N^{(g)}$ is the sample size in the gth group, and N is the total sample size. The pooled sample covariance matrix, S, has as its $p \times q$th element the covariance of x_p and x_q:

$$S_{pq} = \frac{1}{N - 1} \sum (x_p - \bar{x}_p)(x_q - \bar{x}_q)$$

For ease of exposition, let us consider a single (generalized) element of S. The covariance matrix in each group, $S^{(g)}$, corrects each element of the covariance matrix to the corresponding group means:

$$S_{pq}{}^{(g)} = \frac{1}{N^{(g)} - 1} \sum (x_p - \bar{x}_p{}^{(g)})(x_q - \bar{x}_q{}^{(g)}) \qquad (1)$$

Since

$$(x_p - \bar{x}) = (x_p - \bar{x}^{(g)}) + (\bar{x}^{(g)} - \bar{x})$$

it can be seen that the cross-products for the pooled covariance matrix S, calculated in the gth group, are:

$$x_{p,q} = \sum^{N^{(g)}} [(x_p - \bar{x}_p^{(g)}) + (\bar{x}_p^{(g)} - \bar{x}_p)][(x_q - \bar{x}_q^{(g)}) + (\bar{x}_q^{(g)} - \bar{x}_q)] \tag{2}$$

Equation (2) clearly shows that the cross-products for the pooled covariance matrix differ from the cross-products for the group covariance matrix as a function of the difference between the sample group mean and the pooled sample mean.

The preceding expression can be used to derive the correction factor needed for estimating the pooled covariance matrix. Since

$$\sum (x_p - \bar{x}_p^{(g)})(\bar{x}_q^{(g)} - \bar{x}_q) = \sum (x_q - \bar{x}_q^{(g)})(\bar{x}_p^{(g)} - \bar{x}_p) = 0$$

expansion of Equation (2) yields the following equation for the adjusted cross-products in the gth group:

$$x_{p,q} = \sum (x_p - \bar{x}_p^{(g)})(x_q - \bar{x}_q^{(g)}) + N^{(g)}[(\bar{x}_p^{(g)} - \bar{x}_p)(\bar{x}_q^{(g)} - \bar{x}_q)] \tag{3}$$

It can be seen from this equation that the sample covariance element must be corrected to the deviation of the group mean from the pooled mean, weighted by the group sample size.

The pooled covariance element is obtained by summing the adjusted cross-products over all G groups and dividing by $N - 1$.

It is more convenient to express the correction formula for the entire matrix, S, in matrix algebra. The pooled covariance matrix is properly calculated as

$$S = \frac{1}{N-1} \sum^{G} [(N^{(g)} - 1)S^{(g)} + N^{(g)}[(\bar{x}^{(g)} - \bar{x})(\bar{x}^{(g)} - \bar{x})']] \tag{4}$$

where $\bar{x}^{(g)}$ is a $p \times 1$ column vector of sample means in the gth group and $\bar{x}$ is a $p \times 1$ vector of pooled means.[3]

As can be seen from Equation (4), the averaged covariance matrices will equal the pooled covariance matrix when there are no sample differences in means. Conversely, if there are large group differences in sample means then the difference between the aver-

[3] A sample listing of a SAS PROC MATRIX program implementing Equation (4) is available upon request.

aged covariance matrix and the pooled covariance matrix may become substantial. Two implications of the preceding are: (1) the pooled covariance matrix, if needed, is not properly calculated from just the weighted average of the sample covariance matrices; and (2) archival reports of group covariance matrices should also include the group means so that a pooled matrix could be calculated without recovering the original data. Obviously, one is well-advised to calculate a pooled covariance matrix directly from the raw data. Otherwise, the sample covariance matrices and sample means for the multiple groups must be input into Equation (4) to calculate the proper pooled covariance matrix indirectly.

REFERENCES

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

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

Jöreskog, K. G. and Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt Associates.

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

(Continued from front cover)

(Continued on inside back cover)

This journal is open to: (1) discussions of problems in the field of the measurement of individual differences, (2) reports of research on the development and use of tests and measurements in education, industry, and government, (3) descriptions of testing programs being used for various purposes, and (4) reports which are pertinent to the measurement field, such as suggestions of new types of items or improved methods of treating test data. Authors are granted permission to have reprints made of their own articles for their own use at their own expense. Manuscripts should be sent in duplicate to Dr. William B. Michael, Box 6856 College Station, Durham, North Carolina 27708. Authors are requested to put tables, footnotes, and abstracts on pages separate from the text and to follow the general directions given in the *Publication Manual of the American Psychological Association (1983 Revision)*. Journal titles should not be abbreviated.

Publication costs to the author are based on the number of pages, cuts, and special composition. Reprint purchase information will be mailed at the time of the printer's galleys.

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT (ISSN0013-1644) is published quarterly, one volume per calendar year, at 3121 Cheek Road, Durham, North Carolina 27704, Telephone (919) 688-3227. Second class postage paid at Durham, North Carolina and other cities. Postmaster send address changes to EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, Box 6856 College Station, Durham, N.C. 27708.

Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second class forwarding, postage. Other claims for undelivered copies must be made within four (4) months of publication.

Subscription rate, $50.00 a year, plus $2.00 postage to all foreign countries except Canada. Single copies, $12.50. Subscription includes microfiche copies which are mailed at the end of each calendar year. Back volumes: Volumes 5 to the present $50.00 each.

Certain issues of this journal are now out-of-print in the original form. These issues are: 5#2,4; 6#1,2,4; 7#1,2,3, parts 1 and 2; 8#3 parts 1 and 2; 9#4; 14#1; 16#1,3; 17#4; 21#1; and 23#3. Volumes 1 through 4 are available in a small-print edition at $12.50 per volume (paper bound).

Orders should be sent to EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, Box 6856, College Station, Durham, North Carolina 27708.

CANADIAN JOURNAL OF PSYCHOLOGY, 1987, 41(2), 193–208

Age Differences in Metamemory: Resolving the Inconsistencies*

David F. Hultsch, *University of Victoria*
Christopher Hertzog, *Georgia Institute of Technology*
Roger A. Dixon, *University of Victoria*

ABSTRACT Research examining adult age differences in individuals' self-reports about their memory has found somewhat inconsistent results. This paper reports findings from a cross-validation study of two questionnaires designed to measure knowledge, affects, and beliefs about memory in two large samples. The results suggest that there are significant age and sex differences in such perceptions. Compared with younger adults, older adults consistently report less memory capacity, more decline in memory functioning, and believe they have less control over their memory ability. Women report more strategy use and greater anxiety associated with memory than men. The results are discussed within the context of other methodological and substantive issues related to metamemory research.

RÉSUMÉ Cet article rapporte des résultats obtenus à partir d'une étude de validation croisée de deux questionnaires construits pour mesurer la connaissance, les affects et les croyances sur la mémoire chez deux grandes populations. Les résultats suggèrent qu'il existe des différences significatives quant au sexe et à l'âge pour de telles perceptions. Par comparaison aux adultes plus jeunes, les plus vieux rapportent de façon consistente une baisse de leur capacité mnésique, un déclin dans le fonctionnement de leur mémoire et ils croient à la perte de contrôle sur leur habileté mnésique. Les femmes rapportent plus d'utilisation de stratégies et plus d'anxiété associée à la mémoire que les hommes. Les résultats sont discutés dans le cadre d'autres questions méthodologiques et substantives reliées aux recherches sur la métamémoire.

The ways in which people exhibit understanding of their own functioning is a fundamental issue for several areas of psychology. In cognitive psychology, this issue is reflected in research on such "meta" behaviours as metamemory. Broadly defined, metamemory refers to individuals' understanding of their own memory functioning. As originally proposed by Flavell and his colleagues (Flavell, 1971; Flavell & Wellman, 1977), emphasis was placed on two classes of knowledge: (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 by researchers interested in aging to include individuals' beliefs about their own memory abilities, and affects and motivations that may be related to memory functioning (Dixon &

*This research 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 of Aging to Christopher Hertzog. Dr. Hertzog's work is also supported by a Research Career Development Award from the National Institute on Aging. Dr. Dixon's work on this research was supported by the Max Planck Institute for Human Development and Education, Berlin. The cooperation of Robert K. Nielsen, M.D., the other physicians, and the members of the Annville Family Practice, Annville, Pennsylvania is deeply appreciated. Address reprint requests to David F. Hultsch, Department of Psychology, University of Victoria, Victoria, BC, Canada V8W 2Y2.

Hertzog, in press; Hultsch, Dixon, & Hertzog, 1985; Zelinski, Gilewski, & Thompson, 1980). One central assumption underlying much of this work is that individuals' knowledge, beliefs, and affects about their own memory are important determinants of their behaviour in memory demanding situations. In the context of work on aging, it has also been assumed that such memory self-perceptions may become particularly salient in later life, contributing significantly to observed age-related declines in performance.

In part, these notions are derived from more general arguments which emphasize the importance of linkages among social, personality, and cognitive processes in life-span development (e.g., Baltes, Dittmann-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 may be shaped not only by their actual skills, but also by their understanding of the cognitive demand characteristics of the situation and their perception of the likely outcomes of their behaviours in such a situation. This perspective does not deny the occurrence of substantial age-related changes in underlying memory processes. Such changes undoubtedly exist. However, it does presume that observed age differences may be influenced by factors other than those defining memory ability per se. It argues that people's perceptions of their own memory may be important factors as well.

Although a variety of approaches have been employed to measure individuals' perceptions of their own memory, 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 also Dixon, in press; Gilewski & Zelinski, 1987, for reviews of available questionnaires). Initial work with these questionnaires suggests that there is evidence for their reliability and factorial validity. In addition, previous work has provided some indication of the presence of age-related differences on some dimensions of metamemory and a number of linkages between individuals' knowledge and beliefs about memory and their actual performance on memory tasks (Chaffin & Herrmann, 1983; Dixon & Hultsch, 1983a; Zelinski et al., 1980). However, it is equally clear that despite these preliminary positive results, several fundamental issues remain unresolved.

First, the robustness of age-related differences in self-perceptions of memory is somewhat unclear. A review of the available literature reveals a number of inconsistencies (e.g., Dixon, in press). 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 (e.g., Weinstein, Duffy, Underwood, MacDonald, & Gott, 1981). It is also unclear 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 than 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. The greatest consistency appears for indicators of perceived memory decline relative to previous levels of functioning. In this case, the findings suggest that older adults perceive their memory has declined more than younger adults (Dixon & Hultsch, 1983b; Gilewski et al., 1983; Perlmutter, 1978; Williams, Denney, & Schadler, 1983).

A second set of concerns relates to the definition and measurement of the metamemory construct itself. Despite the appearance of several instruments which have demonstrated acceptable levels of reliability and factorial validity, additional measurement development work is required.

First, additional evidence is required to demonstrate that memory self-perceptions can be differentiated from more general self-perceptions such as self-esteem, from affective states or traits such as depression, and from dispositions such as social desirability that may be evoked by the questionnaire formats typically used. For example, it has been suggested that older adults' complaints about their memory may be largely a function of depressive affect (Zarit, 1982). Indeed, a number of studies have found the correlation between memory complaints (as measured by a variety of self-report techniques) and depression to be higher than the correlation between memory complaints and actual performance on selected memory tasks (e.g., Kahn, Zarit, Hilbert, & Niederehe, 1975). The magnitude of the relationship between self-reports of poor memory and depression varies depending on the study, but correlations of .30 to .50 are fairly common (e.g., West, Boatwright, & Schleser, 1984). Such relationships do not necessarily call into question the discriminant validity of the self-report measures involved, although they do suggest that depression is an important variable determining individual differences in metamemory. Nevertheless, data sets examining the relationship of many other aspects of self-perceptions of memory to multiple domains of affect and general self-perceptions are not available. As a result, additional information is required before the discriminant validity of various metamemory questionnaires can be resolved.

The second measurement concern is convergent validity. It is now fairly clear that metamemory is a multidimensional construct. However, it is less clear how many dimensions are necessary to define it, what their relationships are, and whether the various questionnaires are tapping the same ones. For example, a recent analysis of data from 750 younger and older adults who had completed the MIA suggested that this instrument may contain at least two higher-order

dimensions that seem to reflect knowledge about memory and perceived self-efficacy with respect to memory (Hertzog, Dixon, Schulenberg, & Hultsch, in press). Similarly, after comparing specific items and the general content of several existing instruments, Dixon (in press) has suggested that some of these may be tapping similar dimensions even though the subscale names are different. In one preliminary empirical demonstration, Cavanaugh and Poon (1985) have presented evidence of substantial correlations in a small sample between the MIA and SIME. For example, for older adults they report a .72 correlation between the MIA Capacity subscale and the total score from the SIME. Additional evidence examining the number and relationship of the various components of metamemory is clearly required.

The third major issue concerns the relationship between measures of memory perception and measures of actual performance on memory tasks. The development of metamemory questionnaires was guided, in large part, by the assumption that this relationship would be relatively strong. However, as Herrmann (1982, 1984) and others have pointed out, evidence for such predictive validity is limited. The general pattern of results suggests that, when it does appear, the relationship between metamemory and performance is relatively weak, with correlations typically in the .20 to .30 range. Such findings have led some writers to reject the use of self-report questionnaires as substitutes for performance measures in clinical settings (e.g., Sunderland, Watts, Baddeley, & Harris, 1986).

However, two caveats should be kept in mind in evaluating the predictive validity of metamemory instruments. First, the relationship between metamemory and memory performance may be domain-specific. For example, several studies have found that various components of metamemory correlate significantly with text recall performance, but not with word recall performance (Sunderland et al., 1983; Zelinski et al., 1980). Other studies have predicted and found some specific patterns of metamemory-memory correlations (Dixon, Hertzog, & Hultsch, 1986). In addition, 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. Further, metamemory may be important for memory-related behaviours other than performance per se. 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. Such measures of memory-related behaviour are virtually unexamined.

Second, if we take a social-cognitive perspective, we should not really expect self-reports about memory to be veridical indicators of actual memory ability. It is likely that individuals differ considerably in the accuracy of their assessments of their memory abilities and the way they use them in various memory-demanding situations. The question is whether these individual differences in accuracy are systematic, and whether they relate to other behaviours relevant to memory-demanding situations. There is sufficient evidence to support Sunderland et al.'s (1986) rejection of metamemory questionnaires as substitutes for measures of memory performance. However, if older adults' perceptions of their memory

TABLE 1
Mean Education, Vocabulary, and Self-Rated Health for Victoria Sample

Age	Sex	*n*	Education	Vocabulary[a]	Health[b]
20–26	M	49	13.98	33.31	1.78
	F	47	14.15	32.94	2.02
55–61	M	26	14.08	43.00	1.96
	F	37	13.03	42.30	1.68
62–68	M	47	14.98	41.60	1.89
	F	54	12.94	41.20	1.87
69–78	M	50	14.40	44.08	2.04
	F	50	13.30	43.72	2.00

[a]ETS V-2(1), V-4(1), V-5(2), maximum score = 54.
[b]Two items rating health relative to perfect state and to others of the same age on a 5-point Likert scale (1 = very good).

prove to be one link in a process relating the social and cognitive domains, then the construct is of interest even if it is not a substitute for performance measures.

In sum, there are a number of crucial issues that require additional examination before we can determine whether measures of self-perceptions about memory will enhance our understanding of cognitive development in adulthood and old age. In part, what is required is a large-scale validation study that will provide data relevant to all of the issues outlined above within the same sample. This paper is one of several reports describing the results of such a study. Two large samples of subjects were tested in order to permit cross-validation of results. In addition, multiple indicators were used to index the variables of interest in order to permit the estimation of latent constructs. In the present paper, our focus is on the first issue raised earlier: Are there robust age-related differences in metamemory on individual questionnaires which have initial evidence for their reliability and validity? Subsequent reports will address the issues of discriminant, convergent, and predictive validity.

Method

Subjects: Two large samples of subjects were tested in order to permit cross-validation of the findings. The populations from which the samples were drawn differed in city size and nationality. One sample was drawn from a medium-size western Canadian city (Victoria, British Columbia), whereas the other sample was drawn from a semi-rural area in the eastern United States (Annville, Pennsylvania). In addition, the adult age range was represented somewhat differently in the two samples. The Victoria sample was designed to be comparable to the "traditional" cross-sectional sample used in cognitive aging research. In this sample, the younger adults were university students and the older adults were healthy, community-dwelling volunteers. The Annville sample was designed to represent the entire adult age range and consisted of younger, middle-aged, and older healthy, community-dwelling adults, none of whom was enrolled full-time in university at the time of testing. This sampling scheme permitted us to examine whether similar patterns of results occur when age comparisons are made using younger individuals differing in current educational activity.

TABLE 2
Mean Education, Vocabulary, and Self-Rated Health for Annville Sample

Age	Sex	*n*	Education	Vocabulary[a]	Health[b]
20–33	M	13	13.69	27.08	7.00
	F	17	14.06	30.71	7.59
34–40	M	30	15.30	34.47	7.23
	F	45	14.40	34.98	7.18
41–47	M	26	13.73	29.58	7.04
	F	41	13.07	28.89	7.29
48–54	M	18	13.67	31.11	6.94
	F	26	12.96	34.73	7.42
55–61	M	27	13.33	33.93	6.59
	F	46	12.63	32.61	7.02
62–68	M	37	13.65	34.30	6.84
	F	39	12.92	36.87	6.92
69–78	M	25	13.08	35.64	7.00
	F	35	13.17	38.14	6.74

[a]ETS V-2(1), V-4(1), V-5(2), maximum score = 54.
[b]Rated on the Duke "ladder" 9-point Likert Scale (9 = perfect health).

The Victoria sample consisted of 378 individuals. Complete data were available for 360 of these, and it is these data which are reported here. Analyses incorporating the subjects who had some missing data did not result in any alteration of the findings. Subjects were recruited through newspaper advertisements and appeals to university and community groups. They were paid a nominal fee of $15.00 (Cdn) for their participation. The subjects were divided into four age groups: 20-26 years, 55-61 years, 62-68 years, and 69-78 years. Males and females were distributed roughly evenly within each age group. Analysis of the demographic characteristics of the sample, presented in Table 1, suggests that the sample was well above average compared with the general population in terms of education, verbal ability, and self-rated health.

The Annville sample consisted of 447 individuals. Complete data were available for 415 of these, and it is these data that are reported here. Analyses incorporating the subjects who had some missing data did not result in any alteration of the findings. Subjects were recruited through their membership in a large family medical practice. They were paid a nominal fee of $15.00 (US) for their participation. Analysis of the demographic characteristics of the sample, presented in Table 2, suggests that the subjects were somewhat less selected in terms of education and verbal ability than the Victoria sample, but still above average compared with the general population. On average, subjects rated their health as good.

Measures and Procedures: Subjects from both samples completed a common battery of questionnaires and tasks including indicators of metamemory, social desirability, personal control (agency and causality), personality state (anxiety, arousal, fatigue, depression), verbal comprehension, and memory performance (free recall of texts and word lists). These instruments were administered to small groups of 5-15 subjects during two separate sessions. In the present report, we are concerned only with the two metamemory questionnaires which were given during the first session prior to any assessments of memory performance.

Metamemory in Adulthood (MIA) Instrument. This instrument, developed by Dixon and Hultsch (1983b, 1984), measures multiple dimensions of adults' self-perceptions of their

TABLE 3
The 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. Anxiety	Feelings of stress related to memory performance. (+ = high anxiety)	I do not get flustered when I am put on the spot to remember new things.
6. 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.
7. Locus	Perceived personal control over remembering abilities. (+ = internality)	Even if I work on it, my memory ability will go downhill.

Note. Based on Dixon & Hultsch, 1983b.

everyday memory functioning using a five-point Likert scale. We used the 108-item version of this instrument, dropping the original Activity subscale as suggested by Hultsch et al. (1985). The remaining seven subscales, their definitions, and sample items are presented in Table 3. Prior work with multiple samples has suggested that these subscales are internally consistent (Cronbach's alpha range across multiple samples = .61 to .92.) and factorially valid (e.g., Dixon & Hultsch, 1983b).

Memory Functioning Questionnaire (MFQ). This 64-item questionnaire, developed by Gilewski et al. (1983; see also Gilewski & Zelinski, 1987), taps multiple dimensions of metamemory using a seven-point Likert scale. This form is a shortened version of a 92-item instrument originally developed by Zelinski et al. (1980). The a priori subscales and sample items associated with them are shown in Table 4. In the original sample, Cronbach's alpha ranged from .82 to .93, and 3-year test-retest reliabilities of the subscales ranged from .22 to .64 (Zelinski et al., 1980).

Results

Analysis of the data from the two samples was conducted using different techniques in order to take advantage of the different age sampling strategies used in each instance. In the case of the Victoria sample, a traditional groups approach

TABLE 4
A Priori Subscales of the Memory Functioning Questionnaire (MFQ)

Subscale	Sample Item
1. General Rating	1. How would you rate your memory in terms of the kinds of problems you have? (+ = no problems)
2. Retrospective Functioning	2. How is your memory compared to what it was ... (a) one year ago? (+ = much better)
3. Frequency of Forgetting	3. How often do these present a memory problem for you ... (a) names? (+ = never)
4. 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)
5. 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)
6. Remembering Past Events	6. How well do you remember things which occurred ... (a) last month? (+ = very good)
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. (+ = not serious)
8. Mnemonics	8. How often do you use these techniques to remind yourself about things ... (a) keep an appointment book. (+ = never)

Note. Based on Gilewski et al., 1983.

was used. MANOVAs followed by univariate and post hoc tests were used to explore sex and age differences on both the MIA and the MFQ. In the Annville sample, observations were sampled across a continuum of chronological age. Consequently, polynomial regression analysis was conducted on the data to identify age-related trends. Order 4 polynomials (linear through quartic terms) were fit to the data in a multivariate sex by year of birth (YOB) analysis, using hierarchical significance tests to evaluate the significance of increments to R^2 by adding higher order terms (and their interactions with sex) as recommended by Cohen and Cohen (1983). The powers of YOB were taken after centering (subtracting the YOB mean) in order to reduce multicolinearity of the independent variables. Due to the large sample sizes involved, the 1% level of confidence was used as a criterion for all multivariate significance tests. Given the significance of a multivariate test, univariate tests for the various metamemory subscales were evaluated at the 5% level of confidence.

MIA Results: Victoria Sample. The sex by age (2 × 4) MANOVA conducted on the seven subscales of the MIA indicated significant overall effects related to sex, Wilks $\lambda = 0.942$, $F(7, 346) = 3.06$, $p < .004$, and age, Wilks $\lambda = 0.602$, $F(21, 994) = 9.16$, $p < .0001$, and a marginally significant interaction, Wilks $\lambda = 0.907$, $F(21, 994) = 1.63$, $p < .04$.

TABLE 5
Mean MIA Strategy, Anxiety, and Achievement Scores as a Function of Sex for Victoria Sample

Subscales	R^2	Pooled *SD*	Sex	
			Males	Females
Strategy	.035	9.64	62.50	66.20
Anxiety	.011	9.23	40.34	42.31
Achievement	.017	7.15	57.69	59.59

TABLE 6
Mean MIA Strategy, Capacity Change, and Locus Scores as a Function of Age for Victoria Sample

Subscales	R^2	Pooled *SD*	Age Group			
			20–26	55–61	62–68	69–78
Strategy	.043	9.64	61.31	67.21	65.03	63.85
Capacity	.114	9.93	60.02	53.42	52.65	51.84
Change	.274	12.42	63.82	53.15	50.98	49.41
Locus	.054	4.88	33.41	33.70	32.14	30.81

Univariate tests indicated significant sex differences on three subscales: Strategy, $F(1, 352) = 13.51, p < .0003$; Anxiety, $F(1, 352) = 4.11, p < .04$; and Achievement, $F(1, 352) = 6.10, p < .01$. Mean scores for males and females on these subscales are shown in Table 5, along with the pooled standard deviations for the sample and the estimated variance accounted for by each effect.

Univariate tests also indicated significant age differences on four subscales: Strategy, $F(3, 352) = 5.52, p < .001$; Capacity, $F(3, 352) = 15.66, p < .0001$; Change $F(3, 352) = 45.31, p < .0001$; and Locus, $F(3, 352) = 6.72, p < .0002$. Mean scores for the four age groups on these subscales are shown in Table 6, along with the pooled standard deviations for the sample and the estimated variance accounted for by each effect. Follow-up analyses using Bonferroni *t* tests at the 5% level indicated that the youngest group scored significantly higher on Capacity and Change than all three older groups. The youngest group also scored lower than the 55–61-year-old group on Strategy and higher than the 69–78-year-old group on Locus. Finally, the 55–61-year-old group scored higher on Locus than the 69–78-year-old group.

MFQ Results: Victoria Sample. The sex by age (2 × 4) MANOVA conducted on the eight subscales of the MFQ indicated significant overall effects related to sex, Wilks $\lambda = 0.935$, $F(8, 345) = 2.99$, $p < .003$, age, Wilks $\lambda = 0.703$, $F(24, 1001) = 5.37$, $p < .0001$, and a marginally significant interaction, Wilks $\lambda = 0.900$, $F(24, 1001) = 1.54$, $p < .05$.

Univariate tests indicated significant sex differences on two subscales: Remembering Past Events, $F(1, 352) = 4.41, p < .04$, and Mnemonics, $F(1, 352) = 8.60, p < .004$. Mean scores for males and females on these subscales are

TABLE 7
Mean MFQ Mnemonics and Remembering Past Events Scores as a Function of Sex for Victoria Sample

Subscale	R^2	Pooled *SD*	Sex	
			Males	Females
Mnemonics	.023	10.17	29.94	26.74
Remembering Past Events	.012	4.25	18.70	19.67

TABLE 8
Mean MFQ General Rating, Retrospective Functioning, Reading Novels, and Reading Magazines Scores as a Function of Age for Victoria Sample

Subscales	R^2	Pooled *SD*	Age Group			
			20–26	55–61	62–68	69–78
General Rating	.041	1.24	5.02	4.64	4.78	4.33
Retrospective Functioning	.211	5.84	24.27	18.25	19.02	17.65
Reading Novels	.032	5.47	28.06	27.04	26.46	25.45
Reading Magazines	.035	5.42	28.82	27.61	27.42	26.06

shown in Table 7, along with the pooled standard deviations for the sample and the estimated variance accounted for by each effect.

Univariate tests also indicated significant age differences on four subscales: General Rating, $F(3, 352) = 5.19$, $p < .002$; Retrospective Functioning, $F(3, 352) = 32.00$, $p < .0001$; Reading Novels, $F(3, 352) = 3.97$, $p < .008$; and Reading Magazines, $F(3, 352) = 4.39$, $p < .005$. Mean scores for the four age groups on these subscales are shown in Table 8, along with the pooled standard deviations for the sample and the estimated variance accounted for by each effect. Follow-up analyses using Bonferroni t tests at the 5% level indicated that the youngest group scored higher than all three older groups on Retrospective Functioning. The differences associated with the remaining three subscales were a function of significant differences between the youngest and the oldest groups.

MIA Results: Annville Sample. In the Annville sample, separate polynomial regression analyses were run for the MIA and MFQ. For the MIA, the multivariate tests revealed significant sex differences on the subscales, Wilks $\lambda = 0.903$, $F(7, 406) = 6.24$, $p < .0001$. The results of the polynomial analysis suggested a significant linear trend for year of birth, Wilks $\lambda = 0.836$, $F(7, 406) = 11.40$, $p < .0001$. No other trends reached significance at a 1% level of confidence. There were 5% trends for the quadratic and quartic components. Likewise, no sex by YOB polynomial interaction achieved significance at the 1% level of confidence.

Table 9 gives the univariate regression statistics for each MIA subscale. The significant sex effects were a function of univariate gender differences on the

TABLE 9
MIA Scales: Polynomial Regression (Hierarchical) for Annville Sample

Variable	Intercept a	Sex b_1(se)	YOB b_2(se)	F(model)	R^2
Strategy	59.18	3.98 (0.88)***	0.02 (0.03)	10.69***	.049
Task	59.85	0.47 (0.59)	0.03 (0.02)	1.40	.007
Capacity	52.56	−0.05 (0.98)	0.12 (0.04)***	5.56**	.026
Change	55.50	−2.35 (1.13)*	0.31 (0.04)***	30.78***	.130
Anxiety	36.94	4.11 (0.94)***	−0.01 (0.03)	9.67***	.045
Achievement	58.54	0.79 (0.70)	0.00 (0.03)	0.65	.003
Locus	32.38	−0.18 (0.50)	0.04 (0.02)*	2.78	.013

Model with Quadratic Trend Component

	a	b_1(se)	b_2(se)	b_3(se)	R^2	ΔR^2
Strategy	58.35	4.04 (0.88)***	0.01 (0.03)	0.004 (2.07)	.057	.008
Task	59.91	0.47 (0.59)	0.03 (0.02)	−0.000 (.001)	.007	.000
Capacity	52.05	−0.02 (0.98)	0.11 (0.04)**	0.002 (.003)	.029	.003
Change	55.41	−2.34 (1.12)*	0.31 (0.04)***	0.000 (.003)	.130	.000
Anxiety	36.07	4.17 (0.93)***	−0.01 (0.03)	0.000 (.002)	.052	.007
Achievement	57.80	0.84 (0.70)	−0.01 (0.02)	0.003 (.002)	.012	.008
Locus	32.73	−0.21 (0.50)	0.05 (0.02)*	−0.002 (.001)	.018	.005

*$p < .05$. **$p < .01$. ***$p < .001$.

Strategy and Anxiety subscales. Women reported higher levels of strategy use and greater anxiety about memory functioning. The age effects were associated primarily with the Capacity and Change subscales (with a nonsignificant trend on the Locus subscale). Older persons reported lower memory capacity and greater decline in memory functioning than younger persons. Table 9 also illustrates the validity of ignoring the 5% significance level quadratic term for YOB. None of the univariate regression weights associated with the quadratic term achieved statistical significance (even at the 5% level), and in no case did addition of the quadratic term increment R^2 by 1%.

MFQ Results: Annville Sample. For the MFQ, the multivariate tests revealed no significant sex or YOB effects, although there was a 5% level trend for the linear YOB component. For comparison purposes, however, the univariate effects for the MFQ are given in Table 10. There were trends for age differences on the Retrospective Functioning, Reading Novels, and Mnemonics subscales, but not one of these age trends was substantial. Moreover, on the important Frequency of Forgetting subscale, no hint of a significant age difference could be seen.

Finally, in order to permit comparison with the Victoria sample, the means and pooled standard deviations of selected subscales for the Annville sample are shown in Table 11. Included in the table are subscales that showed significant age differences in either one or both of the samples. The three oldest groups in Table 11 span age ranges equivalent to those reported in the Victoria analyses.

TABLE 10
MFQ Scales: Polynomial Regression (Hierarchical) for Annville Sample

Variable	Intercept a	Sex b_1(se)	YOB b_2(se)	F(Model)	R^2
General Rating	4.86	−0.18 (0.13)	0.02 (0.00)	1.57	.008
Retrospective Functioning	19.63	−0.12 (0.55)	0.04 (0.02)*	2.26	.011
Frequency of Forgetting	89.53	−2.32 (1.54)	0.03 (0.06)	1.26	.006
Reading Novels	26.85	−0.58 (0.60)	0.05 (0.02)*	2.61	.013
Reading Magazines	28.78	−1.19 (0.54)	0.03 (0.02)	3.74	.018
Remembering Past Events	17.63	0.37 (0.50)	−0.00 (0.02)	0.27	.001
Seriousness	66.87	−3.55 (2.54)	0.04 (0.09)	1.02	.005
Mnemonics	30.15	−1.42 (1.01)	−0.09 (0.04)*	4.18*	.020

*$p < .05$.

TABLE 11
Mean Scores of Selected MIA and MFQ Subscales as a Function of Age for Annville Sample

Subscales	Pooled *SD*	Age Group 20–33	34–40	41–47	48–54	55–61	62–68	69–78
Strategy	8.74	65.20	67.42	66.64	62.81	64.23	64.40	67.52
Capacity	9.68	56.88	53.37	54.04	52.18	51.58	50.90	50.61
Change	11.05	58.47	57.44	54.42	49.73	49.90	48.32	46.02
Locus	5.05	32.76	32.44	33.42	31.82	32.05	31.35	30.86
General Rating	1.27	4.98	4.64	4.57	4.30	4.53	4.61	4.57
Retrospective Functioning	5.49	21.45	20.12	18.49	19.91	19.39	19.05	18.70
Reading Novels	5.97	27.03	26.89	26.15	26.00	25.01	25.91	24.97
Reading Magazines	5.36	28.43	26.93	27.67	25.91	26.16	27.25	26.33
Mnemonics	9.81	26.46	24.13	27.57	30.06	30.17	29.14	27.39

Discussion

The present results paint a reasonably consistent picture that permits us to unravel some of the confusion present in the literature. In particular, consistencies between the two current samples and between the current samples and earlier studies suggest that there are age differences in adults' self-perceptions of their memory functioning. In addition, there appear to be sex differences in such self-perceptions. In considering these conclusions, it is important to attend to the size of the effects reported in the present data set, as well as their level of statistical significance. Because of the large sample sizes involved, a number of effects reach statistical significance even though they account for little variance.

Age Differences: There appear to be a number of congruous differences between younger and older adults on several metamemory dimensions. Specifically, compared with younger adults, older adults see themselves as having less memory capacity and report that their memory has declined over the years. Such differences, as indexed by the Capacity and Change subscales of the MIA, emerged in both the Victoria and Annville samples. In particular, the Change subscale accounted for substantial portions of the variance in both samples (Victoria 27%, Annville 13%). A similar pattern of age differences was observed in our earlier work (Dixon & Hultsch, 1983b).

In addition, there was a suggestion that, compared with younger adults, older adults believe that there is little that they can do to enhance their memory or prevent its deterioration. This difference, as indexed by the Locus subscale of the MIA, was significant in the Victoria sample, accounting for an estimated 5% of the variance, although it only approached significance in the Annville sample. Age differences between young and middle-aged to older adults were also observed on the Locus subscale in our earlier work (Dixon & Hultsch, 1983b).

The present results also show age differences on the MIA Strategy subscale in the Victoria sample and on the MFQ Mnemonics subscale in the Annville sample. These two subscales correlate highly (Victoria $-.76$, Annville $-.71$). In both cases, older adults report more use of strategies than younger adults. Our earlier work did not show such an age difference. However, the earlier factor analysis of the MIA suggests that the Strategy subscale may have two subcomponents: use of internal mnemonic devices such as imagery and use of external memory aids such as making lists (Dixon & Hultsch, 1983b). These analyses of the two types of items suggested that younger adults may be more likely to rely on mnemonic devices, whereas older adults may be more likely to rely on external aids. A majority of items on the MFQ Mnemonics subscale reflect the use of external aids. Further analyses are planned in order to investigate whether the presence or absence of age differences on strategy use measures, then, may reflect the mix of items referring to internal versus external memory strategies.

There are several qualifications that should be discussed with respect to these findings related to age differences in metamemory. First, 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. In the Victoria sample, age differences between the youngest group and the remaining groups were generally significant, whereas contrasts among the three older groups were generally not significant. This observation, however, should be qualified by the results of the hierarchical regression analysis of the Annville sample, which indicated linear trends consistent with the Victoria data. However, a similar analysis conducted on the data from the 55- to 78-year-olds from the Victoria sample did not show any significant linear trends. This suggests that age-related differences at the mean level within the middle to older age ranges are fairly fragile. As a result, discrepant findings in the literature may be partially due to the nature of the subjects sampled in this portion of the life span. With this in mind, we conducted a MANCOVA on the Victoria data using education,

vocabulary, and self-rated health scores as covariates. The results indicated that the sex and age effects observed in the earlier analyses remained virtually unchanged.

Second, it also appears that the MIA and the MFQ are differentially sensitive to detecting mean age differences. Such differences are more likely to be found with the MIA than with the MFQ. In the Victoria sample, significant age differences were observed on the General Rating, Retrospective Functioning, Reading Novels, and Reading Magazines subscales of the MFQ. However, the magnitude of the effects was generally smaller than those observed in the case of the MIA. In the case of the Annville data where a nonstudent sample was examined, only trends in the direction of age differences were observed on the MFQ. In contrast, as noted above, several significant age differences emerged on the MIA.

Related to this, the present results suggest that the phrasing of the questions may be of significance. No age differences were observed 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 (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 perceive this loss as a problem, either because their current level of functioning conforms to what they expect or because the incidents of forgetting are not perceived as particularly serious for their everyday functioning. Recent data reported by Sunderland et al. (1986) are consistent with this latter notion.

Sex Differences: There do appear to be sex differences in metamemory that are consistent across measures and samples, although they do not account for large amounts of the variance. Our prior work did not address the issue of sex differences because those samples were predominantly female. Similarly, previous work by other investigators does not suggest any consistent pattern of sex differences. In the present study, we found some evidence that women consistently reported more strategy use and greater anxiety associated with memory-demanding situations than men. Differences were observed on the MIA Strategy subscale for both samples and on the MFQ Mnemonics subscale in the Victoria sample. Differences were observed on the MIA Anxiety subscale in both samples. There were also significant sex differences on the MIA Achievement subscale and the MFQ Remembering Past Events subscale, with women scoring higher than men.

Conclusion: In sum, there appear to be sex- and age-related differences on some dimensions of metamemory. In particular, women show a small but consistent tendency to report more strategy use and greater anxiety about memory-demanding situations than men. Compared with younger adults, older adults

report less memory capacity, more decline in memory functioning, and believe that they have less control over their memory ability. The central features of these perceptions appear to cluster around perceptions of self-efficacy and changes in memory over time. There is little evidence for age-related differences in the knowledge aspects of metamemory that were originally of interest to developmentalists (Flavell, 1971; Perlmutter, 1978). The age differences that are observed are reasonably consistent, but particularly within the older age ranges they are somewhat subtle. Similarly, the extent to which age differences are observed appears to be partly a function of the particular way the questions are phrased. As a result, some of the inconsistency observed in previous studies may be related to sampling and measurement problems. Nevertheless, it appears that age-related differences in metamemory constitute a reliable phenomenon that is appropriate for further investigation. As mentioned at the outset, several avenues of research remain to be considered including issues of discriminant, convergent, and predictive validity. Analyses examining these issues within the current data sets are currently in progress.

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.

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 & Clinical Psychology*, **19**, 177–188.

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

Cavanaugh, J.C., & Poon, L.W. (1985, August). *Patterns of individual differences in secondary and tertiary memory performance*. Paper presented at the 93rd Annual Meeting of the American Psychological Association, Los Angeles, CA.

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

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

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, NJ: 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.

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, NJ: Lawrence Erlbaum.

Gilewski, M.J., & Zelinski, E. (1987). 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 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.

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

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., Dixon, R.A., Schulenberg, J., & Hultsch, D.F. (in press). On the differentiation of memory beliefs from memory knowledge: The factor structure of the Metamemory in Adulthood scale. *Experimental Aging Research*.

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., & 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, NJ: Lawrence Erlbaum.

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.

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

Sunderland, A., Harris, J.E., & Baddeley, A.D. (1983). Do laboratory tests predict everyday memory? A neuropsychological study. *Journal of Verbal Learning & 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.

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.

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.A., Denney, N.W., & Schadler, M. (1983). Elderly adults' perception of their own cognitive development during the adult years. *International Journal of Aging & Human Development*, **16**, 147–158.

Zarit, S.H. (1982). Affective correlates of self-reports 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 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 memorial conference* (pp. 519–544). Hillsdale, NJ: Lawrence Erlbaum.

CHAPTER
12

Using Confirmatory Factor Analysis for Scale Development and Validation

Christopher Hertzog

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 the measurement properties of psychological scales. One of the more important approaches has involved the use of confirmatory factor analysis to test properties of individual items and whole scales, examining factorial validity, reliability, and the like [1–4]. These methods are of special interest to gerontologists, because they explicitly provide a way of testing for age group equivalence in scale properties. The methodological foundations for using confirmatory factor analysis to examine measurement properties in scales, and to test age group equivalence in measurement properties, has been discussed in some detail by Schaie and Hertzog [5], and of course, in more technical material on the topic. This review is not designed to duplicate this material, but rather, to add to it by emphasizing practical issues associated with applications of the method. I avoid a mathematical treatment of the topic, and instead attempt to keep the focus at the level of discussing existing work that has used confirmatory factor analysis for scale validation. The topic is unfortunately somewhat complex, and we will wade into deep water occasionally. My hope is that the emphasis on discussing empirical applications will help the reader keep his or her head above the water line!

There are several different types of confirmatory factor analysis designs appropriate for scale development and validation. They fall into two very broad classes: 1) those that analyze the factor structure of sets of individual items; and 2) those that analyze sets of scales (often, summative scales of items using Likert-type ratings). Schaie and Hertzog discuss in some length the analysis of scales for the purpose of evaluating reliability and validity in gerontological research

pplications [5]. Here I shall focus more attention to the analysis of items, although here will be a brief discussion of the analysis of scales as well.

ITEM FACTOR ANALYSIS

GENERAL ISSUES

Confirmatory factor analysis has been employed with increasing frequency to erform item factor analysis of psychological scales, especially measures of subective well-being (life satisfaction, morale) in older populations [6, 7]. Reviews f the subjective well-being concept are beyond the scope of this review [7–9]. his literature can be somewhat overwhelming, in that complex models are resented with little justification or explanation. This chapter illustrates some of he basic features of confirmatory factor models for item factor structure with pecial reference to an analysis of a leading depression scale. With this discusion as background, I shall then review the literature analyzing scales of subjecve well-being in adult populations, focusing primarily on methodological issues.

The advantage of confirmatory factor analysis for analyzing scale properties that it is often the case that the items have been selected in advance to measure ypothesized dimensions. In some cases, it is assumed that all the items in a scale easure one latent variable. Cronbach refers to such scales as homogeneous scales, nd discusses the merits and disadvantages of scale heterogeneity [10]. The ssumption that all items are determined by a single latent variable (one, hopefully, at the scale is designed to measure) has been termed *unidimensionality* by cDonald [11]. It is a little appreciated fact that the calculation of Cronbach's , or of other internal consistency estimates of reliability, depends crucially upon e validity of the unidimensionality assumption. In fact, α is a lower bound stimate of the reliability, and a scale that is not unidimensional will have a reliabili- larger than that estimated by internal consistency methods. Thus, one reason perform item factor analysis on a scale is to test the assumption of idimensionality.

Related reasons for using confirmatory factor analysis for scale validation arise hen a scale is hypothesized to contain multiple subscales (i.e., multidimensionali-). In particular, one may be interested in whether as many factors are required account for item correlations as were originally hypothesized, and whether, some instances, it makes sense to collapse or combine subscales on the basis factor analysis results in order to achieve parsimony. In this case, it is useful specify an item factor model that tests whether the items actually factor as pothesized by the investigator. Depending upon the results of the analysis, it ay be judged necessary and/or appropriate to alter the number of scales calculated on the questionnaire. Confirmatory approaches are appropriate in this context cause the focus is on validating a model for item factors that has been specified *priori*. Thus it is possible to test the hypothesized item factor structure directly ithout resort to interpretation of exploratory factor analyses.

Use of exploratory factor analysis for assessing item factor structure is often appropriate and informative. It is *not* my purpose to castigate previous work on the grounds that exploratory factor analysis was used. Nevertheless, use of exploratory factor analysis can lead to unnecessary interpretive ambiguity, especially if the scales (and their corresponding item factors) correlate at moderate to high levels. Why is this a potential problem? On one extreme, most computer programs use an orthogonal rotation (usually, varimax) as a default option. Use of an orthogonal rotation when the item factors are truly correlated can distort factor pattern matrices and lead to erroneous conclusions about relations of items to factors. On the other hand, use of an oblique rotation to get correlated factors is not necessarily an adequate solution to the problem. There are, in exploratory factor analysis, an infinite number of rotational solutions. With oblique rotation, there are multiple, legitimate rotations that may vary dramatically in the degree of factor correlation permitted by the rotation algorithm. For example, the promax rotation constant controls the maximum degree of factor intercorrelation, and changes in the constant can result in dramatic changes in the estimated factor correlations. This fact should cause discomfort, for it is precisely these factor correlations that are the basis upon which one must decide whether multiple scales can be combined with minimal loss of information.

Even if the scales are expected to be orthogonal, there are distinct advantages of confirmatory factor analysis for an item set, including 1) ability to test the hypothesized configuration of item factor structure, and 2) direct testing of the hypothesis of orthogonality. Confirmatory factor analysis provides a formal basis for testing hypotheses, because it is possible to take the difference in the χ^2 goodness-of-fit test for competing models as a test of the restrictions contained in the more restricted model (for gerontological examples, see refs. [12–15]). Thus, if one wanted to test the hypothesis of orthogonal scales, is is a simple matter to specify two alternative models for a set of items. One model allows the item factors to intercorrelate freely. The second model forces the factor correlations to be fixed to zero. The difference in χ^2 between the two models is a test of the hypothesis that the factor correlations are, indeed, zero in the population. This sort of approach is general and quite powerful, permitting the use of clever psychometric designs to test a number of hypotheses about item and scale interrelationships.

The advantages of confirmatory factor analysis listed above relate to the *factorial validity* of a multidimensional scale. Factorial validity implies that the items form item factors as predicted by their hypothesized relation to an underlying psychological construct [10, 16]. Factorial validity is an important part of demonstrating the construct validity of a scale. The confirmatory factor analysis approach also enables the researcher to address other aspects of construct validity. With item factor analysis, this can be accomplished by modeling the relationships of the item factors to other variables. Latent variable models for *convergent validity* (do the multiple item factors interrelate, and do they converge with item

factors from other scales to measure the same construct), *discriminant validity* (are converged latent variables less than perfectly correlated with cognate constructs [17]), and *predictive validity* (does a laent variable predict other variables in a manner consistent with the assumption that it is a valid measure of the construct defined by theory) can easily be formulated provided that other latent variables have been measured as part of the design [3, 17]. Use of confirmatory modeling approaches to demonstrate evidence for factorial, convergent, discriminant, and predictive validity would in principle constitute compelling evidence for the construct validity of a scale.

AN EMPIRICAL EXAMPLE

As an illustration of item factor analysis, I use data collected by David Hultsch, Roger Dixon, and myself as part of a validation study of the Dixon and Hultsch [19] Metamemory in Adulthood questionnaire. The data were collected on two samples: 1) 437 adult volunteers, ages twenty to eighty from a family practice n Annville, Pennsylvania (hereafter, the Annville sample), and 2) 270 adult volunteers ages fifty five to seventy-seven from Victoria, B.C. (hereafter, the Victoria sample). Participants rated themselves on items from the Center for Epidemiological Studies Depression Scale (CES-D; [20]). The 20-item CES-D scale was specifically designed to measure degree of depressive symptoms in the population at large. Subjects are asked to respond how frequently during the last week a list of statements apply, using a 4-level rating scale (scored 0–3). Scores of 16 or higher are considered above the cutoff for mild depression [20]. The CES-D has become an increasingly popular measure of depression [21] and recently Gatz, Hurwicz, and Weicker reported large cross-sectional data on age and depression using the CES-D [22]. The original validation study by Radloff [20], and subsequent work by Aneshensel [23], suggest that there may be four factors contained in the CES-D: (depressive) Affect, (lack of) Well-Being, Somatic Symptoms (also labeled Psychomotor Retardation), and Interpersonal Problems.

The analysis summarized here was designed to test the four factor model using confirmatory factor analysis. Further information on the CES-D item analysis may be found in ref [24]. Before discussing the confirmatory factor analysis, it is instructive to ask what an exploratory factor analysis tell us about the item factor structure of the CES-D. For illustrative purposes, the Annville sample data were analyzed by the principal factor method, with squared multiple correlations on the diagonal as communality estimates. The least squares solution was then rotated by varimax to an orthogonal solution, and also by promax to two different oblique solutions (one with the rotation constant set at 3, the other at 10). Table 1 reports the factor pattern weights for all three solutions, and the factor correlation matrices for the two oblique rotations. Each column labeled R contains the varimax factor loadings estimated by Radloff [20]. The columns labeled contain the varimax rotation from the Annville data. The columns labeled P3

Table 1. Comparison of Alternative Exploratory Factor Models for the CES-D Factor Loadings[a]

	Factors							
	Affect				Well-Being			
Item	R	V	P3	P10	R	V	P3	P10
Bothered	23	11	−09	−25	−09	*−31*	−28	−28
Appetite	12	12	−01	−11	00	−17	−11	−06
Blues	*60*	*36*	27	26	−15	*−35*	−23	−20
Good	11	−05	16	*35*	*68*	*42*	*48*	*57*
Mind	24	21	08	−02	−10	−03	15	*37*
Depress	*64*	*55*	*53*	*62*	−18	*−44*	−28	−26
Effort	15	*30*	17	10	−07	−18	−01	14
Hopeful	−10	−21	−06	03	*68*	*59*	*66*	*82*
Failure	*44*	29	16	08	−28	*−38*	*−31*	−29
Fearful	*31*	*38*	*40*	*48*	−19	−05	17	*35*
Sleep	21	13	−04	−20	01	−12	00	14
Happy	*−38*	−25	00	18	*62*	*63*	*63*	*72*
Talk	00	23	16	15	−10	−22	−14	−11
Lonely	*72*	*50*	*49*	*57*	−06	*−38*	−22	−17
Unfriendly	15	11	01	−08	−07	−22	−20	−18
Enjoy	*−35*	−26	−04	12	*68*	*61*	*61*	*69*
Cry	*65*	*52*	*66*	*91*	−01	−12	08	17
Sad	*78*	*64*	*69*	*88*	−09	*−37*	−16	−09
Dislike	15	15	−01	−17	−04	−23	−16	−07
Getgoing	14	21	07	−03	−11	−12	02	17

	Factors							
	Somatic				Interpersonal			
Item	R	V	P3	P10	R	V	P3	P10
Bothered	*51*	*51*	*52*	*75*	10	01	−14	−27
Appetite	*50*	*44*	*46*	*66*	−13	01	−10	−21
Blues	*41*	*47*	*33*	*40*	13	13	−06	−15
Good	−01	−16	−04	−03	−11	−28	−22	−27
Mind	*59*	*48*	*48*	*66*	11	*33*	27	*31*
Depress	*43*	*41*	14	06	15	23	00	−07
Effort	*64*	*63*	*61*	*83*	06	14	−01	−11
Hopeful	−06	−15	08	18	01	−17	−02	00
Failure	07	28	09	04	11	*38*	26	32
Fearful	26	*33*	20	19	13	29	20	23
Sleep	*55*	*49*	*51*	*72*	−07	24	16	16
Happy	−25	*−40*	−21	−24	−05	−29	−11	−09
Talk	*54*	*38*	*33*	*43*	20	−01	−14	−26
Lonely	18	*34*	08	−03	09	29	10	09

Table 1. (continued)

	Factors							
	Somatic				Interpersonal			
em	R	V	P3	P10	R	V	P3	P10
nfriendly	07	-03	-19	*-34*	*84*	*47*	*48*	*67*
njoy	-14	*-33*	-11	-07	02	*-35*	-19	-21
ry	15	22	01	-12	-04	09	-07	-13
ad	20	*32*	01	-17	15	29	07	05
islike	08	18	05	00	*83*	*56*	*55*	*73*
etgoing	*66*	*57*	*59*	*83*	07	14	02	-05

[a]Decimals omitted.

Note: Comparison of Radloff [20] four-factor Varimax solution (Column R) and three ferent four factor solutions on Annville Validation study data: Varimax rotation (V), promax ation with constant at 3 (P3), and promax rotation with constant = 10 (P10). Loadings .3 are italicized.

l P10 contain the factor loadings for the two promax solutions for this data . As can be seen, the varimax-rotated factor loadings for the two samples are ilar. The difference in the varimax and promax solutions is predominantly he number of nonzero loadings in the varimax solution—a pattern that would expected if an orthogonal solution were inappropriately imposed on oblique tors.

lote the difference in the factor correlations estimated in the three solutions the Annville sample, as reported in Table 2. In the varimax rotation the factors relations are zero, by fiat. In the promax solution with default values of the stant set at 3, the correlations are substantial and predominantly in the .5 to nge. With the promax constant set at 10, the factors are highly oblique, with t of the correlations .7 or higher. The well-being items have not been reverse ed, so the negative correlation of Well-Being with the other factors is expected. that the semantically polar opposites, *happy* and *sad*, have very high loadings

Table 2. CES-D Item Analysis: Factor Correlations[a]

	Promax (constant = 3)					Promax (constant = 10)			
t	1				Affect	1			
being	-56	1			Well-being	-80	1		
atic	68	-66	1		Somatic	88	-83	1	
ersonal	42	-46	53	1	Interpersonal	75	-75	79	1

cimals omitted.

e: Factor correlations for two oblique rotations on the Annville Validation Study data set.

on Well-Being and Affect, respectively, but the factor correlations are not sufficiently close to −1.0 to conclude that Well-Being and Affect are opposite poles of the same dimension. While this result is evident in both promax solutions, the differences in the magnitudes of the factor correlations between the two solutions presents an important interpretive problem. Which one is "right?" How should the factor correlations be treated, given that their magnitude is dependent upon the constant used in the promax rotation? Given that there are an infinite number of possible rotated solutions, under what rotational transformations would we find the correlation of Affect and Well-Being sufficiently close to −1.0 to alter our conclusions in favor of considering the factors bipolar opposites? In sum, the dependence of estimated factor correlations upon the selection of a particular rotation constant renders a critical research question ambiguous and arbitrarily dependent upon methodological criteria.

In the confirmatory approach, there is no ambiguity about the factor correlation matrix. It has been uniquely identified by the specification of many non-zero factor loadings. In fact, the large number of fixed zero loadings overidentifies the model, and supplies surplus degrees of freedom for evaluating the model's goodness of fit. There is still a methodological and substantive issue, of course. It is whether the relationships of items to scale factors are indeed those specified in the model. In particular, one could be concerned about the accuracy of the assumptions of *lack* of relationships between items and factors—as represented in the fixed 0 factor loadings. A different specification of item-factor relationship would lead to different scale factors and different factor correlations, with the degree of variation depending upon the differences between specifications and their relative deviation from the "correct" model. However, in confirmatory analysis the model specification is clear, and the approach allows us to assess the adequacy of the model in terms of its fit to the sample data.

Table 3 gives the LISREL estimates of the regression coefficients (factor loadings) for a model postulating the isolated configuration of item factors suggested by the Radloff analysis [20]. The solutions for both the Annville and Victoria samples are reported. The model fares well in both samples, with significant factor pattern weights for all items. The correlational patterns are similar, although of smaller magnitude in the Victoria sample.

It is critically important to replicate results in multiple samples. Given that the four factor solution proposed by Radloff fare well in both the samples studied here [20], there is reason to have greater confidence in its validity. *Replication* should not, however, be confused with *confirmation*. Our model may be consistently misspecified, and the misspecified model may be replicable, even if incorrect. Confirmation is more critically important than replication, and is attained when 1) additional predictions of a theory lead to *new* predictions about the behavior of the factors identified by Radloff [20], and 2) these predictions are upheld by new, independent data (see refs. [12, 25, 26] for further discussion of this issue). Nevertheless, the ability to replicate results is critically important.

Table 3. LISREL Model of CES-D Items for Annville (AVS) and Victoria (VIC) Samples

	Factor Loadings[a]							
	Affect		*Well-being*		*Somatic*		*Interpersonal*	
	AVS	*VIC*	*AVS*	*VIC*	*AVS*	*VIC*	*AVS*	*VIC*
Bothered	0	0	0	0	56	54	0	0
Appetite	0	0	0	0	46	52	0	0
Blues	69	75	0	0	0	0	0	0
Good	0	0	47	35	0	0	0	0
Mind	0	0	0	0	56	63	0	0
Depress	85	80	0	0	0	0	0	0
Effort	0	0	0	0	76	82	0	0
Hopeful	0	0	64	54	0	0	0	0
Failure	65	69	0	0	0	0	0	0
Fearful	52	62	0	0	0	0	0	0
Sleep	0	0	0	0	55	46	0	0
Happy	0	0	87	88	0	0	0	0
Talk	0	0	0	0	48	43	0	0
Lonely	77	77	0	0	0	0	0	0
Unfriendly	0	0	0	0	0	0	52	55
Enjoy	0	0	85	77	0	0	0	0
Cry	52	52	0	0	0	0	0	0
Sad	83	85	0	0	0	0	0	0
Dislike	0	0	0	0	0	0	75	71
Getgoing	0	0	0	0	64	69	0	0

Factor Correlations[a]

	AVS					*VIC*			
Affect	1				Affect	1			
Well-being	−85	1			Well-being	−76	1		
Somatic	83	−73	1		Somatic	71	−55	1	
Interpersonal	65	−63	48	1	Interpersonal	54	−41	47	1

Goodness of fit:

AVS: $\chi^{2}_{(164)} = 343.84$ GFI = .93 AGFI = .90

VIC: $\chi^{2}_{(164)} = 280.79$ GFI = .91 AGFI = .86

[a]Decimal omitted.

Note: All 0 loadings and standardized factor variances were fixed by hypothesis. All nonzero parameter estimates were significantly different from 0 beyond the .1 percent level of confidence.

Replication is also crucial for lending credence to any modifications in the model that are based upon the original model's fit to sample data. In the present analysis we modified the model for the Annville sample, using LISREL's modification indices and other diagnostics to free several parameters that were fixed to 0 in the original model. These additional parameters reduced the χ^2 and increased the LISREL goodness-of-fit index. But the critical issue is deciding whether these model modifications are pointing to an improved model for the population, or alternatively, simply capitalizing upon chance to maximize fit to the particular sample. The best way to assure oneself of the broader applicability of the modified model is to cross-validate it in a separate sample. When we cross-validated these modifications in the Victoria sample, we found that the new factor loadings were not statistically different from zero. This result forced us to conclude that the modifications were merely improving fit to the Annville sample, not the general population. Given the importance of replication, it is always advisable either 1) to collect data on more than one sample, or 2) to collect enough data in a single sample to be able to randomly assign subjects to half-samples. In confirmatory factor analysis sample sizes of 200 or greater are preferred, so an overall sample size of 400 or more is optimal [27].

Another issue illustrated in the analysis of the CES-D involves the proper interpretation of a model's fit to sample data. A model can be accepted as useful even if its likelihood ratio χ^2 test is statistically significant. A significant χ^2 implies that we reject the model as fully adequate in accounting for the sample correlation matrix. It is often (indeed, usually) the case that this test statistic will be significant in samples of moderate size. With sample size of 250 or greater, the likelihood ratio test is very powerful and will be significant even when a model is *stable* (in the sense that parameter estimates do not change greatly if new parameters are added to the model) and when it accounts for most of the information in the correlation matrix. Given the power of the likelihood ratio test with large samples, it is a good idea to calculate an index of fit that is independent of sample size. Bentler and Bonett [28] and James, Mulaik, and Brett [29] describe some alternative fit indices. The LISREL program provides its own, alternative, relative fit indices. For the CES-D item analyses, the LISREL goodness-of-fit index in both samples is around .9, indicating a fairly satisfactory level of fit.

The fit of the CES-D model is actually quite good for an item factor analysis, for two different reasons. First, the assumptions of multivariate normality are violated in data consisting of ordinal rating scales. Huba and Harlow showed that this violation did not greatly affect the LISREL parameter estimates, but it did inflate the standard errors and the likelihood ratio χ^2 fit test [30]. The second reason to be satisfied with the fit of the model concerns the nature of inter-item relationships. It is often the case in item analysis that individual items may correlate with each other to a degree not fully accounted for by the item factors. Two items may have very similar wordings, and therefore have a residual relationship with each other even if both load on the same item factor. Such specific

elationships will cause lack of fit to the factor model, but such deviations may ›e relatively trivial, so long as the estimates of the factor loadings and factor ·orrelations have stabilized and are replicable. In our CES-D analysis, the primary ›arameter estimates did not shift greatly when additional modifications were made o the model. It seems safe to conclude that Radloff's model for the CES-D is ı reasonable representation of the underlying factor structure [20].

The principal pragmatic question after the confirmatory factor analysis is what ı implies for calculating CES-D scale scores. Is it appropriate to combine all wenty items into a single scale score measuring depression? Is one better off :stimating scale scores for each of the four factors separately? These issues can ›nly be summarized briefly here (see ref. [21] for a fuller discussion). A deci-ion to use the Radloff model as a basis for calculating four subscales of the CES-D s predicated upon the validity of the four factor model [20]. Clearly, it has sur-/ived the confirmation test, and can be accepted as a useful model for the item ·actors. This finding lends credence to the four factor representation as a possi-›le basis for creating subscales for the CES-D. However, there may be practical ›roblems with calculating four separate subscales. A decision to use the four scales nust face two issues: 1) the scales will be highly intercorrelated, and 2) two of .he four scales are based upon an unacceptably low number of items, minimizing ·cale reliability. The fourth factor, Interpersonal Problems, is defined by two ıtems—too few to be considered an adequate basis for a separate subscale. The Well-Being factor is defined by four items, and similar concerns apply.

Gatz et al. used unit weighting to combine the items using the assignments im-›lied by the Radloff model [22]. Their internal consistency estimates of reliabili-ıy were relatively low for the Interpersonal Problems scale. On the other hand, Gatz et al. did find quite different mean age differences on the four subscales, with larger age differences in Well-Being than on the other CES-D subscales [23]. Their results may indicate that the Well-Being scale is measuring something dif-ferent than the other CES-D subscales, a notion consistent with the factor cor-relations estimated in the Hertzog et al. analysis [24]. Certainly, it does not ap-pear that Well-Being and Depressive Affect are polar opposites, even though they have a strong negative correlation. This finding is consistent with studies on psychological well-being and distress [31, 32] suggesting that Distress (a factor marked by depressive affect and other indicators) and Well-Being should be con-sidered independent, negatively corrrelated dimensions. Apparently there is merit in separating positive and negative affect in well-being measures [8, 9]. Note also that, unlike the exploratory factor analyses, both confirmatory analyses sug-gest that the fourth factor, Interpersonal Problems, has lower correlations with the other three factors. The four factors are substantially correlated, with the lowest correlation beween Interpersonal Problems and the other three scales.

What case can be made for the single CES-D scale score? The relatively high intercorrelations among the scales indicate that there is justification for summing all items into a single CES-D depression scale score. More definitive justification

for the single scale may be assessed by performing second-order factor analysis on the four first-order item factors [6, 33, 34]. Figure 1 shows the Liang model for a second-order factor in a measure of life-satisfaction [33] (see below). The critical concept is that the second-order factor determines the first-order factors.

Let us assume for the moment that the primary issue driving the research is whether subscale scores should be combined into a single scale (given the high

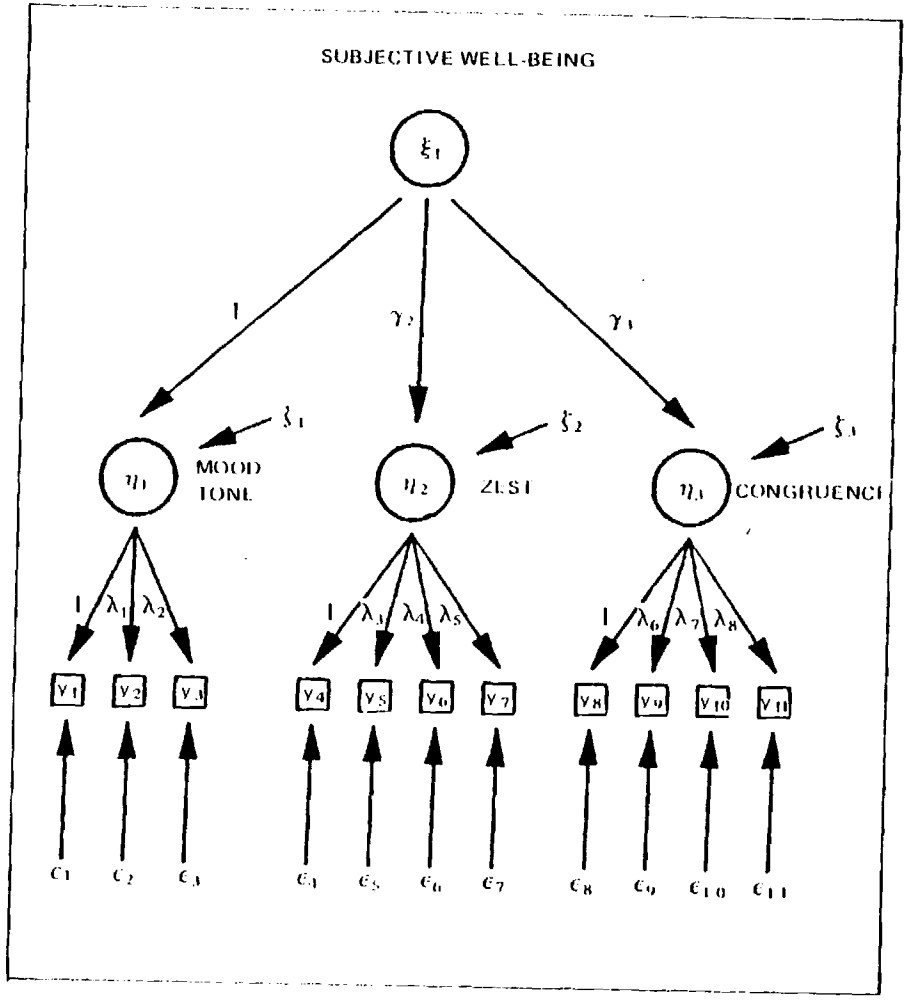


Figure 1. Second-order factor model of Liang [33] relating Subjective Well-Being to three first-order factors (Mood Tone, Zest, Congruence). Items are represented by squares, first-order and second-order factors by circles.

ubscale correlations). The second-order factor analysis enables a full evaluation of this issue. There are three separate questions to be addressed. The first question is: are the first-order factors properly specified? This issue has already been addressed for the CES-D data set. The second question is: how well does the second-order factor model account for the first-order factors? There are two different aspects of this question to be considered. First, can the first-order factors be represented by some (possibly multidimensional) second-order factor model? Second, how much of the variance in each first-order factor is determined by the second-order factors? The first question is essentially asking whether it is plausible to model higher order factors, and if so, whether the second-order model specified fits the data well. This issue of *fit* is conceptually distinct from the issue of how much of the variance in the first-order factors is predicted by the second-order factors. *It is possible for a second-order factor to fit well but still account for relatively little variance in the first-order factors.*

The third question to be addressed is: do the first-order factors covary with other important constructs independent of the second-order factor? This question is, in a sense, the most critical one to be addressed in deciding whether to combine the scales, yet it is one generally neglected in the emergent literature on second-order factors. Of course, if the second-order factor accounts for all the variance in the first-order factor, the third question is moot. That outcome would indicate that there is no residual component at the first-order level left to correlate with other variables. However, if there is substantial residual variance at the first-order level, then it is possible (and perhaps likely) that some variables of interest may covary with the first-order factor independent of the second-order factor. If these variables are of specific interest in a research problem, combining the subscales into a single overall scale would mask this relationship.

Hertzog et al. ran a second-order factor analysis of the CES-D data, simultaneously estimating the first and second-order solution in LISREL [24]. The model specified a single second-order Depression factor, and restricted the residual covariance matrix to be diagonal (modeling the second-order factor as the sole determinant of covariance among the first order factors). Table 4 reports the second-order factor loadings and residual variances. The second-order loadings were substantial, and significant for all four first-order factors. In the Annville data, Depression accounted for 70 percent or more of the variance in three of the four factors. The pattern was similar in the Victoria study, although less variance was determined by Depression. In both studies, Interpersonal Problems had the weakest relationship to Depression.

How do we test whether Depression adequately accounts for the covariances among the first order factors? As is so often the case in confirmatory modeling, this question may be addressed by examining differences in fit between two alternative models. In this case, we have imposed additional restrictions on the first-order factor model by specifying the Depression factor as the sole source of the variances among the four first-order factors. The difference in χ^2 between these

Table 4. Models with a Second-Order Depression Factor for the CES-D in Annville (AVS) and Victoria (VIC) Samples (Standardized Solution)

	Factor Loadings[a]	
	Depression	
First-Order Factor	*AVS*	*VIC*
Affect	98	98
Well-Being	−87	−77
Somatic	84	72
Interpersonal	65	55
	Unique Variances[a]	
	AVS	*VIC*
Affect	03	05
Well-Being	25	40
Somatic	30	48
Interpersonal	58	69
Goodness of fit:		
AVS: $\chi^{2\,(166)} = 350.55$	GFI = .92	AGFI = .90
Loss of fit from first order:	$\chi^{2\,(2)} = 6.71$, $p < .05$;	Δ AGFI = .01
VIC: $\chi^{2\,(166)} = 282.43$	GFI = .91	AGFI = .88
Loss of fit:	$\chi^{2\,(2)} = 1.64$ (N.S.),	Δ AGFI = .00

[a]Decimals omitted.

two models tests the loss of fit caused by adding the second-order factor to the model. In both samples, the loss of fit is small (see Table 4). So the Depression factor does a good job of accounting for the covariance among the factors. However, the goodness-of-fit test does not tell the whole story. In both samples, the covariance of Interpersonal Problems with Somatic Symptoms is *overfit* (a higher predicted than observed covariance), whereas the covariance of Affect and Well-Being is *underfit*. The consistency of this difference suggests there may be subtle, additional relationships that cannot be detected because of omitted factors. On the other hand, it is clear that the second-order factor is a very useful approximation to the relationships among the first-order factors. We can conclude that the combined scale—as reflected in the second-order factor—does have *factorial validity*.

Of course, the third question—the adequacy of predictive validity by the second-order factor—cannot be addressed merely by analysis of the CES-D. It can only be addressed by research that measures the CES-D and other constructs (and outcome measures) of theoretical interest.

These results, then, support the factorial validity of the single CES-D score. They also suggest that the four first-order factors relate differentially to the second-order factor. Thus, although one is justified in using the single (combined) CES-D depression scale, there may be research applications in which maintaining the separate factors is important and useful. Further evidence of predictive and construct validity of both the overall CES-D scale score and the four subscales is needed, however, for this information will be the critical determinant of whether (and when) one is best served by using the overall scale or the four subscales.

A few additional conclusions are warranted by the item analysis. First, if a maximally homogeneous measure of self-report depression is desired, it would be appropriate to form a 14-item scale combining only the Depressive Affect and Somatic Symptoms subscales. These two factors seem closest to a face-valid definition of depression and have strong correlations with each other. Second, the fact that Interpersonal Problems does not relate as strongly to the second-order Depression factor in the CES-D leaves open the possibility that it may be associated with other personal characteristics (e.g., introversion, neuroticism; [35]) in addition to depression. The issue of discriminant validity for this subscale needs to be addressed in further research. The same issue seems important for the Well-Being subscale, for it seems, at the level of face validity, highly related to items contained in Subjective Well-Being scales. Finally, if these latter two scales are found to have convergent and discriminant validity with respect to the domain of depression, and if they have predictive validity of important other constructs independent of the Depression factor, then these subscales would be of interest in their own right. In that event additional items measuring these dimensions should be developed and added to the scale.

CONFIRMATORY FACTOR MODELS OF SUBJECTIVE WELL-BEING

With the preceding discussion of item factor analysis, we are now poised to discuss the burgeoning literature on the factorial validity of measures of subjective well-being in adult populations. This literature has shown a decided progression from the use of exploratory factor analysis to the use of confirmatory methods. One of the first uses of confirmatory factor models was reported by Hoyt and Creech [36]. Using LISREL models, they were unable to confirm the original model for the original Neugarten versions of the Life Satisfaction Index (LISA) [37]. Hoyt and Creech subsequently used exploratory factor analysis to arrive at a reduced three-factor model for eight of the LSIA items [36]. Liang and colleagues [6, 33, 34, 38, 39] have conducted several SEM investigations of the item factor structure of measures of subjects well-being, including the Philadelphia Geriatric Center Morale Scale (PGC) [40], the LSIA, and Bradburn's Affect Balance Scale (ABS) [41]. A central feature of the Liang work has been the simultaneous estimation of first and second-order factors. For example, Liang

and Bollen reported an item factor analysis of the PGC that specified three first-order factors: Agitation, Dissatisfaction, and Attitudes Toward Aging [34]. However, they also estimated a single second-order factor (Subjective Well-Being), and found that the loadings on the second-order factor were high. They argued from these results that the multidimensionality of the PGC was at question. Subsequent analyses by Liang and Bollen [38] and Liang et al. [39] have supported the second-order factor model and indicated invariance of the first and second-order factors across age (young-old versus old-old) and sex groups. The Liang analysis of the LSIA [33] (depicted in Figure 1) posited the three first-order factors (Zest, Mood Tone, and Congruence) suggested by Neugarten et al. [37] that had been previously replicated by Hoyt and Creech [36]. This model was fit to eleven of the seventeen LSIA items using responses from a large national sample, divided into four subsamples for model replication. The model fit acceptably well and replicated across the four subsamples. As in the Liang and Bollen model for the PGC [34], Liang estimated a single second-order factor of Subjective Well-Being [33]. Subjective Well-Being was marked primarily for Zest for Life, and (Positive) Mood Tone, with a much smaller loading for Congruence.

There is much to appreciate in Liang's work. First, the rationale for item selection is clear and explicitly stated. Second, the model specification is fully delineated and relatively complete results reported. This allows the reader to evaluate the results and the models carefully and critically, a feature not present in all reported work in this domain [42]. Third, the models fit well and are parsimonious. Finally, the results are replicated on multiple subsamples, increasing confidence in their generality.

There can be substantive and methodological concerns regarding the analyses however, and these studies should not be considered definitive closure on the appropriate models for subjective well-being (a point noted by Liang and his colleagues in their own papers). What kinds of concerns can be raised? Some are minor methodological points that, in the long run, probably will not vitiate the general conclusions drawn. For example, Liang and his colleagues routinely rely on pair-wise deletion of missing data in order to create (Pearson and polychoric) correlation matrices based upon maximum sample sizes. If items are not missing at random, then this practice can lead to distortions in the solutions, although the replication across multiple subsamples eases some of the concern. A more substantial concern specific to the Liang analysis of the LSIA is that items were deleted for multiple reasons, including "cross-construct error covariances" [32]. Apparently, items with relationships not fully accounted for by the model were deleted from the analysis. This approach suggests that the utility of the model was achieved to some degree by elimination of some of the complex interrelations among items. One might well be suspicious of this approach, however, in that a good solution is achieved in part through elimination of items that do not behave according to theory. On the other hand, the Liang and Bollen model for the PGC specified multiple residual item covariances that seem to form two

relatively large clusters [34]. These clusters seem to involve negative affect (anger, frustration) and perception of negative change in the last year (eg., loss of pep). There is therefore some question as to whether additional item factors could have been extracted, and what that would have implied for the adequacy of Subjective Well-Being as a second-order factor.

Third, although estimation of the second-order Subjective Well-Being factor addresses some useful questions (as illustrated above for the CES-D), one should not prematurely close on the idea that well-being can be adequately measured by a single subjective well-being scale score. Liang's results demonstrate that these scales can have well-defined, multiple dimensions, and still measure a higher-order construct. However, it is important to point out that the models estimated by Liang do *not* provide a definitive test of the single second-order factor model in one important sense. The PGC and LSIA models specify only three first-order factors. In these cases, the Subjective Well-Being factor is *just-identified*. Just-identification is a technical term I cannot fully define here. It means, in essence, that a solution for the parameter estimates may be calculated but that this solution places no restrictions on the model whatsoever. In this case, just-identification of the second-order factor loadings means that the second-order factor fits the covariances among the first-order factors perfectly (and trivially so). Thus it is not possible to use model restrictions to test the adequacy of the fit of the second-order factors (as we did in the example with the CES-D given above). It is true that Liang's estimated factor loadings are large enough to warrant the conclusion that Subjective Well-Being is a valid second-order factor. However, we cannot use the logic of χ^2 to test the adequacy of the second-order factor model as a representation of the covariances among first-order factors. Moreover, the most critical issue regarding the Subjective Well-Being factor—whether it mediates relationships between first-order factors and other constructs—is not in any way addressed in the Liang analyses. The fact that the coefficients of determination for some first-order factors are about 50 percent suggests that there is at least the possibility that the first-order factors (e.g., Congruence) will relate to other constructions independent of Subjective Well-Being.

The problem of a just-identified second-order factor was avoided in Liang's study combining data from the LSIA and the ABS [6]. Liang identified four item factors (Congruence, Happingess [formerly, Mood Tone], Positive Affect, and Negative Affect) on the basis of fifteen items from the two scales [6]. The Positive Affect and Negative Affect factors derive from Bradburn's conceptualization of these two factors of well-being [40] and are marked by eight items from the ABS. In this analysis, Liang *dropped* the Zest item factor defined by the LSIA from the model. Thus, seven of the original LSIA items remain, marking the Congruence and Happiness factors. Liang found good fits to his model, which included a single second-order Subjective Well-Being factor [6]. Given that the four first-order factors overidentify the SWB factor, it is possible to test the fit of the single-factor model relative to an unconstrained first-order factor structure. Unfortunately, this was not done (directly). Instead, Liang demonstrated that the

second-order model fit better than an orthogonal first-order factor model, and that adding residual covariances among first-order residuals improved the fit, but not by much [6]. Since this latter model is just-identified at the level of the second-order analysis, its fit would equal that of a model specifying only first-order factors. Therefore, we can conclude that there is a slight loss of fit for the model with Subjective Well-Being determining all first-order factors. However, Liang appears to be correct in his argument that the loss of fit is not substantial [6]. The most important contribution of the Liang [6] analysis is to show that item factors from two separate scales relate strongly to the second-order factor, justifying the argument of convergent validity for well-being.

Conclusions that appear justified for the Subjective Well-Being factor parallel those outlined above for the CES-D Depression factor. In some cases, it may be preferable to estimate a single scale. However, there is sufficient residual variance for some first-order factors to leave open the question of whether Subjective Well-Being mediates relationships to other constructs. Moreover, Liang's careful approach to item selection may have pushed the analyses in the direction of validating the single second-order factor model. Liang found a small but significant residual for the Positive and Negative affect factors, and one wonders if there might not have been positive residuals for Zest and Happiness (given the strong relationship of the two in the Liang analysis [33]) had the Zest factor been included in the analysis. The conservative approach to item selection was undoubtedly justified, given the confusion in the literature on the structure of these scales. At this point, however, the emphasis should be placed upon risking the model by adding more items and factors and testing the single second-order factor model. This may require new items rather than analysis of the remaining items in the existing scales. Nevertheless, closure on the single Subjective Well-Being factor as adequate to account for the first-order item factors may be premature.

Stock, Okun, and Benin [7] have criticized Liang's [6, 33] model for the LSIA and ABS on conceptual grounds, questioning Liang's first-order item factors. They formulated and estimated an alternative model using SEM. It is based primarily upon the Bradburn perspective, emphasizing Positive and Negative Affect, but was also influenced by Andrews and McKennell [43]. Their first-order factors were Positive Affect, Negative Affect, and Cognition (in essence, items reflecting an evaluation or appraisal of the significance and meaning of one's own life; operationally defined as items involving "social comparison, a self-to-self comparison over time or a life review" [7, p. 95]. These factors were specified to account for six ABS items (Liang used 8 [6]) and the eleven LSIA items used by Liang [33]. The model was fitted to the same data set used by Liang [6, 33]. In addition to their own model, Stock et al. fitted an SEM model based upon the Liang model for the LSIA (including Zest, Mood Tone, and Congruence as factors) [33]. They argued that their model fit as well or better as one based upon the Liang model [33], and championed their own as being more soundly based upon a theory of subjective well-being.

The Stock et al. [7] analysis obviously was conducted before the Liang [6] model

was known to them, for the Liang model does include Positive and Negative Affect (see above). As such, the conceptual differences between the Stock et al. [7] model and the Liang [6] model have narrowed, relative to the Liang model [33]. As the Liang [33] model was based upon Hoyt and Creech [36], I shall attempt to avoid confusion by labeling it the Hoyt-Creech model. In any event, the Stock et al. analysis does not appear to be definitive with respect to which approach is the most appropriate basis for a model for subjective well-being [7]. Although their three factor model fits somewhat better than the Hoyt-Creech model specifying Zest, Mood Tone, and Congruence), it does so by adding more parameters. They do achieve a small but appreciable gain in the relative fit index, but their theory specifies many factor loadings that, empirically, were either not statistically reliable or of small magnitude. Their original model encountered empirical identification problems that were solved by dropping an item from the analysis. Moreover, their model does not appear to cross-validate across two samples as well as the Hoyt-Creech model they estimate (or as well as the Liang model [6]). There is little to differentiate their Cognition factor from the Hoyt-Creech Congruence factor. The items that mark one factor saliently also mark the other, and the loadings that differ, by specification, are not large in magnitude. Moreover, the factor correlations Stock and colleagues report for their Hoyt-Creech model are somewhat more consistent across the two samples than are the correlations in their preferred model.

On the other hand, the pattern of correlations in the Stock et al. model has intuitive appeal [7]. They report a negative correlation of Negative Affect with Positive Affect and Cognition (consistent with Liang's [6] report of a negative second-order factor loading, and consistent with other work on subjective well-being; see above). Their factor correlations are also lower than those among the factors from the Hoyt-Creech model, which may have empirical advantages (e.g., independence of prediction to outcome measures).

At this point, it would appear that both Stock et al. [7] and Liang [6] have offered important alternatives to the Hoyt-Creech formulation. Liang's model can be viewed as merely an extension of the basic Hoyt-Creech model to add ABS factors, whereas the Stock et al. model is a legitimate reformulation. It is too early to tell which model will ultimately prove more useful. Since Liang [6] deleted LSIA items marking Zest, and Stock et al. [7] specified three factors instead of Liang's four, it is difficult to make direct comparisons between the models. At this time, it is not known whether a five-factor model based upon Liang's [6] extension of Hoyt-Creech, but reintroducing Zest, would fit appreciably better than the Stock et al. three factor model [7]. It is also unclear whether a single second-order Subjective Well-Being factor could account for the relationships among the five factors implied by the Liang [6] extension of the Hoyt-Creech model. What is clear, however, is that the debate so far has centered exclusively on the issue of factorial validity for alternative item models, and that it has not yet been extended to an additional focus upon the issues of discriminant and predictive validity of those factors. Additional insights will probably require moving beyond analysis of existing scales from large data archives to the creation of additional item pools formed on theoretical grounds [7]. With respect to the specific issues of selecting an optimal item model (Hoyt-Creech, Liang, Stock et al., or some as-yet-unspecified alternative), any definitive verdict will probably require 1) extension of the item factors to account for new items constructed on theoretical grounds to load on the different factors, and 2) the ability of the different factors to model relationships to other constructs (e.g., depression) and to meaningful outcome measures (e.g., social isolation, morbidity).

One methodological lesson to be learned from this literature is that use of confirmatory factor analysis—in and of itself—does not necessarily resolve disputes about the empirical behavior and proper interpretation of latent variables. Resolution is dependent upon a well chosen design that clearly establishes the alternative models and grounds them in meaningful empirical tests. Thus the discriminative power of confirmatory factor models rests primarily on the measures selected and populations studied, and not upon the statistical procedures *per se* [12]. However, the orientation to use the logic of model falsification inherent in confirmatory analysis establishes the possibility that critical tests of the alternative models can be devised and investigated empirically. Thus one can perceive that a context of discovery has been created in the domain of subjective well-being assessment, fueled largely by the introduction of confirmatory logic and method into the area. The work just reviewed appears to have set the stage for new research that offers a more definitive test of alternative conceptions of well-being scales. There is a justification for optimism that this process will ultimately lead to more valid measures of well-being in adult populations.

MODELS FOR MEASUREMENT PROPERTIES OF SCALES

The type of item analysis just discussed is important for demonstrating that individual scale items map into a summative scale (or subscales) as predicted by the measurement theory. One can also use confirmatory methods to test a variety of hypotheses about the measurement properties of the scales themselves. For this type of analysis, what was a first-order factor in item analysis is converted into an observed variable (scale), and interest focuses on the covariances among these variables.

Schaie and Hertzog discuss in great detail the literature on using confirmatory factor analysis to estimate reliability and to assess equivalence of measurement properties across multiple populations [5]. Here I shall focus on one recent application of these methods, for it nicely illustrates two important concepts: 1) the conceptual distinction between *scale reliability* and *stability of individual differences* and 2) the use of alternate forms to reveal information about the

ıeasurement properties of scales. The data to be discussed were originally col-
ected by Nesselroade, Mitteness, and Thompson [4], and consist of self-ratings
f elderly individuals of two mood state factors: Anxiety and Fatigue. The design
ıvolves a short-term retest, so that individuals were given the mood state ques-
ionnaires twice, with approximately one month intervening between administra-
ons. The three measures of state Anxiety included Spielberger's State Anxiety
ale and Forms A and B of Curran and Cattell's Eight State Questionnaire [44].
he three measures of Fatigue were actually sets of items from the Eight State
atigue scale. Our interest here is on the measurement properties of the Anxiety
ales. In particular, Nesselroade et al. demonstrated that the Anxiety and Fatigue
ctors could be identified using confirmatory factor analysis, and that the stability
individual differences in Anxiety was substantial, but not perfect, over the one-
ionth period [45]. Hertzog and Nesselroade [13] reanalyzed the Nesselroade et
. data [4], focusing on estimating the measurement properties of the Forms A
ıd B of the Eight State Questionnaire.

The model used by Hertzog and Nesselroade is shown in Figure 2 [13]. This
odel specified that the three scales of state anxiety loaded on an Anxiety factor,
ıd that there was a residual covariance for the Spielberger scale across the two
easurement occasions. In a series of models, Hertzog and Nesselroade [13] tested
hether the Cattell Forms A and B could be considered parallel forms [46]. Parallel
rms are interchangeable, because they have equal reliabilities and equal
riances. The hypothesis of parallelism was tested by constraining factor loadings
ıd residual variances (error variances of measurement) to be equal for Forms
and B [1, 2].

Hertzog and Nesselroade also tested whether the measurement properties of
rms A and B were equivalent across the two measurement occasions [13]. The
t of equivalence over time was made by constraining these parameters equal
ross the first and second administrations of the questionnaires. The results
owed clearly that 1) Forms A and B were parallel, and 2) that the measure-
nt properties of Forms A and B were identical over the two occasions of
easurement. The estimated reliability for Forms A and B was .89. Clearly, the
ttell Eight State Anxiety scales have excellent measurement properties in older
pulations.

The high reliabilities for the scales contrast with the moderate (but lower)
ıbilities of individual differences in the latent variable, Anxiety. The stability
individual differences is an important issue in gerontological research, for a
ıjor question is whether individuals maintain their relative differences on
ychological attributes as they grow older [5, 47]. In the Hertzog and Nesselroade
alysis [13], the stability of individual differences is reflected in the covariance
tween the latent Anxiety factors over the two measurement points ($\phi_{3,4}$ in
gure 2). Using the parameter estimates from the Hertzog and Nesselroade
alysis [13], an estimate of .72 is obtained for the correlation of the Anxiety
tor with itself over a one-month period. The advantage of using a latent variable

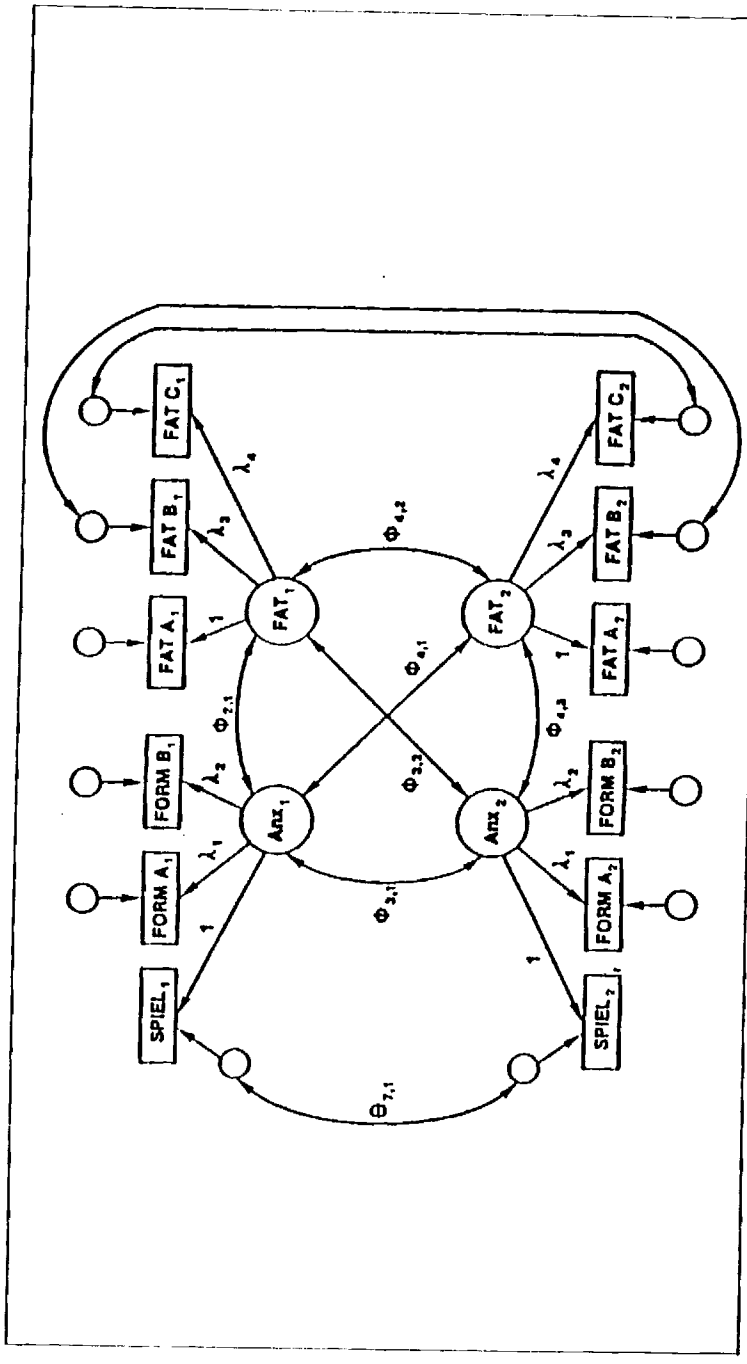


Figure 2. Model for two mood state factors. Anxiety (ANX) and Fatigue (FAT) measured at two longitudinal occasions. For Anxiety, three measures were available, Spielberger's State Anxiety Scale (SPIEL) and two alternate forms of the 8-State Anxiety scale (FORM A, FORM B). A series of models tested the measurement properties of these alternate forms and their relationship to SPIEL at the two occasions (see text). Reprinted with permission from Hertzog and Nesselroade [13].

nodel here is that the factor correlation is not attenuated by measurement error as are the correlations among the scales themselves). The implication of this disat-enuation is that the correlation could be as high as a perfect 1.0 if individual ifferences in Anxiety were perfectly consistent across the one-month period. he estimated correlation was certainly greater than zero, but less than 1.0, ndicating individual differences in mood state change between the two times. stands in marked contrast to the long-term stability of intelligence and per-nality constructs. For example, Hertzog and Schaie found that a general intelli-ence factor correlated .9 or greater with itself over a seven-year longitudinal st interval [14].

From the Hertzog and Nesselroade analysis [13], we can conclude that those ho are anxious at time 1 are likely to be anxious at time 2, but only about 50 rcent of the variance in self-reported anxiety at time 2 can be predicted from xiety levels at time 1. The most important feature of the analysis, however, that this stability of individual differences has been estimated in a way that sentangles it from reliability. We can reject the hypothesis that the less-than-rfect stability in Anxiety is a function of attenuation due to measurement error. onversely, the analysis shows that the lability in mood states does not imply at the mood state measures are unreliable. Indeed, the Eight State Anxiety scale pears to have very good reliability in older populations. Given that one would pect mood states to fluctuate, the lability of Anxiety, and the excellent measure-ent properties of Forms A and B, actually argue indirectly for the construct lidity of the scales, and suggest that they measure something different from e personality trait of Anxiety, which has been shown to exhibit a high degree stability of individual differences [35].

CONCLUSIONS

As is always the case with a complex topic, this chapter has not covered a mber of additional applications that illustrate important other features of con-matory factor analysis for scale validation. One important area, alluded to above, the use of confirmatory techniques for testing convergent and discriminant lidity of multiple scales [25]. For example, Hertzog, Hultsch, and Dixon recently orted results of a series of confirmatory factor analyses that show convergent lidity of two questionnaires measuring metamemory (an individual's knowledge beliefs about memory functioning) [24]: the Dixon and Hultsch Metamemory Adulthood questionnaire [19] and the Memory Functioning Questionnaire [48]. e analysis by Hertzog, Hultsch, and Dixon also demonstrated the discriminant idity of a memory self-efficacy beliefs factor, taken from both questionnaires, l other psychological constructs [e.g., neuroticism, internal local of control, l subject well-being) [24].

Space did not permit an extended discussion of one of the chief advantages

of confirmatory factor analysis for scale validation: the use of simultaneous multiple groups analysis to test the equivalence of factor structures across multiple age groups. In a sense, this chapter provides a foundation that can be generalized to the case of multiple groups analysis. Technical (but rewarding) reading on this topic may be found in several sources [1, 4, 5, 25, 49]. Hertzog reviews a number of empirical applications of confirmatory factor analysis and structural equation models in gerontological research, including the use of multiple groups analysis [50].

There are some disadvantages to confirmatory factor analysis for scale development, but I will leave it to others to point them out in detail! Part of the rationale for doing so is the rapid technological advances that are currently underway. Techniques that avoid some of the assumptions of standard maximum likelihood estimation procedures are now generally available, both in LISREL and competing programs such as EQS [51]. Specialized methods for dealing with categorical and ordinal variables are also appearing [52]. The general principles of confirmatory factory analysis illustrated in this chapter hold for these newer techniques as well.

The thesis of this chapter has been that confirmatory factor analysis provides a powerful method for evaluating the measurement properties of psychological scales. I have sought to show that confirmatory models can be profitably used for item factor analysis and scale validation. As this technique becomes more widely understood and available, it is likely that the current literature using confirmatory approaches to scales of subjective well-being, as reviewed above, will be mirrored in other measurement domains crucial for the study of adult development and aging. This is an exciting prospect, for it suggests that, as a field, we will be making considerable progress in evaluating the reliability and validity of our measures.

REFERENCES

1. D. F. Alwin and D. J. Jackson, Measurement Models for Response Errors in Surveys: Issues and Applications, in *Sociological Methodology 1980*, K. F. Schuessler (Ed.), Jossey-Bass, San Francisco, pp. 68–119, 1979.
2. K. G. Jöreskog, Simultaneous Factor Analysis in Several Populations, *Psychometrika*, *36*, pp. 409–426, 1971.
3. K. G. Jöreskog, Analyzing Psychological Data by Structural Analysis of Covariance Matrices, in *Contemporary Developments in Mathematical Psychology*, D. H. Krantz, R. C. Atkinson, R. D. Luce, and P. Suppes (Eds.), W. H. Freeman, San Francisco, Vol. 2, pp. 1–56, 1974.
4. D. A. Rock, C. E. Werts, and R. L. Flaugher, The Use of Analysis of Covariance Structures for Comparing the Psychometric Properties of Multiple Variables Across Populations, *Multivariate Behavioral Research*, *13*, pp. 403–418, 1978.

5. K. W. Schaie and C. Hertzog, Measurement in the Psychology of Adulthood and Aging, in *Handbook of the Psychology of Aging*, J. W. Birren and K. W. Schaie (Eds.), Van Nostrand Reinhold, New York, 2nd Edition, 1985.
6. J. Liang, A Structural Integration of the Affect Balance Scale and the Life Satisfaction Index A, *Journal of Gerontology, 40*, pp. 552–561, 1985.
7. W. A. Stock, M. A. Okun, and M. Benin, Structure of Subjective Well-Being Among the Elderly, *Psychology and Aging, 1*, pp. 91–102, 1986.
8. E. Diener, Subjective Well-Being, *Psychological Bulletin, 95*, pp. 542–575, 1984.
9. M. P. Lawton, The Varieties of Well-Being, *Experimental Aging Research, 9*, pp. 65–72, 1983.
0. L. J. Cronbach, *The Essential of Psychological Testing*, Harper and Row, New York, Third Edition, 1970.
. R. P. McDonald, The Dimensionality of Tests and Items, *British Journal of Mathematical and Statistical Psychology, 34*, pp. 100–117, 1981.
. C. Hertzog, On the Utility of Structural Regression Models for Developmental Research. Chapter to appear in *Life-Span Development and Behavior*, P. B. Baltes, D. L. Featherman and R. M. Lerner (Eds.), Lawrence Erlbaum Associates, Hillsdale, NJ, Vol. 10, in press.
C. Hertzog and J. R. Nesselroade, Beyond Autoregressive Models: Some Implications of the Trait State Distinction for the Structural Modeling of Developmental Change, *Child Development, 58*, pp. 93–109, 1987.
C. Hertzog and K. W. Schaie, Stability and Change in Adult Intelligence: 1. Analysis of Longitudinal Covariance Structures, *Psychology and Aging, 1*, pp. 159–171, 1986.
J. L. Horn and J. J. McArdle, Perspectives on Mathematical/Statistical Model Building (MASMOB) in Research on Aging, in *Aging in the 1980's: Psychological Issues*, L. W. Poon (Ed.), American Psychological Association, Washington, DC, pp. 503–541, 1980.
S. Messick, Constructs and Their Vicissitudes in Educational and Psychological Measurement, *Psychological Bulletin, 89*, pp. 575–588, 1981.
L. J. Cronbach and P. E. Meehl, Construct Validity in Psychological Tests, *Psychological Bulletin, 52*, pp. 281–302, 1955.
P. M. Bentler, The Interdependence of Theory, Methodology, and Empirical Data: Causal Modeling as an Approach to Construct Validation, in *Longitudinal Research on Drug Abuse, Empirical Findings and Methodological Issues*, D. B. Kandel (Ed.), Hemisphere, Washington, pp. 267–302, 1978.
R. A. Dixon and D. F. Hultsch, Structure and Development of Metamemory in Adulthood, *Journal of Gerontology, 38*, pp. 682–688, 1983.
L. S. Radloff, The CES-D Scale: A Self-Report Depression Scale for Research in the General Population, *Applied Psychological Measurement, 1*, pp. 385–401, 1977.
P. M. Lewinsohn, D. S. Fenn, A. K. Stanton, and J. Franklin, Relation of Age at Onset to Duration of Episode in Unipolar Depression, *Psychology and Aging, 1*, pp. 63–68, 1986.
M. Gatz, M. Hurwicz, and W. Weicker, *Are Old People More Depressed?* Cross-Sectional Data on CES-D Factors, Paper presented at the meeting of the American Psychological Association, Washington, DC, 1986.
C. S. Aneshensel, Race, Ethnicity and Depression: A Confirmatory Analysis, *Journal of Personality and Social Psychology, 44*, pp. 385–398, 1983.

24. C. Hertzog, D. F. Hultsch, and R. A. Dixon, What Do Metamemory Questionnaires Measure? A Construct Validation Study. Paper presented at the 95th Annual Convention of the American Psychological Association, New York, NY, 1987.
25. C. Hertzog, Applications of Confirmatory Factor Analysis to the Study of Intelligence. in *Current Topics in Human Intelligence*, D. K. Detterman (Ed.), Ablex, Norwood, NJ, pp. 59–97, 1985.
26. S. A. Mulaik, Toward a Conception of Causality Applicable to Experimentation and Causal Modeling, *Child Development, 58*, pp. 18–32, 1987.
27. A. Boomsma, The Robustness of LISREL Against Small Sample Sizes in Factor Analysis Models, in *Systems Under Indirect Observation: Causality, Structure, Prediction*, K. G. Jöreskog and H. Wold (Eds.), North Holland, Amsterdam, Volume 1, pp. 149–173, 1982.
28. P. M. Bentler and D. G. Bonett, Significance Tests and Goodness of Fit in the Analysis of Covariance Structures, *Psychological Bulletin, 88*, pp. 588–606, 1980.
29. L. R. James, S. A. Mulaik, and J. M. Brett, *Causal Analysis: Assumptions, Models, and Data*, Sage, Beverly Hills, CA, 1982.
30. G. J. Huba and L. L. Harlow, Robust Estimation for Causal Models: A Comparison of Methods in Some Developmental Data Sets, in *Life-Span Development and Behavior*, R. M. Lerner and D. L. Featherman (Eds.), Academic Press, New York, Vol. 6, 1986.
31. J. S. Tanaka and G. J. Huba, Confirmatory Hierarchical Factor Analysis of Psychological Distress Measures, *Journal of Personality and Social Psychology, 46*, pp. 621–635, 1984.
32. C. Veit and J. E. Ware, Jr., The Structure of Psychological Distress and Well-Being in General Populations, *Journal of Consulting and Clinical Psychology, 51*, pp. 730–742, 1983.
33. J. Liang, Dimensions of the Life Satisfaction Index A: A Structural Formulation, *Journal of Gerontology, 39*, pp. 613–622, 1984.
34. J. Liang and K. A. Bollen, The Structure of the Philadelphia Geriatric Center Morale Scale: A Reinterpretation, *Journal of Gerontology, 38*, pp. 181–189, 1983.
35. P. T. Costa, Jr. and R. R. McCrae, Influence of Extraversion and Neuroticism on Subjective Well-Being: Happy and Unhappy People, *Journal of Personality and Social Psychology, 38*, pp. 668–678, 1980.
36. D. R. Hoyt and J. C. Creech, The Life Satisfaction Index: A Methodological and Theoretical Critique, *Journal of Gerontology, 38*, pp. 111–116, 1983.
37. B. L. Neugarten, R. Havighurst, and S. Tobin, The Measurement of Life Satisfaction, *Journal of Gerontology, 16*, pp. 134–143, 1961.
38. J. Liang and K. A. Bollen, Sex Differences in the Structure of the Philadelphia Geriatric Center Morale Scale, *Journal of Gerontology, 40*, pp. 468–477, 1985.
39. J. Liang, R. H. Lawrence, and K. A. Bollen, Age Differences in the Structure of the Philadelpia Geriatric Center Morale Scale, *Psychology and Aging, 1*, pp. 27–33, 1986.
40. M. P. Lawton, the Philadelphia Geriatric Center Morale Scale: A Revision, *Journal of Gerontology, 30*, pp. 85–89, 1985.
41. N. M. Bradburn, *The Structure of Psychological Well-Being*, Aldine, Chicago, 1969.
42. G. A. Wilson, J. W. Elias, and L. J. Brownlee, Factor Invariance and the Life Satisfaction Index, *Journal of Gerontology, 40*, pp. 344–346, 1985.

43. F. M. Andrews and A. McKennell. Measures of Self-Reported Well-Being: Their Affective, Cognitive, and Other Components. *Social Indicators Research*, *8*, pp. 127–155, 1980.
44. J. P. Curran and R. B. Cattell. *Handbook for the 8-State Questionnaire*, Institute for Personality and Ability Testing, Champaign, IL. 1976.
45. J. R. Nesselroade, L. S. Mitteness, and L. K. Thompson. Short-Term Changes in Anxiety, Fatigue, and Other Psychological States in Older Adulthood. *Research on Aging*, *6*, pp. 3–23, 1984.
46. F. M. Lord and M. R. Novick. *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading, MA. 1968.
47. P. B. Baltes, H. W. Reese, and J. R. Nesselroade. *Life-Span Developmental Psychology: Introduction to Research Methods*. Brooks-Cole. Monterey, CA. 1977.
48. M. J. Gilewski and E. M. Zelinski. Questionnaire Assessment of Memory Complaints. in *Handbook for Clinical Memory Assessment of Older Adults*. L. W. Poon (Ed.). American Psychological Association. Washington. DC. pp. 93–107. 1986.
49. K. G. Jöreskog. Statistical Analysis of Sets of Congeneric Tests. *Psychometrika*. *36*. pp. 109–133, 1971.
50. C. Hertzog. Applications of Structural Equation Models in Gerontological Research. in *Annual Review of Gerontology and Geriatrics*. K. W. Schaie (Ed.). Vol. 7. pp. 265–293. 1987.
51. P. M. Bentler. *Theory and Implementation of EQS: A Structural Equations Program*. BMDP Statistical Software. Los Angeles. 1985.
52. R. J. Mislevy. Recent Developments in the Factor Analysis of Categorical Variables. *Journal of Educational Statistics*, *11*. pp. 3–31. 1986.

Developmental Psychology
1989, Vol. 25, No. 5, 687-700

0012-1649/89/$00.75

Evidence for the Convergent Validity of Two Self-Report Metamemory Questionnaires

Christopher Hertzog
Georgia Institute of Technology

David F. Hultsch and Roger A. Dixon
University of Victoria, British Columbia, Canada

Examined the convergent validity of two metamemory questionnaires: the Metamemory in Adulthood questionnaire (MIA) and the Memory Functioning Questionnaire (MFQ). Confirmatory factor analysis showed that each instrument yields a higher-order factor labeled *Memory Self-Efficacy* (MSE) with approximately a .9 factor correlation. The analysis also showed convergence of the two questionnaires' scales measuring self-reported strategy use and perceived change in MSE. Simultaneous factor analysis in multiple cross-sectional age groups indicated that MSE has age-invariant factor loadings, although there was an age-related increase in the correlation of the MSE and Change factors. Additional models suggested (a) age differences in metamemory scales are primarily produced by age differences in MSE and (b) a minor method factor in the MFQ, producing both the less-than-perfect correlation of the two MSE factors and a reduced sensitivity of the MFQ to age differences.

Cognitive psychologists interested in age-related changes in memory have hypothesized that metamemory, defined as knowledge and beliefs about one's own memory functioning, may represent a key to understanding both age changes in laboratory memory task performance and the use of memory in everyday life (e.g., Cavanaugh, Kramer, Sinnott, Camp, & Markley, 1985; Dixon & Hertzog, 1988; Hultsch, Hertzog, & Dixon, 1985; Perlmutter, 1978; Zelinski, Gilewski, & Thompson, 1980). One approach to metamemory assessment has been the use of self-report questionnaires. There is currently a plethora of questionnaires in the literature (see Dixon, in press; Gilewski & Zelinski, 1986). However, four questionnaires have found the most frequent usage in aging studies: the Short Inventory of Memory Experiences (SIME; Herrmann & Neisser, 1978), the Everyday Memory Questionnaire (EMQ; Sunderland, Harris, & Baddeley, 1983), the Memory Functioning Questionnaire (MFQ; Gilewski, Zelinski, Schaie, & Thompson, 1983), and the Metamemory in Adulthood instrument (MIA; Dixon & Hultsch, 1983, 1984). Each of these questionnaires has apparent virtues and some apparent limitations (Gilewski & Zelinski, 1986).

What is the basis, then, for selecting one or more of these questionnaires for research and assessment? Answering this question is difficult, given that the principals have attended primarily to research validating their own instruments. For example, Dixon and Hultsch (1983) showed that the MIA's multiple subscales have good reliability and factorial validity. However, further work investigating the construct validity (e.g., Cronbach & Meehl, 1955; Messick, 1981) of metamemory questionnaires for adult populations is critically needed (Dixon, in press; Dixon & Hertzog, 1988; Gilewski & Zelinski, 1986; Herrmann, 1982).

The dimensionality of the metamemory construct is a central issue in evaluating the validity of metamemory questionnaires. How many differentiable constructs exist in the broad domain of metamemory? How ought the general domain and the specific constructs be conceptualized? It appears clear that there are multiple dimensions of metamemory. One organizing principle for conceptualizing the domain is the distinction between knowledge about memory mechanisms, processes, and failures and beliefs about one's own memory abilities, strengths, and weaknesses (see Dixon, in press; Hertzog, Dixon, Schulenberg, & Hultsch, 1987; Hultsch, Hertzog, Dixon, & Davidson, 1988). Our approach to this issue is informed by Bandura's (1986) concept of self-efficacy. This concept provides a fruitful theoretical framework for conceptualizing the metamemory domain (Cavanaugh et al., 1985; Hultsch et al., 1985; Lachman, 1986; West, Berry, & Dennehy, 1987). The self-efficacy perspective is consistent with the distinction between memory knowledge and memory beliefs. Consistent with Bandura (1986), memory self-efficacy can be defined as beliefs about one's own capability to use memory effectively in various situations. The differentiation of knowledge about memory from memory self-efficacy allows for the possibility that an older individual may have extensive and accurate knowledge about how memory functions but may also believe that his or her ability to remember in a given context is poor. The concept of memory self-efficacy also makes it possible to entertain questions concerning the accuracy of

This research was supported by a research grant to Christopher Hertzog from the National Institute on Aging (NIA; R01-AG06162) and by a grant to David F. Hultsch from the Social Sciences and Humanities Research Council of Canada (492-84-002). Christopher Hertzog was also supported by a Research Career Development Award from the NIA (K04-AG00335). Roger A. Dixon was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The cooperation of Robert K. Nielsen, the other physicians, and the members of the Annville Family Practice, Annville, Pennsylvania, is deeply appreciated. We also thank Paul Usala, Laurie Saylor, and Judy Van Alstine for their assistance in data preparation and table construction.

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

memory beliefs (Langer, 1981; Sunderland, Watts, Baddeley, & Harris, 1986; West, Boatwright, & Schleser, 1984) and to develop intervention techniques designed to identify and ameliorate negative self-efficacy beliefs so as to enhance the effective functioning of the individual (Bandura, 1986; Zarit, 1982).

The broad distinction between knowledge and beliefs is not sufficient to capture the full range of constructs commonly subsumed under metamemory. Hultsch et al. (1988) have suggested four broad subdomains of metamemory: memory knowledge, memory monitoring, memory self-efficacy, and memory-related affect. Each, in turn, can be subdivided into multiple specific constructs, more akin to the specific scales contained in the multidimensional instruments like the MIA and the MFQ. Gilewski and Zelinski (1986) identify 10 dimensions of metamemory, the most important of which they labeled *frequency of forgetting*. This label is consistent with the format of items from the SIME and the MFQ, which ask individuals to rate how often they forget specific types of information.

Incidents of forgetting are undoubtedly a source of information that individuals use to create and update memory self-efficacy beliefs. In a sense, then, the actual frequency of forgetting is likely to be a proximal determinant of memory self-efficacy. However, it seems highly probable that the proximal cause of self-reported frequency of forgetting is *not* true frequency of forgetting but, rather, memory self-efficacy beliefs. This hypothesis is based on evidence that individuals' self-reports are often based on access to generalized beliefs and self-schemata—the stable representations of self that are products of a self-appraisal process (e.g., Hastie & Park, 1986; Wyer & Srull, 1986)—rather than access to specific, discrete episodes in memory. It is improbable that individuals base a frequency of forgetting estimate on an exhaustive retrieval search of memory for incidents of forgetting. A more plausible representation of frequency report behavior is that individuals access beliefs about their memory self-efficacy and then convert these beliefs into a frequency estimate.

This perspective on self-reported frequency of forgetting predicts substantial correlations between superficially different questions about remembering and forgetting. At least two MIA scales appear to measure related aspects of memory self-efficacy: beliefs about current levels of memory ability (Capacity) and beliefs about the degree of change in capacity from early adulthood (Change). We hypothesize that the MFQ also contains multiple scales that may be subsumed under the self-efficacy category, especially including those related to the frequency of forgetting dimension.

Appreciation of the knowledge/beliefs distinction can alter inferences regarding metamemory scale validity. For example, Gilewski and Zelinski (1986) criticized the EMQ and the Inventory for Memory Experiences (IME) for their lack of sensitivity to age differences (e.g., Chaffin & Herrmann, 1983) and recommended the MIA and the MFQ instead. The implicit basis for their concern seems to be the premise that, given age differences in memory functioning, a valid measure of metamemory ought also to display age differences. This assumption presumes, however, that memory self-ratings are accurate. Alternatively, the IME may be a valid measure of memory self-efficacy, yet fail to show age differences because older persons' self-efficacy beliefs do not change in accordance with changes in their actual memory capacity. This perspective suggests the hypothesis that the IME validly measures the memory self-efficacy construct, but that this construct has little predictive utility for memory task performance.

Ironically, Hultsch, Hertzog, and Dixon (1987) recently reported finding no significant age differences on the MFQ Frequency of Forgetting scale in samples also yielding significant age differences on the MIA measure of memory capacity. Hultsch et al. (1987) did find significant mean differences between a group of older adults and university students for other MFQ scales. However, a different cross-sectional sample of adults, in which age varied continuously from 20 to 78, yielded only marginally significant differences on these same MFQ scales, but robust age differences on the MIA Capacity, Change, and Locus scales. The MIA's greater sensitivity to age differences does not necessarily imply that the MIA is more valid than the MFQ as an instrument for measuring metamemory. The larger age differences could reflect the influence of an age-related source of systematic measurement error. However, the pattern of the Hultsch et al. (1987) results raises an interesting question. Does differential sensitivity to age effects refute the hypothesis that the MIA and MFQ scales converge to measure memory self-efficacy?

This study is part of an ongoing research project designed to examine the construct validity of the MIA and the MFQ. We use confirmatory factor analysis and structural regression models to test the convergent, discriminant, and predictive validity of related MIA and MFQ subscales. Although a complete assessment of construct validity requires attention to all of these aspects of validity, this report focuses on a complex set of analyses examining the convergence of the MIA and MFQ. Convergent validity refers to the degree to which scales from the two questionnaires actually measure the same underlying constructs. Although their construct domains do not overlap completely (e.g., the MIA, but not the MFQ, measures affect about memory), they do appear to overlap in (a) measures of memory self-efficacy, (b) measures of perceived change in memory functioning, and (c) measures of self-reported memory strategies (principally, use of external memory aids).

Campbell and Fiske (1959) argued that convergent validity implied that measures of the same construct correlate more highly with each other than with measures of different constructs (discriminant validity; see also Cook & Campbell, 1979). Although this implication is correct, the convergence hypothesis is inadequately assessed by inspection of zero-order correlations for several reasons. First, higher correlations do not necessarily imply that two variables measure the same construct; instead, they may measure different but correlated constructs. Second, magnitudes of correlations are attenuated by the reliability and validity of the measures. Hence, low correlations may reflect low reliability, low validity, or both, rather than the influence of different constructs on each measure. Although inferences that are based on differential magnitudes of pairs of correlation may often be accurate, the possible biasing effect of combination of differential reliabilities and validities can lead to errors of inference regarding discriminant validity.

Confirmatory factor analysis and structural equation models provide an attractive set of methods for addressing construct validity (e.g., Anderson & Gerbing, 1988; Bentler, 1978; Jöre-

skog, 1974). A principle advantage is that the latent variable model is fit to the entire set of correlations, providing a more direct representation of the convergent and discriminant validity hypothesis. This in turn provides a more stringent falsification test of the model for the latent constructs. The constructs determining a set of measures are operationally defined as latent variables (factors) in these models. Figure 1 illustrates two alternative methods for representing convergence with latent variables, both of which are used in the present study. In Panel A, two measures (X_1 and X_4) are each determined by the same latent construct (Memory Self-Efficacy). In Panel B, two latent Memory Self-Efficacy constructs are each defined by three measures, and the issue of convergence is assessed by evaluating the magnitude of the correlation between the constructs.

When using confirmatory factor models for construct validity, relations among constructs are represented at the level of factor correlations (as represented by the curved arrow in Figure 1). Estimates of factor correlations are disattenuated for measurement error in the individual measures (Anderson & Gerbing, 1988). Because the upper bound of a disattenuated correlation between two factors is 1.0, it becomes meaningful to test whether the correlation of two latent variables is 1.0 and to use this test as a test of convergence (see Jöreskog, 1974). If the factor correlation does not differ from 1.0, there will be no cost in reformulating the model to have a single latent variable determining all measures. The advantage of the two-factor approach is that it does not assume convergence but directly tests the convergence hypothesis.

Rejection of the hypothesis that the factor correlation is 1.0 implies either that (a) the factors are determined by related but distinct constructs or (b) the factors are measures of the same construct, but that systematic sources of variance influence measures of the same factor (i.e., method variance). The latter possibility can be assessed by using a confirmatory factor analysis of a multitrait/multimethod design (Jöreskog, 1974; Long, 1983; Widaman, 1985). That approach is not possible in this study, because the constructs measured by the MIA and the MFQ do not completely overlap, and, thus, the trait and method factors are underdetermined.

Previous factor analyses of the MFQ and the MIA support the contention that there are higher-order dimensions in both questionnaires. Gilewski et al. (1983) found that several MFQ scales loaded on their Frequency of Forgetting factor. Hertzog et al. (1987) reported that the MIA scales of Capacity, Change, Anxiety, and Locus all loaded on a dimension interpreted as Memory Self-Efficacy. In the present study, the hypothesis that multiple scales of the MIA and the MFQ are related to memory self-efficacy is represented in a model with separate memory self-efficacy factors for each questionnaire. This specification enables a test of the hypothesis that these latent variables have a 1.0 correlation.

Method

Subjects

Two samples were included in the study. One sample was drawn from a medium-size western Canadian city (Victoria, British Columbia). The second sample was drawn from a semirural area in the eastern United States (Annville, Pennsylvania). We included only those cases with complete data on the metamemory scales (see below) in this report. The Victoria sample consisted of 360 individuals, 96 university students (age range = 20–26 years; M = 22.11, SD = 1.85) and 264 older adults (age range = 55–78 years; M = 65.88, SD = 5.45), whereas the Annville sample included 415 adults (age range = 20–78 years; M = 52.33 years, SD = 13.67). There were slightly more women than men in both samples. All subjects were paid volunteers. Subjects generally reported themselves to be in good to excellent health. Years of education ranged from 6 to 22 years in the Annville sample (M = 13.47, SD = 2.95) and from 2 to 24 years in the Victoria sample (M = 13.74, SD = 3.32). The Victoria area is a prime retirement location in Canada and tends to produce relatively select samples of older adult volunteers. A comparison of Annville adults, 56 to 78 years of age, with Victoria adults showed Victoria adults to have somewhat greater mean years of education (t = 2.49, $p < .05$). Moreover, the Victoria adult sample scored, on average, one standard deviation above the Annville older adults in a measure of recognition vocabulary (see Hultsch et al., 1987). Mean vocabulary in the Annville older adults was 35.11 (SD = 10.65), whereas the mean vocabulary score for the Victoria adults was 43.34 (SD = 7.71), which

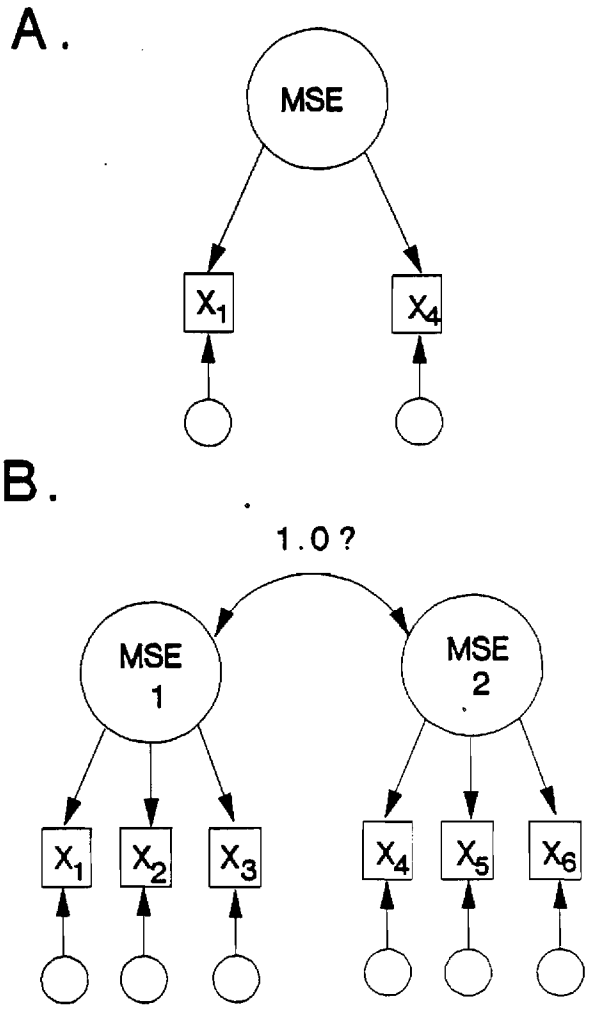


Figure 1. Two alternative methods of representing convergent validity in latent variable models. (In Panel A, two indicators load on a single factor. In Panel B, multiple indicators load on two different factors, with convergence reflected in the magnitude of the factor correlation.)

was a highly significant difference ($t = 100.37$, $p < .001$). Additional details regarding the samples may be found in Hultsch et al. (1987).

Measures and Procedures

Participants from both samples completed a set of questionnaires and tasks measuring metamemory, social desirability, personal control, affective state, verbal comprehension, and memory performance (free recall of word lists and texts). These instruments were administered in two 2-hr sessions to small groups of 5 to 15 participants. In this report, we analyze data from the two metamemory questionnaires, which were given during the first session. The questionnaires were given in invariant order, with the MIA first and the MFQ second. Three affective state questionnaires were administered between the MIA and the MFQ: (a) the Spielberger State Anxiety Scale (Spielberger, Gorsuch, & Lushene, 1969), (b) a reduced version of the Eight State Questionnaire (Curran & Cattell, 1976), and (c) a mood adjective rating scale, using adjectives from the Profile of Mood States (McNair, Lorr, & Droppelman, 1971) and from a scale developed by Lebo and Nesselroade (1978).

The Metamemory in Adulthood (MIA) scale (Dixon & Hultsch, 1983, 1984) contained 108 items, in a 5-point Likert response format, from seven subscales. The Activity subscale was dropped from the original 120-item scale. The remaining seven subscales are presented in Table 1. Prior work with multiple samples has reported internal consistency reliabilities ranging from .74 to .93 for the subscales (Hultsch et al., 1988). The Memory Functioning Questionnaire (MFQ), developed by Gilewski et al. (1983; see Gilewski & Zelinski, 1986) contains 64 items, in a 7-point Likert format, tapping multiple dimensions of metamemory. Gilewski et al. (1983) reported internal consistency reliability estimates ranging from .82 to .93 across MFQ scales. Table 2 presents descriptive information about the MFQ subscales. The MFQ focuses on memory problems and frequency of forgetting in specific domains (e.g., forgetting appointments). This results in primarily negatively worded items. The MIA requires ratings of a mixture of specific and general statements, using both positive and negative wording. The MIA Capacity scale, found by Hertzog et al. (1987) to be the best indicator of memory self-efficacy, uses mostly items with positive wording.

Scoring for all scales was accomplished by summing Likert responses with reverse scoring of MIA items where appropriate (Dixon & Hultsch, 1984). For this analysis, cases with missing data on item responses were handled in the following way. If the scale was missing roughly 20% of item responses, its scale score was coded as missing. If there were fewer missing responses, the item was assigned the sample item mean prior to scale score calculation. Specific criteria for missing value assignment for each scale are reported in Tables 1 and 2. Only cases with complete data on the scale scores were included in the analysis.[1]

An examination of the MFQ items suggested that its General Rating, Frequency of Forgetting, and Remembering Past Events scales measure interrelated aspects of Memory Self-Efficacy. A separate section of the MFQ measures self-reported memory problems for reading materials. The items in this section also seemed to be related to Memory Self-Efficacy, although it seemed plausible that their specificity would result in a less-than-perfect correlation between these items and Memory Self-Efficacy. In their analysis, Gilewski et al. (1983) found one common factor with high loadings on the Frequency of Forgetting, Remembering Past Events, and General Rating indicators. In that analysis, the two separate reading sections of the MFQ (frequency of forgetting books, frequency of forgetting magazines) were combined into a single variable.

Independent of the Memory Self-Efficacy construct, there may be convergence between other pairs of scales from the two questionnaires. Two cases are important. First, both questionnaires contain a scale that asks individuals to assess perceived change in memory (the MIA Change and the MFQ Retrospective Functioning scales). Measures of perceived change in memory show larger age differences than other measures of metamemory (e.g., Hultsch et al., 1987). Therefore, although perceived change indicators may be primary markers of Memory Self-Efficacy, it is plausible that they would form a factor that is independent of Memory Self-Efficacy. Second, both questionnaires contain a self-reported strategy use scale (the MIA Strategy and the MFQ Mnemonics scales). Our expectation on the basis of Hertzog et al. (1987) was that these scales would form a Memory Knowledge factor along with the MIA Task and Achievement scales.

Statistical Procedures

We evaluated convergent validity by conducting factor analysis using the LISREL VI program of Jöreskog and Sörbom (1984). (Readers interested in an introduction to the use of LISREL for confirmatory factor analysis should consult Long, 1983, or Hayduk, 1987.) LISREL contains the following three relevant parameter matrices in its measurement model: (a) a factor pattern matrix, containing regressions of variables on factors (factor loadings), (b) a factor covariance matrix, and (c) a residual covariance matrix, containing residual (unique) variances and residual covariances. All models reported here fixed all residual covariances to zero. In the analyses with entire samples, the model used a correlation matrix and standardized all three parameter matrices. The simultaneous factor analyses with multiple age groups analyzed covariance matrices in order to get appropriate comparisons of model parameters across groups (Jöreskog, 1971; Schaie & Hertzog, 1985).

The LISREL program provides maximum likelihood estimates for factor loadings, factor correlations, and unique variances, as well as standard errors for these estimates. It also computes a likelihood ratio chi-square test of the goodness of fit of the model to the sample data. The null hypothesis for the chi-square test is that the sample covariance (or correlation) matrix is drawn from a population matrix that is determined by the factor model specified. With large samples, the chi-square test may be significant even when discrepancies between the fitted and observed covariance matrices are small. For this reason, LISREL also reports goodness-of-fit indices that are less heavily influenced by sample size. Our multiple groups analyses also report an analagous goodness-of-fit statistic that is based on the Bentler and Bonett (1980) normed-fit indices. These indices reflect the proportion of information in the covariance matrix that is accounted for by the model (see Mulaik et al., 1989, for more detail). Fit indices above .9 are often treated as acceptable, although proper evaluation of a model also requires attention to the salience and interpretability of parameter estimates and the fit to each sample variance and covariance.

Any application of LISREL or related techniques usually involves some combination of model testing and model development. As argued elsewhere (Hertzog, in press), it is appropriate to use LISREL for model development as long as (a) a clear distinction is maintained between confirmatory and exploratory research purposes, (b) tests of statistical hypotheses using LISREL are not necessarily treated as disconfirmation tests of substantive hypotheses formulated a priori, (c) models developed on the basis of sample data are cross-validated in one or more independent samples, and (d) substantive inferences from exploratory modeling are treated as provisional hypotheses to be subjected to later disconfirmation tests.

Consistent with this perspective, the present analysis, conducted in three phases, combined exploratory and confirmatory approaches. First, we developed a model using LISREL to specify and test alternative models for the data from the Annville sample data. The purpose of this phase was to determine (a) if the MIA Memory Self-Efficacy and

[1] The sole exception was the single item General Memory Rating in the MFQ. Twenty-eight Annville and 23 Victoria participants who were missing this single item were assigned their respective sample mean.

Table 1
Dimensions of the Metamemory in Adulthood Instrument

Dimension	Description	Sample item
Strategy (18, 3)	Knowledge and use of information about one's remembering abilities so that performance in given instances is potentially improved (+ = *high use*)	Do you write appointments on a calendar to help you remember them?
Task (15, 2)	Knowledge of basic memory processes, especially that are interesting 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 (17, 3)	Perception of memory capacities as evidenced by rating of performance on given tasks (+ = *high capacity*)	I am good at remembering names.
Change (18, 3)	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 (14, 2)	Feelings of stress related to memory performance (+ = *high anxiety*)	I do not get flustered when I am put on the spot to remember new things.
Achievement (16, 2)	Perceived importance of having a good memory and performing well on memory tasks (+ = *high achievement*)	It is very important that I am very accurate when remembering names of people.
Locus (9, 2)	Perceived personal control over remembering abilities (+ = *internality*)	Even if I work on it, my memory ability will go downhill.

Note. Based on Dixon and Hultsch, 1983. The first number in the parentheses after each dimension represents the total items; the second number represents the missing items allowed (with item mean substitution).

Table 2
A Priori Subscales of the Memory Functioning Questionnaire

Subscale	Sample item
General rating (1, 1)	How would you rate your memory in terms of the kinds of problems you have? (+ = *no problems*)
Retrospective functioning (5, 1)	How is your memory compared to what it was . . . (a) 1 year ago? (+ = *much better*)
Frequency of forgetting (18, 2)	How often do these present a memory problem for you . . . (a) names? (+ = *never*)
Frequency of forgetting when reading novels (5, 1)	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, 1)	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 (4, 1)	How well do you remember things that occurred . . . (a) last month? (+ = *very good*)
Seriousness (18, 2)	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, 2)	How often do you use these techniques to remind yourself about things . . . (a) keep an appointment book. (+ = *never*)

Note. Based on Gilewski, Zelinski, Schaie, & Thompson, 1983. The first number in the parentheses after each dimension represents the total items; the second number represents the missing items allowed (with item mean substitution).

Table 3
Correlations Among Metamemory in Adulthood Scales for the Two Samples

Scale	1	2	3	4	5	6	7
1. Strategy	—	0.20	−0.17	−0.17	0.32	0.29	0.03
2. Task	0.08	—	0.10	0.02	0.10	0.15	0.02
3. Capacity	−0.07	0.00	—	0.65	−0.47	0.12	0.40
4. Change	−0.10	−0.15	0.62	—	−0.47	−0.05	0.42
5. Anxiety	0.20	0.16	−0.51	−0.57	—	0.34	−0.15
6. Achievement	0.28	0.28	0.16	−0.09	0.27	—	0.28
7. Locus	0.10	0.02	0.31	0.43	−0.25	0.25	—

Note. Correlations for the Annville sample are above the diagonal; those for the Victoria sample are below the diagonal.

Knowledge factors that were identified by Hertzog et al. (1987) would be replicated, (b) if the hypothesized MFQ Memory Self-Efficacy factor would be found, and (c) if the convergent Strategy and Change factors could be estimated. Second, the accepted model for the Annville data was replicated (cross-validated) in the Victoria adult sample. The Victoria student sample was excluded to avoid possible effects of pooling over a discontinuous age variable. Finally, a multiple-groups factor analysis was run on all participants to determine the age-related invariance in the joint factor structure of the two scales.

Results

Tables 3, 4, and 5 report zero-order correlations between the MIA and the MFQ scales in both samples. The MFQ Seriousness scale had near-zero correlations with all other scales and, hence, was eliminated from further factor analysis. An examination of the other correlations led to a reassessment of the hypothesis that we would obtain the Memory Knowledge factor that was identified by Hertzog et al. (1987). The zero-order correlations among the MIA Task, Strategy, Locus, and Achievement scales were lower than in the original Dixon and Hultsch (1983) validation samples. In fact, the low correlation of MIA Task with these MIA subscales and the MFQ Mnemonics subscale was a telling finding, for it indicated that simply stripping off the affect-related MIA scales, such as Achievement, from the factor would not result in a well-defined Knowledge factor.

This outcome made it necessary to explore a few alternative models that did not specify the Memory Knowledge factor. The revised model specified two Memory Self-Efficacy factors, one for each questionnaire, an MFQ Reading Self-Efficacy factor (marked by problems remembering 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 Functioning, and MIA Locus), and MIA Task (treated as a separate, single-indicator "factor"). During estimation of the revised model, we encountered a negative unique variance for MIA Strategy. This phenomenon, termed a *Heywood case,* is a familiar problem in latent variable modeling, especially when only two indicators are available for a factor (e.g., Van Driel, 1978). The residual variance for MIA Strategy was subsequently fixed to zero. Empirical identification of the Change factor, independent of Memory Self-Efficacy, was achieved by fixing correlations of Change with factors other than Memory Self-Efficacy to zero.

The model fit the Annville sample data well. The likelihood ratio $\chi^2(57, N = 415)$ was 91.44 ($p = .003$), but the LISREL goodness-of-fit index (GFI) was .970. Given that the model had been modified on the basis of fit to the Annville sample, this excellent level of fit may have been spurious. Therefore, we cross-validated the model in the Victoria older adult sample. Again, the fit was good: $\chi^2(57, N = 264) = 60.34, p = .36$ (GFI = .969). Given that the LISREL GFI indices exceeded .9 in both samples, it seemed clear that the significance of χ^2 in the Ann-

Table 4
Correlations Among Memory Functioning Questionnaire Scales for the Two Samples

Scale	1	2	3	4	5	6	7	8
1. General rating	—	0.24	0.39	0.26	0.28	0.25	0.04	−0.03
2. Retrospective functioning	0.20	—	0.22	0.07	0.10	0.19	0.02	−0.00
3. Frequency forgetting	0.45	0.29	—	0.56	0.57	0.55	0.11	−0.08
4. Forgetting novels	0.37	0.13	0.55	—	0.73	0.40	0.05	−0.03
5. Forgetting magazine	0.37	0.10	0.52	0.72	—	0.44	0.05	−0.05
6. Past events	0.41	0.24	0.62	0.50	0.44	—	0.06	−0.06
7. Mnemonics	0.08	−0.04	0.07	0.10	0.13	0.06	—	0.04
8. Seriousness	−0.15	−0.20	−0.24	−0.08	−0.09	0.04	−0.04	—

Note. Correlations for the Annville sample are above the diagonal; those for the Victoria sample are below the diagonal.

Table 5
Correlations Between the Metamemory in Adulthood (MIA) and the Mental Functioning Questionnaire (MFQ) Scales for the Two Samples

	MIA scales						
MFQ scales	Strategy	Task	Capacity	Change	Anxiety	Achievement	Locus
			Annville sample				
General rating	−0.12	0.04	0.35	0.35	−0.33	−0.12	0.12
Retrospective functioning	−0.09	−0.08	0.26	0.36	−0.18	0.01	0.25
Frequency forgetting	−0.20	0.10	0.67	0.56	−0.60	−0.07	0.26
Forgetting novels	−0.06	0.20	0.52	0.33	−0.39	−0.05	0.16
Forgetting magazines	−0.11	0.11	0.49	0.28	−0.41	−0.06	0.14
Past events	−0.14	0.03	0.58	0.43	−0.38	0.07	0.20
Mnemonics	−0.72	−0.18	0.05	0.04	−0.17	−0.20	−0.03
Seriousness	−0.03	0.02	−0.07	−0.02	0.03	−0.07	−0.03
			Victoria sample				
General rating	−0.05	0.04	0.35	0.39	−0.34	−0.05	0.22
Retrospective functioning	0.04	−0.04	0.27	0.38	−0.14	0.04	0.24
Frequency forgetting	−0.09	0.02	0.64	0.57	−0.54	0.03	0.32
Forgetting novels	−0.09	0.00	0.51	0.34	−0.40	−0.03	0.20
Forgetting magazines	−0.10	0.01	0.43	0.32	−0.38	−0.08	0.14
Past events	−0.11	−0.01	0.60	0.47	−0.37	0.11	0.28
Mnemonics	−0.70	−0.01	0.07	−0.02	−0.10	−0.17	−0.14
Seriousness	0.01	0.06	−0.13	−0.07	0.13	−0.05	−0.07

ville sample was a function of the larger sample size (and higher power).

Table 6 reports the factor loadings and associated standard errors from this model for both samples. The solutions were quite similar. Capacity had the highest loading on the MIA Memory Self-Efficacy factor, but Anxiety and Change displayed substantial loadings as well. MIA Locus loaded about equally on MIA Memory Self-Efficacy, Memory-Related Affect, and Change factors. Frequency of Forgetting produced the largest loading on MFQ Memory Self-Efficacy, followed by the Remembering Past Events and General Rating variables. The MFQ Reading Self-Efficacy factor was well-defined by the two MFQ Reading Frequency of Forgetting variables. The Memory-Related Affect factor was defined almost exclusively by MIA Achievement.

Table 7 reports the estimated factor correlations and standard errors for both samples. These factor correlations are disattenuated for measurement error, so they have a true possible range of −1 to 1 and may be larger than would commonly be expected on the basis of empirically obtained zero-order correlations among individual measures. Table 7 contains the crucial information regarding the question of convergent validity—the correlation between the MIA and the MFQ Memory Self-Efficacy factors. If it were the case that the scales from the two questionnaires both measure the same Memory Self-Efficacy construct, then the correlation between the two factors ought to be 1. For the Annville sample, the estimated correlation was .99 and was within one standard error of 1.0. For the Victoria sample, the 99% confidence interval's upper bound was .97, so the hypothesis that the correlation of the two Memory Self-Efficacy factors was 1.0 was rejected.

The hypothesis of convergence can be put to a formal likelihood ratio test by fixing the correlations between the two Memory Self-Efficacy factors to 1 and constraining the factor correlations between each Memory Self-Efficacy factor and all other factors to be equal (Jöreskog, 1974). The difference in chi-square, between this more restricted model and the model that freely estimates the correlation (with no equality constraints on other factor correlations), tested the null hypothesis that the two factors were equivalent. For the Annville sample, the restricted model yielded a chi-square ($df = 63, N = 415$) of 107.55. Therefore, the hypothesis test of equivalent factors was rejected, $\chi^2(6, N = 415) = 16.15, p < .05$. For the Victoria sample, the same difference was highly significant, $\chi^2(6, N = 264) = 30.50, p < .01$. Rejection of the equivalence hypothesis in the Annville sample was surprising, given the estimated correlation of .99 between the factors (see above). Inspection of the factor correlations in Table 7 suggested that questionnaire differences in

Table 6
Factor Loadings of Metamemory Scales for the Two Samples

	MSE_{MIA}				MSE_{MFQ}				MSE_{RD}			
	Annville		Victoria		Annville		Victoria		Annville		Victoria	
Scale	λ	*SE*	λ	*SE*	λ	*SE*	λ	*SE*	λ	*SE*	λ	*SE*
Strategy	0	—	0	—	0	—	0	—	0	—	0	—
Task	0	—	0	—	0	—	0	—	0	—	0	—
Capacity	84	05	84	06	0	—	0	—	0	—	0	—
Change	51	06	61	08	0	—	0	—	0	—	0	—
Anxiety	−63	04	−70	06	0	—	0	—	0	—	0	—
Achievement	0	—	0	—	0	—	0	—	0	—	0	—
Locus	25	06	28	08	0	—	0	—	0	—	0	—
General rating	0	—	0	—	45	05	54	06	0	—	0	—
Retrospective functioning	0	—	0	—	15	06	22	08	0	—	0	—
Frequency forgetting	0	—	0	—	85	04	85	05	0	—	0	—
Forgetting novels	0	—	0	—	0	—	0	—	86	04	88	06
Forgetting magazines	0	—	0	—	0	—	0	—	85	04	82	06
Past events	0	—	0	—	65	05	74	06	0	—	0	—
Mnemonics	0	—	0	—	0	—	0	—	0	—	0	—

	Strategy				Affect				Change			
	Annville		Victoria		Annville		Victoria		Annville		Victoria	
Scale	λ	*SE*	λ	*SE*	λ	*SE*	λ	*SE*	λ	*SE*	λ	*SE*
Strategy	1.0	—	1.0	—	0	—	0	—	0	—	0	—
Task	0	—	0	—	0	—	0	—	0	—	0	—
Capacity	0	—	0	—	24	07	31	08	0	—	0	—
Change	0	—	0	—	0	—	0	—	50	07	51	10
Anxiety	0	—	0	—	33	07	17	07	0	—	0	—
Achievement	0	—	0	—	85	07	94	09	0	—	0	—
Locus	0	—	0	—	34	05	30	06	34	07	34	09
General rating	0	—	0	—	0	—	0	—	0	—	0	—
Retrospective functioning	0	—	0	—	0	—	0	—	33	07	35	09
Frequency forgetting	0	—	0	—	0	—	0	—	0	—	0	—
Frequency novels	0	—	0	—	0	—	0	—	0	—	0	—
Forgetting magazines	0	—	0	—	0	—	0	—	0	—	0	—
Past events	0	—	0	—	0	—	0	—	0	—	0	—
Mnemonics	−72	04	−70	05	0	—	0	—	0	—	0	—

Note. MSE = Memory Self-Efficacy; MIA = Metamemory in Adulthood scale; MFQ = Memory Functioning Questionnaire; RD = Reading. Decimals are omitted for *SE*s. Fixed parameters have no standard errors, which is indicated by a dash in the *SE* columns. Lambdas denote factor loadings.

the correlation of Memory Self-Efficacy with Change may have been a primary source of the poorer fit of the factor-equivalence model. Either (a) the two factors are correlated, but different, or (b) the scales are, indeed, measuring the same construct, except for the influence of questionnaire-specific sources of systematic measurement error (e.g., method variance). Given the high correlations between the factors in both samples, the latter alternative seemed more plausible.

Other factor correlations reported in Table 7 merit discussion. Change correlated significantly with both Memory Self-Efficacy factors, and in the expected direction (high Change with low Memory Self-Efficacy). The MFQ Reading Self-Efficacy factor correlated highly with both Memory Self-Efficacy factors. Strategy had a modest but significant correlation with Memory Affect. The correlation of the other factors with Memory Self-Efficacy was relatively low. In particular, the MIA Task variable had nonsignificant correlations with both Memory Self-Efficacy factors.

Table 7
Factor Correlations in the Two Samples

Factor	1	2	3	4	5	6	7
1. MSE_{MIA}	—	99 (03)	72 (04)	−29 (06)	−11 (11)	05 (06)	49 (08)
2. MSE_{MFQ}	87 (04)	—	77 (03)	−22 (05)	−08 (07)	11 (05)	31 (07)
3. MSE_{RD}	67 (05)	75 (04)	—	−09 (05)	−07 (06)	19 (05)	0 (—)
4. Strategy	−22 (07)	−12 (07)	−11 (07)	—	34 (05)	20 (05)	0 (—)
5. Affect	−17 (10)	−06 (07)	−06 (07)	30 (06)	—	19 (06)	0 (—)
6. Task	−12 (07)	04 (07)	01 (07)	08 (06)	30 (06)	—	0 (—)
7. Change	34 (11)	31 (09)	0 (—)	0 (—)	0 (—)	0 (—)	—

Note. MSE = Memory Self-Efficacy; MIA = Metamemory in Adulthood scale; MFQ = Memory Functioning Questionnaire; RD = Reading. Correlations for the Annville sample are above the diagonal; those for the Victoria sample are below the diagonal. Decimals are omitted for all values. Standard errors appear in parentheses. Dashes denote fixed parameters.

Multiple-Age-Group Analyses

The next phase consisted of a simultaneous, multiple-groups analysis that enabled us to test the hypothesis of age differences in factor loadings. The two samples were each divided into two age groups. We split the Annville sample into an older group (56 to 78 years) and a younger group (20 to 55 years). In the Victoria sample, the student group was analyzed separately from the older adults. Thus, there were four age groups for the analysis. We analyzed covariance matrices so that between-groups tests of invariance in factor pattern weights would be meaningful (Jöreskog, 1971; Schaie & Hertzog, 1985).

The first model used the same specification reported above for all four age groups. It placed no between-groups constraints on the model parameters. The second model tested the hypothesis that all age groups had equivalent factor loadings by constraining this matrix equally across the four groups. As can be seen from the first model comparison in Table 8, Model 2 did not fit significantly worse than the Model 1, implying that the groups did not differ significantly in factor loadings. This result was somewhat surprising, for it disagreed with previous work where we found significant age group differences in the loadings of the MIA scales on a Memory Self-Efficacy factor (Hertzog et al., 1987). Model 3 suggested an explanation of the discrepancy. It forced age-group equivalence in the factor covariance matrices. Comparison 2 in Table 8 shows that this model fit significantly worse than the model with constraints on factor pattern weights alone (Model 2). We then tested the hypothesis that these differences were associated primarily with the Change factor by examining the group equivalence in Change factor variances and the covariances of Change with the Memory Self-Efficacy factors. Model 4, allowing these variances and covariances to differ across groups, fit better than Model 3. Thus, we concluded that there were significant age-group differences in the relationship of the Change factor to the Memory Self-Efficacy factor. As might be expected, Change covaried more highly with Memory Self-Efficacy in the old groups. Inspection of the remaining factor variances from Model 2 suggested that the other significant differences were primarily associated with the factor variances in the Victoria younger group. They were much less variable in all three self-efficacy factors, but more variable in Strategy and Task.

Table 9 reports the rescaled factor correlations for the four different age groups. Given that the Annville sample spanned a larger age range than the Victoria adult sample, it was possible that the near-perfect correlation between the two Memory Self-Efficacy factors had been produced by age heterogeneity. As can be seen in Table 9, the factor correlations in the two Annville groups remained quite high. In general, the magnitude of correlations was similar across the four groups and consistent with those reported in Table 7 for the entire sample, with the exception of the Change factor correlations with Memory Self-Efficacy. The correlations involving MIA Memory Self-Efficacy and MFQ Memory Self-Efficacy ranged from .13 to .64 and from .11 to .49, respectively. Although there was a trend for the older groups to have higher correlations than the two younger groups, it was also the case that the correlations in the Annville old group were greater than the Victoria old group.

Age and Sex Differences in Metamemory Factors

As noted above, the analysis by Hultsch et al. (1987) suggested that the MIA and the MFQ scales show differential sensitivity to chronological age differences, with the MIA being more likely to show age differences than the MFQ. This result seems surprising in light of the high estimated correlation in this study between the two Memory Self-Efficacy factors. Although models specifying perfect 1.0 correlations between the two factors had been previously rejected, correlations in the high .8 to middle .9 range would seem, on the surface, to predict similar mean patterns. As a possible reconciliation, we examined the possibility that an MFQ Method factor caused the less-than-perfect correlation of the two Memory Self-Efficacy factors and that this source of method variance also caused the discrepant pattern of age relationships between the MIA and the MFQ scales that was found by Hultsch et al. (1987). We reasoned that the weak age differences on MFQ scales (such as Frequency of Forgetting) was caused by combining negative age differences (old lower than young) in Memory Self-Efficacy with positive age differences on the MFQ Method factor. Given that Hultsch et al. (1987) had already analyzed the data for mean age differences, this analysis did not constitute an a priori hypothesis test. Nevertheless, the ex post facto hypothesis would be supported if the following could be shown: (a) that a common Memory Self-Efficacy factor, including both MIA and MFQ scales, could be successfully estimated; (b) that a MFQ method factor could

Table 8
Goodness-of-Fit Statistics for Multiple-Groups Factor Analyses

Model	χ^2	*df*	*p*	δ^a
1. No group constraints	287.97	224	.003	.936
2. Equal factor loadings	324.33	263	.006	.928
3. Equal factor loadings and equal factor covariance matrices	436.10	335	.000	.903
4. Equal factor loadings and equal factor matrices (except Change factor)	412.46	326	.001	.908

Model comparisons	Source[b]	χ^2	*df*	*p*
1. H_0: Equal factor loadings	2 and 1	36.36	39	*ns*
2. H_0: Equal factor covariance matrices	3 and 2	111.77	72	<.01
3. H_0: Equal variances and covariances for Change factor	3 and 4	23.64	9	<.01

Note. H_0 = null hypothesis.
[a] Bentler–Bonett normed fit index. [b] Two models used to calculate differences in χ^2.

be identified independent of this Memory Self-Efficacy factor; (c) that this specification gave a good fit to the covariances among the MIA and the MFQ scales, as well as to their covariances with Age; and (d) that there would be large age differences on Change and Memory Self-Efficacy, but reversed age differences on the MFQ Method factor. The model specification forced the MFQ Method factor to be uncorrelated with the other metamemory factors. To identify the MFQ Method factor, it was also necessary to maintain the MFQ Reading Self-Efficacy factor as distinct from general Memory Self-Efficacy, although it was expected that its correlation with Memory Self-Efficacy would be high.

We ran a structural regression model on the factor solution in the Annville data, where the discrepant age differences had been most pronounced (Hultsch et al., 1987). The regression model used Age and Sex as independent variables, with the metamemory factors serving as dependent variables. The polynomial regression analysis reported by Hultsch et al. (1987) produced linear age relationships to the MIA Capacity, Change, and Locus scales, so the regression approach that we used here was adequate to represent the age trends. Specification was based on the results reported by Hultsch et al. (1987). Age predicted Memory Self-Efficacy, the MFQ Method factor, MFQ Reading Self-Efficacy, and Change. Gender predicted the Strategy Use factor. On the basis of the modification indices, we added Age and Gender effects on the MIA Anxiety residual. This final regression model fit the data about as well as the original factor model omitting Age and Gender, $\chi^2(77, N = 415) = 124.81$ (GFI = .964).

Table 9
Metamemory Factor Correlations in Four Groups

Factors	Victoria (55–78 years)	Annville (56–78 years)	Annville (20–55 years)	Victoria (20–26 years)
MSE_{MIA}, MSE_{MFQ}	88	96	1.04[a]	84
MSE_{MIA}, MSE_{RD}	67	73	74	59
MSE_{MFQ}, MSE_{RD}	74	75	81	78
MSE_{MIA}, Strategy	−19	−37	−26	−18
MSE_{MFQ}, Strategy	−12	−21	−23	−15
MSE_{RD}, Strategy	−11	−07	−12	−14
MSE_{MIA}, Affect	−10	−21	−13	−17
MSE_{MFQ}, Affect	09	−15	−01	11
MSE_{RD}, Affect	−06	−16	01	−03
Strategy, Affect	32	29	27	38
MSE_{MIA}, Task	09	10	00	10
MSE_{MFQ}, Task	05	20	06	19
MSE_{RD}, Task	01	23	14	24
Strategy, Task	08	09	27	12
Affect, Task	32	18	18	27
MSE_{MIA}, Change	37	64	34	13
MSE_{MFQ}, Change	32	49	17	11

Note. MSE = Memory Self-Efficacy; MIA = Metamemory in Adulthood scale; MFQ = Memory Functioning Questionnaire; RD = Reading. Decimals are omitted.
[a] Estimated covariance, when rescaled, was greater than 1.0.

Table 10
Unstandardized Factor Loadings for Memory Self-Efficacy and Memory Functioning Questionnaire Method Factors

	MSE		MFQ method		MSE_{RD}		Change	
Factor	λ	*SE*	λ	*SE*	λ	*SE*	λ	*SE*
Capacity	1.0	—	0	—	0	—	0	—
Change	0.700	0.074	0	—	0	—	1.0	—
Anxiety	−0.716	0.047	0	—	0	—	0	—
Locus	0.165	0.034	0	—	0	—	0.183	0.055
General rating	0.067	0.007	0	—	0	—	0	—
Retrospective functioning	0.123	0.043	0.036	0.444	0	—	0.202	0.079
Frequency forgetting	1.504	0.076	1.0	—	0	—	0	—
Forgetting novels	0	—	0.693	0.538	1.0	—	0	—
Forgetting magazines	0	—	0.704	0.493	0.893	0.052	0	—
Past events	0.381	0.027	0.270	0.091	0	—	0	—
Mnemonics	0	—	0.430	0.174	0	—	0	—

Note. Dashes denote fixed parameters. MSE = Memory Self-Efficacy; MFQ = Memory Functioning Questionnaire; RD = Reading. Factor loadings that exceed their standard errors by a ratio of 2:1 or greater are different from zero at approximately a 95% level of confidence.

Table 10 reports the factor loadings for the Memory Self-Efficacy and MFQ Method factors. Table 11 reports the regression coefficients for Age and Gender. The MFQ Method loadings were small but significant for several variables, and the regression of MFQ Method on Age was significant and positive. Conversely, the regression of Memory Self-Efficacy on Age was significant and negative. The strongest relationship was for Age and Change, although in this case the relationship is independent of the relationship of Age to Memory Self-Efficacy (where the MIA Change scale also loads). The fit to the covariances of the metamemory scales with Age was good, as reflected in low normalized residuals. This pattern indicated that it was reasonable to model age differences in the metamemory scales in terms of age differences in the Memory Self-Efficacy and Change factors. The one exception was MIA Anxiety, given the significant regression of the Anxiety residual on Age. In sum, this model lends credence to the notion that there is a unique component to the MFQ scales that attenuates the observed age differences on these scales and produces the slight attenuation in the correlation between the MFQ Memory Self-Efficacy with MIA Memory Self-Efficacy factors.

Discussion

This study found that, despite differences in conceptualization, wording, and response format between the two questionnaires, the Memory Self-Efficacy factor in the MIA that was found by Hertzog et al. (1987) and the Frequency of Forgetting factor in the MFQ that was found by Gilewski et al. (1983) correlate nearly perfectly when the two scales are factored simultaneously. This finding, replicated in two independent samples, supports the hypothesis that scales from both questionnaires are measures of the Memory Self-Efficacy construct as discussed by Cavanaugh et al. (1985), by West et al. (1987), and by us (Hertzog et al., 1987; Hultsch et al., 1985). There was a small, but statistically significant, loss of fit when the two factors were forced to be equivalent by setting their correlation to be 1.0. However, we were able to fit successfully a model that loaded scales from the two questionnaires on a single Memory Self-Efficacy factor, provided that a separate MFQ Method factor was modeled that had an inverse relationship to chronological age.

How might the MFQ Method factor that we identified be determined? One possibility is that this factor reflects a tendency by older persons to be less willing to endorse items that indicate memory problems. For example, the reverse age differences on the MFQ Method factor may represent a cohort difference in willingness to complain that is correlated with lower self-ratings of Neuroticism (see Costa & McCrae, 1980). Another possibility is that, regardless of item content, older individuals are loathe to use extremes of frequency rating scales. Moreover, we cannot rule out the possibility that the Method factor arose because of reactive effects of responding to the mood state questionnaires prior to the MFQ. Nevertheless, the success of the model with Memory Self-Efficacy and the MFQ Method factor in accounting for age differences in the MIA and the MFQ scales enhances the argument for convergent validity. That is, one cannot argue that the two scales do not measure the same construct solely because they show differential sensitivity to age (Gilewski & Zelinski, 1986).

Cavanaugh and Poon (in press) reported zero-order correlations between the SIME Forgetting scale and the MIA Capacity and Change scales that were similar in magnitude to the zero-order correlations of the MFQ Frequency of Forgetting scale and the MIA scales reported in Table 5. Therefore, it is

Table 11
Regression Coefficients of Metamemory Factors on Age and Sex

Dependent variable	Independent variable: Age			Sex		
	β	*SE*	β^a	β	*SE*	β^a
MSE	−0.115	0.034	−0.182	0	—	0
MFQ Method	0.151	0.039	0.621	0	—	0
MSE_{RD}	0.148	0.086	−0.390	0	—	0
Strategy	0	—	0	−1.825	0.403	−0.201
Achievement	0	—	0	0		0
Task				0		0
Change	−0.223	0.032	−0.400	0	—	0
Anxiety (residual)	−0.078	0.027	−0.167	−1.480	0.335	−0.227

Note. MSE = Memory Self-Efficacy; MFQ = Mental Functioning Questionnaire; RD = Reading. β denotes LISREL estimate. Estimates exceeding their standard errors by a ratio of 2:1 or greater are significantly different from zero at a 95% level of confidence.
[a] These values are standardized estimates.

tempting to speculate that joint analysis of the three questionnaires would also show that a general Memory Self-Efficacy construct operates in all three instruments.

As discussed earlier, there are limitations to the use of zero-order correlations in making inferences regarding convergent validity. In the present study, we were able to demonstrate that the observed correlations between measures of memory self-efficacy reflected a near-perfect relationship between the underlying latent variables, although that need not have been the case. The advantage of the modeling procedures that we used here is that the convergence hypothesis is evaluated on the basis of the pattern of relationships among the entire set of metamemory scales. Our model fit this pattern well. Similarly, a strong test of convergent validity of the SIME, MIA, and MFQ requires evaluation of the factor equivalence hypothesis.

We also found evidence for convergent validity of MIA and MFQ measures of strategy use and perceived change in memory. Scales from both questionnaires had significant loadings on Strategy and Change factors. However, there are qualifications regarding convergence for these scales. In the case of the Strategy factor, there was a problem with a negative unique variance estimate for the MIA Strategy scale that was addressed by fixing the unique variance to zero. This Heywood case was probably a function of the fact that the MIA Strategy scale covers both external memory aids (e.g., use of lists and calendars) and mnemonic strategies (e.g., rehearsal and use of imagery), whereas the MFQ Mnemonics scale measures primarily use of external aids.[2] Convergent validity of the MIA and MFQ measures of perceived change was evident in the pattern of zero-order correlations in Table 5, but was ambiguous given the low (less than .4) correlations of MIA Change and MFQ Retrospective Functioning in both samples. The significant loading of Retrospective Functioning on the Change factor, however, permits the inference of convergence, qualified by the additional conclusion that the validity of Retrospective Functioning for measuring the Change construct is relatively modest.

The correlations among metamemory factors support the conceptualization that there are multiple dimensions of metamemory (Dixon, in press; Dixon & Hultsch, 1983). Across samples, there was a consistently small, negative relationship between Memory Self-Efficacy and Strategy Use, with individuals low in self-efficacy likely to report more strategy use. Both Memory Self-Efficacy factors had virtually zero correlations with the MIA Task scale. This latter finding supports the expectation that it is possible to differentiate knowledge about how memory functions from memory self-efficacy beliefs. It should be noted, however, that we found no evidence for a higher-order memory knowledge factor in the MIA and the MFQ scales. Given Hertzog et al. (1987), we had expected to identify such a memory knowledge factor involving MIA Task, MIA Strategy, and MFQ Mnemonics. The correlations among these scales were lower in the present study than in Hertzog et al. (1987). Therefore, it now appears the MIA may contain only a single, robust indicator of the memory knowledge dimension (i.e., Task).

The analysis of multiple age groups showed that the Memory Self-Efficacy factor was defined equivalently, in terms of factor loadings, across age levels. This finding suggests that quantitative comparisons of derived Memory Self-Efficacy scores from the MIA or the MFQ across age are appropriate and justified (Labouvie, 1980; Schaie & Hertzog, 1985). The results also showed that the variances and covariances for the Change factor differed across groups, with the highest correlation of Change and Memory Self-Efficacy in the Annville sample's old group. The finding of equivalent factor loadings suggests that the previous finding by Hertzog et al. (1987) of differences in Memory Self-Efficacy loadings across age groups actually was a function of age differences in the correlation of the Change factor with the Memory Self-Efficacy factor. In the Hertzog et al. (1987) analysis, only the MIA scales were available, and, consequently, MIA Change loaded only on the MIA Memory Self-Efficacy fac-

[2] An additional analysis showed that redefining the MIA Strategy scale to include only the external-aids items eliminated the Heywood case. Because the new solution did not alter the general pattern of results, we do not present it here.

tor. In the present analysis, the availability of the MFQ Retrospective Functioning scale enabled us to find that Change is defined equivalently in all age groups. As individuals grow older and perceive more change in memory, however, they are also more likely to report lower memory self-efficacy, resulting in an increased correlation between these factors. Some caution in interpretation is warranted here, however, because there appeared to be differences between the two old groups in the magnitude of this factor correlation.

What are the implications of the present results for metamemory scales and questionnaires? The convergent validity of the MIA and the MFQ in measuring Memory Self-Efficacy reinforces earlier recommendations (Dixon, in press; Gilewski & Zelinski, 1986) that use of either questionnaire seems appropriate. Given that there is not complete overlap in subscales, the selection might be made on the basis of matching specific research questions to the available scales (e.g., the MIA explicitly measures affect regarding memory that is not assessed by the MFQ). However, if the goal is to obtain multiple indicators of Memory Self-Efficacy, it is advantageous to use scales from both the MIA and the MFQ. Use of the MIA by itself is limited by the fact that two of the three best indicators of Memory Self-Efficacy in that scale are Change (which also relates, obviously, to the perceived Change factor) and Anxiety (which is arguably a distal outcome of low self-efficacy beliefs; see Bandura, 1986). Use of the MFQ by itself may be limited, given the results suggesting an MFQ Method factor that prevented perfect convergence of the two Memory Self-Efficacy factors. Therefore, using the MFQ scales alone might absorb a construct-irrelevant component of variance into the latent variable. Considering also the benefits of breadth of definition of the Memory Self-Efficacy construct, a good latent variable could be formed by using MIA Capacity and MFQ Frequency of Forgetting (and, perhaps, a third indicator, such as MFQ General Rating or MFQ Remembering Past Events).

In conclusion, we have reported strong evidence in favor of the hypotheses (a) that the MIA and the MFQ questionnaires converge to measure a construct that we have labeled *Memory Self-Efficacy,* and (b) that this factor has a very modest relationship to Memory Knowledge, at least as measured by the MIA Task scale. As discussed above, convergent validity is only one aspect of construct validation. We are currently conducting analyses of (a) the discriminant validity of multiple dimensions of metamemory from related constructs (e.g., general self-efficacy, personality, and affective states) and (b) the predictive validity of the Memory Self-Efficacy factor for actual memory performance. The self-efficacy perspective argues that predictive validity of Memory Self-Efficacy for memory performance is an empirical question rather than a criterion for judging the construct validity of metamemory questionnaires. To be sure, modest predictive validity may argue against use of metamemory scales as proxies for memory assessment in clinical settings (e.g., Sunderland et al., 1986) but does not necessarily imply that metamemory questionnaires are invalid. It is possible that the predictive validity of Memory Self-Efficacy for actual memory performance, although probably a function of a number of factors (Dixon & Hertzog, 1988), is limited because self-efficacy beliefs are not necessarily veridical. In particular, the negative self-efficacy beliefs of some older individuals may be inaccurate. Indeed, as Zarit (1982) suggested, one major benefit of memory training with the elderly may be to improve memory self-efficacy, even if objective memory performance and effective use of memory skills in everyday life do not improve.

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling practice: A review and recommended two-step approach. *Psychological Bulletin, 103,* 411–423.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice Hall.

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 findings and methodological issues* (pp. 267–302). Washington, DC: Hemisphere.

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

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Cavanaugh, J. C., Kramer, D. A., Sinnott, J. D., Camp, C. J., & Markley, R. P. (1985). On missing links and such: Interfaces between cognitive research and everyday problem solving. *Human Development, 28,* 146–168.

Cavanaugh, J. C., & Poon, L. W. (in press). Patterns of individual differences in secondary and tertiary memory performance. *Psychology and Aging.*

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

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

Costa, P. T., Jr., & McCrae, R. R. (1980). Influence of extraversion and neuroticism on subjective well-being: Happy and unhappy people. *Journal of Personality and Social Psychology, 38,* 668–678.

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

Curran, J. P., & Cattell, R. B. (1976). *Handbook for the eight-state questionnaire.* Champaign, IL: Institute for Personality and Ability Testing.

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., & Hertzog, C. (1988). 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* (pp. 293–330). Hillsdale, NJ: Erlbaum.

Dixon, R. A., & Hultsch, D. F. (1983). 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.

Gilewski, M. J., & Zelinski, E. M. (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 91st Annual Meeting of the American Psychological Association, Anaheim, CA.

Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. *Psychological Review, 93,* 258–268.

Hayduk, L. A. (1987). *Structural equation modeling with LISREL: Essentials and advances.* Baltimore, MD: Johns Hopkins University Press.

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

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). New York: Academic Press.

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

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

Hultsch, D. F., Hertzog, C., & Dixon, R. A. (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.

Hultsch, D. F., Hertzog, C., Dixon, R. A., & Davidson, H. (1988). In M. L. Howe & C. J. Brainerd (Eds.), *Cognitive development in adulthood: Progress in cognitive development research* (pp. 65–92). New York: Springer.

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

Jöreskog, 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.

Jöreskog, K. G., & Sörbom, D. (1984). *LISREL VI user's guide.* Mooresville, IN: Scientific Software.

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

Labouvie, E. W. (1980). Identity versus equivalence of psychological measures and construct. In L. W. Poon (Ed.), *Aging in the 1980s: Psychological issues* (pp. 493–502). Washington, DC: American Psychological Association.

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.

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

Long, J. S. (1983). *Confirmatory factor analysis.* (Sage University Paper Series on Quantitative Application in the Social Sciences, Series No. 07–033). Beverly Hills, CA: Sage.

McNair, D. M., Lorr, M., & Droppelman, L. F. (1971). *Profile of mood states.* San Diego, CA: Educational and Industrial Testing Service.

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

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stillwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105,* 430–445.

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

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.

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. (1969). *The State–Trait Anxiety Inventory (STAI) test manual, Form X.* Palo Alto, CA: Consulting Psychologists 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.

Sunderland, A., Watts, K., Baddeley, A. D. M., & Harris, J. E. (1986). Subjective memory assessment and test performance in elderly adults. *Journal of Gerontology, 41,* 376–384.

Van Driel, O. P. (1978). On various causes of improper solutions in maximum likelihood factor analysis. *Psychometrika, 43,* 225–243.

West, R. L., Berry, J. M., & Dennehy, D. (1987). *Self-efficacy and memory performance: Measurement issues.* Paper presented at the 95th annual convention 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.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait–multimethod data. *Applied Psychological Measurement, 9,* 1–26.

Wyer, R. S., Jr., & Srull, T. K. (1986). Human cognition in its social context. *Psychological Review, 93,* 322–359.

Zarit, S. H. (1982). Affective correlates of self-reports 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 tests relate to self-assessment 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 direction in memory and aging: Proceedings of the George A. Talland Memorial Conference* (pp. 519–544). Hillsdale, NJ: Erlbaum.

Received April 19, 1988
Revision received December 13, 1988
Accepted January 16, 1989 ■

Measurement of Affective States in Adults

Evaluation of an Adjective Rating Scale Instrument

PAUL D. USALA
CHRISTOPHER HERTZOG
Georgia Institute of Technology

A cross-sectional sample of adults, ages 20-79, were administered a adjective rating scale instrument measuring multiple affective states, including items from the Profile of Mood States (POMS) instrument. Confirmatory item factor analysis supported, for the most part, a priori assignments of items to scales based upon prior research, but revealed a few small, additional item factors that were cross-validated in a second sample. Items measuring different aspects of psychological distress, including anxiety and depressive affect, showed appreciable skew and kurtosis, with a substantial proportion of respondents indicating no perceived distress. Items measuring psychological well-being tended to show more normal response distributions. Tests of age-related invariance in item factor structure indicated that the unstandardized factor pattern weights (loadings) were not fully equivalent across two age groups, but showed that the same configuration of items loading on factors was supported. The scales perform well enough to justify continued use in older populations, but further research on the contributions of item distributions to age differences in factor loadings is needed.

MOOD STATE AND THE ELDERLY

The relationship of aging, depression, and related affective disorders has been increasingly studied in recent years, with advances in our understanding of its epidemiology, correlates, and clinical manifestations (e.g., Gallagher 1986; Zarit et al. 1985). In contrast, relatively little is currently known about the relationship between normal variations in mood and affect in older populations, and the role affective states play in influencing behavioral patterns (Schulz 1985).

RESEARCH ON AGING, Vol. 11 No. 4, December 1989 403-426

The most frequently researched question regarding the affective status of older persons is whether they are happy and satisfied with life, especially when compared to younger adults (e.g., Cameron 1975; Campbell et al. 1976; George 1981; Lawton 1983). Most of the research has focused specifically on the elderly and their adjustment to aging. Studies of life satisfaction, morale, happiness, and subjective well-being, using instruments such as Neugarten's Life Satisfaction Index (Hoyt and Creech 1983; Wilson et al. 1985), the Bradburn Affect Balance Scale (Lawton 1975), and the Philadelphia Geriatric Center Morale Scale (Lawton 1975; Morris and Sherwood 1975), have consistently found significant correlations between well-being and health, socioeconomic status, social activity, and other variables for persons 60 and older (Larson 1978; Lawton 1983).

Some studies have also suggested age-related decreases in endorsement of items reflecting feelings of well-being and happiness (e.g., Bradburn and Caplovitz 1965; Clark et al. 1981). However, Cameron (1975) conducted a large-scale investigation of positive moods in various settings, and found that age did not predict variation in mood. One interesting possible explanation of the discrepancy is that, unlike more traditional measures of perceived well-being, Cameron explicitly limited his questions to current or very recent affective state (e.g., "How would you characterize your mood over the last half hour?").

There is ample evidence that affective states can be differentiated from more enduring dispositions to experience specific affect across different situations (affective states; Fridhandler 1986; Nesselroade and Bartsch 1977). In the psychometric tradition, many personality traits have a state analogue, as with trait and state anxiety (Cattell 1966; Spielberger 1966). Trait anxiety refers to the disposition of an individual to be anxious in multiple environmental contexts. State anxiety, in contrast, reflects the degree of anxious affect actually experienced in a given situation, which may not be perfectly predictable from the degree to which an individual usually experiences anxiety in that type of situation, or more generally, within the normal range of anxiety-eliciting conditions.

The issue of time frame for assessing affective status is relevant for measurement of both affective disorders and nonclinical variations in

mood. For example, some screening measures for depression use an explicit time frame (e.g., the CES-D, Radloff 1977, asks for symptom endorsement over a one-week period) but others do not. Gurland (1976), in reviewing the early research on depression and aging, noted a discrepancy between estimated incidence rates for clinically diagnosed depressions and rates of depressive symptoms in older populations. The pattern of decreasing depressive affect from young adulthood through age 65, with an increase in depressive symptoms in late life, has been suggested by a number of epidemiological studies (e.g., Ensel 1986; Murrell et al. 1983). Gurland (1976) noted that rates of psychiatric diagnoses of depression did not mirror the late-life increase in depressive symptoms, and cited the hypothesis that depressive symptoms in the elderly may be transient and recurrent, thus often not leading to clinical diagnosis and hospitalization.

In general, moods and emotions of older persons may significantly influence their behavior in particular situations. For example, cognitive gerontologists have been concerned for some time that older persons' anxiety in assessment situations may lead to invalid inferences about age changes in cognition (Botwinick 1984). Affective states are not necessarily just a methodological nuisance; understanding their influences may help clarify relationships between age and multiple attributes, such as metacognition and cognitive performance (Cavanaugh, forthcoming; Dixon and Hertzog 1988). For example, several studies have reported significant correlations between state anxiety and perceived self-efficacy regarding memory in older samples (Cavanaugh and Murphy 1986; West et al. 1984). Rodin et al. (1988) demonstrated a relationship between depressive affect and the presence of certain sleep disturbances in the elderly.

MEASUREMENT ISSUES

Studies specifically demonstrating reliability and validity of state measures in older populations are a logical prerequisite to use of such scales in gerontological research. Nesselroade et al. (1984) administered the state subscale of the Spielberger State-Trait Anxiety Inventory (STAI), alternate forms of the precursor to the Cattell 8-SQ

Anxiety subscale, and four other subscales of the 8-SQ (stress, regression, depression, and fatigue) to 111 persons from a senior citizens club. With two occasions of measurement (two to four weeks apart), high internal consistency and only moderate stability of individual differences was found. Moderate stability may actually be a favorable finding, given that one plausible criterion for construct validity of a state measure is that it demonstrate reliable fluctuation (lability) over time, at least under conditions where variation in state is expected. In addition, measurement equivalence over time for the fatigue and anxiety measures was indicated by the equivalence of state measures' factor loadings, estimated by maximum likelihood confirmatory factor analysis. A further analysis of these data by Hertzog and Nesselroade (1987) demonstrated that the two 8-SQ measures had equal within-occasion factor loadings and error variances, establishing them as parallel forms (Jöreskog 1971). The analysis also demonstrated that the forms maintained invariant error variances across occasions, indicating little reactive effects on measurement properties at the second occasion. Such findings support the argument that state anxiety can be reliably and validly measured in older populations.

One prerequisite to using a psychological instrument to compare and contrast different populations is the demonstration of measurement equivalence of the instrument across such groups (Drasgow and Kanfer 1985; Schaie and Hertzog 1985). Studies like Nesselroade et al. (1984) support the utility of the affective state measures in an older population, but one cannot, in general, assume that the affective state measures have equivalent reliability and validity in younger and older populations (e.g., Labouvie 1980; Schaie and Hertzog 1985). The validity of comparative studies of affective states across age groups can be assured only by demonstration of invariant measurement properties via simultaneous analysis of item properties in multiple-age groups, showing that items are similarly meaningful in their representations of the latent constructs (Hertzog 1987; 1989).

The present study investigates the factor structure of an adjective rating scale for multiple mood states in an age-heterogeneous sample. It addresses the following questions: Do the item factors converge to the same mood dimensions for such samples? Would the subscale item-scoring designations need to be revised?

Method

SUBJECTS

The principal sample consisted of 447 volunteers who were each paid $15 for their participation. They were obtained through a large family medical practice in the semi-urban community of Annville, Pennsylvania during the summer and fall of 1985. The subjects' ages ranged from 20-78, although the 20-35 age range was underrepresented in the sample. The sample was, like its parent population, heterogeneous in education, socioeconomic status, and vocabulary test scores. Participants' self-rated health was above average. Analyses were based on cases with complete item data on the mood questionnaire, resulting in a reduced sample of 352 subjects.

Data for an independent cross-validation sample, collected in 1987, was composed of additional subjects assessed at the time of a longitudinal retesting for the larger study. Subject enlistment procedures were similar to 1985, and measurement scales and procedures were the same. Complete item data was found for 287 of these subjects. Although the gender distribution was nearly identical to the 1985 sample, the age distribution differed somewhat, with the 1987 sample generally being younger. Other demographic characteristics of the two samples were comparable. Table 1 reports relevant sample sizes and demographic statistics in specified age brackets.

MEASURES

The scales measuring affective states were drawn from a larger study designed to investigate the construct validity of self-reported metamemory questionnaires (Hertzog et al., forthcoming). The mood adjective rating scale questionnaire was constructed from the 65 adjectives measuring anxiety, fatigue, depression, vigor (energy), and psychological well-being. In all, 38 of the items had been used in the POMS and 27 adjectives had been used by Lebo and Nesselroade (1978) in their Pregnancy Mood Checklist (PMcL). Table 2 shows the original source and order of administration for each adjective. All 10 well-being adjectives were from the Lebo and Nesselroade study; the other subscales contain item descriptors from both sources. These

TABLE 1
Mean Education, Vocabulary, and Self-Rated Health
(standard deviations in parentheses)

Age	Sex	N	Education	Health[a]	Vocabulary
			1985 Sample		
20-29	M	13	13.69(3.47)	7.00(1.15)	27.77(7.69)
	F	12	13.33(2.19)	7.50(1.31)	30.75(7.41)
30-39	M	22	16.50(3.42)	7.27(1.58)	37.48(7.88)
	F	34	14.47(2.44)	7.56(1.28)	36.29(7.95)
40-49	M	28	13.86(3.49)	7.14(1.56)	32.67(10.99)
	F	46	13.50(2.82)	6.98(1.69)	32.87(10.19)
50-59	M	24	13.70(3.14)	6.96(1.68)	34.46(10.19)
	F	46	12.40(2.04)	7.37(1.25)	34.02(9.32)
60-69	M	47	13.96(3.04)	6.63(1.72)	35.66(9.66)
	F	43	14.02(2.95)	7.12(1.69)	38.60(7.41)
70-79	M	18	12.94(2.96)	7.12(1.32)	35.67(10.83)
	F	19	13.89(2.11)	6.68(2.19)	40.79(7.03)
			1987 Sample		
20-29	M	18	12.47(1.84)	6.88(1.32)	24.44(9.09)
	F	20	13.25(2.10)	7.00(1.81)	27.80(9.58)
30-39	M	34	13.74(3.17)	7.45(1.35)	29.24(9.08)
	F	43	14.40(2.37)	7.60(1.53)	35.52(6.91)
40-49	M	22	14.55(3.35)	7.29(1.27)	34.18(7.04)
	F	28	14.25(3.16)	7.64(1.10)	34.89(8.11)
50-59	M	12	14.58(2.71)	7.18(1.08)	34.92(10.27)
	F	23	13.23(2.29)	7.27(1.20)	37.82(8.35)
60-69	M	17	13.24(2.68)	7.76(1.03)	35.88(10.22)
	F	31	13.13(3.60)	6.27(1.96)	33.13(11.25)
70-79	M	20	12.40(3.27)	7.32(1.34)	35.32(10.28)
	F	15	12.07(2.58)	6.60(2.13)	33.20(9.07)
80-89	M	2	19.00(1.41)	8.50(0.71)	49.50(4.95)
	F	2	18.00(1.41)	6.50(3.54)	42.00(12.73)

a. Rated on a 9-point Likert scale (OARS ladder), with 9 being excellent health.

subscales were chosen because of their relevance for the construct validation of metamemory. Subjects were asked to rate the extent to which each adjective reflected their current mood on a 5-point Likert scale. The instructions were as follows:

Below is a list of words that describe feelings people have. Please read each one carefully. We would like you to decide how well each word describes your feelings at this moment. Don't answer according to how you usually feel, but rather how you feel right here and now. Read each word, and then circle the letter on the right that best describes how you are feeling now.

PROCEDURE

The questionnaire was administered as part of a larger assessment battery made up of measures for metamemory, social desirability, personality, affective states, personal control, memory performance, and vocabulary. Small groups of 5-15 subjects participated in two sessions of approximately two hours each. The mood adjective rating scale was the fifth questionnaire of the first session.

Confirmatory factor analytic techniques were used to test item factor models, obtaining parameter estimates by using the maximum likelihood procedure of LISREL VI (Jöreskog and Sörbom 1984). Indices of the fit of a model reported include the χ^2 statistic and Goodness of Fit Index (GFI) provided by the LISREL program, as well as the Normed Fit Index (NFI) developed by Bentler and Bonett (1980). A model that fits the data (adequately reproduces the variance/covariance matrix) will have a low χ^2 and high GFI and NFI. The χ^2 statistic is more sensitive to sample size than either the GFI or NFI. These indices can range from 0.0 to 1.0 with fits above 0.9 often considered to be excellent (see Marsh et al. 1988; Mulaik et al. 1989).

Simultaneous multiple group comparisons between the younger Annville group and the older Annville group were made using the procedure outlined in Jöreskog (1971). The multiple-groups analyses were conducted using covariance matrices in order to maintain equivalent metrics across groups (Jöreskog 1971; Schaie and Hertzog 1985).

Hypotheses about invariance in factor structure between groups can be evaluated by using a nested sequence of models and calculating

TABLE 2
Item Origins

ITEM #	ITEM	SUBSCALE	SOURCE
1	listless	Fatigue	POMS
2	comfortable	Well-Being	PMCL
3	unhappy	Depressed	POMS
4	shakey	Anxiety	POMS
5	glad	Well-Being	PMCL
6	jittery	Anxiety	PMCL
7	sorry for things done	Depressed	POMS
8	vigorous	Vigor	POMS
9	discouraged	Depressed	POMS
10	lonely	Depressed	POMS
11	elated	Well-Being	PMCL
12	subdued	Depressed	PMCL
13	carefree	Vigor	POMS
14	blue	Depressed	POMS
15	happy	Well-Being	PMCL
16	relaxed	Anxiety	POMS
17	bushed	Fatigue	POMS
18	ecstatic	Well-Being	PMCL
19	exhausted	Fatigue	POMS
20	active	Vigor	POMS
21	tired	Fatigue	PMCL
22	hopeless	Depressed	POMS
23	dull	Fatigue	PMCL
24	panicky	Anxiety	POMS
25	worn out	Fatigue	POMS
26	anxious	Anxiety	POMS
27	calm	Well-Being	PMCL
28	depressed	Depressed	PMCL
29	full of pep	Vigor	POMS
30	weary	Fatigue	POMS
31	aroused	Vigor	PMCL
32	sad	Depressed	POMS
33	restless	Anxiety	POMS
34	regretful	Depressed	PMCL
35	pleased	Well-Being	PMCL
36	enthusiastic	Vigor	PMCL
37	miserable	Depressed	POMS
38	overjoyed	Well-Being	PMCL
39	fearful	Anxiety	PMCL
40	unworthy	Depressed	POMS
41	uneasy	Anxiety	POMS
42	glum	Depressed	PMCL
43	sluggish	Fatigue	POMS
44	alert	Vigor	POMS
45	helpless	Depressed	POMS
46	drowsy	Fatigue	PMCL
47	gloomy	Depressed	POMS
48	frightened	Anxiety	PMCL
49	fatigued	Fatigue	POMS
50	cheerful	Vigor	POMS
51	contented	Well-Being	PMCL
52	worthless	Depressed	POMS
53	cautious	Anxiety	PMCL
54	inadequate	Depressed	PMCL
55	tense	Anxiety	POMS
56	excited	Vigor	PMCL
57	sleepy	Fatigue	PMCL
58	lively	Vigor	POMS
59	terrified	Depressed	POMS
60	on edge	Anxiety	POMS
61	forceful	Vigor	PMCL
62	guilty	Depressed	POMS
63	at ease	Well-Being	PMCL
64	nervous	Anxiety	POMS
65	energetic	Vigor	POMS

appropriate χ^2 tests. The χ^2 obtained from fitting the base model can be compared to one (e.g., factor loadings) constraining a set of parameters to be equal across groups. If this difference in χ^2, which is also distributed χ^2, is significant, it would indicate that the additional restrictions have substantially increased or decreased the fit of the model. Issues of age-related measurement equivalence can be directly translated into test of invariance in factor structure for items or scales (Rock et al. 1978; Schaie and Hertzog 1985). Nested models relevant to the issue of measurement equivalence include (1) a model where no constraints except the same factor pattern are specified across the groups, (2) specification of invariant factor loadings, (3) adding the specification of invariance for the factor variance-covariance matrix, and (4) adding the specification of equality for the unique variances. Test of each subsequent model is contingent upon acceptance of the previous, less constrained model (Alwin and Jackson 1981). The most crucial issues for measurement equivalence are invariance in the configuration of factors and equality of the unstandardized factor pattern weights (factor loadings) (Horn et al. 1984; Meredith 1964; Rock et al. 1978).

The χ^2 tests identify sets of parameters that differ significantly between groups. Following a significant χ^2, interest centers on identifying parameters that are responsible for the group differences. In the present study, principal interest is on group differences in factor loadings. A group comparison of each loading may be calculated to identify significant group differences via a t-statistic (given large sample size, calculated as $z = (\lambda_1 - \lambda_2/(se_1^2 + se_2^2)^{1/2}$ where λ represents the factor loading of an item for each of the two groups, and se denotes the associated standard error).

Results

OVERVIEW

The single group analysis, performed on the entire age range of subjects, began with an exploratory factor analysis to determine a general factor structure for the items. A series of confirmatory factor analyses of a model derived from an interpretation of the exploratory

analysis was then executed. Next, a multiple-groups analysis comparing young versus old subsamples was performed. The total sample was divided into two age groups: a young group (age = < 54; n = 181) and an old group (age = >55; n = 171). Although finer gradations of age might have been desirable, the cut point at age 55 seemed appropriate and preserved sufficient sample size to enable meaningful LISREL analysis. Simultaneous confirmatory factor analysis was used to evaluate the instrument's factor structure properties concurrently for the separate groups. The factor model was then tested for cross-validation in a new, comparable sample obtained in 1987. Revised subscales were created based on the single-group factor analysis results and an inspection of item-total correlations and internal consistency estimates, and age differences in mean levels of affective state in the assessment situation were analyzed.

SINGLE GROUP ANALYSIS

Exploratory factor analysis. An exploratory factor analysis using a promax rotation was performed. Seven-, 8-, and 9-factor solutions were examined, with the 9-factor achieving the most interpretable results. Depression, Fatigue, and Anxiety factors fell out quite clearly. One factor seemed to be a combination of Vigor and Well-Being items, while another factor consisted of a subset of Well-Being items. The solution suggested some more narrowly defined factors not originally anticipated. One new factor was composed of a few Depression and Anxiety items (sorry for things done, regretful, fearful, unworthy, frightened, worthless, and guilty), and was labeled Guilt. The last three factors did not appear to correspond to any meaningful affect concepts, and all but one of the items forming them were excluded from the final revision of the instrument.

Confirmatory factor analysis. The purpose of this analysis was to clarify ambiguities in item/scale assignments that were implied by the exploratory factor analysis. A model was constructed using both the exploratory results and a priori subscale assignments. The Guilt and Anxiety factors were specified as indicated by the exploratory results. A Depression factor was formed from the five items with the highest exploratory loadings, plus the item "worthless." This latter item was also loaded onto Guilt, per the exploratory results. Because the explor-

atory Vigor factor displayed a confusing combination of a priori Vigor and Well-Being items (considering item face validity), the Vigor and Well-Being factors were specified according to the a priori assignment of items to factors. The Fatigue factor was not included in this phase of confirmatory analysis because the Fatigue factor identified in the exploratory analysis so closely matched the a priori scale assignments that it was felt unnecessary to include it in the first phase of confirmatory analysis.

The first LISREL model did not fit the data particularly well, with a goodness-of-fit index of 0.711 (χ^2 = 2621.86, df = 808). Tests of subsequent models compelled significant changes in the Well-Being factor, most notably the extraction of a five-item "Calm" factor (a combination of Vigor and a reversed scored Anxiety items), and an "Elation" factor composed of three items: elated, ecstatic, and overjoyed.[1] In addition, a three-item "Fear" factor was extracted from the Guilt and Anxiety factors. There were also six items that are loaded on more than one factor. The model also allowed for five pairs of correlated residuals, primarily among Well-Being items. The fit of this 8-factor model was an improvement over the initial model (GFI = 0.822, χ^2 = 1668.98, df = 821).

A final factor model of nine factors, involving all the items of the revised factor arrangement plus the Fatigue items, was tested. The Fatigue factor was constructed based on the results from the exploratory factor analysis, which matched almost perfectly the a priori assignments. The goodness-of-fit index was 0.726 (χ^2 = 3961.38, df = 1732), with only one correlated residual ("sleepy" and "drowsy"). A few modifications, in the forms of double-loadings and correlated residuals, were indicated and were theoretically sound. The final model tested allows five items to be double-loaded and has three pairs of correlated residuals. This model fit moderately well (GFI = 0.747, χ^2 = 3606.44, df = 1725). Table 3 exhibits factor loadings and unique variances of the items. Most of the factor loadings range from 0.6 to a high of 0.878 for "energetic" on the Vigor factor.

Factor correlations are reported in Table 4. Item factors can be conceptualized as representing latent scale true scores; therefore, the factor correlations are attenuated for measurement error and are somewhat higher than subscale correlations (see below). The correlations of Well-Being with Vigor and with Calm, and of Depression with

TABLE 3
Confirmatory Model: Factor Loadings (λ) and Unique Variances (θ)

ITEM	FACTOR	λ_1	λ_2	(Factor)[a]	θ^2
nervous	Anxiety	.861			.258
on edge	Anxiety	.861			.259
tense	Anxiety	.858			.264
uneasy	Anxiety	.790			.375
jittery	Anxiety	.697			.514
shaky	Anxiety	.694			.518
anxious	Anxiety	.580			.663
panicky	Anxiety	.306	.516	(fear)	.384
worn out	Fatigue	.868			.247
fatigued	Fatigue	.861			.259
tired	Fatigue	.852			.273
exhausted	Fatigue	.846			.285
bushed	Fatigue	.823			.322
weary	Fatigue	.798			.363
sleepy	Fatigue	.744			.446
sluggish	Fatigue	.740			.452
drowsy	Fatigue	.687			.528
listless	Fatigue	.526			.724
sad	Depressed	.848			.281
blue	Depressed	.841			.293
depressed	Depressed	.836			.300
gloomy	Depressed	.828			.315
glum	Depressed	.806			.350
hopeless	Depressed	.803			.355
discouraged	Depressed	.802			.357
miserable	Depressed	.788			.379
unhappy	Depressed	.787			.381
lonely	Depressed	.706			.501
helpless	Depressed	.350	.398	(fear)	.531
happy	Well-being	.848			.281
cheerful	Well-being	.828			.314
pleased	Well-being	.795			.368
content	Well-being	.793			.372
carefree	Well-being	.707			.500
glad	Well-being	.680			.537
enthusiastic	Well-being	.421	.412	(vigor)	.386
energetic	Vigor	.878			.229
lively	Vigor	.845			.285
full of pep	Vigor	.845			.286
vigor	Vigor	.793			.372
active	Vigor	.720			.481
excited	Vigor	.592	.324	(anxiety)	.632
alert	Vigor	.462	.295	(calm)	.536
forceful	Vigor	.438			.808
aroused	Vigor	.396			.843
worthless	Guilt	.763			.418
guilty	Guilt	.743			.448
inadequate	Guilt	.734			.461
regretful	Guilt	.723			.477
unworthy	Guilt	.688			.527
sorry for things done	Guilt	.577			.667
relaxed	Calm	.838			.297
calm	Calm	.833			.306
at ease	Calm	.777			.397
comfortable	Calm	.681			.536
frightened	Fear	.796			.367
fearful	Fear	.794			.369
terrified	Fear	.792			.373
ecstatic	Elation	.825			.319
elated	Elation	.809			.345
overjoyed	Elation	.833			.307

a. Indicates secondary loading, factor in parentheses.

Guilt, are the highest except for ones involving Elation (with Well-Being) and Fear (with Anxiety and Guilt).

MULTIPLE-GROUPS ANALYSIS

The total exploratory sample was divided into two subgroups: a "young" group containing the subjects whose ages ranged from 20 to 54 (n = 181), and an "old" group of subjects aged 55 to 78 (n = 171).

The first model specified in this phase applied the final measurement model achieved in the single group analysis to a simultaneous factor analysis problem between the two age subgroups. The model specified the same factor pattern but allowed different estimates of factor loadings in each age group. The overall χ^2 statistic was 6471.66 with 3450 degrees of freedom. The second model constrained the factor loadings to be invariant across the subgroups, generating a χ^2 statistic of 6616.64 with 3507 degrees of freedom. The change in GFI was, however, relatively small. The older group's GFI was 0.635 in the unconstrained model and 0.628 in the constrained; the younger group's GFI indices were 0.651 and 0.647. The χ^2 difference between the two models was significant (144.98 for a difference of 57 degrees of freedom). Nevertheless, the hypothesis of complete measurement equivalence across all factors was rejected. A post hoc series of models consecutively releasing factor loading constraints on single subscales (suspected from the pattern of estimates) was conducted to isolate the source of group differences in the factor loadings. The sequential procedure terminated following achievement of a nonsignificant χ^2 difference. This approach identified the Fatigue, Fear, Anxiety, and Guilt subscales as having significant group differences.

Many of the instrument items' response distributions were heavily skewed, which may have had a distorting effect on a model's χ^2 fit statistic. Table 5 exhibits some sample Fatigue and Anxiety items and their mean response by age group. The post hoc *t*-tests of group differences in individual loadings revealed significant age group differences on some but not all factor loadings for the offending items with the most salient degree of skew. The Fatigue subscales yielded the greatest degree of consistency in all age differences in both skew and factor loading. However, while skew may affect estimation of

TABLE 4
Factor Correlations (decimals omitted)

	1	2	3	4	5	6	7	8
1 Anxiety								
2 Fatigue	436							
3 Depression	631	456						
4 Well-being	-421	-440	-616					
5 Vigor	-228	-616	-376	769				
6 Guilt	677	348	774	-387	-212			
7 Calm	-705	-462	-505	798	605	-516		
8 Fear	780	296	674	-339	-134	826	-526	
9 Elation	-177	-303	-317	778	743	-139	512	-149

factor loadings (Olsson 1979), it appears that at least some real age differences in relations of items to factors were uncovered.

CROSS-VALIDATION

The final single-group model was applied to the new sample of 287 subjects in a LISREL VI confirmatory factor analysis for cross-validation. The model did not fit this sample as well ($\chi^2 = 3931.35$; df = 1728; GFI = 0.697). Table 6 displays factor loadings based on covariance matrix analysis for both the 1985 and 1987 samples. The Fear and Guilt factors demonstrate the largest differences in loadings between the two groups, with sizable differences also for the Anxiety factor. An examination of the factor correlations for the 1987 analysis (Table 7) shows Guilt and Fear to be less differentiated from their parent factors (Depression and Anxiety, respectively) than in the 1985 sample. Because the Guilt, Fear, and Anxiety factors were found to have some differences in item responding between old and young subgroups, univariate statistics for the 1987 data were examined. For several of the items for these factors, it was found that responding in the "old" subgroup for 1987 more closely resembled "young" responding for both years than "old" responding for 1985. The poorer model fit for the 1987 sample may be due to less age-related differences in responding than in the 1985 sample. For those items where

TABLE 5
Univariate Statistics of Selected Fatigue and Anxiety Items

Item	Age Group	N	Mean	SD	Skew	Kurtosis
Fatigue:						
sleepy	old	171	1.39	.72	1.90	3.04
	young	181	1.90	.97	1.14	.92
Anxiety:						
jittery	old	171	1.25	.63	3.11	11.14
	young	181	1.30	.69	2.80	8.52
shaky	old	171	1.19	.46	2.41	5.25
	young	181	1.24	.59	2.94	9.43
anxious	old	171	1.54	.78	1.61	2.77
	young	181	1.60	.91	1.86	3.53

the differences were reproduced, the difference in age distribution between the two samples may be another contributing factor.

ITEM-TOTAL CORRELATIONS AND INTERNAL CONSISTENCY

Items were summed to form subscales using item/scale assignments implied by factor structure from the single-group confirmatory factor analysis. Items that had been excluded from the model were assigned on a theoretical basis. In all, 61 of the original 65 items were retained in the revised scale. Items that had been loaded on more than one factor were assigned to single factors based on face validity. Because the Elation factor scale has only three items and correlated highly with the Well-Being factor (.79), its items were subsumed in the Well-Being subscale for purposes of scale assignment. Item-total correlations by subscales and internal consistency reliabilities (Cronbach's alpha) were calculated. Table 8 displays the number of items per scale and the unstandardized coefficient alpha for both the a priori and revised scales. Most of the coefficients increased after scale revision, and all remained acceptably high. The only exception was the new Fear subscale, which had only three items and poor internal consistency. Dropping this subscale's items from scale scoring should be considered as an option when using the instrument.

TABLE 6
1985 and 1987 Unstandardized Factor Loadings
Based on Analysis of Covariance Matrices
(standard errors in parentheses)

FACTOR	ITEM	1985 Sample	1987 Sample
ANXIETY	nervous	1.257(.106)	1.130(.088)
	on edge	1.231(.104)	1.065(.088)
	tense	1.362(.115)	1.309(.101)
	uneasy	.983(.087)	1.069(.087)
	jittery	.936(.090)	.766(.072)
	shaky	.751(.072)	.542(.062)
	anxious	1.000(.000)	1.000(.000)
	panicky	.313(.074)	.581(.096)
	(excited)	.696(.113)	.309(.100)
FATIGUE	worn out	.951(.044)	.939(.043)
	fatigued	.891(.042)	.880(.042)
	tired	1.000(.000)	1.000(.000)
	exhausted	.964(.047)	.880(.043)
	bushed	1.015(.052)	.973(.048)
	weary	.813(.044)	.775(.042)
	sleepy	.781(.047)	.809(.050)
	sluggish	.736(.044)	.745(.041)
	drowsy	.665(.045)	.745(.043)
	listless	.495(.047)	.560(.046)
DEPRESSED	sad	.915(.059)	.987(.070)
	blue	.979(.064)	.988(.077)
	depressed	1.000(.066)	1.100(.081)
	gloomy	.835(.055)	.807(.063)
	glum	.786(.054)	.908(.068)
	hopeless	.775(.053)	.926(.071)
	discouraged	1.028(.070)	.984(.076)
	miserable	.786(.055)	.947(.069)
	unhappy	.906(.063)	.868(.078)
	lonely	1.000(.000)	1.000(.000)
	helpless	.341(.061)	.397(.078)
WELL-BEING	happy	1.107(.072)	1.259(.109)
	cheerful	1.067(.071)	1.282(.106)
	pleased	.953(.066)	1.226(.108)
	content	1.036(.072)	1.204(.106)
	carefree	1.000(.000)	1.000(.000)
	glad	.901(.073)	1.086(.104)
	enthusiastic	.554(.085)	.773(.122)
VIGOR	energetic	1.123(.059)	1.002(.059)
	lively	1.015(.056)	.927(.061)
	full of pep	1.113(.062)	1.157(.065)
	vigor	1.000(.000)	1.000(.000)
	active	.854(.058)	.826(.061)
	excited	.713(.063)	.603(.069)
	alert	.468(.058)	.528(.069)
	forceful	.492(.059)	.470(.069)
	aroused	.458(.062)	.547(.067)
	(enthusiastic)	.515(.079)	.423(.088)

TABLE 6 Continued

FACTOR	ITEM	1985 Sample	1987 Sample
GUILT	worthless	.839(.079)	1.652(.239)
	guilty	.839(.080)	1.186(.182)
	inadequate	1.137(.109)	1.476(.227)
	regretful	.942(.091)	1.628(.245)
	unworthy	.902(.090)	1.718(.251)
	sorry for..	1.000(.000)	1.000(.000)
CALM	relaxed	1.061(.057)	.976(.068)
	calm	1.000(.000)	1.000(.000)
	at ease	.961(.057)	1.049(.080)
	comfortable	.811(.058)	.802(.072)
	(alert)	.296(.056)	.355(.073)
FEAR	frightened	.732(.046)	1.028(.069)
	fearful	1.000(.000)	1.000(.000)
	terrified	.532(.033)	1.034(.072)
	(panicky)	.521(.074)	.226(.098)
	(helpless)	.517(.083)	.100(.094)
ELATION	ecstatic	1.015(.060)	1.099(.079)
	elated	1.000(.000)	1.000(.000)
	overjoyed	1.019(.060)	1.034(.078)

TABLE 7
Factor Correlations 1987 Cross-Validation Sample
(decimals omitted)

	1	2	3	4	5	6	7	8
1 Anxiety								
2 Fatigue	459							
3 Depression	772	544						
4 Well-being	-531	-553	-679					
5 Vigor	-272	-659	-420	801				
6 Guilt	779	510	908	-596	-339			
7 Calm	-715	-542	-603	815	646	-588		
8 Fear	846	324	755	-460	-211	721	-523	
9 Elation	-205	-389	-338	769	767	-246	538	-189

TABLE 8
Coefficient Alphas (unstandardized) for the A Priori and Revised Subscales (number of items in parentheses)

Subscale	A Priori	Revised
Anxiety	.901 (13)	.916 (8)
Fatigue	.929 (11)	.935 (10)
Depression	.935 (19)	.939 (11)
Well-being	.905 (10)	.923 (10)
Vigor	.908 (12)	.881 (9)
Guilt	-	.836 (6)
Calm	-	.858 (4)
Fear	-	.757 (3)

AGE DIFFERENCES IN AFFECT

The revised affect scale scores for the original 1985 sample were analyzed in a 2 × 2 (Age Group × Sex) MANOVA. In order to maximize *N*, subjects with less than 20% missing item responses on all scales were included in the analysis by assigning the sample item mean to the missing value prior to summing item responses. This procedure increased the *N* for the MANOVA to 438.

There was a significant multivariate Age effect ($F[7,428] = 6.48$, $p < .001$), with significant ($p < 0.01$) univariate effects for Fatigue, Vigor, Well-Being, and Depression. The age difference on Anxiety did not approach statistical significance ($F < 1$). Older adults (age > 54) reported less fatigue, more vigor, greater well-being, and less depressive affect than younger adults. The multivariate test for Sex failed to reach statistical significance ($F[7, 428] = 1.94$, $p = 0.06$), as did the Age × Sex interaction ($F[7, 428] = 1.14$, $p > 0.25$).

Discussion

Several important issues have been addressed by the present research. One is the degree to which adjective self-ratings form coherent scales with acceptable psychometric properties. The present factor analytic study isolated nine distinct factors—Anxiety, Fatigue, De-

pression, Well-Being, Vigor, Guilt, Calm, Elation, and Fear – from 61 items collected from the Profile-of-Mood-States (POMS) instrument (McNair et al. 1971) and the PMcL (Lebo and Nesselroade 1978), which were the basis for seven subscales with acceptable internal consistencies and sensible interscale correlations. A sound factor structure was supported. Of further importance is that these results were achieved on a sample consisting of young, middle-aged, and elderly persons from the general population, with an age range of 21 to 78. Previously, the POMS had been validated on psychiatric patients and college students, while the PMcL was administered to young pregnant women. Similar work by the present authors on five subscales of the Cattell Eight-State Questionnaire likewise demonstrates acceptable subscaling of mood adjectives administered to this sample (although subscale assignments require extensive revision).

Issues regarding the resultant factor structure include the formation of two monopolar scales – Anxiety and Calm – rather than a commonsense bipolar scale subsuming the two. This suggests that the two concepts, while highly negatively correlated, are construed by persons as related but distinct and not merely opposites. Another possibility for such a finding, though, is put forth by Meddis (1972), who claims that unbalanced Likert-format options (i.e., more acceptance than rejection categories) tend to suppress negative correlations and artificially produce monopolar factors. The current study involves an instrument with one rejection option and four acceptance options of varying degree. Additional research involving varied response formatting using the present instrument is needed to investigate this issue further.

Because the analyses here suggest some modification of item-to-subscale assignments from the a priori designations of the POMS and the PMcL, one may ask which subscale structure should be used. Although the original two instruments were not validated on samples using elderly from the general population, the present analysis of this hybrid instrument using such a sample shows the original subscale structure to be basically retained. High correlations among the revised subscales (e.g., Well-Being with Vigor and Calm), which generally share items originally from one subscale, suggest that collapsing the appropriate items into their original scales is a viable alternative. In part, the revised structure may result from enhanced concept representation over the POMS due to the inclusion of the PMcL items. Overall,

the results support the original item structure of the instrument; adoption of the suggested modifications are appropriate for investigations concerned with still finer sensitivity to mood states, particularly to those scales uniquely formed by the revisions.

The results of the multiple-groups analysis are cause for some caution in using the instrument. Measurement equivalence of the instrument was not fully supported in the comparison of young and old subgroups. This lack of measurement equivalence creates interpretational problems for analysis of mean age differences in the summated scales (Baltes and Nesselroade 1973; Labouvie 1980). For example, the analysis of age differences in affect during the assessment session showed that older persons reported less fatigue than young persons. However, the Fatigue item factor showed marked age differences in item factor loadings that appeared to be influenced by age differences in skew and kurtosis of the response distributions. The issue is whether the lack of metric invariance in the item factor loadings is principally a function of the nonnormal distributions, and in turn, whether the distributional differences reflect differences in the underlying affect construct. Response to an item such as "sleepy" may actually reveal that the elderly are less sleepy than are younger persons. However, the skew may also be determined by substantive influences on fatigue; younger subjects generally were employed and were more likely to participate after work hours, whereas elderly subjects were more available to participate earlier in the day. The presence of such differences, then, does not necessarily indicate a problem with the instrument. On the other hand, the skewed distributions may reflect age differences in willingness to endorse items measuring negative affect. An adjective such as "anxious" may have negative connotations for the elderly as symptomatic of old age, undermining their willingness to endorse it even if they are experiencing anxiety. Another possibility is that of age differences in influences on affect-related manifestations: "shaky" may reflect an anxiety response for the young, but it may be influenced by age-related somatic health changes for some older persons, attenuating its validity as an indicator of anxiety per se in an age-heterogeneous population. Clearly, resolution of measurement equivalence problem will require resolution of the issue of item skew and its effects on factor analysis parameter estimates (see Muthén and Kaplan 1985).

The pattern of age differences in means (greater self-reported well-being, vigor; less fatigue and depression) might be construed as implausible—and as a basis for questioning the validity of the adjective rating scales. However, epidemiological studies of depressive affect using the CES-D instrument have consistently shown older persons to report lower levels of depressive affect than do younger individuals (e.g., Ensel 1986)—a pattern consistent with CES-D responses in this sample, which we are reporting elsewhere (Hertzog et al. forthcoming). Nevertheless, the present volunteer sample may be positively selected for older persons with higher psychological well-being. Also, as noted above, age differences in time of assessment may have influenced the results. There may also be age differences in response criteria mapping Likert rating scale points to subjectively experienced affect. Nevertheless, the current study offers little support for the hypothesis that, on average, older persons experience lower levels of psychological well-being than do young to middle-aged adults. The present work leaves open the possibility that differences in measurement properties and the scales influence observed age differences in affect as further research questions. Until resolved, caution is suggested in use of the questionnaire to compare different age groups.

The present research raises some relevant issues for the construction of scales appropriate for group score comparisons. We have identified relatively robust mood state factors in an adult population, and have shown in two independent samples that the basic configuration of adjectives loading on affect to factors is replicable. On the basis of these factor analyses, a slightly revised set of mood adjective scales has been proposed. Further studies sensitive to the issues delineated above is required to determine whether (1) the scales can be constructed so as to yield equivalent psychometric properties across the adult life span, and (2) the scales have acceptable construct validity.

NOTE

1. The finding of an Elation factor replicates results from the Lebo and Nesselroade [1978] P-technique study using exploratory factor analysis; they designate it a secondary Well-Being factor.

REFERENCES

Alwin, Duane F. and David J. Jackson. 1981. "Applications of Simultaneous Factor Analysis to Issues of Factorial Invariance." Pp. 249-78 in *Factor Analysis and Measurement*, edited by D. J. Jackson and E. F. Borgatta. London: Sage.

Baltes, Paul B. and John R. Nesselroade. 1973. "The Developmental Analysis of Individual Differences on Multiple Measures." Pp. 219-52 in *Life-Span Developmental Psychology: Methodological Issues*, edited by J. R. Nesselroade and H. W. Reese. New York: Academic Press.

Bentler, Paul M. and Douglas G. Bonett. 1980. "Significance Test and Goodness of Fit in the Analysis of Covariance Structures." *Psychological Bulletin* 88: 588-606.

Botwinick, J. 1984. *Aging and Behavior*, 3rd ed. New York: Springer.

Bradburn, Norman M. and David Caplovitz. 1965. *Reports on Happiness: A Pilot Study of Behavior Related to Mental Health*. Chicago: Aldine.

Campbell, Angus, Philip E. Converse, and Williard L. Rodgers. 1976. *The Quality of American Life*. New York: Russell Sage.

Cameron, Paul. 1975. "Mood as an Indicant of Happiness: Age, Sex, Social Class, and Situational Factors." *Journal of Gerontology* 30: 216-24.

Cattell, Raymond B. 1966. "Anxiety and Motivation: Theory and Crucial Experiments." Pp. 23-62 in *Anxiety and Behavior*, edited by C. D. Spielberger. New York: Academic Press.

Cavanaugh, Jonn C. Forthcoming. "The Importance of Awareness in Memory Aging." In *Everyday Cognition in Adult and Late Life*, edited by L. W. Poon, D. C. Rubin, and B. A. Wilson. New York: Cambridge University Press.

Cavanaugh, John C. and Nancy Z. Murphy. 1986. "Personality and Metamemory Correlates of Memory Performance in Younger and Older Adults." *Educational Gerontology* 12: 387-96.

Clark, Virginia A., Carol S. Aneshensel, Ralph R. Frerichs, and T. M. Morgan. 1981. "Analysis of Effects of Sex and Age in Response to Items on the CES-D Scale." *Psychiatry Research* 5: 171-81.

Dixon, Roger A. and Christopher Hertzog. 1988. "A Functional Approach to Metamemory Development in Adulthood." Pp. 293-330 in *Memory Development: Universal Changes and Individual Differences*, edited by F. E. Weinert and M. Perlmutter. Hillsdale, NJ: Lawrence Erlbaum.

Drasgow, Fritz and Ruth Kanfer. 1985. "Equivalence of Psychological Measurement in Heterogeneous Populations." *Journal of Applied Psychology* 70: 662-80.

Ensel, Walter M. 1986. "Measuring Depression: The CES-D Scale." Pp. 51-70 in *Social Support, Life Events, and Depression*, edited by N. Lin, A. Dean, and W. Ensel. New York: Academic Press.

Fridhandler, Bram M. 1986. "Conceptual Note on State, Trait, and the State-Trait Distinction." *Journal of Personality and Social Psychology* 50: 169-74.

Gallagher, Dolores. 1986. "Assessment of Depression by Interview Methods and Psychiatric Rating Scales." Pp. 202-12 in *Clinical memory assessment of older adults*, edited by L. W. Poon. Washington, DC: American Psychological Association.

George, Linda K. 1981. "Subjective Well-Being: Conceptual and Methodological Issues." Pp. 345-82 in *Annual Review of Gerontology and Geriatrics*, Vol. 2, edited by C. Eisdorfer. New York: Springer.

Gurland, Barry J. 1976. "The Comparative Frequency of Depression in Various Adult Age Groups." *Journal of Gerontology* 31(3): 283-92.

Hertzog, Christopher. 1987. "Application of Structural Equation Models in Gerontological Research." Pp. 265-93 in *Annual Review of Gerontology and Geriatrics*, Vol. 7, edited by K. W. Schaie. New York: Springer.

———. (1989). "Using Confirmatory Factor Analysis for Scale Development and Validation." Pp. 281-306 in *Special Research Methods for Gerontology*, edited by M. P. Lawton and A. R. Herzog. New York: Baywood.

Hertzog, Christopher, David F. Hultsch, and Roger A. Dixon. Forthcoming. "Evidence for the Convergent Validity of Two Self-Report Metamemory Questionnaires." *Developmental Psychology.*

Hertzog, Christopher and John R. Nesselroade. 1987. "Beyond Autoregressive Models: Some Implications of the Trait-State Distinction for the Structural Modeling of Developmental Change." *Child Development* 58: 93-109.

Hertzog, Christopher, Judy Van Alstine, Paul Usala, David Hultsch, and Roger Dixon. Forthcoming. "Measurement Properties of the Center for Epidemiological Studies Depression Scale (CES-D) in Older Populations." *Psychological Assessment: A Journal of Consulting and Clinical Psychology.*

Horn, Jonn L., Jack J. McArdle, and Ralph Mason. 1984. "When Is Invariance not Invariant: A Practical Scientist's Look at the Ethereal Concept of Factor Invariance." *Southern Psychologist* 1: 179-188.

Hoyt, Danny R. and James C. Creech. 1983. "The Life Satisfaction Index. A Methodological and Theoretical Critique." *Journal of Gerontology* 38: 111-116.

Jöreskog, Karl G. 1971. "Statistical Analysis of Sets of Congeneric Tests." *Psychometrika* 36: 109-33.

Jöreskog, Karl G. and Dag Sörbom. 1984. *LISREL VI User's Guide.* Mooresville, IN: Scientific Software.

Labouvie, Erich W. 1980. "Measurement of Individual Differences in Intraindividual Changes." *Psychological Bulletin* 88: 54-9.

Larson, Reed. 1978. "Thirty Years of Research on the Subjective Well-Being of Older Americans." *Journal of Gerontology* 33: 109-25.

Lawton, M. Powell. 1975. "The Philadelphia Geriatric Center Morale Scale: A Revision." *Journal of Gerontology* 30: 85-89.

Lawton, M. Powell. 1983. "The Varieties of Well Being." *Experimental Aging Research* 9: 65-72.

Lebo, Michael A. and John R. Nesselroade. 1978. "Intraindividual Differences Dimensions of Mood Change During Pregnancy Identified in Give P-technique Factor Analyses." *Journal of Research in Personality* 12: 205-24.

McNair, Douglas M., Maurice Lorr, and Leo F. Droppleman. 1971. *Manual for the Profile of Mood States.* San Diego, CA: Educational and Industrial Testing Service.

Marsh, Herbert W., John R. Balla, and Roderick P. McDonald. 1988. "Goodness-of-Fit Indexes in Confirmatory Factor Analysis: The Effect of Sample Size." *Psychological Bulletin* 103: 391-410.

Meddis, Ray. 1972. "Bipolar Factor in Mood Adjective Checklists." *British Journal of Social and Clinical Psychology* 11: 178-84.

Meredith, William. 1964. "Notes on Factorial Invariance." *Psychometrika* 29: 177-85.

Morris, John N. and Sylvia Sherwood. 1975. "A Retesting and Modification of the Philadelphia Geriatric Center Morale Scale." *Journal of Gerontology* 30: 77-84.

Mulaik, S. A., L. R. James, J. Van Alstine, N. Bennett, S. Lind, and C. D. Stilwell. 1989. "Evaluation of Goodness-of-Fit Indices for Structural Equation Models." *Psychological Bulletin* 3: 430-445.

Murrell, Stanley A., Samuel Himmelfarb, and Katherine Wright. 1983. "Prevalence of Depression and Its Correlates in Older Adults." *American Journal of Epidemiology* 117(2): 173-85.

Muthén, Bengt and David Kaplan. 1985. "A Comparison of Some Methodologies for the Factor Analysis of Non-Normal Likert Variables." *British Journal of Mathematical and Statistical Psychology* 38: 171-89.

Nesselroade, John R. and Thomas W. Bartsch. 1977. "Multivariate Experimental Perspectives on the Construct Validity of the Trait-State Distinction." In *Handbook of Modern Personality Theory*, edited by R. B. Cattell and R. M. Dreger. Washington, DC: Hemisphere/Halstead.

Nesselroade, John R., L. S. Mitteness, and L. K. Thompson. 1984. "Older Adulthood: Short-Term Changes in Anxiety, Fatigue, and Other Psychological States." *Research on Aging* 6: 3-23.

Olsson, Ulf. 1979. "On the Robustness of Factor Analysis Against Crude Classification of the Observations." *Multivariate Behavioral Research* 14: 485-500.

Radloff, Lenore S. 1977. "The CES-D Scale: A Self-Report Depression Scale for Research in the General Population." *Applied Psychological Measurement* 1: 385-401.

Rock, Donald A., Charles E. Werts, and Ronald L. Flaugher. 1978. "The Use of Analysis of Covariance Structures for Comparing the Psychometric Properties of Multiple Variables Across Populations." *Multivariate Behavioral Research* 13: 403-18.

Rodin, Judith, Gail McAvay, and Christine Timko. 1988. "A Longitudinal Study of Depressed Mood and Sleep Disturbances in Elderly Adults." *Journal of Gerontology: Psychological Sciences* 43(2): 45-53.

Schaie, K. Warner and Christopher Hertzog. 1985. "Measurement in the Psychology of Adulthood and Aging." Pp. 61-92 in *Handbook of the Psychology of Aging*, edited by J. E. Birren and K. W. Schaie. New York: Van Nostrand Reinhold.

Schulz, Richard. 1985. "Emotion and Affect." Pp. 531-43 in *Handbook of the Psychology of Aging*, 2nd ed., edited by J. E. Birren and K. W. Schaie. New York: Van Nostrand Reinhold.

Spielberger, Charles D. 1966. "Theory and Research on Anxiety." Pp. 3-20 in *Anxiety and Behavior*, edited by C. D. Spielberger. New York: Academic Press.

West, Robin L., Lynn K. Boatwright, and Robert Schleser. 1984. "The Link Between Memory Performance, Self-Assessment, and Affective Status." *Experimental Aging Research* 10: 197-200.

Wilson, Gail A., Jeffrey W. Elias, and Leonard J. Brownlee. 1985. "Factor Invariance and the Life Satisfaction Index." *Journal of Gerontology* 40: 344-46.

Zarit, Steven H., John Eiler, and Marla Hassinger. 1985. "Clinical Assessment." Pp. 725-54 in *Handbook of the psychology of aging*, 2nd ed., edited by J. E. Birren and K. W. Schaie. New York: Van Nostrand Reinhold.

Paul D. Usala is a doctoral student in the School of Psychology of the Georgia Institute of Technology, specializing in industrial/organization psychology.

Christopher Hertzog is Associate Professor of Psychology at Georgia Institute of Technology, with interests in aging, cognition, personality, and their interrelationships. He co-edited a special issue of Research on Aging *with Edgar F. Borgatta in 1985, contributing to it, "An Individual Differences Perspective: Implications for Cognitive Research in Gerontology."*

Psychological Assessment:
A Journal of Consulting and Clinical Psychology
1990, Vol. 2, No. 1, 64-72

1040-3590/90/$00.75

Measurement Properties of the Center for Epidemiological Studies Depression Scale (CES-D) in Older Populations

Christopher Hertzog, Judith Van Alstine, Paul D. Usala
Georgia Institute of Technology

David F. Hultsch and Roger Dixon
University of Victoria

Two cross-sectional samples of adults were administered the 20-item Center for Epidemiological Studies-Depression Scale (CES-D). Confirmatory item factor analysis showed that Radloff's (1977) four factor model fit the data well, but that the four factors were highly intercorrelated. A simultaneous second-order factor model fitting a single second-order Depression factor also fit well. Multiple group analyses of the first-order solution yielded invariant unstandardized item factor loadings across samples and age groups. A Cohort (Age) × Sex ANOVA on the total and subscale scores revealed lower total CES-D and subscale (Well-Being and Depressive Affect) scores for older persons. The Somatic subscale showed no significant age differences. The results support the measurement validity of the CES-D for depression screening in older adult populations.

Although it is widely recognized that depression is the most common mental health problem in the elderly, there is some confusion regarding whether there is also an age-related increase in the prevalence of depression. Recent studies have not supported the hypothesis of an age-related increase in either clinically diagnosed depression (Blazer, Hughes, & George, 1987) or depressive symptoms in adulthood (e.g., Radloff & Teri, 1986).

A major issue involves the measurement properties of instruments assessing depression (Gallagher, Thompson, & Levy, 1980; Zarit, Eiler, & Hassinger, 1985). For example, several studies suggest that self-report measures of depression that include items measuring somatic manifestations of depression (e.g., fatigue, poor sleep, listlessness) may be artifactually elevated in the elderly because of somatic effects of physical illness, side effects of medications, and the like (Berry, Storandt, & Coyne, 1984; Blumenthal, 1975; Downes, Davies, & Copeland, 1988; Steuer, Bank, Olsen, & Jarvik, 1980).

The Center for Epidemiological Studies-Depression Scale (CES-D; Radloff, 1977) has been widely used in research on depressive affect in community populations, including the elderly (e.g., Krause, 1986; Lewinsohn, Fenn, Stanton, & Franklin, 1986). Validation studies have shown that the CES-D correlates significantly with clinical ratings of depression (Roberts & Vernon, 1983; Weissman et al. 1977), suggesting its utility as a screening instrument. Several cross-sectional studies using the CES-D suggest higher prevalence rates of depression in young adults than in old adult samples (e.g., Craig & VanNatta, 1979; Clark, Aneshensel, Frerichs, & Morgan, 1981). However, Murrell, Himmelfarb, & Wright (1983) report a late-life increase in prevalence of depression, with individuals 75 years old and older having higher mean CES-D scores than individuals of ages 55-74. Age may ultimately be shown to interact with a number of other factors in determining risk for depression (Phifer & Murrell, 1986), but it appears that old age is not necessarily accompanied by increases in depressive affect.

There remains some concern about the psychometric properties of the CES-D in older populations (Yesavage, 1986). One issue is that the CES-D is a multidimensional instrument. Virtually all the studies cited above have used a single, summated score from the 20-item CES-D to measure depression, in spite of the fact that exploratory factor analyses of the instrument have consistently found four separate factors: Depressive Affect, Somatic Symptoms, Well-Being, and Interpersonal Relations (Clark et al., 1981; Ensel, 1986; Radloff, 1977; Ross & Mirowski, 1984). As Gatz and Hurwicz (in press) pointed out, differential age patterns on subscales of the CES-D may be obscured by use of the total score. Given the literature on other depression scales showing age-related increases in somatic symptoms, this issue seems of special concern. Gatz and Hurwicz (in press) found a curvilinear age pattern of CES-D total scores, but the primary source of the age differences was lower positive well-being in the elderly, not higher endorsement of somatic symptoms.

There are two possible arguments against partitioning the CES-D into multiple subscales for screening and research

This research was supported by a research grant to Christopher Hertzog from the National Institute on Aging (R01-AG06162) and by a grant to David F. Hultsch from the Social Sciences and Humanities Research Council of Canada (492-84-002). Christopher Hertzog was also supported by a Research Career Development Award from the National Institute on Aging (K04-AG00335). Roger Dixon was also supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The cooperation of Robert K. Nielsen, the other physicians, and the members of the Annville Family Practice, Annville, Pennsylvania, is deeply appreciated.

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

purposes. First, the different scales may all be conceptualized as subdimensions of a higher-order depression construct (e.g., Radloff, 1977). Second, there may be little empirical differentiation among the four subscales, so that correct classification of persons is not materially enhanced by partitioning the scale into multiple subscales of lower reliability. Indeed, there are only two Interpersonal Relations items and four Well-Being items in the CES–D, which is probably too few items to have confidence in the corresponding subscales unless these items were supplemented with additional items.

A second methodological issue regarding the CES–D is its measurement equivalence across different age populations. Is the CES–D measuring the same construct(s), with equivalent quantitative relations of depression to scale scores, at different ages (Liang, Van Tran, Krause, & Markides, 1989)? As pointed out by several developmental methodologists, quantitative comparisons of scale scores across different age groups is justified only if the underlying scales have equivalent measurement properties in the groups (Baltes & Nesselroade, 1970; Labouvie, 1980; Schaie & Hertzog, 1985).

Empirical leverage on these issues can be gained by applying confirmatory factor analysis to the item data. Previously, Aneshensel, Clark, & Frerichs (1983) used confirmatory analysis to factor the CES–D subscales and other scales measuring aspects of psychological distress and well-being. It is also possible to use higher-order confirmatory factor analysis to simultaneously evaluate (a) the validity of the four factor structure for the CES–D items and (b) the extent to which these four item factors map onto a single higher-order depression factor (e.g., Hertzog, 1989; Marsh, 1985).

The first goal of this research was to test the four factor model for the CES–D in a cross-sectional sample of adults. The second goal was to determine the adequacy of a single second-order Depression factor in fitting correlations among the four first-order item factors. The third goal was to determine whether the CES–D factor structure is invariant across different age groups. The fourth and final goal was to estimate the cross-sectional age pattern in CES–D subscale means.

Method

Subjects

The two samples of subjects participated in a study designed to validate two metamemory questionnaires (Hertzog, Hultsch, & Dixon, 1989). The subjects in both samples were paid for their participation. The first sample consisted of 447 community-dwelling adult volunteers, ages 20–80 who belonged to a large family medical practice in Annville, Pennsylvania. Although questionnaire data were obtained on all participants, a few subjects omitted one or more item responses. Only the 437 subjects with complete item data on the CES–D were used in the item factor analyses. The second sample consisted of a group of older, community-dwelling, Canadian adults from Victoria, British Columbia. The 278 persons in the Victoria sample ranged in age from 55 to 78. Of these participants, 270 had complete item data on the CES–D.

Data regarding the demographic characteristics of both samples may be found in Hertzog et al. (1989). Both samples contained individuals with a range of background education and cognitive abilities, although the majority of individuals reported themselves to be in good to excellent physical health. Both samples are therefore probably more select than their parent populations.

Measures

Subjects from the two samples completed questionnaires and tasks designed to assess metamemory, locus of control, affective states, personality, and memory performance. The tests were administered in two separate sessions to small groups of subjects. The CES–D scale was administered during the first session.

The 20-item CES–D scale is designed to measure depression in the general population (Radloff, 1977; Weissman et al., 1977). Respondents rate the frequency with which they have experienced particular depressive symptoms during the past week. The possible responses range from 0 (less than 1 day) to 3 (5–7 days). The Appendix gives the specific items and the labels used to report results.

Statistical Procedures

Confirmatory factor analysis parameter estimates were obtained using maximum likelihood estimation in LISREL VI (Jöreskog & Sörbom, 1984). The LISREL measurement model regresses observed (or empirically measured) variables, x, on latent variables (factors), ξ, through the regression parameter matrix, Λ, with regression residuals, δ:

$$x = \Lambda\xi + \delta$$

The model assumes that the covariance matrix of the x, Σ, is

$$\Sigma = \Lambda\Phi\Lambda' + \Theta,$$

where Λ is the matrix of factor pattern coefficients, or factor loadings, Φ is the covariance matrix of ξ, and Θ is the covariance matrix of δ. Indexes of a model's fit include the likelihood ratio χ^2 statistic, LISREL's goodness-of-fit index (GFI) and the normed fit index (NFI; Bentler and Bonett, 1980). A model that fits the data well will have a low χ^2 and high GFI and NFI. The χ^2 statistic is more sensitive to sample size than the GFI and NFI. These indexes can range from 0.0 to 1.0, with fits above .9 often considered to be excellent (Marsh, Balla, & McDonald, 1988).

Simultaneous factor analyses for multiple age groups were also conducted in LISREL, using covariance matrixes to maintain equivalent metrics across groups (Jöreskog, 1971; Schaie & Hertzog, 1985). The relative fit of the multiple groups model can be evaluated using a nested sequence of model tests. A significant difference in χ^2 between nested models indicates that the hypothesized specification of the more restricted model (e.g., age group equivalence in factor loadings) would be rejected (Jöreskog, 1971).

A second-order factor analysis allows one to investigate the relationship between hierarchically nested factors (Rindskopf & Rose, 1988). In the present case, our first-order models tested the four factor solution for CES–D items proposed by Radloff (1977). The second-order models evaluated the ability of a single second-order Depression factor to account for the covariances among the four first-order factors. The fit of a second-order model is relative to the fit of the first-order model on which it is based (Hertzog, 1989). In Bentler & Bonett's (1980) terms, the first-order model is the saturated model for the second-order model, which attempts to fit the first-order factor covariance matrix by means of second-order factor loadings, second-order factor variances, residual variance for first-order factors, and (if necessary) residual covariances among first-order factors. Any second-order factor model can therefore be tested for lack of fit to the second-order model specification by computing the difference in χ^2 between first- and second-order models (see Hertzog, 1989; Rindskopf & Rose, 1988).

It is also possible to calculate a relative fit index that reflects the degree of fit of the second-order model to the first-order covariance matrix. The null second-order model is a model fixing all first-order factor covariances to equal zero (i.e., a model specifying no association between the first-order factors). Then the relative normed fit index (RNFI) of the second-order model is

Table 1
Factor Loadings of CES–D Items for Annville (ANN) and Victoria (VIC) Samples

	Depressive affect		Well-Being		Somatic		Interpersonal	
CES–D item	ANN	VIC	ANN	VIC	ANN	VIC	ANN	VIC
Bothered	0	0	0	0	56	54	0	0
Appetite	0	0	0	0	46	52	0	0
Blues	69	75	0	0	0	0	0	0
Good	0	0	47	35	0	0	0	0
Mind	0	0	0	0	56	63	0	0
Depress	85	80	0	0	0	0	0	0
Effort	0	0	0	0	76	82	0	0
Hopeful	0	0	64	54	0	0	0	0
Failure	65	69	0	0	0	0	0	0
Fearful	52	62	0	0	0	0	0	0
Sleep	0	0	0	0	55	46	0	0
Happy	0	0	87	88	0	0	0	0
Talk	0	0	0	0	48	43	0	0
Lonely	77	77	0	0	0	0	0	0
Unfriend	0	0	0	0	0	0	52	55
Enjoy	0	0	85	77	0	0	0	0
Cry	52	52	0	0	0	0	0	0
Sad	83	85	0	0	0	0	0	0
Dislike	0	0	0	0	0	0	75	71
Getgoing	0	0	0	0	64	69	0	0

Note. Decimals omitted. All 0 loadings and standardized factor variances were fixed by hypothesis. All nonzero parameter estimates were significantly different from 0 beyond the .1% level of confidence.

$$RNFI = (F_{n2} - F_2)/(F_{n2} - F_1)$$

where F_{n2} is the null second-order model's LISREL fitting function value, F_1 is the LISREL fitting function value for the corresponding first-order (saturated) model, and F_2 is the fitting function value for any second-order model to be evaluated.

Results

Item Factor Analyses

The first series of analyses attempted to validate the four factor model described in Radloff (1977). This model, depicted in Figure 1, fit well in the Annville sample, $\chi^2(164) = 343.84$, GFI = .925, and NFI = .908, and was replicated in the Victoria sample, $\chi^2(164) = 280.79$, GFI = .909, and NFI = .886. The standardized factor loadings and the factor correlations for the two samples are provided in Table 1. The item factor loadings were very similar in the two samples, with perhaps some salient but small differences on the Well-Being factor. The factors were more highly intercorrelated in the Annville sample than in the Victoria sample, although the pattern of correlations was quite similar (see Table 2).

A second-order confirmatory analysis, as depicted in Figure 2, was performed to determine whether or not the four first-order dimensions could be modeled using a single second-order depression factor. The fit of the second-order model on the Annville sample was very good, $\chi^2(166) = 350.55$, GFI = .923. The loss of fit from the first-order model with unconstrained factor correlations was not significant at the 1% level of confidence, and the RNFI of .995 indicated that most of the information in the first-order factor correlations was accounted for by the second-order factor loadings. In the Victoria sample, the fit of the second-order model was also very good, $\chi^2(166) = 282.43$, GFI = .909, with a nonsignificant difference in χ^2 and an RNFI = .990.

As can be seen in Figure 2, the standardized second-order factor loadings relating the first-order factors to the general Depression dimension were quite similar in the two groups, with the Interpersonal Affect factor loading considerably less than the Affect, Well-Being, and Somatic factors.

Item and Scale Distributions

The distribution of responses to the CES–D items were highly skewed and kurtotic, with a substantial number of respondents failing to endorse any CES–D item. Such a pattern may reflect an underlying non-normal distribution of depressive affect in the general population. However, it could also be the case that the scale may not be sensitive to the lower levels of depressive affect, such that the distribution of depression scale scores is truncated. The CES–D's use of frequency ratings (how often the subject has felt depressed in the past week) may contribute to the skew.

We therefore reestimated the second-order factor model, excluding subjects who had used the lowest frequency category on all the items. The solutions were very similar to the original results, indicating that the substantial number of persons at or near the minimum scale score did not materially affect the factor solutions. The entire dataset was therefore used in the subsequent analyses.

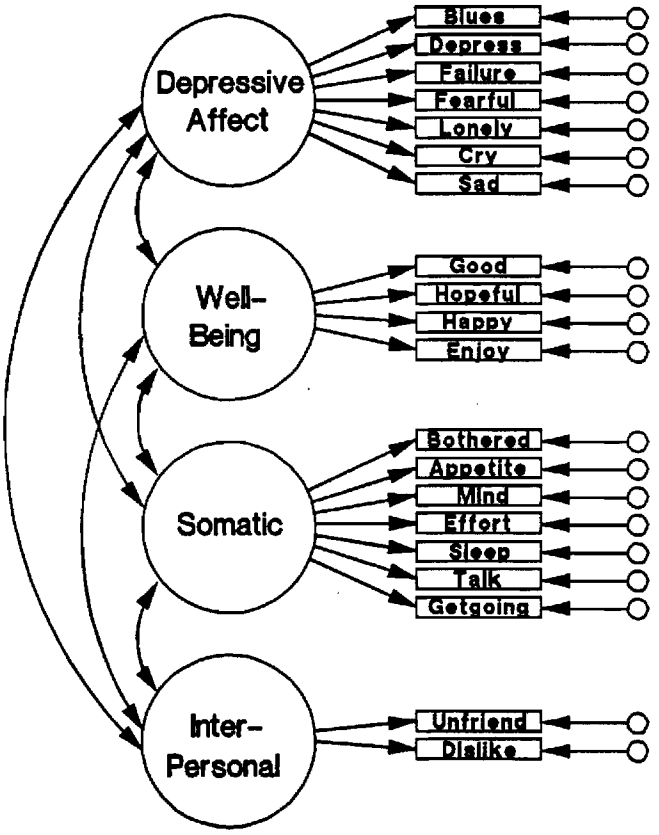


Figure 1. Four factor model for the CES–D.

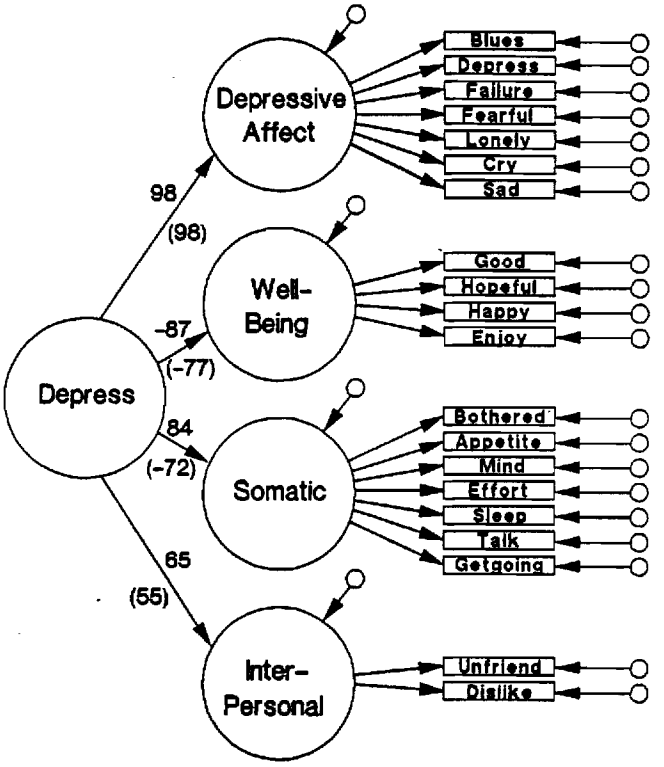


Figure 2. Second-order factor model for the CES–D with estimated loadings for the Annville and Victoria (in parentheses) samples.

Multiple Groups Analyses

To evaluate both age and sample differences, we split the Annville sample into two age groups. Individuals age 20–54 formed one group (N = 217), and those age 55 and over formed the second group (N = 230). This second group represented the same range of ages present in the Victoria sample, thus allowing for sample comparisons based on individuals of the same age. Comparisons were then made between these three groups (Annville younger, Annville older, Victoria older) to determine whether or not the groups had equivalent factor structures.

A nested series of hierarchically organized models tested group equivalence in multiple parameter matrixes (see Jöreskog, 1971; Schaie & Hertzog, 1985). The first model, M_1, specified the same first-order factor model in all three groups, but did not impose any between-group equality constraints on the parameters. The second model, M_2, forced the elements of the factor pattern matrix to be equal across the three groups. The differences in fit between M_2 and M_1 tested whether or not the groups had invariant factor loadings of items on factors. The nonsignificant change in χ^2 indicated that the hypothesis that the groups were equivalent in terms of their factor pattern was not rejected (see Tables 3 & 4).

The next set of models tested whether or not the factor covariance matrix was equivalent for the three groups. First, model M_3, which constrained the entire factor covariance matrix for the two groups of older individuals to be equal, was tested. The significant loss of fit from model M_2 indicated that the two older groups were not equivalent. Next, model M_4 constrained the two Annville samples to be equal, testing the hypothesis that the two age group subsamples constructed from the total Annville sample had equivalent factor covariances. The model also fit significantly poorer than did model M_2.

Two additional models determined whether the lack of fit in the factor covariance matrix across the three groups reflected only group differences in factor variances. Model M_5 constrained the entire factor covariance matrix to be equal across

Table 2
CES–D Item Factor Correlations

Factor	Affect	Well-Being	Somatic	Interpersonal
		Annville sample		
Affect	—			
Well-Being	−85	—		
Somatic	83	−73	—	
Interpersonal	65	−63	48	—
		Victoria sample		
Affect	—			
Well-Being	−76	—		
Somatic	71	−55	—	
Interpersonal	54	−41	47	—

Note. Decimals omitted.

Table 3
Models Testing Group Equivalence

	Model	χ^2	*df*	NFI[a]
M_1	Λ_x unconstrained	954.46	492	.841
M_2	$\Lambda_1 = \Lambda_2 = \Lambda_3$	991.54	524	.835
M_3	$\Lambda_1 = \Lambda_2 = \Lambda_3$ $\Phi_{ANOLD} = \Phi_{VICOLD}$	1027.67	534	.829
M_4	$\Lambda_1 = \Lambda_2 = \Lambda_3$ $\Phi_{ANYOUNG} = \Phi_{ANOLD}$	1073.78	534	.821
M_5	$\Lambda_1 = \Lambda_2 = \Lambda_3$ $\Phi_1 = \Phi_2 = \Phi_3$	1097.28	544	.818
M_6	$\Lambda_1 = \Lambda_2 = \Lambda_3$ covariances Φ =	1058.18	536	.824

[a] Normed fit index (see Method section).

all three groups; Model M_6 constrained the factor covariances to be equal across the three groups, while allowing the variances to differ. Model M_6 fit more poorly than did model M_5, suggesting that neither the factor variances nor the covariances were equivalent across the three groups.

Table 5 shows the differences in depression factor variances and covariances across the three groups. The younger group had greater variance than both older groups on all but one of the factors. The Annville and Victoria older groups differed significantly in variance on the Depression and Well-Being factors, but not on the Somatic and Interpersonal Affect factors.

Age Differences in CES-D Scale Scores

To examine age differences in the CES-D subscales, we classified participants into six age and birth year cohort groups (birth years 1906–1915 [ages 70–79]; 1916–1925 [60–69]; 1926–1935 [50–59]; 1936–1945 [40–49]; 1946–1955 [30–39]; and 1956–1965 [20–29]. A nonorthogonal Cohort × Sex ANOVA of the total CES-D score revealed only a significant Cohort effect, $F(5, 427) = 3.90$, $p < .01$. Figure 3 plots the least squares adjusted marginal means for Cohort across the six groups. Bonferroni-adjusted post-hoc comparisons showed that the 20–29 year-old and 40–49 year-old group had significantly higher depression scores than did the two oldest cohorts. Other contrasts did not achieve the 1% level of confidence.

Using a cutoff of CES-D scale scores of 16 or greater for possible depression (Weissman et al., 1977), 19% of the men and 23% of the women in the total Annville sample were classified as depressed. However, only 28 of the 204 individuals 60 and older (13.7%) had CES-D scores of 16 or greater.

A multivariate analysis of variance (MANOVA) on the four subscales produced a significant multivariate Cohort effect, which was manifested in significant univariate effects for Depressive Affect, $F(5, 428) = 4.17$, $p < .001$; Well-Being, $F(5, 428) = 3.96$, $p < .01$; and Interpersonal Affect, $F(5, 428) = 3.25$, $p < .01$. There were no significant cohort differences on the Somatic subscale ($p > .25$). As shown in Figure 4, the three scales with significant effects followed the same general pattern as the overall CES-D scale.

Discussion

This study is encouraging regarding the measurement properties of the CES-D in age-heterogeneous samples. These results replicate the four factor solution for the CES-D items originally proposed by Radloff (1977) on the basis of exploratory factor analyses. The present confirmatory factor analysis differs, however, in that it conclusively demonstrates that the four item factors are highly intercorrelated, a fact obscured by past studies using orthogonal rotations. Moreover, a model with a single second-order Depression factor fit the first-order item factor covariances relatively well in both samples. This pattern of results justifies use of the total CES-D score on empirical as well as conceptual grounds, although there is sufficient variance in the item factors not determined by the second-order factor to warrant concern about whether items from subscales like Interpersonal Affect ought to be used for initial screening for depression.

This study also found invariant factor pattern weights across two different older samples as well as between the middle-aged group and the two old groups. The first finding—invariance in raw score regression weights across two older samples—suggests that the minor differences in standardized factor loadings reported in Table 2 reflect unimportant sampling variability in the weights, combined with sample differences in variances of items and factors. The second finding—age-related invariance in the unstandardized factor loadings—is crucially important. It indicates probable age equivalence in the measurement properties of the CES-D, justifying use of the scale for quantitative comparisons across age levels. These results are inconsistent with recent findings of Liang et al. (1989), who found generational differences in CES-D item factor loadings in a Mexican-American population. Given the present findings of invariance in factor loadings for predominantly Caucasian, semiurban Pennsylvania, and urban Canadian populations, the lack of age-related equivalence in factor structure found by Liang et al. (1989) may be specific to their Mexican-American population.

There were significant age differences in the variances and covariances among CES-D factors. These differences were not a simple function of age (cohort) group membership, since the Victoria old and Annville old groups differed significantly. Given the skewed distribution of CES-D scores, the greater variance in the young Annville group may reflect a higher prevalence of depressed persons. Application of alternative techniques for analyzing non-normal ordinal rating scale data (e.g., Muthén & Kaplan, 1985) may be needed to determine whether group differences in factor covariances reflect influences of non-normal distributions or differences in relations among the four CES-D item factors.

Table 4
Model Comparisons

Comparison	$\Delta\chi^2$	Δdf	*p*
$M_2 - M_1$	37.08	32	*ns*
$M_3 - M_2$	38.13	10	<.001
$M_4 - M_2$	82.24	10	<.001
$M_5 - M_2$	105.74	20	<.001
$M_5 - M_3$	67.61	10	<.001
$M_5 - M_4$	23.50	10	<.01
$M_6 - M_2$	66.64	12	<.001
$M_5 - M_6$	39.10	8	<.001

Note. Models in column 1 are defined in Table 3.

Table 5
CES–D Variance/Covariance Matrixes for the Three Groups[a]

Depression	Well-Being	Somatic	Interpersonal
	Annville (age 20–54)		
612 (068)[c,d]			
565 (065)[c,d]	723 (083)[c,d]		
428 (051)[c,d]	393 (053)[c,d]	415 (055)	
250 (040)[c,d]	259 (043)[c,d]	152 (033)[c]	259 (051)[c,d]
	Annville (age 55+)		
210 (026)[b,d]			
197 (025)[b,d]	277 (035)[b,d]		
201 (026)[b]	210 (029)[b]	286 (040)	
072 (013)[b]	077 (015)[b]	063 (015)[b]	051 (015)[b]
	Victoria (age 58+)		
346 (037)[b,c]			
293 (034)[b,c]	432 (047)[b,c]		
239 (029)[b]	209 (031)[b]	335 (039)	
089 (015)[b]	078 (017)[b]	076 (015)	087 (017)[b]

Note. Decimals omitted for clarity. Standard errors in parentheses.
[a] Significance tested with a normal deviate, computed as $z = (\phi_1 - \phi_2)/\sqrt{(se_1^2 + se_2^2)}$.
[b] Significantly different from the Annville (age 20–54) group.
[c] Significantly different from the Annville (age 55+) group.
[d] Significantly different from the Victoria (age 58+) group.

This study found only minor differences in age-related patterns on the four CES–D subscales. There were age-related decreases from middle to old age in Depressive Affect, Well-Being, and Interpersonal Affect subscales, but there were no significant age differences on the Somatic scale. These results show more coherence in subscale means than found by Gatz and Hurwicz (in press), although we also found no evidence of age-related increases in Somatic scale scores. This finding may indicate that the biasing influence of somatic symptomatology is reduced in the CES–D relative to other self-report depression scales. How-

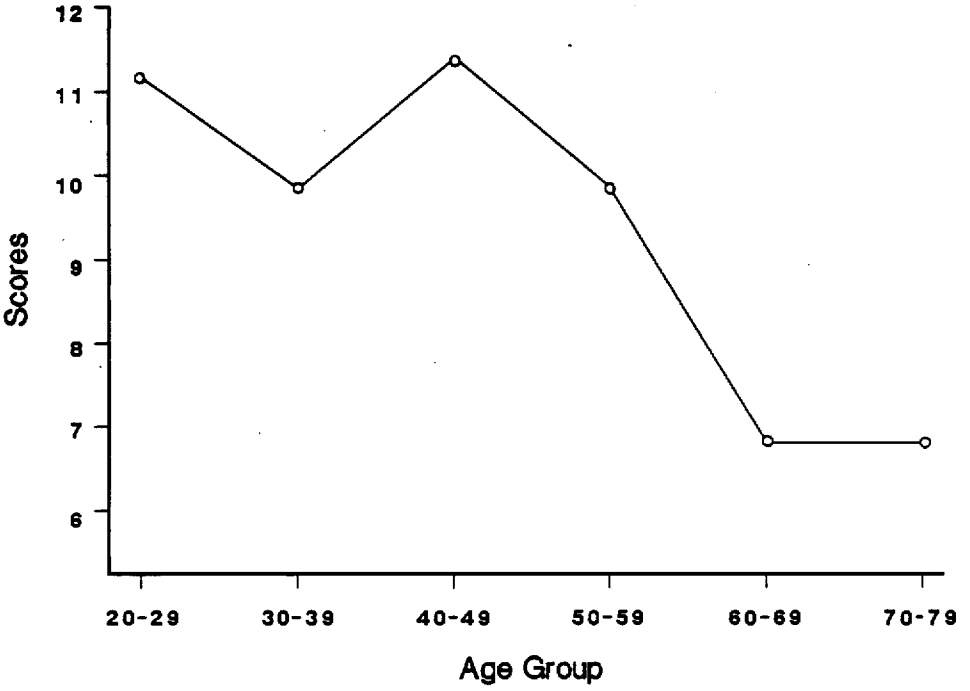


Figure 3. Mean CES–D total scale score across Age (Cohort) groups.

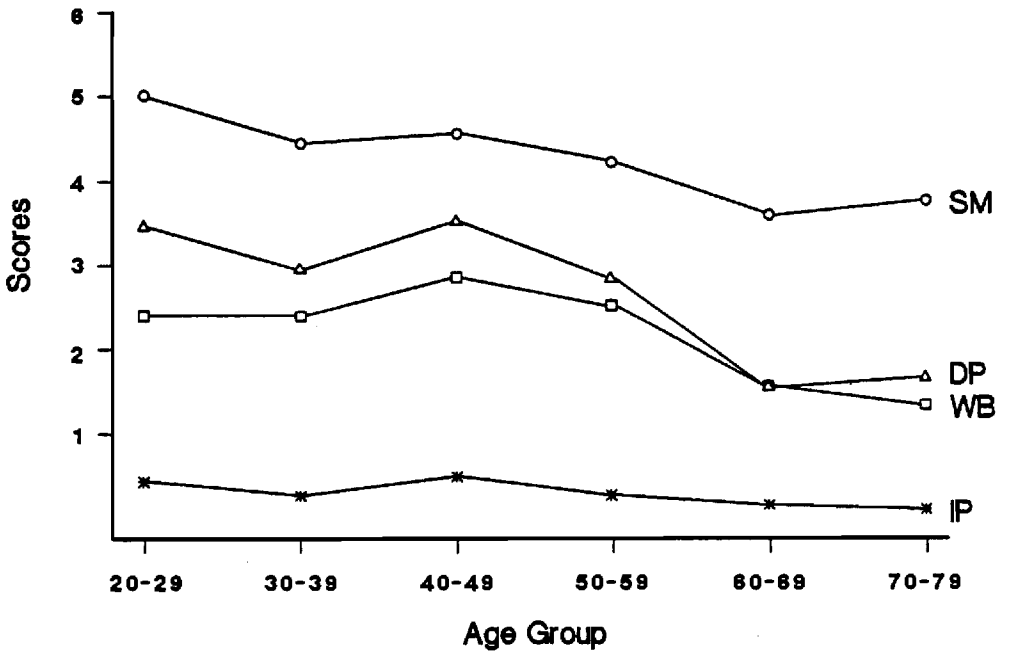


Figure 4. Mean CES–D subscale scores across Age (Cohort) groups (SM = Somatic; DP = Depression; WB = Well-Being; IP = Interpersonal Relations).

ever, it is still possible that elevated Somatic scale scores could produce false positive initial depression diagnoses when the total CES–D score is used. For example, one of our elderly men had a total CES–D score of 17 that was heavily influenced by a Somatic subscale score of 14. It is possible that classifications requiring the use of cutoffs on both the total CES–D and Depressive Affect subscale score would improve the accuracy of initial classification of unipolar depression in the elderly.

It is interesting to note that this study replicates recent findings on age differences in the CES–D in spite of the fact that it uses a volunteer (nonprobability) sample. Moreover, despite the likelihood of reduced volunteering behavior in depressed individuals, over 20% of the present sample had CES–D scores greater than the cutoff of 16. This pattern of results is encouraging to psychologists seeking to use volunteer samples to study relationships between affective status and other variables in adulthood, although the observed prevalence of depression in older persons in this sample is lower than would be expected from data reported by Murrell et al. (1983). Moreover, we cannot rule out the possibility that there are subtle selection effects (e.g., differences in the type of persons with depressed affect who volunteer versus those who do not) that may affect relationships of CES–D scores with other variables (see Nesselroade, 1988).

Although the preceding discussion has treated the age differences in CES–D scores as if they reflect aging effects, it is crucial to note that such age differences might reflect true generational (birth cohort) differences in (a) prevalence of depression and depressive affect, (b) willingness to endorse items measuring negative affect, or (c) both. Use of long-term sequential studies would be required to address this issue.

There is some question as to whether the second-order factor from the CES–D is best interpreted as a measure of depression or more broadly as a measure of general psychological distress. There is strong evidence that individual differences on self-report depression scales relate highly to other affective dimensions such as Anxiety (Aneshensel et al., 1983; Tanaka & Huba, 1984; Veit & Ware, 1983). The literature on subjective well-being also shows that measures of negative and positive affect are highly and negatively related, but that these two dimensions are not perfect polar opposites (Diener, 1984; Lawton, 1983). One might therefore be justified in treating the second-order CES–D factor combining Well-Being and Depression subscales as a composite scale combining aspects of positive and negative affect (Aneshensel et al., 1983; Roberts and Vernon, 1983). In any event, the present study suggests that both the CES–D total scale and its subscales have age-invariant measurement properties, justifying their use for assessment purposes with older populations (Radloff & Teri, 1986).

References

Aneshensel, C. S., Clark, V. A., & Frerichs, R. R. (1983). Race, ethnicity, and depression: A confirmatory analysis. *Journal of Personality and Social Psychology, 44,* 385–398.

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

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

Berry, J. M., Storandt, M., & Coyne, A. (1984). Age and sex differences in somatic complaints associated with depression. *Journal of Gerontology, 39,* 465–467.

Blazer, D., Hughes, D. C., & George, L. K. (1987). The epidemiology of depression in an elderly community population. *The Gerontologist, 27,* 281–287.

Blumenthal, M. D. (1975). Measuring depressive symptomatology in a general population. *Archives of General Psychiatry, 32,* 971–978.

Clark, V. A., Aneshensel, C. S., Frerichs, R. R., & Morgan, T. M. (1981). Analysis of effects of sex and age in response to items on the CES–D scale. *Psychiatry Research, 5,* 171–181.

Craig, T. J., & VanNatta, P. A. (1979). Influences of demographic characteristics on two measures of depressive symptoms. *Archives of General Psychiatry, 36,* 149–154.

Diener, E. (1984). Subjective well-being. *Psychological Bulletin, 95,* 542–575.

Downes, J. J., Davies, A., & Copeland, J. (1988). Organization of depressive symptoms in the elderly population: Hierarchical patterns and Guttman scales. *Psychology and Aging, 3,* 367–374.

Ensel, W. M. (1986). Measuring depression: The CES–D scale. In N. Lin, A. Dean, & W. Ensel (Eds.), *Social support, life events, and depression* (pp. 51–70). New York: Academic Press.

Gallagher, D., Thompson, L. W., & Levy, S. M. (1980). Clinical psychological assessment of older adults. In L. W. Poon (Ed.), *Aging in the 1980s* (pp. 19–40). Washington, DC: American Psychological Association.

Gatz, M., & Hurwicz, M. (in press). Are old people more depressed? Cross-sectional data on CES–D factors. *Psychology and Aging.*

Hertzog, C. (1989). Using confirmatory factor analysis for scale development and validation. In M. P. Lawton & A. R. Herzog (Eds.), *Special research methods for gerontology* (pp. 281–306). New York: Baywood.

Hertzog, C., Hultsch, D. F., & Dixon, R. A. (1989). Evidence for the convergent validity of two self-report metamemory questionnaires. *Developmental Psychology, 25,* 687–700.

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

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

Krause, N. (1986). Stress and sex differences in depressive symptoms among older adults. *Journal of Gerontology, 41,* 727–731.

Labouvie, E. W. (1980). Identity versus equivalence of psychological measures and constructs. In L. W. Poon (Ed.), *Aging in the 1980's: Psychological Issues* (pp. 493–502). Washington, DC: American Psychological Association.

Lawton, M. P. (1983). The varieties of well being. *Experimental Aging Research, 9,* 65–72.

Lewinsohn, P. M., Fenn, D. S., Stanton, A. K., & Franklin, J. (1986). Relation of age at onset to duration of episode in unipolar depression. *Psychology and Aging, 1,* 63–68.

Liang, J., Van Tran, T., Krause, N., & Markides, K. S. (1989). Generational differences in the structure of CES–D in Mexican Americans. *Journal of Gerontology: Psychological Sciences, 44,* 110–120.

Marsh, H. W. (1985). The structure of masculinity/femininity: An application of confirmatory factor analysis to higher-order factor structures and factorial invariance. *Multivariate Behavioral Research, 20,* 427–480.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103,* 391–410.

Murrell, S. A., Himmelfarb, S., & Wright, K. (1983). Prevalence of depression and its correlates in older adults. *American Journal of Epidemiology, 117,* 173–185.

Muthén, B., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38,* 171–189.

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

Phifer, J. F., & Murrell, S. A. (1986). Etiologic factors in the onset of depressive symptoms in older adults. *Journal of Abnormal Psychology, 95,* 282–291.

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

Radloff, L. S., & Teri, L. (1986). Use of the Center for Epidemiological Studies-Depression scale with older adults. *Clinical Gerontology, 5,* 119–136.

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research, 23,* 51–67.

Roberts, R. E., & Vernon, S. W. (1983). The center for epidemiologic studies depression scale: Its use in a community sample. *American Journal of Psychiatry, 140,* 41–46.

Ross, C. E., & Mirowski, J. (1984). Components of depressed mood in married men and women. *American Journal of Epidemiology, 119,* 997–1004.

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.

Steuer, J., Bank, L., Olsen, E., & Jarvik, L. F. (1980). Depression, physical health and somatic complaints in the elderly: A study of the Zung self-rating depression scales. *Journal of Gerontology, 35,* 683–688.

Tanaka, J. S., & Huba, G. J. (1984). Confirmatory hierarchical factor analysis of psychological distress measures. *Journal of Personality and Social Psychology, 46,* 621–635.

Veit, C. T., & 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.

Weissman, M. M., Sholomskas, D., Pottenger, M., Prusoff, B. A., & Locke, B. Z. (1977). Assessing depressive symptoms in five psychiatric populations: A validation study. *Journal of Epidemiology, 106,* 203–214.

Yesavage, J. A. (1986). The use of self-rating depression scales in the elderly. In L. W. Poon, T. Crook, K. L. Davis, C. Eisdorfer, B. J. Gurland, A. W. Kaszniak, & L. W. Thompson, *Handbook for clinical memory assessment* (pp. 213–225). Washington, DC: American Psychological Association.

Zarit, S. H., Eiler, J., & Hassinger, M. (1985). Clinical assessment. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (2nd ed., pp. 725–754). New York: Van Nostrand Reinhold.

(Appendix follows on next page)

Appendix

Items from Center for Epidemiologic Studies Depression Scale (CES–D)

1. I was bothered by things that usually don't bother me. (Bothered)
2. I did not feel like eating; my appetite was poor. (Appetite)
3. I felt that I could not shake off the blues even with help from my family or friends. (Blues)
4. I felt that I was just as good as other people. (Good)
5. I had trouble keeping my mind on what I was doing. (Mind)
6. I felt depressed. (Depress)
7. I felt that everything I did was an effort. (Effort)
8. I felt hopeful about the future. (Hopeful)
9. I thought my life had been a failure. (Failure)
10. I felt fearful. (Fearful)
11. My sleep was restless. (Sleep)
12. I was happy. (Happy)
13. I talked less than usual. (Talk)
14. I felt lonely. (Lonely)
15. People were unfriendly. (Unfriend)
16. I enjoyed life. (Enjoy)
17. I had crying spells. (Cry)
18. I felt sad. (Sad)
19. I felt that people dislike me. (Dislike)
20. I could not get "going." (Getgoing)

Received March 6, 1989
Revision received May 11, 1989
Accepted June 22, 1989 ■

Psychology and Aging
1990, Vol. 5, No. 2, 215-227

0882-7974/90/$00.75

Relationships Between Metamemory, Memory Predictions, and Memory Task Performance in Adults

Christopher Hertzog
Georgia Institute of Technology

Roger A. Dixon and David F. Hultsch
University of Victoria
Victoria, British Columbia, Canada

A cross-sectional sample of adults recalled categorized word lists and narrative texts. Subjects gave performance predictions before each of 3 recall trials for each task. Older subjects had poorer memory performance and also predicted lower performance levels than did younger subjects. The LISREL models suggested (a) direct effects of memory self-efficacy (MSE) on initial predictions; (b) upgrading of prediction–performance correlations across trials, determined by direct effects of performance on subsequent predictions; (c) significant effects of a higher order verbal memory factor on MSE; and (d) an independent relationship of text recall ability to initial text recall performance predictions. These results lend support to the theoretical treatment of predictions as task-specific MSE judgments.

Many psychologists interested in age-related changes in memory have focused their attention on the role that metamemory plays in memory task performance and everyday memory behaviors (Cavanaugh & Perlmutter, 1982; Dixon & Hertzog, 1988; Zelinski, Gilewski, & Thompson, 1980). *Metamemory*, defined broadly as cognitions about memory (e.g., Wellman, 1983), is a multifaceted domain that includes such constructs as strategy selection and utilization, knowledge about how memory functions, and memory self-efficacy (beliefs about one's own ability to remember; Berry, West, & Dennehey, 1989; Dixon, 1989; Hertzog, Hultsch, & Dixon, 1989).

A common technique for operationally defining metamemory in the early developmental literature was prediction of memory task performance (Schneider, 1985). The original theorists treated metamemory as knowledge and awareness of memory processes (Flavell & Wellman, 1977; Wellman, 1983), and memory predictions were often conceptualized as an index of knowledge about one's own memory (Cavanaugh & Perlmutter, 1982; Schneider, 1985). Cavanaugh (1989) interpreted predictions as an aspect of awareness of memory functioning, a construct closely tied to the concept of memory monitoring (knowledge about current memory use, contents, and states).

This research was supported by a grant (R01-AG06162) and a Research Career Development Award (K04-AG00335) from the National Institute on Aging to Christopher Hertzog. Roger A. Dixon and David F. Hultsch were also supported by operating grants from the Natural Sciences and Engineering Research Council of Canada. These results were presented at the 41st Annual Scientific Meeting of the Gerontological Society of America.

The cooperation of Robert K. Nielsen, the other physicians, and the members of the Annville Family Practice of Annville, Pennsylvania, is deeply appreciated. Thanks also to Paul Usala, Judith Van Alstine, Lori Blank, Laurie Saylor, and Michele Mobley, who assisted with text recall scoring, data preparation, and data analysis.

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

Much of the early literature on adult age differences in metamemory involved some type of memory prediction paradigm (e.g., Bruce, Coyne, & Botwinick, 1982; Murphy, Sanders, Gabriesheski, & Schmitt, 1981). A common procedure is to give subjects a description of a task with examples, or limited experience with the task, followed by a request to predict performance. The central question is usually whether there are age differences in the accuracy of performance predictions. Several studies have suggested that older subjects overestimate their performance on cognitive tasks, although not necessarily in all conditions (Bruce et al., 1982; Coyne, 1985; M. E. Lachman & Jelalian, 1984; Lovelace & Marsh, 1985; Murphy et al., 1981). In contrast, other studies have found relatively accurate memory task predictions by older adults (Camp, Markley, & Kramer, 1983; M. E. Lachman, Steinberg, & Trotter, 1987).

Recently, M. E. Lachman et al. (1987) reported results suggesting that older persons can upgrade the accuracy of their predictions after task experience. They found a significant relationship between prior recall performance and predictions of a second recall trial. Moreover, there seems to be relatively little age difference in either (a) accuracy of predictions of future performance in the form of item-by-item ratings of memorability (Lovelace & Marsh, 1985; Rabinowitz, Ackerman, Craik, & Hinchley, 1982) or (b) feeling-of-knowing judgments collected during or after the study phase of a memory task but prior to recall (e.g., Butterfield, Nelson, & Peck, 1988; J. L. Lachman, Lachman, & Thronesbery, 1979). The different pattern of age effects suggests that performance predictions after instructions or practice, but prior to the study of memory task materials, may tap a different aspect of metamemory than other prediction paradigms (Lovelace & Marsh, 1985).

What, then, is the relationship between performance predictions and metamemory? Along with others (e.g., Berry et al., 1989; Rebok & Balcerak, 1989), we have argued that memory self-efficacy represents an important unifying construct for understanding metamemory (Hultsch, Hertzog, Dixon, & Davidson, 1988). Hertzog et al. (1989) recently showed that a subset

of scales from both the Metamemory in Adulthood questionnaire (MIA; Dixon, Hultsch, & Hertzog, 1988) and the Memory Functioning Questionnaire (MFQ; Gilewski & Zelinski, 1986) converge on a strong memory self-efficacy factor. A major goal for further empirical research is to examine the relationship of memory self-efficacy, as measured by metamemory questionnaires, with memory task predictions.

To accomplish this goal, one needs to distinguish general or global memory self-efficacy beliefs from local efficacy judgments in a particular context (i.e., a particular memory task, the testing environment, and the concurrent physiological and psychological state of the rememberer; Dixon & Hertzog, 1988). Memory self-efficacy is probably a highly schematized system of beliefs regarding one's ability to use multiple types of memory in various contexts. Questionnaires such as the MIA and MFQ measure memory self-efficacy by aggregating ratings of memory capacity and forgetting across several specific memory functions (e.g., remembering names, faces, and telephone numbers). The items are generalized in that they are divorced from a specific temporal and physical context.

Performance predictions, on the other hand, may be conceptualized as self-efficacy judgments in the context of a specific memory task (e.g., Berry et al., 1989). We propose that a performance prediction is based on (a) global and local memory self-efficacy, (b) an appraisal of the memory task, and (c) an as-yet unspecified set of processes translating one's memory self-efficacy into a prediction estimate by using a representation of the distribution of task performance derived from the task assessment (see also Cavanaugh, 1989). A performance prediction will be based on a global memory self-efficacy belief in the absence of specific, local experience in the memory domain assessed by the task or on more specific beliefs about one's memory self-efficacy in familiar situations. In either case, however, the belief system must be combined with the task appraisal to produce a performance estimate.

This conceptualization of beliefs and efficacy judgments has important implications for explaining prediction behavior as well as prediction accuracy. It suggests three classes of possible reasons for inaccurate performance predictions: (a) inaccurate memory self-efficacy, either global or local; (b) inaccurate appraisal of the memory task and, by implication, inaccurate representation of the distribution of task performance; and (c) faulty mapping of memory self-efficacy beliefs onto the performance distribution. Shaw and Craik (1989) recently reported results that can be interpreted as indicating faulty memory task appraisal by subjects. They found that performance predictions of both older and younger subjects were insensitive to experimental manipulations known to affect cued-recall performance, even though subjects were informed of the manipulations. As was the case in several studies cited earlier, Shaw and Craik's (1989) older subjects overestimated their recall performance, predicting recall performance approximately equivalent to that of the younger subjects, but performing more poorly than the younger adults. Thus, overestimation of performance by older persons might not reflect overconfidence in memory ability, but rather, inaccurate task appraisal—for example, an inaccurate estimate of how well the average individual performs on that memory task.

This study used structural regression models to address the nature and degree of relationships between metamemory, memory predictions, and memory performance. The primary focus was on modeling individual differences in these variables, although hypotheses about mean age differences more typical of the existing literature were also examined. Adult participants performed three times on two different memory tasks: free recall of words (nouns from multiple taxonomic categories) and free recall of narrative texts. They also predicted performance levels before each recall task. Two types of memory task were used because the issue of whether there are differential relationships of performance on different memory tasks to memory self-efficacy and to task-specific predictions is still unresolved. Some have argued that memory self-efficacy beliefs will relate more highly to tasks that assess memory as it is used in everyday life (Berry et al., 1989). Although significant metamemory–memory task performance relationships are not always found, most studies using text recall tasks have detected significant relationships of text recall with questionnaire measures of memory self-efficacy (e.g., Cavanaugh & Poon, 1989; Dixon & Hultsch, 1983; Sunderland, Watts, Baddeley, & Harris, 1986; Zelinski et al., 1980). Perhaps the need to recall information from text materials occurs relatively often in everyday life, enhancing the accuracy of self-efficacy beliefs. However, Zelinski et al. (1980) and Cavanaugh and Poon (1989) reported significant memory self-efficacy relationships with both text recall and word-list recall.

The prediction paradigm was designed to minimize the influence of individual differences in task assessments on memory task predictions by giving subjects prior information about average performance levels on the task. Assuming multiple influences on memory predictions, we reasoned that the correlation between general memory self-efficacy beliefs and predictions would be maximized when individuals already possess, or are explicitly given, information about normative levels of task performance. Moreover, given task experience, individuals should use memory-monitoring skills to form task-specific performance evaluations and self-efficacy beliefs (Cavanaugh, 1989; Cavanaugh & Perlmutter, 1982). This process should improve prediction accuracy over recall trials (Herrmann, Grubs, Sigmundi, & Grueneich, 1986; M. E. Lachman et al., 1987) but lower the correlations of performance predictions on later trials with more general memory self-efficacy beliefs. We therefore hypothesized (a) a significant relationship between memory self-efficacy, as measured by the MIA and MFQ, and memory task performance predictions; (b) significant relationships between memory self-efficacy and memory task performance; (c) higher relationships of text recall to memory self-efficacy and text predictions than of word recall to memory self-efficacy and word-list predictions; and (d) significant increases in the correlation of predictions with performance across trials. In addition, we assigned subjects to prediction and no-prediction conditions to assess possible reactive effects of making predictions on recall performance.

Method

Subjects

The sample consisted of adults, aged 20 to 79, drawn from the greater Lebanon Valley of Pennsylvania. The total sample included 422 com-

Table 1
Demographic Data for Prediction and No-Prediction Groups

Group	*n*	Age		Education		Vocabulary		Self-rated health[a]	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Prediction									
Younger									
Men	34	35.9	7.2	14.4	3.6	32.8	9.2	6.9	1.0
Women	51	37.3	6.0	13.7	2.5	33.8	10.2	7.6	1.2
Middle-aged									
Men	22	52.6	4.6	13.9	3.8	32.9	8.4	6.6	1.0
Women	31	54.1	4.1	12.3	2.1	31.4	11.1	6.7	1.5
Older									
Men	41	65.6	5.2	13.1	2.8	34.1	10.9	6.0	1.3
Women	46	67.0	5.5	12.7	2.4	36.2	9.4	5.8	1.7
No prediction									
Younger									
Men	24	38.0	5.6	14.5	4.0	32.4	11.7	6.8	1.3
Women	40	36.8	5.8	14.4	2.9	34.1	8.4	7.5	1.2
Middle-aged									
Men	19	52.2	4.5	13.4	2.7	34.8	10.4	6.7	1.5
Women	38	53.2	4.1	12.6	2.4	34.2	8.6	6.7	1.5
Older									
Men	32	68.4	5.0	13.7	3.4	35.8	9.8	6.0	1.3
Women	44	67.6	5.0	13.4	3.3	38.3	8.9	5.7	1.9

[a] Rated on a 1–9 scale, with 1 = *poor* and 9 = *excellent*.

munity-dwelling adults. The sample was recruited by mail from the membership of a large medical family practice located in Annville, Pennsylvania, and was supplemented by a snowballing technique. The recruitment letter and telephone script informed potential participants that tests of memory would be part of the study. Subjects were paid a nominal fee of $15 for their participation.

Subjects were assigned to either a memory task prediction or no-prediction condition (see below). For some analyses, the sample was divided into three age groups (20–45, 46–59, and 60–79). Descriptive statistics on age, education, vocabulary scores, and self-rated health for subjects in the Prediction × Age Group × Sex factorial design are shown in Table 1. The sample was relatively heterogeneous in vocabulary and educational attainment. Participants generally reported themselves to be in good health. There were more women than men, and there were relatively few young adults (under age 30) in the sample.

Measures

Metamemory. Metamemory was measured with multiple scales from the MIA and MFQ. The construct of memory self-efficacy was the primary aspect of metamemory emphasized in this study. Memory self-efficacy was measured by the Capacity scale of the MIA and by the Frequency of Forgetting scale from the MFQ. Both have high reliability and have been shown to be convergent measures of a higher order memory self-efficacy factor (see Hertzog et al., 1989). The MIA Capacity scale requires Likert scale ratings of statements like, "I am good at remembering names." The MFQ Frequency of Forgetting scale requires frequency judgments on how often one forgets specific types of information (e.g., names and appointments; Gilewski & Zelinski, 1986).

The remaining metamemory scales from the 108-item version of the MIA were also administered: Strategy (use of mnemonics and external memory aids in everyday life), Task (knowledge of memory processes and functions), Achievement (achievement motivation regarding memory), Anxiety (degree of anxiety involving memory), and Locus (perceived control over memory). Additional information regarding these scales may be found in Dixon et al. (1988).

Vocabulary. A vocabulary test was formed by pooling 54 items from the ETS Kit of Factor-Referenced Cognitive Tests (Ekstrom, French, Harman, & Dermen, 1976), measures of V2, V4, and V5. Subjects were given 15 min to take the test.

Categorized free recall. Three lists of 30 nouns were constructed, with 6 words from each of five taxonomic categories. All exemplars were highly typical of their category, according to norms established by Howard (1980). Categories were assigned to lists so as to maximize category distinctiveness (i.e., avoiding joint testing of categories such as trees and fruits). The List 1 categories were metals, animals, trees, sports, and flowers. The List 2 categories included relatives, fruits, birds, furniture, and weapons. List 3 was vegetables, insects, fish, jobs, and toys. The order of words in each list was randomized, under the constraint that exemplars from the same category could not be adjacent in the list. The same word lists were presented to all subjects.

Two measures were used for each list: the total number of words recalled and the adjusted ratio of clustering (ARC) developed by Roenker, Thompson, and Brown (1971). The ARC reflects degree of clustering of words by categories during recall. Random clustering is reflected by a score of 0; perfect clustering is indicated by a score of 1.0. For purposes of data analysis, undefined ratios were treated as missing data.

Text recall. Three narrative stories were constructed and analyzed according to a hierarchical propositional system developed by Kintsch (1974). Stories 1, 2, and 3 contained 149, 154, and 158 propositions, respectively. Recall protocols were scored for presence of each proposition, with the criterion that a proposition was scored as present if the gist of its meaning was expressed correctly, irrespective of the words used to express it (Turner & Greene, 1977). Although differential recall of propositions from different organizational levels of texts is a well-established phenomenon (e.g., Hultsch & Dixon, 1984; van Dijk &

Kintsch, 1983), this study measured text memory by proportion of total propositions recalled. There were several reasons for this decision. First, this study is primarily concerned with prediction–performance relationships instead of text recall per se. Second, an analysis of variance (ANOVA) differentiating levels as a within-subject factor showed typical levels effects but no interactions of levels with other factors (including age and prediction condition). Third, examination of correlations of text recall and text prediction suggested little variation in correlations across levels.

Recall protocols were stratified by age of respondent and then assigned at random to one of three raters. Rater reliabilities were obtained by having all three raters score data for 10 subjects (30 stories), distributed over the entire age range. Interrater reliability, estimated by using repeated measures ANOVA to estimate reliable and unreliable variance components, was .93.

Predictions. Before each memory task, subjects in the memory prediction condition were given a brief description of the task. They were then told: "Before you actually do the task described above, we would like to ask you to tell us how well you think you will do. It may help you to know that on this task the average person is able to remember about [15 of 30 words on the categorized word recall task or 25 of 50 ideas on the text recall task]." These values were arbitrarily selected to anchor predictions in the middle of the range of possible responses.

Procedure

Subjects were tested in small groups consisting of between 5 and 15 persons. The data were collected as part of a larger construct validation study of the MIA and MFQ (Hertzog et al., 1989). All questionnaires and tasks were given in two separate sessions lasting approximately 2 hr. Subjects were scheduled in sessions separated by exactly 1 week, with few exceptions, although individuals unable to attend the scheduled second session were retested whenever possible. The metamemory questionnaires were given in the first session, along with several other questionnaires. The second session began with the vocabulary test, followed by the memory task booklet.

Memory task booklets either did or did not request predictions and confidence levels. Groups of subjects were assigned in advance to receive one type of booklet. Assignment of groups was made randomly under the constraint that an attempt was made to get a similar age distribution in each condition. To achieve this goal, deliberate assignment of some groups to prediction conditions was necessary during the latter stages of data collection.

Recall test booklets were arranged with memory tasks in invariant order: Word List 1, Story 1, Word List 2, Story 2, Word List 3, and Story 3. Memory tasks were paced by the experimenter. Subjects were allowed 2 min for study and 5 min for recall of each word list. They were given 3 min to read and 7 min to recall each story. Word lists and texts were presented in invariant order to facilitate individual differences analyses. This design has the disadvantage of confounding list (or text) with trial, but counterbalancing order across subjects would have allowed order effects to influence the correlations of recall measures with other variables.

Statistical Procedures

Analysis of (a) reactive effects of predictions on memory task performance and (b) age and sex differences in predictions used repeated measures multivariate analysis of variance (MANOVA) with orthogonal polynomial contrasts to analyze the within-subject trials factor of the multiple recall trials (Hertzog & Rovine, 1985).

Relationships among metamemory, predictions, and performance were analyzed with the LISREL VI program (Jöreskog & Sörbom, 1984). LISREL uses full-information maximum likelihood methods to estimate parameters of structural regression models (Hayduk, 1987). Identification is achieved by placing sufficient restrictions on the parameters, often in the form of fixed coefficients (e.g., fixing the regression of two unrelated variables to 0). LISREL produces parameter estimates, standard errors, and associated fit statistics. The overall fit of the model is evaluated by a likelihood ratio chi-square test and by a goodness-of-fit index (GFI). The GFI ranges from 0 to 1.0, with values greater than 0.9 traditionally considered good (Anderson & Gerbing, 1988). Competing models may be compared in terms of the fit indexes, plausibility and parsimony of parameter estimates, and additional model diagnostics, such as residuals (differences between predicted and observed variances and covariances) and modification indexes indicating fixed parameters that are possible sources of poor model fit.

Results

Reactive Effects of Predictions

Reactive effects of predictions were analyzed in a 2 × 3 × 2 × 3 (Prediction × Age × Sex × Trial) MANOVA, with repeated measures on the trial variable. Table 2 provides the means and standard deviations for the 417 subjects with complete data on all three dependent measures: word recall, ARC scores, and text recall.

The MANOVA showed neither significant prediction effects nor significant interactions involving Prediction × Trial effects for any of the dependent measures. In particular, the between-groups effect of prediction and age was not significant (Wilks's $\Lambda = 0.975$), multivariate $F(6, 806) = 1.69$, $p > .10$, indicating no significant impact of making predictions on age differences in memory performance. Although all individuals had been assigned to either the prediction or no-prediction condition, the group testing environment allowed some individuals to omit usable predictions. Sixty-five of the 225 persons assigned to the prediction group were missing at least one prediction (usually the first word recall prediction). Age-group membership related to the frequency of missing data, $\chi^2(2, N = 225) = 18.90$, $p < .001$; 60% of the cases with missing predictions were age 60 or older. Subjects omitting predictions also had significantly fewer years of education, as well as significantly lower word and text recall performance than those subjects making all predictions. However, the two groups did not differ significantly in metamemory scales from the MIA and MFQ. An analysis excluding subjects with missing data produced a modest univariate Prediction × Age interaction for word recall, $F(2, 341) = 3.75$, $p < .03$, but not for text recall ($p > .10$). The marginal means for word recall showed larger age differences in the no-prediction condition (older subjects recalled 4.25 fewer words than did younger subjects) than in the prediction condition (older subjects recalled 3.29 fewer words). Although this result might indicate a true differential impact of predictions (e.g., making predictions increases motivation in older persons), it is also consistent with the hypothesis that persons omitting predictions are less cognitively able and that exclusion because of missing data positively biased the older prediction group (presumably, low-education and low-memory-ability subjects who would have been at risk for omitting predictions remained in the no-prediction group).

In contrast to the lack of salient prediction effects, the analysis produced strong effects of age (Wilks's $\Lambda = 0.751$), $F(6,$

Table 2
Recall Performance for Prediction × Age × Sex Groups

Age group (years)	Dependent variable: Word recall *M*	Word recall *SD*	ARC *M*	ARC *SD*	Text recall *M*	Text recall *SD*
Prediction						
20–45						
Men (n = 33)						
Trial 1	19.27	3.50	.489	.371	.306	.119
Trial 2	19.97	4.56	.722	.224	.306	.112
Trial 3	20.03	4.39	.592	.332	.335	.129
Women (n = 51)						
Trial 1	20.63	3.96	.531	.341	.349	.112
Trial 2	21.71	4.20	.771	.243	.324	.103
Trial 3	20.90	4.80	.583	.316	.347	.135
46–59						
Men (n = 22)						
Trial 1	16.32	4.40	.364	.395	.257	.111
Trial 2	15.96	4.46	.521	.353	.238	.098
Trial 3	15.82	4.10	.291	.295	.243	.107
Women (n = 31)						
Trial 1	17.00	4.97	.419	.426	.277	.086
Trial 2	18.65	4.39	.685	.268	.252	.078
Trial 3	19.16	5.11	.577	.274	.249	.081
60–79						
Men (n = 40)						
Trial 1	15.78	4.74	.523	.316	.240	.071
Trial 2	14.46	4.56	.598	.442	.199	.059
Trial 3	14.28	4.57	.421	.431	.188	.081
Women (n = 44)						
Trial 1	17.09	3.48	.542	.289	.283	.093
Trial 2	17.25	4.07	.710	.237	.254	.091
Trial 3	16.64	4.41	.533	.325	.241	.098
No prediction						
20–45						
Men (n = 24)						
Trial 1	18.96	3.95	.450	.353	.286	.113
Trial 2	20.92	5.05	.618	.491	.237	.108
Trial 3	19.88	5.76	.514	.384	.267	.109
Women (n = 40)						
Trial 1	19.80	3.61	.470	.348	.373	.111
Trial 2	21.18	4.60	.618	.328	.320	.125
Trial 3	20.55	4.21	.493	.381	.358	.153
46–59						
Men (n = 19)						
Trial 1	17.79	5.95	.584	.370	.272	.088
Trial 2	18.58	6.18	.579	.362	.251	.106
Trial 3	17.47	6.02	.430	.374	.269	.119
Women (n = 38)						
Trial 1	19.63	4.50	.572	.266	.310	.088
Trial 2	20.61	3.98	.724	.225	.285	.092
Trial 3	20.34	4.36	.605	.268	.272	.100
60–79						
Men (n = 31)						
Trial 1	14.94	4.70	.553	.359	.247	.079
Trial 2	15.74	5.45	.661	.351	.210	.081
Trial 3	14.97	4.87	.531	.360	.212	.106
Women (n = 44)						
Trial 1	16.27	4.49	.510	.369	.270	.079
Trial 2	17.09	4.38	.650	.257	.227	.073
Trial 3	16.75	4.24	.546	.329	.233	.094

Note. ARC = adjusted ratio of clustering.

806) = 20.67, $p < .001$, and sex (Wilks's $\Lambda = 0.952$), $F(6, 678) = 6.75$, $p < .001$. There was an age-related decrease in word recall and text recall performance, but there were no age differences in clustering behavior. Women performed significantly better than men on both recall tasks, and women clustered responses more during word recall, $F(1, 405) = 4.92$, $p < .05$.

The MANOVA tests for the within-subject effects of trial, $F(6, 400) = 30.63$, $p < .001$, and the Age × Trial interaction, $F(6, 400) = 3.53$, $p < .001$, were also significant. Polynomial trend contrasts showed salient quadratic effects for all three dependent variables. The quadratic trend for ARC scores was significant, $F(1, 341) = 106.19$, $p < .001$. All groups showed greater clustering for List 2 than for List 1 or 3, which did not differ, and there was a matching, weaker trend in word-list recall, with highest performance on List 2. Significant linear, $F(1, 405) = 31.91$, $p < .001$, and quadratic, $F(1, 405) = 43.30$, $p < .001$, effects for text recall reflected higher overall performance on the first story. The overall Age × Trial interaction was isolated to a significant difference between younger and older participants in the linear trend across trials for text recall, $F(1, 405) = 10.81$, $p < .001$. Older subjects showed significantly lower recall than did younger subjects on the third text relative to the first text.

Given the focus of the study on individual differences in memory and metamemory, we tested the homogeneity of covariance matrices for the nine performance measures (word recall, ARC, and text recall) between the prediction and no-prediction groups. Box's test approximated its expected value under H_0, $\chi^2(45, N = 353) = 44.41$, *ns*, justifying acceptance of the null hypothesis of equality of the prediction-group covariance matrices.

Age Differences in Predictions

Word and text recall predictions for the 157 subjects with complete data were analyzed in a 3 × 2 × 3 (Age × Sex × Trial) MANOVA. There were significant multivariate age differences, $F(4, 306) = 4.98$, $p < .001$, with increasing age associated with lower predicted performance levels for both text and word recall.

The analysis also revealed significant trial and Sex × Trial effects. The trial effect was restricted to a significant linear increase in the word recall predictions, $F(1, 154) = 13.20$, $p < .001$, and a weak linear trend for increase in text recall predictions, $F(1, 154) = 4.11$, $p < .05$. The Sex × Trial interaction involved only the linear trend for word recall predictions. Figure 1 shows the unweighted marginal means for word recall predictions separately for men and women, as well as the average

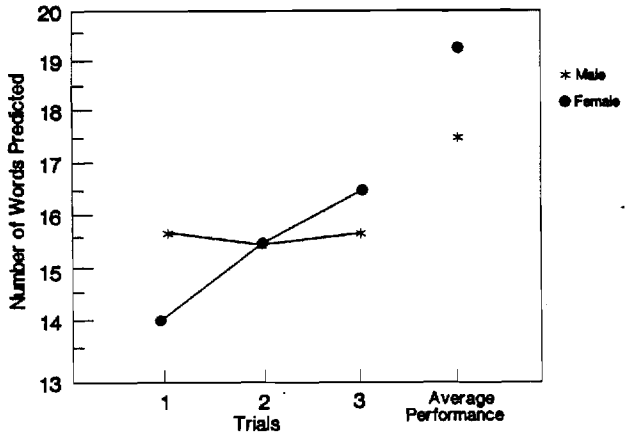


Figure 1. Word-list predictions and performance for men and women.

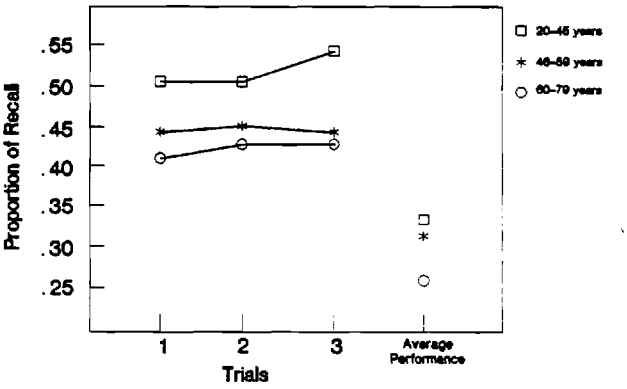


Figure 3. Text predictions and performance for three age groups.

word recall performance (pooled over trials). The pooled standard deviation across all cells was 4.1 words. The significant interaction reflected a disordinal increase in women's predictions across trials. Women and men alike underestimated their performance on List 1, probably because they had been told that average performance was 15 words recalled, although the empirically observed mean recall was approximately 18 words. Women predicted lower performance than did men for List 1 but actually outperformed them. Subsequently, women's predictions showed significant increases over time, presumably because they were able to monitor their performance and update their predictions.

Figure 2 provides the unweighted marginal means for the three age groups in word recall predictions, along with the average word recall performance (pooled over trials). All three age groups increased their performance estimates slightly, but the age differences in the amount of increase were not significant, in spite of the fact that younger adults' performance levels were considerably higher than their initial prediction.

Figure 3 shows the unweighted marginal means of the text recall predictions for three age groups. To facilitate comparisons with recall performance, predictions were rescaled as proportion of the 50 idea units given to subjects as total number of ideas in the text.[1] The pooled standard deviation across cells was 0.14 (14%). Text recall predictions showed virtually no change across the three trials. Indeed, even though average performance levels were lower than predicted levels, text recall predictions increased slightly over trials.

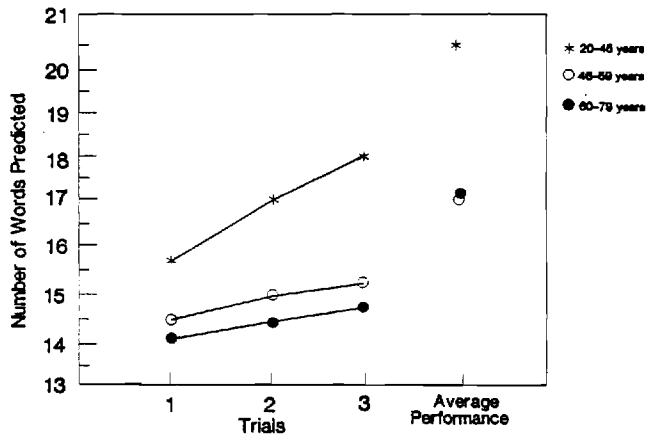


Figure 2. Word-list predictions and performance for three age groups.

Prediction–Performance Differences

Most previous studies of prediction accuracy have analyzed difference scores between predictions and actual performance (e.g., Bruce et al., 1982; cf. Shaw & Craik, 1989). An ANOVA of recall–prediction difference scores for the 164 cases with complete data for word-list recall produced significant effects for sex, $F(1, 158) = 9.21$, $p < .01$; age, $F(2, 158) = 3.47$, $p < .05$; and the Age × Sex interaction, $F(2, 158) = 4.02$, $p < .05$. There were no interactions involving the trial variable, and the main effect for trial did not achieve statistical significance ($p > .05$).

The difference scores on Trial 1 for word recall for all Age × Sex groups are shown in Table 3. Women underestimated performance more than men did, but the differences were inconsistent across age groups. Older women's predictions were significantly closer to actual performance than were younger women's predictions, but the same comparison was not significant for men (and indeed, the trend was in the opposite direction). The provision of prior normative information that was graded by neither age nor sex, given significant age and sex effects in recall, complicates interpretation of the difference scores, as discussed later. The corresponding analysis of recall–prediction difference scores for text recall produced a different pattern of results. There were no effects associated with age, but women were more accurate than men, $F(1, 172) = 5.18$, $p < .05$. Accuracy increased linearly across trials, $F(1, 172) = 5.55$, $p < .05$, but the modest improvements did not interact with either age or sex.

[1] The performance benchmark in Figure 3 is somewhat arbitrary, given the lack of perfect correspondence between text propositions as defined by the Kintsch system and the term *idea* provided to subjects.

Table 3
Recall–Prediction Difference Scores (Proportion Correct) for Initial Recall Trials

Group	Word recall (n = 164)		Text recall (n = 178)	
	M	*SD*	*M*	*SD*
	Men			
Young	.09	.17	.21	.12
Middle-aged	.02	.17	.20	.10
Old	.11	.13	.18	.10
	Women			
Young	.19	.18	.14	.14
Middle-aged	.15	.19	.15	.16
Old	.09	.13	.15	.13

Structural Models of Performance

Zero-order correlations. Correlations of metamemory scales and the memory performance predictions for the 155 subjects with complete data on all metamemory, prediction, and performance variables are shown in Table 4. Note the significant correlations of MIA Capacity, MIA Change, and MFQ Frequency of Forgetting scales with word and text recall predictions. These scales relate to the higher order memory self-efficacy factor identified by Hertzog et al. (1989). In contrast, the correlations of other MIA scales (Strategy, Task, and Achievement) with performance predictions were small and generally nonsignificant. This finding suggested that predictions were related primarily to the memory self-efficacy factor. The correlations between memory self-efficacy measures were largest for the first prediction but then declined slightly across trials.

Table 5 shows the correlations of memory performance predictions and actual memory performance for word and text recall. There was a significant but modest correlation of the first word recall task with the first word recall prediction, but subsequent correlations between predictions and word recall performance increased dramatically. The highest sample correlations were associated with the immediately prior task performance with the subsequent prediction. For example, performance on Trial 1 correlated .58 with the prediction for Trial 2, whereas predicted Trial 2 performance correlated .52 with actual Trial 2 performance. Correlations of text recall predictions and text recall task performance were initially higher than the correlations between word prediction and recall, but those between text prediction and recall showed less increase across trials.

Initial regression analyses. Hierarchical regression analysis was performed to determine whether there were interactions between age, metamemory, and predictions. The analysis entered (a) the MIA Capacity, Task, and Strategy scales; (b) age and sex; (c) the three MIA Scale × Age interactions; and (d) the three MIA Scale × Sex interactions into an equation using the first word recall prediction as the dependent variable. In Stage 1, the three MIA scales related significantly to the first word recall prediction (R^2 = .18), $F(3, 151) = 10.93$, $p < .001$. The MIA Capacity scale produced the only significant regression coefficient ($\beta = .42$, $t = 5.43$, $p < .001$).

Although significant age differences in the word recall predictions had been observed (see earlier text), age was not a significant predictor of word recall predictions when added to the equation in the second hierarchical stage, $F(1, 150) = 1.07$, $p > .05$. This outcome suggested that age differences in predictions may be mediated by age differences in memory self-efficacy. There were no significant interactions involving age or sex, suggesting no age differences in the relationships of MIA scales to predictions. Similar analyses showed no interactions of age and metamemory measures in relation to word-list recall, as well as no interactions of age and recall in determination of the second word-list performance prediction. Subsequent LISREL models were therefore conducted without using multiplicative interaction terms associated with age.

The LISREL model for word prediction and recall. The first model used the word recall data to identify the relationship between age, sex, predictions, ARC measures of clustering, and word recall. The major issue was whether predictions reflect an upgraded efficacy judgment on the basis of past performance or whether predictions reflect upgrading of an efficacy judgment on the basis of both past performance and concurrent evaluation of present state (e.g., level of fatigue prior to memory performance). Cross-lagged regression models were used to deter-

Table 4
Correlations Among Metamemory and Memory Predictions

Metamemory scale	Prediction 1		Prediction 2		Prediction 3	
	Word	Text	Word	Text	Word	Text
Strategy	.06	.16*	.17*	.18*	.25**	.13
Task	.14*	.20**	.07	.22**	.14*	.17*
Locus	.14*	.06	−.04	.01	−.01	−.07
Achievement	.03	.08	−.03	.09	.01	.02
Anxiety	−.25**	−.14*	−.10	−.11	−.08	−.10
Capacity	.39**	.37**	.20**	.32**	.33**	.27**
Change	.41**	.38**	.25**	.31**	.27**	.26**
Frequency of forgetting	.36**	.30**	.24**	.26**	.27**	.20**

* $p < .05$. ** $p < .01$.

Table 5
Correlations of Recall Predictions With Recall Performance

Trial	Performance	Trial 1	Trial 2	Trial 3
	Word recall prediction			
1	Recall	.24*	.58*	.62*
1	ARC	.08	.35*	.40*
2	Recall	.24*	.52*	.71*
2	ARC	.11	.30*	.40*
3	Recall	.29*	.52*	.62*
3	ARC	.07	.20*	.35*
	Text recall prediction			
1	Recall	.44*	.54*	.51*
2	Recall	.50*	.54*	.54*
3	Recall	.58*	.58*	.58*

Note. ARC = adjusted ratio of clustering.
* $p < .01$.

mine whether the lagged effect of past performance on the next prediction was sufficient to account for the correlations among predictions and memory performance (Schaie & Hertzog, 1985).

We initially specified a first-order autoregressive model in which each variable determined itself over time. For example, word recall for List 1 determined recall for List 2, which in turn determined recall for List 3. However, we found that we needed to add second-order lagged regressions for all three sets of performance measures (e.g., recall of List 1 predicting recall of List 3 independent of the mediated effect through recall of List 2) in order to fit the data. The model's standardized regression coefficients clearly demonstrated upgrading of performance estimates on the basis of prior performance (Figure 4). The model yielded roughly equal stability for prediction and lagged determination of prediction by prior recall performance (e.g., amount of recall of List 1 determining prediction for List 2). However, an alternative model adding concurrent predictions of recall by the immediately preceding prediction did not improve the fit.

Given the relative absence of relationship between clustering and all variables except recall performance, the clustering measures were dropped from subsequent LISREL models.

Structural model for word and text recall. The next set of models used (a) three word recall predictions, (b) three word recall performances, (c) three text recall predictions, (d) three text recall performances, and (e) two measures of memory self-efficacy (i.e., MIA Capacity scale and MFQ Frequency of Forgetting scale). A central question was whether text recall influenced subsequent text recall performance predictions, as was the case for word recall, and if so, would it be statistically independent of the upgrading of word recall predictions? An initial model specifying simultaneous autoregressive sequences for text and word recall measures was rejected because it provided very poor fit to the correlations between text and word recall measures and text and word predictions. This result led us to reconsider the conceptual basis for modeling relationships among prediction and performance measures.

One possibly faulty assumption of the autoregressive model was that each recall performance can be treated as a separate variable with a direct influence on subsequent performance. It was more parsimonious to assume no causal relation between recall over trials, and instead we assumed that recall of each word list reflects an underlying latent ability for free recall of words from a categorized list and that recall of each text measures a latent text recall ability. If each performance is an observed measure of an underlying memory ability construct, then any attempt to fit a first-order autoregressive model to covariances actually determined by a latent variable would result in residual covariances between the Trial 1 and Trial 3 measures. Similarly, the strong residual relationships between text and word recall measures suggested that the sizable correlation of the text and word recall latent variables could not be fit adequately by an autoregressive model, which mediates the relationship between text and word recall through a correlation of recall of the first word list with recall of the first story.

This alternative perspective led us to test a model with only latent variables for predictions and recall performance, ignoring trial-specific relationships and changes in individual differences across trials. We anticipated that such a pure latent-variable model would inadequately represent the upgrading of performance predictions so clearly evident in the zero-order correlations, especially for word recall. At issue, however, was (a) how well the latent variable model would fit the recall performance correlations and (b) whether, subsequently, more complex models would provide an even better fit to the data. The first pure latent-variable model specified five factors: Word Recall Prediction, Word Recall Performance, Text Prediction, Text Performance, and Memory Self-efficacy. This model actually fared relatively well, $\chi^2(67, N = 155) = 142.94$, $p < .001$, GFI = 0.884. However, the factor loadings on the Word Recall Prediction factor differed substantially, with the lowest loading for predictions of Trial 1 performance, and there were some large modification indexes. One obvious source of stress was a

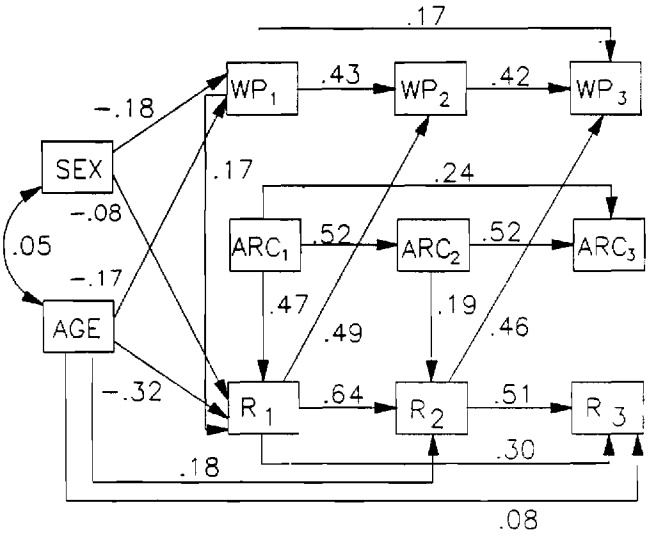


Figure 4. Autoregressive model for word recall prediction and performance. (WP = word recall prediction; ARC = adjusted ratio of clustering; R = word recall.)

high residual between the first word recall prediction and the metamemory measures. A model adding factor loadings of the first word recall prediction and the first text recall prediction on the Memory Self-efficacy factor still contained an implausible pattern of factor loadings and some large residual correlations between predicted and actual recall measures, confirming the expectation that memory ability and memory prediction latent variables alone could not fully account for the data.

Another possible problem with the original autoregressive models involved the direction of effects specified in the model. The regression approach had been driven implicitly by the predictive validity question: Are recall predictions statistically associated with recall? However, we thought it more parsimonious to assume that latent text and word recall abilities cause memory self-efficacy beliefs, which in turn influence task performance predictions. A second problem derived from the need for a latent-variable specification for recall performance. How should the phenomenon of upgraded performance predictions across trials be modeled in the context of a latent ability framework for memory? We reasoned that the proximal cause of the predictions in the later trials of a multitrial design is not the latent ability to recall text or words per se, but rather the actual performance on the task. One evaluates the performance on the current task and revises subsequent estimates accordingly. This perspective suggested that latent memory ability determines memory self-efficacy and, indirectly, initial performance predictions, but that actual performance on a particular recall trial determines changes in predictions between trials.

We executed a series of models implementing this alternative perspective. The first model, M_1, was actually a measurement model similar to the pure latent-variable models previously estimated, but it allowed each prediction variable to be a separate factor. This model formed the baseline for evaluating any structural equation model using it as the base measurement model (Anderson & Gerbing, 1988). The model fit well, $\chi^2(47, N = 155) = 72.13$, $p = .01$, GFI = 0.938, confirming the hypothesis that much of the poor fit of the pure latent-variable models derived from between-trials differences in the correlations of predictions and performance variables.

The second model, M_2, specified an endogenous Memory Self-efficacy factor, determined jointly by latent Text Memory and Word Memory factors. Memory Self-efficacy, in turn, influenced performance predictions. M_2 also specified both autoregressive relations among text and word predictions and lagged influences of word predictions on immediately following text predictions (e.g., first word recall prediction influencing first text recall prediction). M_2 fit slightly better than the pure latent-variable model but fit significantly poorer than M_1, difference in $\chi^2(19, N = 155) = 59.95$, $p < .001$. A salient improvement in fit was achieved by adding a direct path of the latent Text Memory factor to the first text recall prediction in the third model, M_3, difference in $\chi^2(1, N = 155) = 33.59$, $p < .001$. Indeed, the difference in fit between M_3 and M_1 became small relative to the difference in degrees of freedom. A major problem with M_3, however, was apparent multicollinearity between the Text Memory and Word Memory factors in their determination of Memory Self-efficacy, evident in the low regression weights but high correlations among the three factors. The estimated correlation of .68 between the Text Memory and Word Memory factors indicated that a higher order Verbal Memory factor may have been determining both memory factors and mediating the relationship of latent memory ability to memory self-efficacy.

Model M_4 added a higher order Verbal Memory factor but maintained an independent path from Text Memory to the first text recall prediction. M_4 fit approximately the same as M_3, $\chi^2(67, N = 155) = 100.38$, $p < .01$, GFI = 0.913, suggesting little cost in specifying the higher order Verbal Memory factor. We next attempted an alternative specification, in which an orthogonal Prediction factor was added to the model. This factor was intended to account for systematic individual differences in prediction behavior (for example, individual differences in the tendency to under- or overestimate performance). The model still allowed autoregressive coefficients among the text and word recall predictions but accounted for all correlations between text and word predictions in terms of the latent Prediction factor. The model specified Memory Self-efficacy to influence the first word recall prediction and first text recall prediction. It also specified an independent, direct effect of Text Memory on the first text prediction. This model, M_5, fit about the same as model M_4. Modification indexes suggested further paths of text recall variables to the immediately following word recall prediction. The paths of the first text recall performance to second word recall prediction and the second text recall performance to third word recall prediction were added, generating model M_6. It fit better than did its predecessors, $\chi^2(61, N = 155) = 83.54$, $p = .03$, GFI = 0.931, and did not differ appreciably from model M_1, difference in $\chi^2(14, N = 155) = 11.41$, *ns*.[2]

The final model included age and sex as exogenous variables. The model fit well, $\chi^2(85, N = 155) = 117.51$, $p = .01$, GFI = 0.92, in spite of a highly restricted specification of relationships of age and sex with the other latent variables. The standardized regression coefficients (factor loadings and structural regression coefficients) are shown in Figure 5. There are several noteworthy features of the model. First, the latent variables were well defined, as shown by (a) high loadings of the Text Memory and Word Memory factors on the second-order Verbal Memory factor and (b) strong relationships of the Memory Self-efficacy factor to the MIA Capacity and MFQ Frequency of Forgetting scales. Second, the model suggests that the modest correlation between the first word recall performance and the first word recall prediction was actually mediated by a salient path from Verbal Memory to Memory Self-efficacy and, in turn, a similarly salient path from Memory Self-efficacy to the first word recall prediction. Third, the relationship of Memory Self-efficacy to the first text recall prediction, although statistically significant, was relatively weak. Instead, there was a direct effect

[2] We attempted to fit two alternative models using the orthogonal prediction factor specification. A model removing the path from text memory to text prediction and forcing this relationship to be mediated by a path from the first word recall performance to the first text recall prediction fit worse than did M_6. A second alternative model mediating the text recall to first text prediction effect through the higher order verbal memory factor was rejected because the estimated factor loading of text memory on the higher order factor increased to 0.95. Thus, there was a specific relationship between text memory and the first text recall prediction independent of word recall performance.

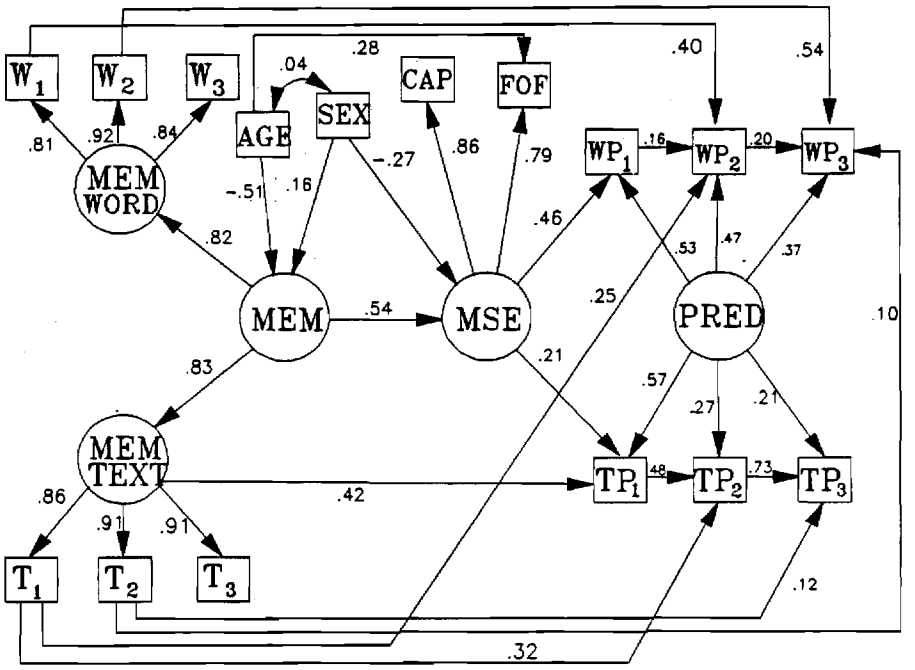


Figure 5. Structural regression model for recall performance and predictions (standardized solution). (W = word recall; CAP = Capacity scale; FOF = Frequency of Forgetting scale; WP = word recall prediction; MEM WORD = Word Memory factor; MEM = higher order Verbal Memory factor; MSE = Memory Self-efficacy factor; PRED = Prediction factor; MEM TEXT = Text Memory factor; TP = text recall prediction; T = text recall.)

of Text Memory to the first text prediction. Fourth, all loadings of the predictions on the Prediction factor were statistically significant, but the first two predictions (word recall and text recall) had the largest loadings. Fifth, there was evidence of upgrading of performance predictions, in the form of low autoregressive coefficients for predictions and salient lagged coefficients from prior recall performance to the next recall prediction, for both the word recall and text recall variables. The lagged effects were much larger for word recall, however. Finally, sex, but not age, had a direct effect on Memory Self-efficacy independent of Verbal Memory, with women demonstrating lower Memory Self-efficacy than that of men. On the other hand, covariances of predictions with both age and sex were mediated by Memory Self-efficacy and Text Memory.

Discussion

Several findings of this study support the thesis that memory performance predictions should be viewed as task-specific memory self-efficacy judgments (Berry et al., 1989; Cavanaugh, 1989). First, metamemory scales known to relate principally to memory self-efficacy (Hertzog et al., 1989) correlated significantly with predictions for both word recall and text recall. Second, measures of memory self-efficacy correlated more with predictions than did metamemory scales largely independent of memory self-efficacy, including Achievement, Strategy, and Task scales. Third, the LISREL models successfully represented the covariances between the MIA Capacity and MFQ Frequency of Forgetting scales as a direct effect of a Memory Self-efficacy latent variable on predictions. Fourth, there was no direct influence of age, sex, and memory performance on the first word recall prediction when Memory Self-efficacy was statistically controlled.

The age differences in memory task predictions merit additional discussion. Unlike many previous studies that showed older subjects predict equal or higher performance than do younger adults, older adults in this study predicted poorer recall than did younger adults. This pattern of mean age differences in predictions covaries with the patterns of age differences seen for questionnaire measures of memory self-efficacy (e.g., Hultsch, Hertzog, & Dixon, 1987). This result may be a direct consequence of this study's alternative method for obtaining performance predictions—providing participants with prior information about normative levels of task performance. The rationale for this procedure was that normative information would reduce the influence of individual differences in memory task appraisal, in terms of likely levels of performance, on predictions. Thus, this study was not designed to determine the relative influences of task appraisal and memory self-efficacy on actual predictions. However, the fact that older persons in this study did not differ from younger persons in the degree of prediction accuracy leads to the speculation that previous findings of overestimation of performance levels by older adults may be influenced by inaccurate task assessments. At a minimum, age differences in overestimation shown by other studies should not necessarily be interpreted as higher task-specific efficacy beliefs on the part of older individuals.

We did not experimentally manipulate the nature and the accuracy of information about task performance provided to subjects. Indeed, we emphasize that the average performance information provided to subjects was neither accurate, in terms of actual performance levels of the sample, nor age graded. This design feature was appropriate for the individual differences analyses that were the primary focus of the study, but it creates problems for interpreting mean age differences in discrepancies between predictions and performance. The degree to which the relationship between memory self-efficacy measures and performance predictions would vary as a function of prior knowledge and amount of experimenter-provided information about the distribution of task performance is not yet known. Nevertheless, our results, when contrasted with those of previous studies, suggest that new experiments manipulating information given to subjects in prediction paradigms may prove useful in teasing apart the influences of memory self-efficacy and task evaluation on task-specific efficacy judgments.

We found no evidence for reactive effects of predictions on individual differences in memory task performance and little evidence of effects of making predictions on mean levels of memory task performance. There was also no evidence of prediction effects on text recall. A marginal trend for an Age × Prediction interaction in word recall was detected when subjects missing one or more valid prediction responses were excluded from the analysis, but the mean differences were small and consistent with the hypothesis that the effect was an artifact of differential subject selection resulting from omission of predictions. These results support the generalizability of previous prediction studies that have opted not to include a no-prediction control group.

The results partially supported the hypotheses regarding relationships of memory self-efficacy and memory performance. One important unanticipated finding was that there was relatively little difference between the two types of memory tasks in the magnitude of their relationship to memory self-efficacy, as measured by the two metamemory questionnaires. We were able to model the relationship of both types of memory tasks to the Memory Self-efficacy latent variable through a higher order Verbal Memory factor that had equally high loadings on both the Text Memory and Word Memory factors. Moreover, there was no direct effect of either Text Memory or Word Memory on Memory Self-efficacy independent of the higher order Verbal Memory factor. This finding is consistent with reports of relationships of both types of recall task with metamemory scales (e.g., Cavanaugh & Poon, 1989). It is inconsistent with the hypothesis that discrepancies in the literature regarding memory–metamemory relationships are attributable to a higher association of text recall with metamemory.

The higher order Verbal Memory factor significantly influenced the latent Memory Self-efficacy factor, with a standardized regression coefficient of .54. This relationship is higher than might be expected from the previous literature (e.g., Dixon & Hultsch, 1983; Sunderland et al., 1986). One reason is that the LISREL estimate of the structural regression coefficient is disattenuated for measurement error in both the recall tasks and the metamemory scales. The salient but moderate relationship between the two latent variables supports the position that memory self-efficacy beliefs are based in part on actual memory ability but are not necessarily veridical (Hultsch et al., 1988).

Despite the equivalent relationship of text and word recall to general memory self-efficacy, the structural equation models demonstrated a salient influence of Text Memory on the first text performance prediction, independent of the relationship of Verbal Memory to text predictions via general Memory Self-efficacy. This specific relationship is consistent with the hypothesis that individuals have more implicit knowledge about their text recall abilities (e.g., Dixon, 1989). Note that in our study, all participants received the word recall measures first. We tested—and rejected—several models that attempted to represent the differential relationship of text recall to text predictions as being mediated by prior exposure to the word recall task (given the high correlation of word and text recall).

The two types of memory tasks also differed with respect to both initial magnitude and subsequent degree of change in their relationship with performance predictions. As in previous studies (Herrmann et al., 1986; M. E. Lachman et al., 1987), predictions of word recall performance showed strong upgrading in degree of accuracy across trials, manifested as substantial increases in the correlations of predictions with performance. The structural equation models showed that this relationship was best represented as a lagged effect of word recall performance on subsequent performance predictions. There was no discernible direct influence of predictions on recall within the same trial, even though recall immediately followed predictions. Text recall produced lagged effects on text predictions that were much weaker in magnitude. Indeed, the effect of the second text recall on the third text prediction was not statistically reliable.

The upgrading of accuracy was more subtle in mean levels of predictions. There was a significant Sex × Trial interaction for word recall predictions but not for text recall predictions. Initially, women predicted lower word recall than did men; women subsequently increased their predicted performance more than men so that by the third trial, the sex differences in predictions were consistent with the sex differences in performance (women recalled more words than did men). All three age groups increased mean levels of predicted word recall performance toward actual performance levels, even though the prior information that average recall performance level is 15 words was repeated at each trial. In contrast, the mean story estimates show little movement toward the actual mean performance.

The most plausible explanation for this pattern of prediction upgrading is that individuals could more easily monitor their own performance levels on the word recall task. This accuracy in memory monitoring could be due to concurrent awareness of the status of the memory system (e.g., Cavanaugh, 1989) or to some kind of postrecall performance evaluation not necessarily associated with conscious awareness of the contents of memory (e.g., counting words recalled before the end of the time allotted for recall). The upgrading observed is consistent with results from other prediction paradigms, in which item-specific predictions of recall after words are studied are more accurate than initial predictions (Lovelace, 1984; Lovelace & Marsh, 1985). Additional studies will be required to identify the psychological mechanisms responsible for the upgrading of prediction accuracy. The reduced level of upgrading for text recall predictions

probably reflects the greater difficulty in gauging how well one is doing on that task.

There was no evidence of age differences in prediction accuracy or in relation of memory self-efficacy to performance predictions, which is generally consistent with lack of age differences in other types of performance prediction and memory-monitoring paradigms (e.g., Butterfield et al., 1988; Lovelace & Marsh, 1985). The inferences that can be drawn from this study are limited, however, by the provision of average performance levels to subjects. The norms provided did not reflect actual mean recall performance in this sample and may have therefore influenced age-group differences in prediction accuracy. The presentation of the same average performance levels to the subjects on each trial may also have reduced the amount of mean upgrading (e.g., by increasing the probability that individuals would dismiss high or low scores as chance occurrences).

The modest sex differences, favoring women, found on both text recall and word recall were not consistent with an absence of sex differences in questionnaire measures of memory self-efficacy found by Hultsch et al. (1987), as well as with the lower initial performance predictions by women in this study. This discrepancy is manifested in the structural equation model shown in Figure 5 as a direct, negative effect of sex on Memory Self-efficacy. When Verbal Memory was modeled as a direct cause of Memory Self-efficacy, women's perceived self-efficacy is seen to be low relative to their actual memory ability.

The set of structural equation models could be criticized on several grounds: (a) We conducted extensive model development without a cross-validation sample, (b) we modeled the relationships between memory self-efficacy and memory performance solely in terms of an effect of Verbal Memory on Memory Self-efficacy, and (c) we modeled the text and word recall measures as stable indicators of underlying memory ability constructs. With respect to Point a, there is nothing inappropriate about developing a new model in the context of analyzing a data set (see Hertzog, in press), but the model ultimately requires replication in a new, independent sample. Regarding Point b, the memory-to-prediction direction of effects might seem inconsistent with the idea that low self-efficacy beliefs adversely influence performance, perhaps indirectly by increasing test anxiety or reducing motivation (Bandura, 1986). This study showed no direct effect of predictions on recall. That is, there was no effect independent of the lagged relationship between past recall and predictions, as would be expected if self-efficacy judgments influence contiguous recall performance via affect, motivation, and the like. With respect to Point c, the specification of stable latent memory factors was consistent with the large correlations of each memory task with itself across trials, as well as the large and stable correlations between text recall and word recall. The stability of individual differences in cognitive performance is not surprising. For example, Hertzog and Schaie (1986) found high stability of individual differences in psychometric intelligence over 7-year retest intervals.

Although our results justify specification of the memory latent variables, one should not rule out the type of self-efficacy and performance relationship hypothesized by Bandura (1986) and others in the domain of memory. Perhaps inhibition of performance would be more readily observed for individuals with very low levels of self-efficacy (or very high levels of negative affect), resulting in a weak degree of association when calculated across the entire range of self-efficacy beliefs and affect. Or perhaps individuals at highest risk for poor performance, and with lowest prior self-efficacy beliefs, are much less likely to volunteer for memory experiments. Finally, a full examination of the inhibitory effects of self-efficacy on memory performance probably requires concomitant measurement of probable intervening variables such as task-specific performance anxiety.

References

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin, 103,* 411–423.

Bandura, A. (1986). *Social foundation of thought & action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice-Hall.

Berry, J. M., West, R. L., & Dennehey, D. M. (1989). Reliability and validity of the memory self-efficacy questionnaire. *Developmental Psychology, 25,* 701–713.

Bruce, P. R., Coyne, A. C., & Botwinick, J. (1982). Adult age differences in metamemory. *Journal of Gerontology, 37,* 354–357.

Butterfield, E. C., Nelson, T. O., & Peck, V. (1988). Developmental aspects of the feeling of knowing. *Developmental Psychology, 24,* 654–663.

Camp, C. J., Markley, R. P., & Kramer, J. J. (1983). Spontaneous use of mnemonics by elderly individuals. *Educational Gerontology, 9,* 57–71.

Cavanaugh, J. C. (1989). The importance of awareness in memory aging. In L. W. Poon, D. C. Rubin, & B. A. Wilson (Eds.), *Everyday cognition in adult and late life* (pp. 416–436). New York: Cambridge University Press.

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

Cavanaugh, J. C., & Poon, L. W. (1989). Metamemorial predictors of memory performance in young and older adults. *Psychology and Aging, 4,* 365–368.

Coyne, A. C. (1985). Adult age, presentation time, and memory performance. *Experimental Aging Research, 11,* 147–149.

Dixon, R. A. (1989). 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 adult and late life* (pp. 394–415). New York: Cambridge University Press.

Dixon, R. A., & Hertzog, C. (1988). A functional approach to memory and metamemory development in adulthood. In F. E. Weinert & M. Perlmutter (Eds.), *Memory development: Universal changes and individual differences* (pp. 293–330). Hillsdale, NJ: Erlbaum.

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

Dixon, R. A., Hultsch, D. F., & Hertzog, C. (1988). The metamemory in adulthood (MIA) questionnaire. *Psychopharmacology Bulletin, 24,* 671–688.

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

Flavell, J. H., & Wellman, H. M. (1977). Metamemory. In R. V. Kail & J. W. Hagen (Eds.), *Perspectives on the development of memory and cognition.* Hillsdale, NJ: Erlbaum.

Gilewski, M. J., & Zelinski, E. M. (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.

Hayduk, L. A. (1987). *Structural equation modeling with LISREL: Es-*

sentials and advances. Baltimore, MD: Johns Hopkins University Press.

Herrmann, D. J., Grubs, L., Sigmundi, R., & Grueneich, R. (1986). Awareness of memory ability before and after relevant memory experience. *Human Learning, 5,* 91–107.

Hertzog, C. (in press). On the utility of structural equation models for developmental research. In P. B. Baltes, D. L. Featherman, & R. M. Lerner (Eds.), *Life-span development and behavior* (Vol. 10). Hillsdale, NJ: Erlbaum.

Hertzog, C., Hultsch, D. F., & Dixon, R. A. (1989). Evidence for the convergent validity of two self-report metamemory questionnaires. *Developmental Psychology, 25,* 687–700.

Hertzog, C., & Rovine, M. (1985). Repeated-measures analysis of variance in developmental research: Selected issues. *Child Development, 56,* 787–810.

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

Howard, D. V. (1980). Category norms for adults between the ages of 20 and 80. *JSAS: Catalog of Selected Documents in Psychology, 10,* 7. (Ms. No. 2009)

Hultsch, D. F., & Dixon, R. A. (1984). Memory for text materials in adulthood. In P. B. Baltes & O. G. Brim, Jr. (Eds.), *Life-span development and behavior* (Vol. 6, pp. 77–108). New York: Academic Press.

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

Hultsch, D. F., Hertzog, C., Dixon, R. A., & Davidson, H. (1988). Memory self-knowledge and self-efficacy in the aged. In M. L. Howe & C. J. Brainerd (Eds.), *Cognitive development in adulthood: Progress in cognitive development research* (pp. 65–92). New York: Springer-Verlag.

Jöreskog, K. G., & Sörbom, D. (1984). *LISREL VI user's guide.* Mooresville, IN: Scientific Software.

Kintsch, W. (1974). *The representation of meaning in memory.* Hillsdale, NJ: Erlbaum.

Lachman, J. L., Lachman, R., & Thronesbery, C. (1979). Metamemory through the adult life span. *Developmental Psychology, 15,* 543–551.

Lachman, M. E., & Jelalian, E. (1984). Self-efficacy and attributions for intellectual performance in young and elderly adults. *Journal of Gerontology, 39,* 557–582.

Lachman, M. E., Steinberg, E. S., & Trotter, S. D. (1987). Effects of control beliefs and attributions on memory self-assessments and performance. *Psychology and Aging, 2,* 266–271.

Lovelace, E. A. (1984). Metamemory: Monitoring future recallability during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 756–766.

Lovelace, E. A., & Marsh, G. R. (1985). Prediction and evaluation of memory performance by young and old adults. *Journal of Gerontology, 40,* 192–197.

Murphy, M. D., Sanders, R. E., Gabriesheski, A. S., & Schmitt, F. A. (1981). Metamemory in the aged. *Journal of Gerontology, 36,* 185–193.

Rabinowitz, J. C., Ackerman, B. P., Craik, F. I. M., & Hinchley, J. L. (1982). Aging and metamemory: The roles of relatedness and imagery. *Journal of Gerontology, 37,* 688–695.

Rebok, G. W., & Balcerak, L. J. (1989). Memory self-efficacy and performance differences in young and old adults: The effect of mnemonic training. *Developmental Psychology, 25,* 714–721.

Roenker, D. L., Thompson, C. P., & Brown, S. C. (1971). Comparison of measures for the estimation of clustering in free recall *Psychological Bulletin, 76,* 45–48.

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* (pp. 61–92). New York: Van Nostrand Reinhold.

Schneider, W. (1985). Developmental trends in the metamemory-memory behavior relationship: An integrative review. In D. L. Forrest-Pressley, G. E. Mackinnon, & T. G. Waller (Eds.), *Cognition, metacognition, and human performance* (Vol. 1, pp. 57–109). New York: Academic Press.

Shaw, R. J., & Craik, F. I. M. (1989). Age differences in predictions and performance on a cued recall task. *Psychology and Aging, 4,* 131–135.

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.

Turner, A., & Greene, E. (1977). *The construction and use of a propositional text base* (Tech. report). Boulder: University of Colorado.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* New York: Academic Press.

Wellman, H. M. (1983). Metamemory revisited. In M. T. H. Chi (Ed.), *Trends in memory development research* (pp. 31–51). Basel, Switzerland: Karger.

Zelinski, E. M., Gilewski, M. J., & Thompson, L. W. (1980). Do laboratory tests relate to self-assessment 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, NJ: Erlbaum.

Received January 9, 1989
Revision received August 24, 1989
Accepted August 25, 1989 ■

Intraindividual Change in Text Recall of the Elderly

Christopher Hertzog

Georgia Institute of Technology

Roger A. Dixon and David F. Hultsch

University of Victoria

Date: August 17, 1990

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

Abstract

Patterns of change and variability in text recall performance were assessed in seven elderly women by testing them weekly for up to two years. Results showed markedly different patterns of intraindividual change in gist recall performance for different participants. The two women whose performance declined were characterized by deteriorating physical health. Texts having female protagonists yielded superior recall performance. There was significant intraindividual variability after adjustment for text effects, which may indicate that weekly fluctuations in psychological states of the participants influenced their memory performance.

Introduction

Memory for prose has played an increasingly important role in research on normative and pathological changes in verbal memory in adults. Cognitive psychologists studying normal aging have found a complicated pattern of age differences in story recall, and have focused on a number of variables that influence text recall performance (Meyer, 1987). In neuropsychological assessment, memory for paragraphs has been successfully used to discriminate normal adults from other populations such as amnesics (Kopelman, Wilson, & Baddeley, 1989), Alzheimer's disease (Flicker, Ferris, Crook, Bartus, & Reisberg, 1986), multiple sclerosis (Grant, McDonald, Trimble, Smith, & Reed, 1984; Rao, Leo, & St. Aubin-Faubert, 1989), and Huntington's disease (Caine, Banford, Schiffer, Shoulson, & Levy, 1986). Several neuropsychologists have recommended inclusion of text memory measures in general neuropsychological assessment batteries (e.g., Egelko et al., 1988; Kopelman, 1986; Wilson, 1987).

Current theories in cognitive psychology posit that memory for stories requires an active interplay between prior knowledge and new information. Reading and understanding text requires access to information held in long-term semantic memory and the

use of this information in constructing an active representation of new information through the processes of verbal comprehension. Comprehension is characterized by the construction of integrated representations of old and new information in the form of new idea units, often termed propositions, about the content of the passage (van Dijk & Kintsch, 1983). These idea units represent abstract concepts that are not necessarily expressed solely in terms of the surface structure of the text. Remembering a story is typically characterized by recall of concepts and relations among concepts even when retention of the actual words used in the passage is limited. For this reason, some cognitive research with both normal and disadvantaged populations has focused on the gist recall of various types of idea units, not just the quantity of word-for-word reproductions. Gist recall performance has been shown to be influenced by a number of characteristics of both the reader (e.g., ability) and the story (e.g., structure; see Hultsch & Dixon, 1984).

The Wechsler Memory Scale (WMS) Logical Memory scale is a frequently used measure of story recall (Erickson & Scott, 1977). It contains two simple paragraphs, each with approximately 20 segments. Performance on this task is measured in terms of the total number of segments recalled verbatim, although some investigators give partial credit for gist recall. The Logical Memory scale has several advantages in clinical applications. It

is easy to administer and score, and its widespread use has generated considerable normative data on performance levels in different populations. However, the scale may have limitations for some applications. The idea units of Logical Memory scale stories are not grounded in any formal theory of text representation in memory. Consequently, rules governing the segmentation of the text are not explicitly stated, the differences across segments in terms of importance for the story's theme are unknown, and the scoring criteria regarding verbatim and gist recall do not necessarily reflect the nature of propositional units stored in memory. Moreover, the brief length and structural simplicity of the stories may not afford differentiation of subtle distinctions in memory dysfunction in special populations.

Several authors have commented on the value of integrating recent trends in cognitive psychology--and, in particular, cognitive discourse analysis--with neuropsychology and cognitive aging (e.g., Kahn, Joanette, Ska, & Goulet, 1990). Recent research in these areas has begun to address the problem of how to best represent story structures and score story recall in special populations. As is known from basic cognitive psychological work with complex stories (e.g., Kintsch, 1974) it is possible to measure not only the quantity of idea units (number of propositions or segments) recalled, but also

qualitative differences in the types of ideas recalled such as level of importance in the story and type of recall errors (Dall'Ora, Della Sala, & Spinnler, 1989; Dixon, Hultsch, Simon, & Von Eye, 1984; Kopelman, 1987). Schultz, Schmitt, Logue, and Rubin (1986) explicitly examined the idea units in Russell's (1975) revision of the WMS Logical Memory scale. They found quantitative differences in recall among multiple patient groups accompanied by similar profiles of variation in recall of each idea unit aggregated within the groups.

Recent research on normal adult populations has employed more complex stories based upon formal theories of text representation. Reliable age differences favoring younger adults have been found for recall of information from multiple kinds of text (Hultsch & Dixon, 1984; Meyer, 1987; Zelinski & Gilewski, 1988). Indeed, several studies have examined age differences in recall as a function of level of thematic importance of the idea units. Adults of all ages show strong influences of level of importance on recall, but Age X Levels interaction effects are inconsistent and may depend upon other variables (Dixon et al., 1984). In general, text recall performance appears to be influenced by a number of individual differences characteristics, such as previous knowledge about the story's subject matter, verbal ability, education, and reading comprehension (Hartley, 1986; Hultsch, Hertzog, & Dixon, 1984, in press). In some cases

age differences have been found to interact with such variables (Hultsch et al., 1984; Meyer, 1987).

Although individual differences in cognitive performance have received a fair amount of study, the issue of intraindividual (within-person) variability over time is rarely addressed, despite informal observations of clinicians that there are fluctuations in the functional competence of elderly individuals. Indeed, as Kaszniak (1990) noted, lack of consistency over time in certain behaviors and constructs can be expected for elderly persons. Such intraindividual variation in performance is not merely a reflection of unreliability in the task, but rather is determined both by imprecision of measurement and lawful but unstable influences on performance. Moreover, fluctuations in intraindividual performance levels will influence individual differences as measured in a cross-sectional design. In the case of memory performance, intraindividual variability may occur lawfully as a function of intraindividual state changes and shifting environmental influences (Dixon & Hertzog, 1988). Individual differences in memory performance within a population (e.g., older normals or patients) will reflect (a) stable individual differences in memory ability, (b) current status of the individual (e.g., health, mood), and (c) random measurement error. It is a largely untested assumption that the magnitude of within-group error in older samples reflects sources (a) and (c),

but not (b).

What if it were the case, however, that a meaningful proportion of between-persons variance in memory was due to intraindividual influences that were unstable over time? Researchers examining mood states have shown that substantial proportions of variance in state measures are in fact unstable, such that test-retest correlations are low even though the reliability of the measures is relatively high (e.g., Nesselroade, Jacobs, & Pruchno, 1986; Spielberger, 1977). Indeed, several studies of intraindividual variation in affect suggest coherent patterns of flux in mood that relate to variables such as perceived stress and social behavior (Watson, 1988; Zevon & Tellegen, 1982). There is evidence for effects of physiological and psychological states on cognition, including mood (e.g., Bower, 1981; Broadbent, Broadbent, & Jones, 1989; Clark & Teasdale, 1982; Eich & Metcalfe, 1989; Sutton, Teasdale, & Broadbent, 1988). Intraindividual variation in mood and other variables may therefore produce intraindividual variation in memory performance.

Substantial intraindividual variability in text memory performance could have implications for clinical memory assessment. It could materially increase the variance within normal control groups, reducing statistical power and requiring larger sample sizes to obtain sensitive tests of patient group

differences. More important, intraindividual variability could harm the validity of cross-sectional, age-stratified norms on tests like the Logical Memory scale for diagnosis of memory impairment in the elderly. Benign senescent forgetting (Kral, 1962) or the more recently proposed syndrome of age-associated memory impairment (AAMI; Crook et al., 1986) is characterized by poor memory performance that is not secondary to a pathological physical process (e.g., Alzheimer's disease or cerebral infarct). One proposed criterion for AAMI is memory task performance that is one SD below norms for young adults (Crook et al., 1986). Intraindividual variation in performance might cause an elderly individual to be classified as having AAMI one day, but to be judged to have normal memory functioning another day.

Use of between-persons group norms for standardized memory tests to aid in the diagnosis of AAMI or other age-related, progressive pathologies of memory requires the assumption that low norm-referenced performance indicates within-person decline. It is true that the expected value of an individual's performance is the age-group mean, and that performance below the cutoff has a higher probability of being influenced by age-related decline, but there are other influences on performance. Reports from the Baltimore Longitudinal Study of Aging show normative age changes in visual and verbal memory performance (Arenberg, 1982), but individual differences in the amount of longitudinal change are

not highly correlated with initial status (Alder, Adam, & Arenberg, in press). Moreover, the estimated magnitude of mean intraindividual changes with increasing age in those studies is relatively small when scaled against the distribution of individual differences. One cannot be certain that an individual exhibiting low performance has actually declined over time if one has not directly or indirectly estimated change from pre-morbid status. Alternatively, age-related decline in an individual with high pre-morbid memory ability might go undetected if assessment were based only on between-subjects norms.

Given these concerns, short- and long-term longitudinal case studies may be a valuable method for assessing both normative and pathological decline in older persons' memory function. In such designs each individual serves as his/her own control, and change over time can be evaluated against a standard error of measurement based upon observed intraindividual variability in performance (Hertzog & Schear, 1989). There are, however, two major potential obstacles to implementation of such designs in assessment contexts. The first problem involves the magnitude of intraindividual variability relative to true, long-term intraindividual change (Nesselroade, 1986). To the extent that variability in performance is large, long periods of observation may be required to detect statistically reliable decrement (Salthouse, Kausler, & Saults, 1986). The second obstacle is the

high likelihood of practice effects on memory task performance, even in the elderly (Hultsch, 1974). This problem would be especially acute if the same instrument were repeatedly administered over short time periods. In the extreme, practice effects could lead to ceiling effects in performance, especially on measures of memory for simple materials, such as in paragraph recall. One could therefore wonder whether longitudinal case-study designs are feasible for evaluating change in cognitive performance in the elderly and whether the pattern of decline across multiple cognitive attributes has diagnostic value (with respect to either etiology or long-term prognosis).

The present study was designed with three major purposes in mind. First, we wished to investigate the extent of intraindividual variability in story recall performance in normal older adults. Second, we wanted to determine whether reliable intraindividual change in text memory could be detected over a relatively short time period (one to two years) given the magnitude of background intraindividual variability. Finally, we wanted to investigate the implications of short-term intraindividual designs for assessment of story memory. In order to accomplish these goals, it was necessary to measure individuals frequently over a relatively extended period of time. Frequent measurement required, in turn, the use of a large number (25) of parallel stories which could be administered in blocks in

order to minimize practice effects, thus enhancing our ability to estimate age-related decline in text recall performance. The stories we used were developed using a propositional coding system that enabled standardization of story recall scoring according to gist criteria (Dixon, Hultsch, & Hertzog, 1989). The goal of evaluating the assessment implications of intraindividual case-study approaches was addressed by obtaining an initial norm-referenced assessment of text recall, using similar texts. Performance was compared to normative data on these stories collected on a cross-sectional sample from the same geographic region. The norm-referenced assessment data was then compared to the data obtained from the intraindividual panel design, thus determining the degree of consistency in initial performance and subsequent change in text recall over time.

Method

Subjects

A panel of seven older women was recruited in two phases from a retirement community in central Pennsylvania. Three older women (IDs P01, P02, and P03) were recruited in the first phase in late 1985; four women (IDs P04 through P07) were recruited in spring, 1986. The women were all over the age of 65 at initial assessment ($M = 75.4$).

A demographic questionnaire and a vocabulary test were

administered to P04 through P07 at the time of intake, and retrospectively to P03. P02 was unavailable for this assessment, and data for P01's interview were lost. The questionnaire provided information about the background characteristics of the participants, self-rated physical health status, and self-reported medication profiles. We used the background information to calculate an estimate of premorbid intelligence (Wilson, Rosenbaum, and Brown, 1979). Table 1 provides some relevant data regarding the background characteristics of all seven participants, including measures of state anxiety (Spielberger, 1983) and depression (Radloff, 1977) collected at the initial testing session. For comparison purposes, Table 1 reports data on the same measures from a similar sample of 64 elderly women from the same geographic area (age range 65 - 78, mean age 70.4) studied by Hertzog et al., 1990).

Insert Table 1 About Here

One subject, P02, was providing care for a spouse afflicted with Alzheimer's Disease. As can be seen in Table 1, P02 functioned at a normal level in vocabulary (and in text memory; see below) but would be classified as possibly depressed on the basis of a CES-D scale score of 28 (Radloff, 1977). Her spouse's relocation to the retirement community's intermediate

care facility -- and its influence on her psychological distress -- appeared to cause her withdrawal from the study after 25 occasions of measurement.

The physician of the retirement community gave all potential participants a physical health examination and an extensive vision screen (including range of visual field, glaucoma, cataracts, visual acuity, and accomodation). The physical exam was designed to maximize the probability that participants could continue over the expected two-year duration of the project. The vision assessment was needed to insure that the participants were not likely to experience difficulties viewing computer screens or written tests and questionnaires. One subject's test sessions were delayed in order to obtain a new set of eyeglasses.

Data on the self-reported health and medication profiles of each subject are contained in Table 2. The participants were relatively healthy older adults, with some normal aging-related symptoms and medications. Overall, the subjects reported that their status had very little impact on their daily lives. Two cases should be noted. First, P03 was under medication for high blood pressure and coronary artery disease. Second, P06 had an active case of tuberculosis, which was under medical management. At the onset of the study, however, both participants reported no health-related impact on their daily lives and no alterations of behavior due to medication.

Insert Table 2 About Here

Measures

Text recall was measured by administration of twenty-five structurally equivalent texts developed by Dixon et al. (1989). The texts were designed to contain approximately equal numbers of words, sentences, and basic idea units. All texts had about 300 words in 24 sentences. The structure of each text was defined according to a hierarchically organized network of propositions, as defined by Kintsch (1974; van Dijk & Kintsch, 1983). A basic proposition contains a predicate (usually a verb) and one or more concepts such as an actor and an object (e.g., DESERVE, PROFESSOR, RAISE). Subordinate propositions may qualify or elaborate upon predicate or word concepts of basic propositions. Propositions are hierarchically organized to reflect centrality to the story theme. The twenty-five stories all contained approximately 160 propositions, ranging from level 1 to level 7 in organizational hierarchy, and were comparable in percentage of propositions at each hierarchical level. They all presented propositions in an approximately linear sequence, that is, in order according to their relevance to the organizational structure of the text. The average reading difficulty of the 25 stories was seventh grade level.

The stories were explicitly designed to include plots thought to be generally relevant and appealing to older persons, and generally described positive outcomes of individuals or couples encountering salient events, such as a family reunion, retirement, and vacation plans. Story attributes such as age and gender of principal characters and geographic location of the event were explicitly varied across the texts. In particular, the stories included 9 plots involving female protagonists, 9 plots involving male protagonists, and 7 plots involving couples as protagonists. Additional details regarding story characteristics may be found in Dixon et al. (1989).

Procedure

General design. The text recall task was part of a larger data collection performed on a weekly basis. Subjects also completed self-ratings of affective states and metamemory, and performed on a computer-administered test of recognition memory for words. These data are not reported here. The text recall task was administered after the self-rating questionnaires and before the recognition memory task.

Subjects were tested weekly, on the same day (although the day occasionally changed to accomodate subject requests) by the same experimenters (which also changed due to personnel changes and needs). In the event of illness, testing days were altered or postponed one or more weeks, as needed. Gaps in testing were

infrequent and short, generally, although one subject (P05) experienced a major illness that caused a 17 week delay in testing and reduced the total number of occasions of measurement. Gaps in testing had no appreciable impact on the temporal pattern of text recall performance, and were hence ignored in data analysis.

Subjects were tested in a laboratory office located above the general health clinic at the retirement community. The women came to the laboratory for testing; during winter periods, or at their request, the participants were transported by retirement community staff from their residence to the laboratory. P06's respiratory disorder worsened midway through the study, and this confined her to her residence in order to be able to use an oxygen tank. She was tested at home thereafter.

Prior to assessment with the twenty-five stories, subjects' text recall performance was assessed using the three texts from Hertzog et al. (1990). The purpose of this assessment was to compare participants' initial text recall performance to norms derived from same-aged peers. Each subject received one of the comparative stories in each of the first three testing sessions, with the order of administration the same as used in the larger validation study. The three baseline stories were also given to the first three subjects (P01 through P03) during occasions 4 through 9, so that assessment with the 25 stories for these

subjects began at the tenth occasion of testing.

Each subject received multiple blocks of the twenty-five stories; most subjects received three complete administrations of the twenty-five story set. Each subject received an independent random order of texts within blocks under the constraint that two stories could not be repeated as the last story of a block and the first story of the subsequent block.

Methods of obtaining text recall data were invariant across both the comparison and parallel story types. Subjects were given a text recall booklet, including instructions, the text, and two pages of lined paper upon which to recall the story. They were instructed to recall as much of the story as they could, but that they could respond in their own words, not necessarily the words used in the actual story. They were given five minutes to read and seven minutes to write what they could recall from the story. Previous work with similar stories had demonstrated that this amount of response time was sufficient to obtain complete recall protocols (Hertzog et al., 1990).

Text scoring. Scoring of the stories was done by using a gist criterion to judge the presence of a proposition in the recall protocol, adapting scoring methods for the Kintsch propositional system developed by Turner and Green (1977). Propositions were scored as present if the idea represented by the proposition was present in the protocol, irrespective of

wording or order of recall. Measures used in the present study were proportion of correctly recalled propositions at each hierarchical level (1, 2, 3, 4, and 5 or greater), as well as the total proportion of propositions recalled. We also obtained three additional measures of recall behavior for verbal responses not directly related to specific propositions: macrostatements (summary statements describing gist-consistent aspects of the text which combined multiple propositions), elaborations (statements consistent with the gist of the text but not explicitly presented in the story), and metastatements (statements concerning the cognitive or affective state of the participant).

Four raters were trained for purposes of text scoring using the twenty-five stories. After training to criterion, the raters were given the first recall protocols from P01, P02, and P03 for a randomly selected set of 5 of the 25 stories, which they scored blind to each other for the purpose of computing interrater reliability. The raters achieved 92% agreement on propositional scoring, averaged over raters.

Results

Representativeness of Sample

In order to test the extent to which these participants were representative of their population in story recall ability, Table

3 reports the aggregated recall performance for the comparison stories administered on the first three occasions, using both (a) raw proportion of propositions correctly recalled at each hierarchical level of text organization, using a gist criterion, and (b) z-scores scaled from the normative subsample of elderly women. Performance by the present sample varied within normal ranges, establishing that the panel is essentially representative of their population in story recall.

Insert Table 3 About Here

Performance on the 25 Stories

Table 4 reports the performance of all seven subjects across the main set of text recall trials, aggregated across occasions (and stories). There were large individual differences in overall text recall performance, with total recall ranging from 56% (P04) to 21% (P03). Table 4 also demonstrates the expected effects of hierarchical idea structure; recall is highest for Level 1 propositions, drops off for Levels 2 and 3, and plateaus thereafter.

Insert Table 4 About Here

Figure 1, panels A through F, plots the data for each

subject against occasion of measurement. Two aspects of the data are evident for all subjects. First, there is substantial intraindividual variation in performance for all subjects. For example, performance by P01 ranged from 14% to 64% of total propositions recalled. Second, the trends of performance over time clearly differ across subjects. This divergence can be summarized by the correlations of total propositions recalled with occasion of measurement (see Table 5). P03 and P06 showed significant negative correlations, indicating decline in performance over time, but P01 and P04 showed significant positive correlations. Correlations of occasion of measurement with propositions recalled from hierarchical levels 2 through 4 were highly consistent with each other and with the total number of propositions recalled. Correlations were much lower for Levels 1 and 5, reflecting a tendency toward ceiling and floor effects, respectively. Given these findings, the remainder of the analyses of change and variability in recall relied solely on total proportion recall as the dependent variable.

Insert Table 5 and Figure 1 About Here

In order to analyze trends in performance over time, we conducted separate polynomial regression analyses for each subject, using linear and quadratic trend components in

regressions of recall on occasion. Subjects P01 and P04 showed robust linear increment over time, and subjects P03 and P06 yielded robust decrement. In addition to significant linear effects, the analysis detected significant curvature for P05, whose plot was mildly concave downward. Other subjects did not produce significant quadratic trends.

Much of the variance in story performance remained unexplained by the polynomial regression on occasion. The remaining variance might represent a number of influences, including mood states, transient changes in motion, and other organismic and environmental influences. An alternative source of variation is differential recall of the 25 stories. We used 24 dummy variables to capture the different story effects using hierarchical multiple regression. Substantial increments in R^2 indicated differences between the stories in average recall, adjusted for time-related trends, for all subjects (see Table 6).

Insert Table 6 About Here

Although the magnitude of story effects varied somewhat across subjects, several stories were consistently recalled better or worse than average. Mean story effects, aggregated across subjects, ranged from .09 (Story 13) to -.08 (Story 5). Thus, Story 13 yielded 9% more propositions recalled than the

occasion-adjusted grand mean. In addition to Story 13, Sto 1, 11, 14, and 16 yielded significant positive story effect .05). In addition to Story 5, Stories 10, 19, and 25 had negative story effects, reflecting below average recall performance. There was no obvious association of structural characteristics of the stories with the empirically obtained differences in recall performance. For example, Flesch-Kincai readability ratings, derived from structural characteristics of the texts (Dixon et al., 1989), correlated -.13 with the estimated story effects. Considering that the texts had been constructed to be approximately parallel with respect to structural characteristics (e.g., number of words, number of propositions at each hierarchical level, linearity of plot sequence), this result was not surprising. However, there was a clear association between gender of the protagonist and the magnitude of story effects. Estimated story effects for the nine stories with female protagonists ($M = .031$, $SD = .042$) were significantly higher than both (a) effects for the nine stories with male protagonists ($M = -.019$, $SD = .036$; $t = 2.71$, $p < .01$), and (b) effects for the seven stories in which couples were protagonists ($M = -.013$, $SD = .022$; $t = 2.70$, $p < .01$). In addition, an analysis of the stories' plots suggested that exceptions to this trend (e.g., poorer recall of a story with a female protagonist) could probably be explained in part by

gender-based relevance of the story line. For example, Story 4, which described an older woman attending a western rodeo, was the only story with a female protagonist producing a negative story effect. Similarly, Story 24 produced a positive story effect for a male protagonist cast in a story about a family holiday. Both stories describe older protagonists in nontraditional (although not exceptional) activities.

Even after accounting for story effects, large proportions of variance in performance remained unexplained for five of the six subjects. Total R^2 for the intraindividual variation was .6 or less for all subjects except P03 (R^2 = .769). These text recall measures have been shown to have high factor loadings and communalities when between-person correlations are factor analyzed (Hertzog, et al., 1990; Hultsch et al., 1984). These communalities reflect high reliability in between-person differences in text recall (between .8 and .9). Assuming these reliabilities generalize to intraindividual consistency in measurement, then between 20 to 30% of the intraindividual variation in story performance found in this study is reliable variance which remains to be explained by other variables. Figure 2 gives some graphic evidence of the partition of variance for subject P01. Panel A plots predicted scores from the polynomial regression, including story effects. Panel B plots the regression residuals. Note that, even after controlling for

story effects, there are interesting fluctuations in performance, even among adjacent occasions (e.g., occasions 34 through 40).

Insert Figure 2 About Here

We hypothesized that intraindividual variability in text performance would correlate with variability in affective states, especially after text recall was de-trended for systematic intraindividual change. However, measures of self-rated depression, anxiety, fatigue, vigor, and well-being exhibited few significant correlations with either overall text recall performance or residualized recall, controlling for systematic time-related trend and story effects.

Initial Status Versus Intraindividual Change

Table 7 presents the fitted linear slopes for occasion of measurement and the fitted intercepts from the intraindividual regression equations, along with the z-scores from the first validation text presented to all subjects. The normal deviates are directly comparable to data that would be obtained in a single assessment of memory functioning, followed by rescaling according to age-stratified norms. There was a relatively high level of rank-order agreement between the three variables, although the opportunity for disagreement given only six individuals with widely separated initial scores is minimal.

Note that disagreements in rank between the intercepts and the normal deviates are small, and occur where the intercepts are virtually equivalent (subjects P01 and P05). There is less agreement between the slopes and the initial text recall scores. Both P03 and P06 show negative initial scores (although neither scores are 2 standard deviations below the age group mean). However, P06 showed roughly comparable linear decline, in spite of an initial .75 SD difference between P03 and P06. P01 showed the most gain, however, even though she was third out of six in initial text recall.

Insert Table 7 About Here

Discussion

In this study we employed a series of stories constructed on the basis of contemporary cognitive psychological theories of discourse to investigate intraindividual cognitive change in seven older women. The value of integrating cognitive psychology with research questions concerning special populations (e.g., older normal adults) has been emphasized in the cognitive neuropsychology and aging literature (e.g., Kahn et al., 1990). Overall, we have shown that long-term intraindividual panel designs (a) can be successfully implemented, and (b) provide

useful information for the purpose of cognitive assessment of normal older persons. All but one of the seven older participants completed at least seventy weekly test occasions, lasting well over one year.

The first major purpose of this study was to investigate the extent of intraindividual variability in text recall performance. We found dramatic variability in performance in all seven participants. In addition, and relevant to the second major aim of the study, there were marked differences between panel members in the long-term pattern of intraindividual change. Two panel members showed significant increases, two showed relative stability, and two showed significant decline over the time interval studied. These change patterns were related to, but were not fully predictable from, initial performance levels. Use of cross-sectional norms to evaluate the initial text recall scores would not have predicted the significant decline in one subject (P06) nor have anticipated the significant curvilinearity, reflecting decline late in the sequence of testing, for another subject (P05).

Although this set of results is not conclusive, the mixed pattern of change over time reinforces the concern that an important benchmark for assessing cognitive decline in the elderly, either for clinical or research purposes, may be the establishment of an intraindividual performance baseline. To be

sure, standard use of cross-sectional norms would have correctly identified P03, the subject experiencing the most dramatic cognitive decline, as performing significantly below age and gender-specific norms. However, the major advantage of an intraindividual panel design with regular assessments is that it becomes possible to detect subtle changes in the trajectory of cognitive performance, as in the case of subject P05 in the present study. The sensitivity of the intraindividual design stems from the power of the repeated measures design to detect small amounts of change. In essence, one is able to compute an intraindividual standard error of measurement for each person, based upon their own baseline variability in premorbid performance, and then to use this standard error as the basis for inferring the onset of significant cognitive decline.

This finding may be contrasted to those of Salthouse et al. (1986). They administered cognitive tasks twice within a single test session to a large cross-sectional sample of adults, and found moderate correlations between the two scores for each task. Given the relative magnitude of (a) mean cross-sectional age differences in the tasks with (b) within-session differences in performance levels for individuals, they argued that it would take decades to detect age-related change at the level of individual subjects. Their argument is based upon assumptions which cannot be fully critiqued here (e.g., that the difference

score between alternative administrations of a task, within a single multiple-task test session, is a valid estimator of intraindividual variability across different sessions). The present study clearly shows that change in at least some individuals can be reliably detected over short periods of time, given sufficient density of observations.

Intraindividual designs may be well suited for the early detection of pathological or nonnormative change in the elderly. Two subjects who experienced significant linear declines in performance may have been experiencing terminal decline. P06's tuberculosis worsened, and she eventually withdrew and died shortly thereafter. Although she completed the study, P03 also died a few months afterwards from her cardiovascular disease. Furthermore, the one subject (P05) showing significant curvature in change (reflecting decline late in the study) experienced health problems that caused her to miss several test sessions in the second year of the study. Such evidence suggests the potential importance of intraindividual panel designs for early detection of cognitive decline in the elderly. It was possible to chart a negative trend in P03's performance over a year in advance of her death.

The present study has several characteristics which limit our inferences regarding intraindividual variability and change in text recall. First, we selected subjects who were initially

in relatively good health. We did not seek to examine cognitive change in well-identified patient populations, and did not obtain data that could definitively establish the causes of cognitive change in our older panel. We recommend the use of such intraindividual panel designs in future studies of cognitive change in clinical populations. Second, we selected female participants, in part because of their longer expected longevity, and also restricted the sample to older adults. Given the nature of our sample, additional research will be needed to address the issue of adult age and gender differences in intraindividual variability and change.

A third issue is the mixing of positive gains due to practice effects with declines in performance due to normal and pathological age-related factors. One expectation, based upon learning to learn research (e.g., Hultsch, 1974), might have been that all older persons would show increments in text recall performance. Indeed, one possible concern about such panel studies is that large practice effects would create problems for interpretation of performance, possibly due to ceiling effects. It appears that our use of the large number of stories, characterized by (a) a wide range of possible recall scores, and (b) sensitivity to recall of both basic and low-level propositions represented in the texts, successfully avoided the risk of ceiling effects and enabled us to detect performance

declines in several panel members. Two panel members (P01, P04) showed significant improvements in performance, suggesting positive learning-to-remember effects. One can hypothesize, then, that such positive effects might counteract small age-related declines, producing stable levels of performance. If so, then declines in panel designs may have considerable clinical significance, but it may also be the case that failure to improve indicates more subtle effects of normal aging or age-related pathology. This hypothesis could be addressed in the future by conducting age comparative studies that feature experimental variation of the amount and frequency of testing (which will affect the magnitude of learning effects). Ideally, this design could be combined with more systematic assessment of physiological and neurological functioning. Such a comprehensive design could establish both: (a) the frequency of observations needed to minimize practice effects while still estimating a sensitive intraindividual baseline (in terms of both level and intraindividual variability in cognition), and (b) the diagnostic value of failure to increase performance as a function of repeated testing.

Systematic variation in text recall by story was sizable, and showed an interesting pattern. For the female subjects in this study, stories with female protagonists and gender-appropriate themes were recalled better, on average, than stories

with males or couples as protagonists. This finding therefore suggests that the 25 stories, although structurally equivalent, may vary in recallability as a function of relevance of material for the individual reader. Several studies have shown that text recall performance is influenced by knowledge and belief structures of the rememberers (e.g., Hultsch & Dixon, 1984), and there is some suggestion in the literature that pre-existing cognitive schemata can influence memory performance (Sherman, Judd, & Park, 1989; Hess, in press). Herrmann and Crawford (1989) had male and female subjects read ambiguous instructions using differentially gender-relevant titles (directions for making a shirt versus directions for making a workbench). They obtained Gender X Title interactions in recall of the instructions, with members of each sex recalling more instructions when paired with the sex role-consistent title. We hypothesize that older men would perform better on the stories with male protagonists and male-relevant themes. Such effects may have implications for assessment inferences based on a single administration of a differentially schema-relevant passage, such as the passage used frequently in the WMS Logical Memory task. An advantage of the present set of 25 structurally equivalent stories (Dixon et al., 1989) is that investigators may select from a variety of stories differing in schema relevance.

Even after accounting for story effects and occasion-

specific trends, substantial proportions of variance in text recall performance remained. Because our interrater reliabilities are high and we typically observe substantial within-session correlations of text recall across different stories (e.g., Hertzog et al., 1990), it is implausible that the magnitude of residual variance is caused solely by measurement error. It appears that there is residual within-person variation that remains to be explained by states of the rememberer (Sutton et al., 1988). Our initial evaluation of the covariation of text recall with concurrently rated mood states was disappointing, in that self-rated mood states did not correlate with performance. There can be a number of reasons for a negative finding here, including limited validity of self-ratings due to stereotyping of responses and lack of self-insight into concurrent affect. It remains to be seen whether residual intraindividual variance in text recall performance will covary with other measures of the concurrent state of the rememberer.

In sum, the present study encourages the use of the intraindividual panel design with elderly populations. Such designs, which are particularly useful for charting the course of cognitive change, may also assist in the assessment of impending cognitive decline.

References

Alder, A. G., Adam, J., & Arenberg, D. (in press). An individual differences assessment of the relationship between change in an initial level of adult cognitive functioning. Psychology and Aging.

Arenberg, D. (1982). Estimates of age changes in the Benton Visual Retention Test. Journal of Gerontology, 37, 87-90.

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

Bower, G. H. (1981). Mood and memory. Amercian Psychologist, 36, 129-148.

Broadbent, D. E., Broadbent, M. H. P., & Jones, J. L. (1989). Time of day as an instrument for the analysis of attention. European of Journal of Cognitive Psychology, 1, 69-94.

Caine, E. D., Bamford, K. A., Schiffer, R. B., Shoulson, I., & Levy, S. (1986). A controlled neuropsychological comparison of Huntington's Disease and Multiple Sclerosis. Archives of Neurology, 43, 249-254.

Clark, D. M. & Teasdale, J. D. (1982). Diurnal variations in clinical depression and accessibility of memories of positive and negative experiences. Journal of Abnormal Psychology, 91, 87-95.

Crook, T., Bartus, R. T., Ferris, S. H., Whitehouse, P., Cohen,

G.D., & Gershon, S. (1986). Age-associated memory impairment: Proposed diagnostic criteria and measures of clincial change-report of a National Institute of Mental Health work group. Developmental Neuropsychology, 2, 261-276.

Dall'Ora, P., Della Sala, S., & Spinnler, H. (1989). Autobiographical memory: Its impairment in amnesic syndromes. Cortex, 25, 197-217.

Dixon, R. A. & Hertzog, C. (1988). A functional approach to metamemory development in adulthood. In F. E. Weinert & M. Perlmutter (Eds.), Memory development: Universal changes and individual differences (pp. 293-330). Lawrence Erlbaum Associates.

Dixon, R. A., Hultsch, D. F., & Hertzog, C. (1989). A manual of twenty-five three-tiered structurally equivalent texts for use in aging research (2nd Ed.; Technical Report No. 2). Victoria/Atlanta: Collaborative Research Group on Cognitive Aging.

Dixon, R. A., Hultsch, D. F., Simon, E. W., & von Eye, A. (1984). Verbal ability and text structure effects on adult age differences in text recall. Journal of Verbal Learning and Verbal Behavior, 23, 569-578.

Egelko, S., Gordon, W. A., Hibbard, M. R., Diller, L., Lieberman, A., Holliday, R., Ragnarsson, K., Shaver, M. S., & Orazem, J. (1988). Relationship among CT scans, neurological exam, and

neuropsychological test performance in right-brain-damaged stroke patients. Journal of Clinical and Experimental Psychology, 10, 539-564.

Eich, E., & Metcalfe, J. (1989). Mood dependent memory for internal versus external events. Journal of Experimental Psychology: Learning, Memory, and Cognition, 15, 443-455.

Erickson, R. C., & Scott, M. L. (1977). Clinical memory testing: A review. Psychological Bulletin, 84, 1130-1149.

Flicker, C., Ferris, S. H., Crook, T., Bartus, R. T., & Reisberg, B. (1986). Cognitive decline in advanced age: Future directions for psychometric differentiation of normal and pathological age changes in cognitive function. Developmental Neuropsychology, 2, 309-322.

Grant, I., McDonald, W. I., Trimble, M. R., Smith, E., & Reed, R. (1984). Deficient learning and memory in early and middle phases of multiple sclerosis. Journal of Neurology, Neurosurgery, and Psychiatry, 47, 250-255.

Hartley, J. T. (1986). Reader and text variables as determinants of discourse memory in adulthood. Psychology and Aging, 1, 150-158.

Herrmann, D. J., & Crawford, M. (1989). Gender-linked differences in everyday memory performance. Unpublished Manuscript.

Hertzog, C., Dixon, R. A., & Hultsch, D. F. (1990). Relationships

between metamemory, memory predictions, and memory task performance in adults. Psychology and Aging, 5, 215-227.

Hertzog, C., & Schear, J. M. (1989). Psychometric considerations in testing the older person. In T. Hunt & C. J. Lindley (Eds.), Testing the older person: A reference guide for geropsychological assessments (pp. 24-50). Austin, TX: PRO-ED.

Hess, T. M. (in press). Aging and schematic influences on memory. In T. M. Hess (Ed), Aging and cognition: Knowledge organization and utilization. Amsterdam: North Holland.

Hultsch, D. F. (1974). Learning to learn in adulthood. Journal of Gerontology, 29, 302-308.

Hultsch, D. F. & Dixon, R. A. (1984). Memory for text materials in adulthood. In P. B. Baltes & O. G. Brim, Jr. (Eds.), Life-span development and behavior (Vol. 6). New York: Academic Press.

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

Hultsch, D. F., Hertzog, C., & Dixon, R. A. (in press). Ability correlates of memory performance in adulthood and aging. Psychology and Aging.

Kahn, H. J., Joanette, Y., Ska, B., & Goulet, P. Discourse analysis in neuropsychology: Comment on Chapman and Ulatowska.

Brain and Language, 38, 454-461.

Kaszniak, A. W. (1990). Psychological assessment of the aging individual. In J. E. Birren & K. W. Schaie (Eds.), Handbook of the psychology of aging (3rd ed.; pp. 427-445). San Diego, CA: Academic Press.

Kintsch, W. F. (1974). The representation of meaning in memory. Hillsdale, NJ: Erlbaum.

Kopelman, M. D. (1986). Clinical tests of memory. British Journal of Psychiatry, 148, 517-525.

Kopelman, M. D. (1987). Two types of confabulation. Journal of Neurology, Neurosurgery, and Psychiatry, 50, 1482-1487.

Kopelman, M. D., Wilson, B. A., & Baddeley, A. D. (1989). The autobiographical memory interview: A new assessment of autobiographical and personal semantic memory in amnesic patients. Journal of Clinical and Experimental Neuropsychology, 11, 724-744.

Kral, V. A. (1962). Senescent forgetfullness: Benign and malignant. Canadian Medical Association Journal, 86, 257-260.

Meyer, B. J. F. (1987). Reading Comprehension and aging. In K. W. Schaie (Ed.), Annual Review of Gerontology and Geriatrics (Vol. 7, pp. 93-115). New York, Springer.

Nesselroade, J. R., Pruchno, R., & Jacobs, A. (1986). Reliability vs. stability in the measurement of psychological states: An illustration with anxiety measures. Psychologische Beitrage,

28, 255-264.

Rao, S. M., Leo, G. J., & St. Aubin-Faubert, P. (1989). On the nature of memory disturbance in multiple sclerosis. Journal of Clinical and Experimental Neuropsychology, 11, 699-712.

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

Salthouse, T. A., Kausler, D. H., & Saults, J. S. (1986). Groups versus individuals as the comparison unit in cognitive aging research. Developmental Neuropsychology, 2, 363-372.

Schultz, K. A., Schmitt, F. A., Logue, P. E., & Rubin, D. C. (1986). Unit analysis of prose memory in clinical and elderly populations. Developmental Neuropsychology, 2, 77-87.

Sherman, S. J., Judd, C. M., & Park, B. (1989). Social cognition. Annual Review of Psychology, 40, 281-326.

Spielberger, C. D. (1977). State-trait anxiety and interactional psychology. In D. Magnusson & N. S. Endler (Eds.), Personality at the crossroads: Current issues in interactional psychology (pp. 173-183). Hillsdale, NJ: Lawrence Erlbaum.

Spielberger, C. D. (1983). Manual for the State-Trait Anxiety Inventory (Form Y). Palo Alto, CA: Consulting Psychologists Press.

Sutton, L. J., Teasdale, J. B., & Broadbent, D. E. (1988). Negative self-schema: The effects of induced depressed mood.

British Journal of Clinical Psychology, 27, 188-190.

Turner, A., & Greene, E. (1977). The construction and use of a propositional text base. Technical Report, University of Colorado.

van Dijk, T. A., & Kintsch, W. (1983). Strategies for discourse comprehension. New York: Academic Press.

Watson, D. (1988). Intraindividual and interindividual analyses of positive and negative affect: Their relation to health complaints, perceived stress, and daily activities. Journal of Personality and Social Psychology, 54, 1020-1030.

Wilson, B. A. (1987). Rehabilitation of memory. New York: Guilford.

Wilson, R. S., Rosenbaum, G., & Brown, G. (1979). The problem of premorbid intelligence in neuropsychological assessment. Journal of Clinical Neuropsychology, 1, 49-53.

Zelinski, E. M., & Gilewski, M. J. (1988). Memory for prose and aging: A meta-analysis. In M. L. Howe & C. J. Brainerd (Eds.), Cognitive development in adulthood (pp. 134-158). New York: Springer.

Zevon, M. A., & Tellegen, A. (1982). The structure of mood change: An idiographic/nomothetic analysis. Journal of Personality and Social Psychology, 43, 111-122.

Authors' Note

This research was supported by a research grant (RC from the National Institute on Aging. The first autho supported by a Research Career Development Award from National Institute on Aging (K04-AG00335). The second authors' work on this project was supported by grants f Natural Sciences and Engineering Research Council of Ca cooperation of Robert K. Nielsen, MD., and the staff and residents of the Cornwall Manor of the United Methodist Cornwall, PA, is greatly appreciated. We also thank Pau and John Smith for their assistance in data management a analysis. Address correspondence concerning this article Christopher Hertzog, School of Psychology, Georgia Insti Technology, Atlanta, GA, USA, 30332-0170.

Table 1

Participant Characteristics on Background Variables

	Subject						
	P01	P02	P03	P04	P05	P06	P07
Variable							
Age	74	76	80	83	73	67	75
CES-D[a]	4	28	16	6	13	2	0
SSA[b]	30	44	28	16	24	34	20
Premorbid IQ	---[c]	---	117	123	125	117	125
Vocabulary[d]	---	---	113	100	116	114	116
Education[e]	---	---	12	12	17	13	18

Abbreviations: CES-D: Community Epidemiological Survey - Depression scale; SSA - Spielberger State Anxiety scale.

[a] Norm sample: M = 7.46 (SD = 6.46).

[b] Norm sample: M = 30.09 (SD = 7.88).

[c] Denotes missing data.

[d] Test based upon ETS Vocabulary (Hertzog et al., 1990). Scores are scaled relative to Norm sample (M = 100, SD = 15).

[e] Norm sample: M = 12.92 (SD = 2.88).

Table 2

Self-reported Health and Medication Profiles

	P03	P04	P05	P06	P07
Self-reported Health					
Perceived Health (1-9 scale; 9 = excellent)	4	8	8	6	7
Total Number of Illnesses (Max. 29)	5	3	2	5	4
Physical Symptoms (Max. 40)	8	11	2	7	7
Perceived Impact on Life	None	None	---[a]	None	None
Number of Bedridden Days (past year)	0	0	0	0	1
Health-related Behaviors					
Regular Health Checkups?	Yes	Yes	Yes	Yes	Yes
Number of Physician Visits (past year)	9-10	10+	3-4	3-4	5-6
Medication Usage					
Number of Medications (Max. 16)	4	4	0	4	3
Alter Behavior?	No	No	No	No	Yes

[a] Denotes missing data.

Table 3

Average Text Recall Performance on Initial Stories for All Participants (Proportion Correct [P(c)] and Normal Deviate [z])

Subject		Total	Level 1	Level 2	Level 3	Level 4+
P01	P(c)	0.26	0.43	0.28	0.22	0.20
	z	0.37	0.28	0.75	0.58	0.10
P02	P(c)	0.45	0.58	0.50	0.43	0.31
	z	2.68	1.37	2.62	2.97	1.29
P03	P(c)	0.17	0.34	0.20	0.13	0.12
	z	-0.79	-0.33	-0.64	-0.73	-0.80
P04	P(c)	0.46	0.77	0.53	0.37	0.32
	z	2.74	2.67	2.93	2.34	1.83
P05	P(c)	0.26	0.49	0.27	0.21	0.21
	z	0.33	0.71	0.11	0.48	0.30
P06	P(c)	0.18	0.31	0.21	0.16	0.10
	z	-0.59	-0.58	-0.46	-0.17	-1.27
P07	P(c)	0.31	0.40	0.38	0.24	0.26
	z	0.96	0.08	1.26	0.63	0.83

Table 4
Proportion of Text Proposition Recalled by each Participant

Subject		Total	L1	L2	L3	L4	L5+	EL	MAC
P01	M	.444	.926	.588	.454	.373	.323	2.22	3.04
	SD	.093	.263	.122	.110	.108	.165	1.64	1.84
P02	M	.555	1.00	.609	.562	.514	.504	2.81	1.81
	SD	.094	.000	.129	.089	.111	.249	1.64	1.72
P03	M	.205	.683	.300	.204	.161	.137	2.32	3.42
	SD	.075	.458	.124	.080	.091	.097	1.38	1.83
P04	M	.556	.856	.636	.558	.516	.493	2.47	2.55
	SD	.083	.340	.126	.102	.101	.158	2.47	1.62
P05	M	.360	.788	.471	.354	.308	.293	4.15	4.17
	SD	.077	.404	.120	.086	.085	.150	1.84	1.99
P06	M	.251	.797	.322	.259	.202	.191	3.27	2.23
	SD	.063	.385	.109	.082	.079	.124	1.64	1.37
P07	M	.329	.671	.407	.334	.291	.252	2.69	3.67
	SD	.079	.465	.130	.086	.085	.132	1.70	1.93

Abrreviations : L1 - Level 1; L2 - Level 2; L3 - Level 3; L4 - Level 4 ;L5+ - Level 5 and higher; EL - Elaborations; MAC - Macrostatements.

Table 5

Correlations of Occasion of Measurement

With Text Recall

Subject	Total	Level 1	Level 2	Level 3	Level 4	Level 5+
P01	.43	-.12	.38	.38	.35	.15
P02	.12	.00	.12	.08	.05	.15
P03	-.61	-.22	-.57	-.56	-.38	-.31
P04	.30	-.18	.25	.26	.29	.10
P05	.05	-.01	.02	.06	.03	.04
P06	-.32	.13	-.31	-.32	-.23	.01
P07	-.06	-.05	-.02	-.04	.03	-.14

Table 6

Hierarchical Regression Analyses:

Increments to R^2

	Step 1 Occasion (Linear)	Step 2 Add Story Effects	Step 3 Add Occasion (Quadratic)
P01	.182**	.357**	.004
P03	.370**	.399**	.002
P04	.093*	.452**	.008
P05	.003	.600**	.029*
P06	.104*	.590**	.006
P07	.004	.497**	.003

*p < .05

**p < .01

Table 7

Initial Text Recall Performance (Story 26), Fitted Slopes, and Fitted Intercepts for Individual Regression Equations

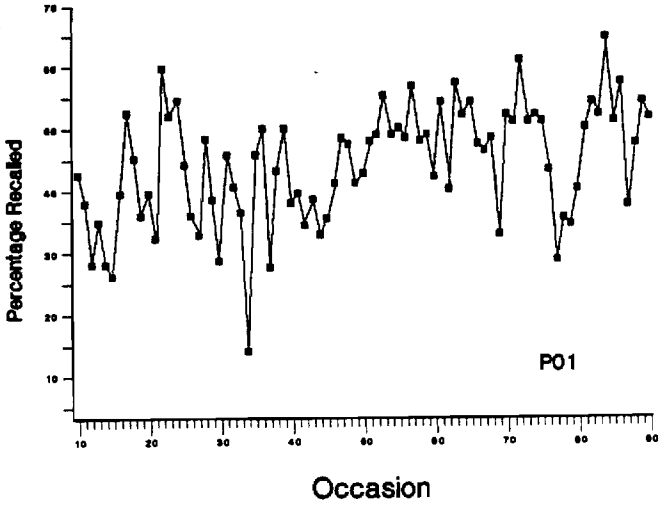
	Story 26		Intercept		Slope	
	z	Rank[a]	Estimate	Rank	Estimate ($\times 10^3$)	Rank
P01	1.13	3	0.36	2	1.7	1
P03	-1.42	6	0.30	5	-2.0	6
P04	3.50	1	0.51	1	1.1	4
P05	1.89	2	0.35	3	1.4	2
P06	-0.65	5	0.29	6	-1.9	5
P07	0.11	4	0.34	4	1.2	3

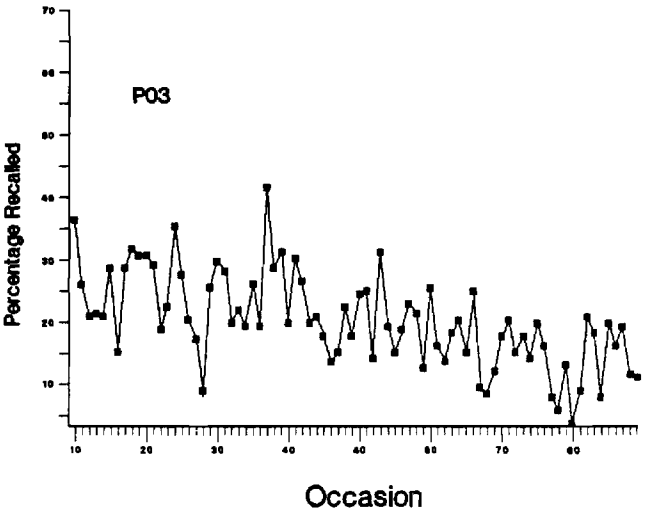
[a]Rank order within the group (range = 1 to 6)

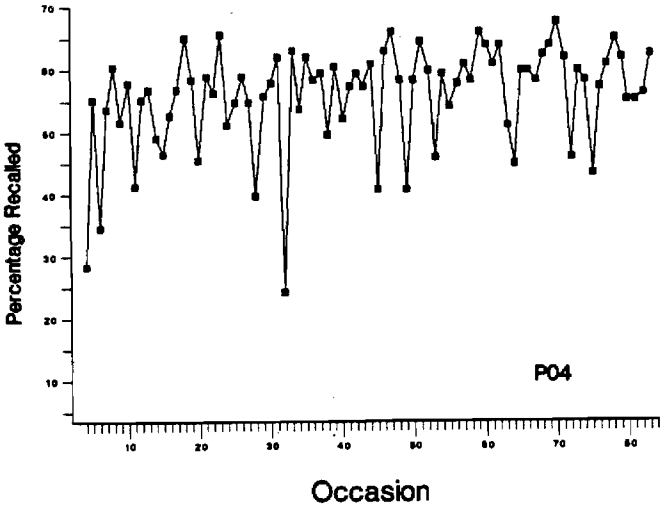
Figure Captions

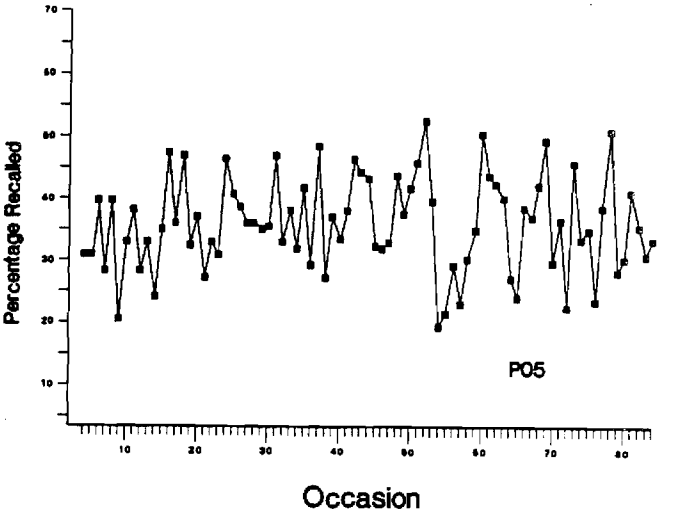
1. Plots of Text Recall over Occasion of Measurement for Each Subject

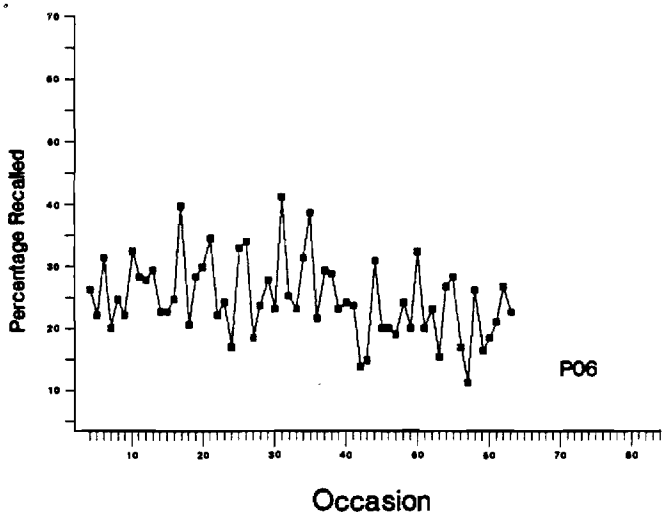
2. Plot of Predicted (Panel A) and Residual (Panel B) Scores for Regression of Text Recall on Occasion and Story for Subject P01.

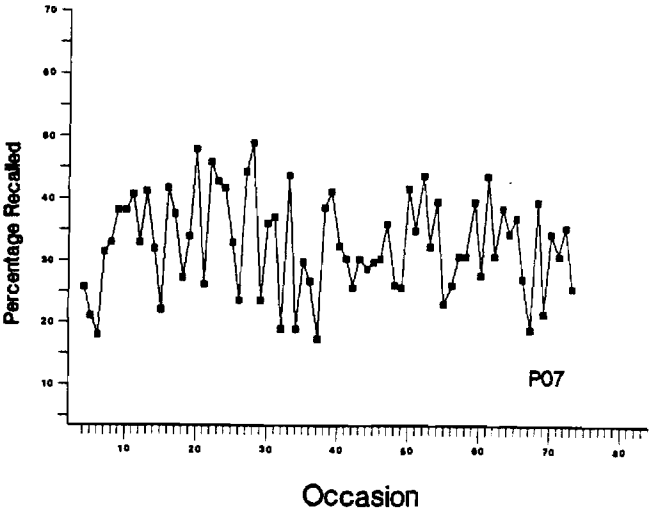

Percentage Recalled
P01
Occasion

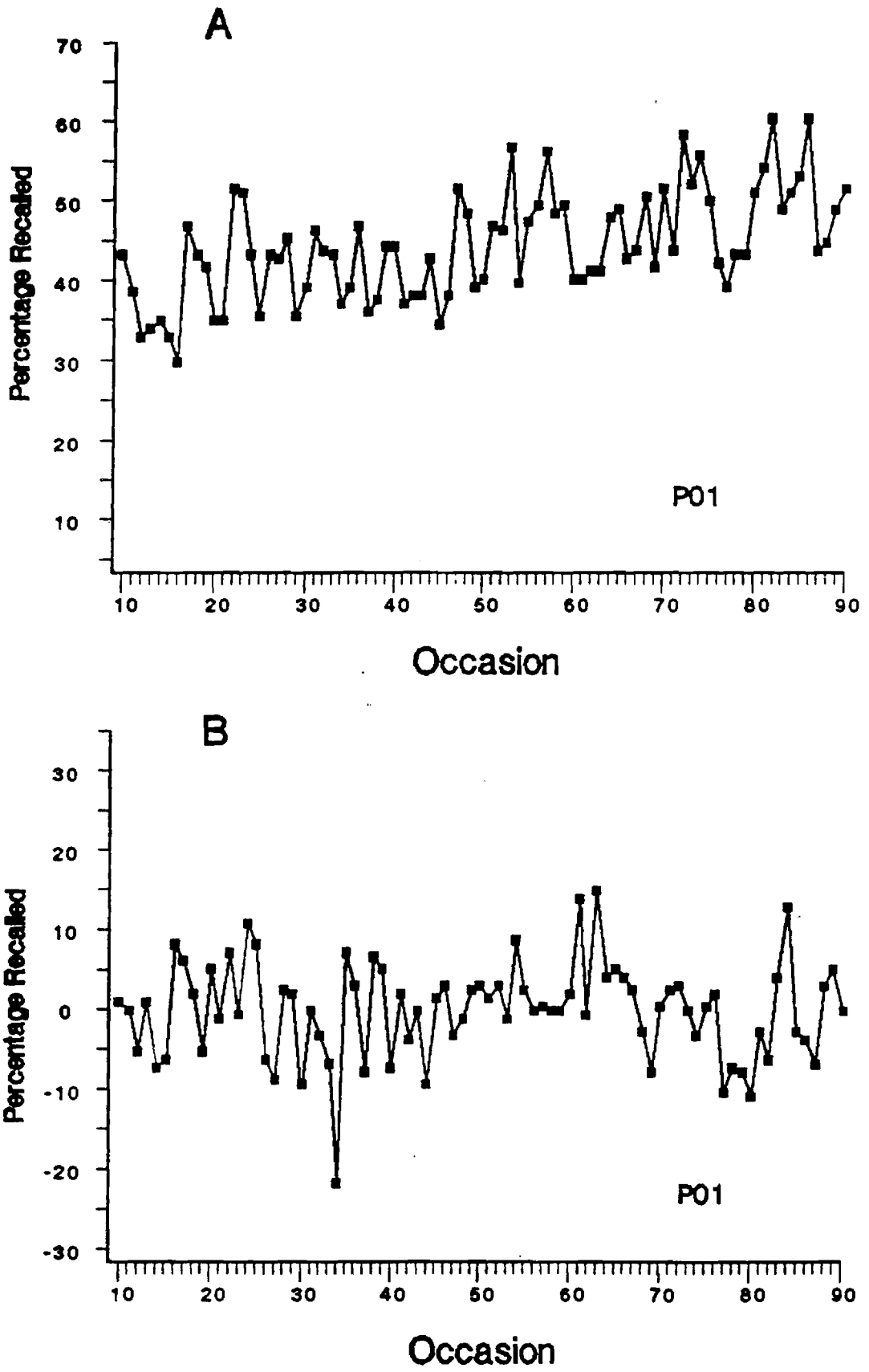

Percentage Recalled
P03
Occasion


Percentage Recalled
P04
Occasion


Percentage Recalled
P05
Occasion


Percentage Recalled
P06
Occasion


Percentage Recalled
P07
Occasion


A
Percentage Recalled
70
60
50
40
30
20
10
P01
10
20
30
40
50
60
70
80
90
Occasion
B
Percentage Recalled
30
20
10
0
-10
-20
-30
P01
10
20
30
40
50
60
70
80
90
Occasion