Journal of Clinical and Experimental Neuropsychology

ISSN: 1380-3395 (Print) 1744-411X (Online) Journal homepage: https://www.tandfonline.com/loi/ncen20

Color-Word Stroop test performance across the adult life span

B. Uttl & P. Graf

To cite this article: B. Uttl & P. Graf (1997) Color-Word Stroop test performance across the adult life span, Journal of Clinical and Experimental Neuropsychology, 19:3, 405-420, DOI: 10.1080/01688639708403869

To link to this article: https://doi.org/10.1080/01688639708403869

Published online: 04 Jan 2008.

Journal of Clinical and Experimental Neuropsychology
1997, Vol. 19, No. 3, pp. 405-420
1380-3395/97/1903-405$12.00

Color-Word Stroop Test Performance Across the Adult Life Span*

B. Uttl and P. Graf
University of British Columbia, Vancouver, Canada

ABSTRACT

In the Color-Word Stroop test (CWST), the basic task is to name the ink color of rows of XXXs, and performance in this condition is compared with performance in naming the ink-color of color words under conditions where word meanings and ink colors mismatch or are incongruent (e.g., the word *red* printed in green ink). The present study investigated whether Stroop test interference, defined as the cost associated with ink-color naming in the incongruous stimulus condition versus in the basic color-naming condition, provides positive evidence for a kind of processing *qualitatively* different than that which is required for color naming or for word reading. Does the pattern of age-related differences in Stroop interference force the conclusion that the incongruous condition taps a qualitatively different kind of processing than that required for color naming or for word reading? We gave the CWST to 310 healthy adults. Their performance in each condition of the test replicates and extends previous findings. Structural equation modeling of the data showed a significant, direct link between age and performance in the latent factor associated with the incongruent condition. However, this direct link with age produced a relatively small increase in the model's fit; it amounted to only a .024 increase in the proportion of variance explained in the incongruent condition. In light of this small direct influence due to age, the most parsimonious explanation of our findings is that age effects in Stroop interference are due to age-related slowing (which is also indexed by color naming and by word reading) primarily; the findings do not provide positive evidence for a qualitatively different kind of processing that declines with age.

Stroop tests come in many different variations, and each of them focuses on performance of a basic task, such as color or picture naming, versus performance of the same task in the presence of conflicting or incongruous stimuli. For the Color-Word Stroop test (CWST), for example, the basic task is to name the ink color of rows of XXXs, and performance in this condition is compared with performance in naming the ink-color of color words under conditions where word meanings and ink colors mismatch or are incongruent (e.g., the word *red* printed in green ink). A recent review of research with Stroop tests by MacLeod (1991) gives a comprehensive picture of the variety of tests that have been used; it also reveals that the most commonly used index of Stroop test performance is an interference score, defined on the CWST as the difference in the amount of time required for ink-color naming in the incongruous stimulus condition versus in the basic color naming condition.

* This research was supported by operating grants from the British Columbia Health Research Foundation and by the Natural Sciences and Engineering Research Council of Canada to P. Graf, and by graduate scholarships to B. Uttl from the Natural Sciences and Engineering Research Council of Canada and from the University of British Columbia.
We thank Jennifer Shapka and Arsalan Fatima Ghani for assisting with the preparation of this manuscript. Special thanks also go to the staff of the Kerrisdale Community Center and of Science World in Vancouver whose cooperation enabled us to carry out this research.
Address correspondence to: Peter Graf, Department of Psychology, University of British Columbia, 2136 West Mall, Vancouver, BC, V6T 1Z4, Canada.
Accepted for publication: December 7, 1996.

The CWST poses a processing challenge, to attend and respond to one stimulus dimension (e.g., ink colors) while ignoring and suppressing responses to a set of highly familiar visual stimuli, printed color words (e.g., red, blue, green). Stroop test interference is often viewed as a general index of cognitive flexibility and control, of the "*ease* [italics added] with which a person can shift his or her perceptual set to conform to changing demands and suppress a habitual response in favor of an unusual one" (Spreen & Strauss, 1991, p. 52). Consistent with this analysis of the requirements of Stroop tests, performance is assumed to depend directly on access to attentional capacity or task appropriate processing resources. For this reason, Stroop test interference scores have been used to track the availability of attentional capacity across different experimental conditions, as well as to speculate about how processing resources are influenced by such factors as age and dementia (for recent reviews see Graf & Uttl, 1995; Graf et al., 1995; MacLeod, 1991; Spreen & Strauss, 1991).

In a recent study, we investigated the relationship between age and CWST performance in older adults (Graf et al., 1995). Individuals beyond their 64th year of age were given the CWST together with a battery of other tests. The results showed that word reading times differed only by a small amount between the youngest (mean = 68 years) and oldest (mean = 86 years) group (421 ms vs. 490 ms), whereas color naming increased more noticeably across the same age groups (592 ms vs. 854 ms), and naming colors in the incongruent condition changed even more substantially (1087 ms vs. 1564 ms). When summarized in the usual manner, as the difference in time required for color naming in the incongruous stimulus condition versus in the basic color-naming condition, these findings show a clear age-related increase in Stroop interference, and this outcome is consistent with the results of previous investigations on the same subject population (reviewed in Graf et al., 1995). An age-increase in Stroop interference is consistent with the view that processing resources decline in old age (Craik, 1986; Hasher & Zacks, 1979; Jorm, 1986; Kausler, 1991; Salthouse, 1988).

The aim of our previous investigation was to ask whether CWST performance differences across conditions are primarily quantitative, for example, by one condition requiring more resources than another, or whether the differences across conditions are primarily (or also) qualitative, with the incongruous condition requiring special resources (e.g., cognitive flexibility, inhibitory processes) that are not engaged for color naming or for word reading. We focused on age changes in CWST performance and reasoned that if age changes in the incongruous condition can be predicted entirely by (if they can be shown to be a direct function of) age changes in color naming or in word reading, then there is no need or basis for postulating that Stroop interference (i.e., the difference score) indexes a special kind of resource that declines with age. Alternatively, if age effects in the incongruous condition are not predicted by color naming, if it can be shown that age also has another, direct influence (i.e., one that is not mediated by color naming) on the incongruous condition, this outcome would be positive evidence that requires postulating an additional, separate factor or cause in order to explain the performance difference between color naming and naming the ink-colors in the incongruous condition.

A quantitative interpretation of age changes in Stroop interference is consistent with an account of cognitive aging proposed by Salthouse and his colleagues (Salthouse, 1980, 1985, 1991, 1993, 1996; Salthouse, Kausler, & Saults, 1988; see also Cerella, 1985; Cerella, Poon, & Williams, 1980). According to this account, a broad range of age effects on perceptual, cognitive, and memory tasks are a direct consequence of a general age-related decline in processing speed. Salthouse and his colleagues have underscored that changes in processing speed can yield complex patterns of effects, even interaction effects between age and tasks or between age and conditions; because for any two subject groups that differ in processing speed, the performance difference between them will be larger on tasks that take longer to complete, for whatever reasons such as difficulty, complexity or novelty. This type of global slowing account is consistent with the finding that on the CWST, age effects are larger in the incongruous and more time-consuming condition than in the color-naming and less time-consuming condition.

The possibility that age changes in Stroop interference reflect something special, that they index something that is qualitatively different from what leads to age changes in color naming or in word reading, is consistent with the 'standard' interpretation of Stroop test performance (Spreen & Strauss, 1991). Recent research has provided many other examples of such selective impairments due to age. It has been shown, for example, that aging is associated with selective loss in working memory capacity (Baddeley, 1986; Baddeley & Hitch, 1974; Daneman & Carpenter, 1980; Graf & Uttl, 1995; Kausler, 1994), and that aging comes with declining explicit memory for episodes while sparing implicit memory for the same episodes (see Graf & Masson, 1993). In light of such findings, it seems plausible that age effects in Stroop interference may index yet another selective breakdown in a specific resource, cognitive skill, or capacity that is not required for color naming or word reading.

In order to decide between these alternatives, we examined our CWST data in a number of different ways. First, we translated the scores from the incongruent condition into proportions so as to highlight the *relative amount* of Stroop interference. Each proportion was defined as the additional time a subject required in the incongruous condition versus in the color-naming condition divided by the time required in the color-naming condition (i.e., (incongruous *minus* color-naming) / color-naming). Our data showed that in terms of proportional scores, the *relative amount* of Stroop interference did not change with age (Graf et al., 1995). Moreover, when the findings from previous investigations are translated into proportions, they also suggest that Stroop interference remains approximately constant in old age (see Graf et al., 1995). We also examined our CWST data by means of hierarchical regression analyses and found that while age was a significant predictor of word reading and color naming speed, it made no direct contribution to color naming with incongruous stimuli. Finally, we fitted our data with a structural equation model and it showed no direct, significant links between age and performance in the incongruous condition. Most critical is that all of these outcomes converge and show that age effects in Stroop interference are strongly correlated with and predicted by color naming, and we concluded that age effects in Stroop interference are most parsimoniously explained by global slowing.

Not all existing evidence seems to fit a global slowing interpretation of age changes in Stroop test interference, however. In our brief review of the literature (Graf et al., 1995), we also discovered a few cases that showed a clear age-related increase in Stroop interference even when interference was expressed as a proportion relative to color naming (see Dulaney & Rogers, 1994; Hartman & Sweeny, 1994; Houx, Jolles, & Vreeling, 1993; Spreen & Strauss, 1991). Why the difference in findings? One possibility is that they reflect the different subject samples used in previous research; especially, samples of older adults may differ by including varying proportions of persons with mild cognitive impairments, perhaps subtle precursors of dementia, and previous work has shown pronounced differences in Stroop test performance of such populations (Graf et al., 1995; Houx et al., 1993; Spreen & Strauss, 1991). Whatever the reasons for the difference in the available findings, they point to a limit in the generalizability of our previous results which were obtained entirely from healthy adults 65 years of age and older. Thus, it is possible that CWST interference is a valid index of decline in a special kind of processing or processing capacity, but one that occurs much earlier in life; this kind of age effect in Stroop interference will be found only in experiments with individuals drawn from a much wider age range than we had in our previous work.

The research reported in this article examined this possibility. We used the same method as for our previous investigation (Graf et al., 1995) in order to examine across a much wider age range whether age changes in Stroop interference are positive proof for a separate cognitive factor or processing resource, one that is not indexed by color naming or by word reading. We used continuous sampling of adults from all age groups (i.e., from the late teenage years to very old age) in contrast to most previous studies that either sampled a much more limited age range or that used a group design with samples drawn from only two or three narrow age ranges. In our project, a large number of community-living healthy volunteers, ranging between 12 and 83 years of age, were given the CWST together with several additional tasks. By contrast with the method used in our earlier work, we used a computer version of the CWST, in part because computer tests have the advantage of controlled, unbiased, and standardized administration. Our second objective was to obtain normative data for a standardized CWST from a large group of healthy adults.

METHOD

Subjects

The subjects were community-living volunteers that were obtained from two different sources. About half of them, 173 individuals between 12 and 81 years of age, participated in the course of a visit to a local science center; the other half, 164 adults between 16 and 83 years of age, volunteered in response to advertisements in community newspapers. For the latter group, 23% were paid a nominal amount to cover commuting costs for attending a single test session that lasted about 2.5 hours. The vast majority of all subjects (83%) were native English speakers. The data from the two subject sources were intended for pooling; the CWST was administered in exactly the same way to everybody.

A total of 337 individuals were given the CWST. However, the data from 12 of these subjects could not be used for various reasons. For 3 subjects, the files containing their demographic data (including years of age) were not available or they were unreadable. Two subjects failed to complete all trials in at least one condition of the CWST. For 3 additional subjects, their performance on all trials of at least one condition was identified as a univariate outlier (i.e., at least 2.58 *SD*s, at $p < .01$, away from the value predicted by the age group). Finally, the data from 4 subjects were identified as multivariate outliers (and thus not usable) by means of the Mahalanobis distance statistic, with the criterion set at χ^2 (9) > 27.87, at $p = .001$.

We checked the combined data for skewness and kurtosis and found that according to Mardia's coefficient of multivariate kurtosis (Mardia, 1970; Mardia & Foster, 1983; Steiger, 1995), the data violated assumptions specified for structural equation modeling. With the Mahalanobis distance statistic set at a stricter criterion, specifically χ^2 (9) > 21.67 at $p = .01$, we identified 15 additional subjects as multivariate outliers. When these subjects were removed from the sample, none of the indexes of skewness and kurtosis exceeded $|1|$, and the normalized Mardia coefficient of multivariate kurtosis was 1.79 and thus acceptable for structural equation modeling.

The data from the remaining sample of 310 subjects were examined for differences due to source, that is, differences between subjects obtained at the science center as opposed to subjects who volunteered in response to community newspaper advertisements. An ANOVA with subject source and age groups (12–29, 30–39, 40–49, 50–59, 60–69, 70+ years of age) as between-subjects factors showed no significant main or interaction effects involving the different subject sources (all *F*s < 1), and thus, the samples were combined for all subsequent analyses.

Demographic data on the screened, final sample appear in Table 1. The data are arranged by age groups, and each group spans 10 years, except for the youngest and oldest. The table shows that nearly two thirds of the subjects were women, and that women were in the majority in all groups. The tabled means also show that the groups were comparable in terms of years of formal education, except for the youngest subjects who averaged significantly fewer years of education than the oldest, next-lowest group (by Newman-Keuls, with α = .05); the remaining groups were not significantly different from each other.

The subjects were also asked for a self-rating of overall health. The specific question that we asked was "How would you rate your overall health at the present time: Excellent, Good, Fair, or Poor?" The rating categories were translated into values from 1 = poor to 4 = excellent, and these values were averaged across subjects. Table 1 shows that all subject groups gave similar self-ratings of their overall health.

Finally, a subset consisting of 181 (58.4 %) of the subjects who volunteered time for an additional task also completed the North American Adult Reading Test (NAART; Spreen & Strauss, 1991). This test, which is assumed to provide a general, reliable estimate of verbal intellectual ability, was administered according to the instructions by Spreen and Strauss (1991). Table 1 shows the

Table 1. Demographic Data.

		Age Group									
		12–19	20–29	30–39	40–49	50–59	60–69	70–83	All	Age[e] $F(1,308)$	ΔAge[2e] $F(1,307)$
n		15	59	67	64	40	32	33	310		
Gender											
Men		5	28	29	17	9	12	11	111		
Women		10	31	38	47	31	20	22	199		
Retired		0	0	0	1	6	20	32	59		
Education (years)	*M*	10.5	14.9	15.4	15.8	14.7	14.5	14.0	14.8	.30	26.66*
	SD	(2.53)	(1.86)	(2.52)	(2.92)	(3.43)	(2.90)	(3.85)	(3.02)		
Health ratings[a]	*M*	3.1	3.2	3.4	3.3	3.1	3.3	3.3	3.3	.13	.01
	SD	(.51)	(.56)	(.60)	(.71)	(.75)	(.72)	(.57)	(.65)		
NAART[b]	*M*	29.0	40.42	40.21	43.59	44.5	43.9	47.2	43.0	15.55*[2]	.77[2]
	SD	(9.84)	(9.09)	(8.61)	(8.49)	(8.77)	(10.2)	(8.03)	(9.22)		
	n	3	26	33	34	33	27	25	181		
Est. Verbal IQ[b,c]	*M*	100.2	110.4	110.2	113.2	114.0	113.5	116.5	112.7		
	SD	(8.76)	(8.09)	(7.66)	(7.55)	(7.80)	(9.10)	(7.14)	(8.21)		
Est. Full Scale IQ[b,d]	*M*	102.8	111.8	111.6	114.2	114.9	114.5	117.1	113.8		
	SD	(7.68)	(7.09)	(6.71)	(6.61)	(6.84)	(7.97)	(6.27)	(7.19)		

[a] Self-reported health ratings: 1 = Poor, 2 = Fair, 3 = Good, 4 = Excellent.
[b] NAART, Estimated Verbal IQ, and Estimated Full Scale IQ were available for only 181 subjects (see text).
[c] Verbal IQ was estimated from NAART scores using the following formula: Est. Verbal IQ = 128.7 – .89 (NAART errors) (Spreen & Strauss, 1991).
[d] Full Scale IQ was estimated from NAART scores using the following formula: Est. Full Scale IQ = 127.8 – .78 (NAART errors) (Spreen & Strauss, 1991).
[e] Age and age^2 effects were computed by regression analysis.
* $p < .05$.

NAART scores, along with the estimated verbal and full scale IQ scores. A regression analysis showed a significant age-related increase in NAART scores (see Table 1), thereby replicating the previous finding of a positive relation between age and verbal intellectual ability (Graf & Uttl, 1995; Spreen & Strauss, 1991).

Procedure

Subjects were tested individually, in one of two settings. The first setting was a small office in a local community center; it housed nothing but the test station – two desks, one with a computer on it, two chairs, and a filing cabinet. At the science center, the same test station (minus one of the desks) plus an advertisement poster was located in an large anteroom, where it was partially shielded by sound-suppressing room dividers. The anteroom was relatively high in people-traffic, and the test station and the room dividers were arranged so that subjects would suffer less from visual and auditory distraction while engaged in their experimental task. The poster invited passers-by to volunteer for the experiment and to learn more about "how the mind works". The experimenter was a young woman.

The experiment was run on an AST Premia 4/66d computer, equipped with a 15" NEC MultiSync FG color monitor. To begin a test session, each subject was asked a set of background questions that focused on formal education, language, and health. At the science center, subjects were then shown a menu of tasks, including the CWST, on the computer monitor, and this article only reports results from those individuals who opted to do the CWST. At the community center, a test session began with the same background questions. After that, however, the subjects completed a battery of 12 different instruments that focused on sensory, perceptual, and higher-level cognitive functions. A full report of the entire set of instruments and their order of administration appears in Graf and Uttl (1995) and in Graf, Uttl, Mori, Birt, and Shapka (1997). The CWST was included in the middle of the sequence, and there was a rest-pause prior to each test.

Color-Word Stroop Test (CWST)

The CWST was administered on a 15" NEC MultiSync FG color monitor[1]. The test required the use of three different stimulus *display-pages*, and each of these was presented for three consecutive trials. Each display-page consisted of 36 items arranged into four columns and nine rows. The word-page (W) showed the color words *red*, *blue*, and *green* printed in white letters against a black background. The color-page (C) showed a series of Xs, either three, four, or five letters long (i.e., the same length as the corresponding color words) printed in *red*, *blue*, or *green* color against a black background. Finally, the incongruent-page (I) showed the color words *red*, *blue*, and *green* printed in incongruent colors (e.g., the word *red* printed in green) against a black background. The W and I pages had an equal number of the three critical colors words (i.e., 12 words in *red*, 12 words in *blue*, 12 words in *green*) and the C and I pages had an equal number of items printed in the three colors (i.e., 12 in red, 12 in blue, 12 in green). The position of the 36 items on each page was chosen randomly for each trial and for each subject. All items were displayed in upper-case letters, in 28-point Helvetica font. For each of the test pages, we also prepared a small practice page that differed from the critical materials only by being smaller and having fewer items, specifically, 12 items arranged in four columns and three rows.

Subjects were tested individually. The CWST was described as measuring attention and reading speed. The test pages were administered in the same order, W, C, and I, to each subject. Each subject received three different consecutive trials on each display-page, with trials separated by about a 20-s pause. For the W page, subjects were instructed to "read out loud the words on the computer monitor as fast as you can". For the C page, subjects were asked to "name out loud the color of the printed XXXs as quickly as you can", and for the I page they were asked to "ignore the written words and name out loud their ink-color as quickly as you can". Subjects received instructions prior to each series of critical trials, and they were allowed to practice each task. When they were ready and able to follow instructions, they completed the three trials with each test page.

Each trial was initiated by the experimenter. When a subject was ready, a key-press triggered the display of a test page. Subjects' then read the words or they named the ink colors (depending on instruction/condition), starting in the upper left corner and proceeding from the left to the right and from top to the bottom of the page. Subjects were instructed and also prompted (if necessary) to correct immediately any errors, and the experimenter recorded the number of errors. When a subject finished reading the last words or finished naming the color of the last item, a key-press terminated the trial and the computer recorded the time required

[1] A DOS version of our Color-Word Stroop Test is available from the authors.

for that trial. After a brief delay, the computer showed a 'Ready?' prompt, and when the subject was ready, the experimenter initiated the next trial.

RESULTS AND DISCUSSION

The critical dependent measures were the time (in seconds) spent on each page of the test, as well as the number of reading/naming errors per page. Errors were rare; they averaged 0.41 items (1.13%) per page. As expected, most of the errors were made in the incongruent condition where the average was 2.08%, while errors were below 1% on the remaining test pages. Consistent with previous findings, error rates were relatively stable across all subjects, except for the youngest and oldest groups (Graf et al., 1995; Hartman & Sweeny, 1994); they averaged 1.75%, 1.09%, 1.00%, 1.09%, 1.03%, 0.87%, and 1.63%, respectively, for the age groups listed in Table 1. Because errors were rare and self-corrected, they were not considered in any further analysis.

Each subject's recorded time per trial and display-page was divided by 36, the number of items on the page, to yield the time-per-item scores that were used for all subsequent analyses. As discussed elsewhere (Graf et al., 1995), time-per-item scores are useful because they allow direct comparison with other investigations that used instruments with either more or fewer items.

The data were screened for uni- and multivariate outliers as described in the Subject section. For subjects with univariate outliers on only some of the trials in a condition, their outlying scores were replaced with values equal to the upper or lower bound (for high or low outliers, respectively) of the 99% regression line confidence interval, thereby minimizing their influence.

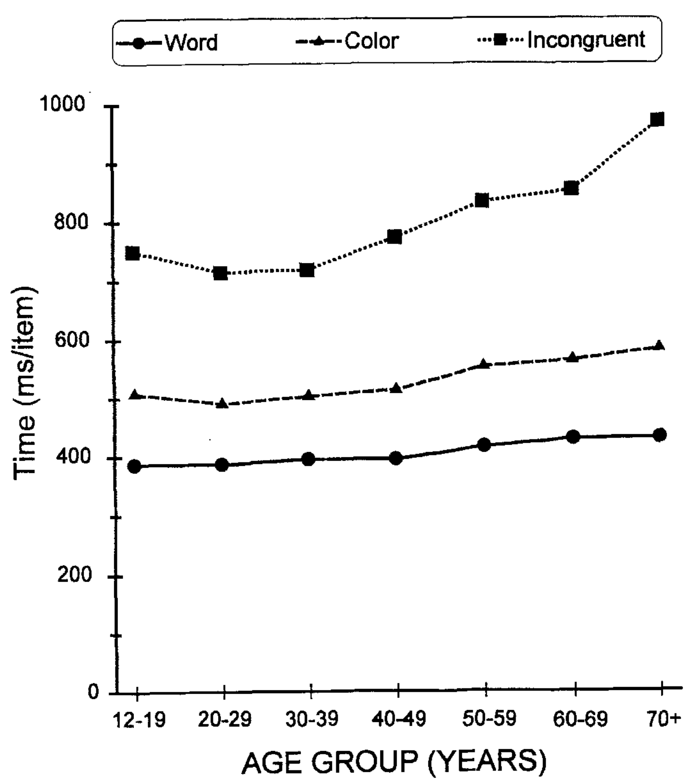

Fig. 1. The average time-per-item required by subjects from different age groups for performance in three different conditions of the Color-Word Stroop test.

Figure 1 shows CWST performance for the different age groups listed in Table 1. Table 2 shows the same data set, but arranged by midpoint overlapping age groups (Pauker, 1988). (Arranging the data in this manner, by midpoint overlapping age groups, highlights the regular changes across age groups; more important, it makes the findings more versatile as a normative set that can be used to make decisions about age-group typical performance.) The time-per-item data revealed: (a) that word reading speeds increased slightly but gradually across age groups, (b) that color naming was slower than word reading and it increased noticeably more with age, and (c), that naming colors with incongruous words was even slower and showed the most pronounced age-related increase. Hierarchical regression analyses that used age and age^2 to predict the time-per-item scores showed significant age effects for all conditions (see Table 2 for results), as well as a significant age^2 effect for the incongruent condition. (The age^2 effects index any non-linear influences due to age.) This general pattern of findings is consistent with and strengthens the results in the literature (Cohn, Dustman, & Bradford, 1984; Graf et al., 1995; Houx et al., 1993).

The data in Table 2 also permit more detailed quantitative comparisons between performance in the present versus previous investigations. Graf, Uttl, and Tuokko (1995) administered the CWST to older adults, and their published data (see their Table 3) show performance times that are only a few ms different from those in Table 2, but only for the word-reading condition. By contrast, the older adults' performance in the color-naming condition as well as in the incongruent condition was significantly higher in Graf et al. than in the present investigation. The same pattern of differences is also apparent in the data by Hartman and Sweeny (1994) who gave the CWST to young adults (M = 19.5 years of age) and to older adults (M = 67.3 years of age). Their subjects' performance in the word-reading condition was within a few ms of the results in Table 2, whereas their scores were significantly higher in the color-naming condition and in the incongruent condition. Performance that is comparable to the results in Table 2 for the word-reading condition, together with higher scores in the color-naming and incongruent conditions, has also been reported by several other investigations (e.g., Comalli, 1965; Houx et al., 1993; Rush, Panek, & Russell, 1990; Spreen & Strauss, 1991).

The pattern of similarities and differences in findings between the present and previous investigations is most easily explained in terms of the different test instruments that were used. Because performance differences occurred only in the conditions that involved color stimuli (i.e., in the color-naming and inçongruous condition), the finding of shorter times in the present study may reflect the fact that we used stimuli that differed in terms of properties, such as contrast or brightness, that facilitate color perception but not word reading.

The data in Table 2 constitute a substantive normative set that can be used for making decisions about normal or deviant CWST performance. In our experiment, each participant had three trials in each test condition, and thus we were able to compute test-retest (i.e., trial to trial) reliabilities as well as estimated reliabilities for averages across the three trials in each condition. The results are summarized in Table 3. Overall, the level of reliability in all conditions was respectable and comparable to that found in previous research (e.g., Graf, Uttl, & Tuokko, 1995; Jensen, 1965; Spreen & Strauss, 1991), as well as to those from many common psychometric tests (Anastasi, 1988; Kaplan & Saccuzzo, 1989). Equally important, all of our measures met the general level of reliability that is recommended for structural equation modeling (Cliff, 1983; Cohen, Cohen, Teresi, Marchi, & Velez, 1990).

Modeling Age Changes in CWST Performance

One objective of the present investigation was to explore the generality of a model of CWST performance reported by Graf et al. (1995) that postulates no direct connection between age and performance in the incongruous condition of the CWST. The first step towards specifying and evaluating a structural model for the present data was to develop a measurement model (An-

Table 2. Color-Word Stroop Test Performance (in ms per item) by Midpoint Overlapping Age Groups.

Condition		Age group midpoint and range[a]												Age[b]		Age[2b]	
		20 12–25 (n = 51)	25 20–30 (n = 63)	30 25–35 (n = 72)	35 30–40 (n = 70)	40 35–45 (n = 75)	45 40–50 (n = 68)	50 45–55 (n = 60)	55 50–60 (n = 41)	60 55–65 (n = 36)	65 60–70 (n = 33)	70 65–75 (n = 38)	75 70–83 (n = 33)	r^2	$F(1, 308)$	Δr^2	$F(1,307)$
Word	M	383	385	390	395	399	396	404	417	416	426	432	431	.06	19.55*	.00	.54
	SD	(58)	(55)	(61)	(64)	(54)	(54)	(59)	(59)	(60)	(64)	(63)	(62)				
Color	M	494	494	500	502	513	515	532	556	555	561	577	583	.11	36.12*	.01	2.17
	SD	(89)	(72)	(82)	(88)	(75)	(79)	(84)	(85)	(90)	(83)	(96)	(103)				
Incongruent	M	719	717	724	722	753	778	805	833	823	852	933	970	.19	72.12*	.02	7.92*
	SD	(155)	(125)	(121)	(129)	(127)	(141)	(149)	(145)	(161)	(159)	(191)	(193)				

[a] The sum of all subjects is greater then 310, due to the manner in which groups were constructed.
[b] Age and age^2 effects were computed by regression analysis.
* $p < .05$

Table 3. Test-Retest Reliabilities and the Estimated Reliabilities for the Average of Three Trials.

Condition	Test-retest R_i[a]	Estimated R_i[b]
Word	.88, .89	.96
Color	.78, .79	.92
Incongruent	.75, .77	.90

[a] Values for successive trials in each condition.
[b] Values for scores that were averaged across all trials per condition.

derson & Gerbing, 1988; Bollen, 1989) to represent the cognitive abilities that are indexed by performance on the various trials and conditions of the CWST (e.g., word reading, color naming, naming colors with incongruent stimuli). The second step was to examine specific predictions made by the global slowing hypothesis (Cerella, 1985; Salthouse, 1980, 1985, 1991, 1993, 1996; Salthouse, Kausler, & Saults, 1988), and by the competing hypothesis that age might also have a unique, selective, or additional effect on Stroop interference. All analysis were carried out with SEPATH, the structural equation module of Statistica/W (Steiger, 1994, 1995; Statistica/W, 1994, 1995). Model fitting was based on the correlation matrix following the general analytical strategy advocated by Cudeck (1989). Because our data were consistent with assumptions for using the standard maximum likelihood estimation procedure we used this method for evaluating possible models[2] (note: the normalized Mardia coefficient of multivariate kurtosis was 1.79; see Subjects section).

A summary of our approach and of our models appears in Table 4. To evaluate each model's fit, we used the Steiger-Lind *RMSEA* (Steiger, 1994, 1995; Steiger & Lind, 1980), an absolute badness-of-fit index (as opposed to the indices that are relative to a null or to an alternative model) that takes into account a model's parsimony, and that is not sample-size dependent (McDonald & Marsh, 1990). The table gives a point estimate and the 90% confidence interval for each *RMSEA*; *RMSEA* values above .10 indicate a poor fit and values below .05 indicate a very good fit (Steiger, 1994, 1995). We also provide Γ (Steiger, 1989), the population estimate of *GFI* (Jöreskog & Sörbom, 1984), which is a goodness-of-fit index reflecting the proportion of variances and covariances accounted for by a model. Table 4 also includes the normed-fit-index (*NFI*) and the non-normed-fit-index (*NNFI*) by Bentler and Bonnett (1980). As a general rule, Γ-, *NFI*-, and *NNFI*-values above .95 are desirable (Marsh, Balla, & McDonald, 1988; Steiger, 1994). Finally, we report χ^2 values (lower values indicate better fits), degrees of freedom, and corresponding p values for all models that were examined.

Measurement models

A measurement model with five correlated factors was fit to the data. Two of these factors (age and age^2) were defined directly by subjects' stated age, whereas the remaining factors (word reading, color naming, and color naming with incongruent stimuli) were defined by the time-per-item scores for the three trials on each display-page. Age^2 was included in order to capture the non-linear influence due to age; this factor was computed by regressing age^2 on age and then saving the residuals (i.e., by finding the quadratic component of age which is orthogonal to the linear component of age). Initially, all five factors were free to correlate with each other.

The fit indexes in Table 4 indicate that the initial measurement model (MM1) was very good. Because age^2 was defined as being orthogonal to (i.e., uncorrelated with) the correlation between age and age^2 was set to zero for MM2. As expected, the fit of this new measurement model was also very good. Finally, in order to examine whether age^2 influences any of the factors involved in CWST performance, the links between age^2 and all other factors were set to zero for MM3. The difference in the χ^2 values associated with MM3 and MM2 was significant, $\Delta\chi^2\,(3) = 10.21, p < .05$, thereby indicating that age^2 is important for explaining the pattern of correlations among the factors.

Table 5 lists the standardized path coefficients of all factor indicators, as well as all correlations among the factors, associated with the most parsimonious measurement model (MM2). In addition, Table 6 shows the correlations among all factor indicators.

[2] It is also possible to model CWST performance by means of path analysis (i.e., modeling with observed variables), by using the average of the three trials per condition. However, structural equation modeling with latent variables, with each trial score serving as an indicator of an underlying factor, is preferable because it focuses only on that part of performance which is common to all trials.

Table 4. Color-Word Stroop Test Performance: Summary of Model-Testing Approach.

Model	Commentary	$\chi2$	df	p	$RMSEA$[a] Point	$RMSEA$[a] 90% C.I.	Γ[b]	NFI[c]	NNFI[d]	$R^2_{Incong.}$[e]
Measurement models (MM)										
MM1	Free correlations among all factors.	53.45	36	.031	.038	.008↔.060	.991	.982	.991	
MM2	Free correlations among all factors except with correlation between age and age^2 set to zero.	53.45	37	.039	.036	0↔.058	.991	.982	.992	
MM3	Same as MM2 but with all correlations between age^2 and other factors set to zero.	63.66	40	.010	.043	.020↔.062	.987	.979	.989	
Structural models (SM)										
SM1	Age → word[f], word → color, color → incongruent.	108.07	43	< .001	.066	.049↔.083	.967	.964	.972	.771
SM2	Same as SM1 but with age^2 → word added.	107.31	42	< .001	.067	.050↔.084	.967	.964	.971	.771
SM3	Same as SM1 but with age → color added.	85.92	42	< .001	.056	.038↔.074	.976	.971	.980	.784
SM4	Same as SM3 but with age^2 → color added.	81.83	41	< .001	.056	.038↔.074	.977	.973	.981	.786
SM5	Same as SM4 but with word → incongruent added.	81.65	40	< .001	.057	.039↔.075	.977	.973	.980	.791
SM6	Same as SM4 but with age → incongruent added.	60.00	40	.022	.040	.014↔.060	.989	.980	.991	.805
SM7	Same as SM6 but with age^2 → incongruent added.	54.48	39	.051	.034	0↔.056	.992	.982	.992	.810
SM8	Same as SM7 but with age^2 → color deleted.	56.76	40	.041	.035	0↔.056	.991	.981	.992	.809

[a] Point estimate and 90% confidence interval for Steiger-Lind RMSEA Index (Steiger & Lind, 1980; Steiger, 1994).
[b] Population Gamma Index (Steiger, 1989).
[c] Bentler-Bonett Normed Fit Index (Bentler & Bonett, 1980).
[d] Bentler-Bonett Non-Normed Fit Index (Bentler & Bonett, 1980).
[e] Proportion variance explained in the latent variable associated with performance in the Incongruent condition.
[f] "→" stands for "linked to" or "affects".
For Evaluating Model Fit: Relevant Critical difference is $\chi2$ (df = 1) = 3.84.

Table 5. Final Measurement Model (MM2): Indicator Loadings and Correlations Among Factors.

Factor Indicators	λ_i
Age	.98[a]
Age^2	.98[a]
Word	
Trial 1	.92*
Trial 2	.96*
Trial 3	.94*
Color	
Trial 1	.90*
Trial 2	.89*
Trial 3	.88*
Incongruent	
Trial 1	.86*
Trial 2	.88*
Trial 3	.86*

Correlations Among Factors	1	2	3	4
1. Age				
2. Age^2	.00[a]			
3. Word Reading	.25*	.04		
4. Color Naming	.35*	.09	.84*	
5. Color Naming w. Incongruent Stimuli	.47*	.16*	.74*	.88*

[a]One parameter was fixed prior to the model estimation (see text).
* $p < .05$.

Table 6. Correlations Among All Stroop Test Scores, Age, and Age^2.

	1	2	3	4	5	6	7	8	9	10
1. Word/Trial 1										
2. Word/Trial 2	.88*									
3. Word/Trial 3	.86*	.89*								
4. Color/Trial 1	.68*	.73*	.72*							
5. Color/Trial 2	.67*	.71*	.69*	.78*						
6. Color/Trial 3	.67*	.69*	.71*	.80*	.79*					
7. Incongruent/Trial 1	.55*	.61*	.59*	.68*	.68*	.62*				
8. Incongruent/Trial 2	.59*	.62*	.60*	.69*	.69*	.65*	.75*			
9. Incongruent/Trial 3	.59*	.63*	.62*	.69*	.72*	.66*	.72*	.77*		
10. Age	.24*	.22*	.24*	.30*	.33*	.27*	.47*	.38*	.33*	
11. Age^{2a}	.05	.06	.01	.10	.10	.04	.13*	.11*	.16*	.00[a]

[a] Age^2 score reflects only a component orthogonal to participants' age (see text).
* $p < .05$.

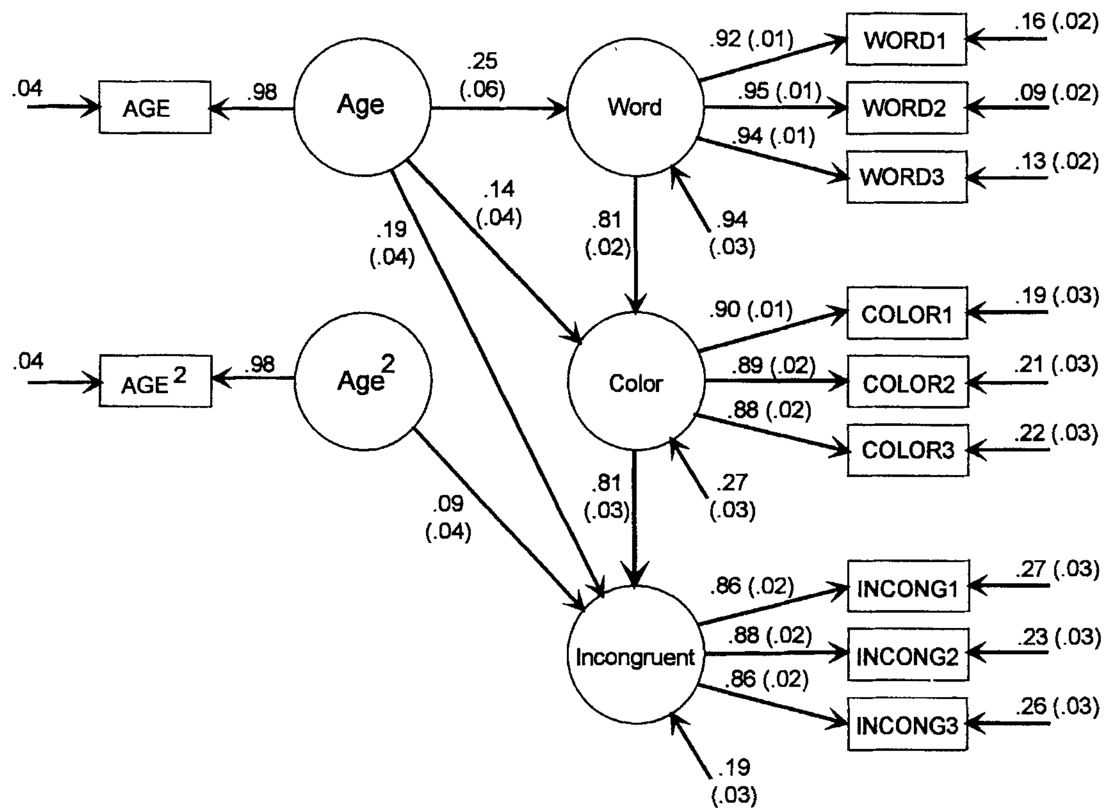

Fig. 2. The final structural model (SM8) showing the relationship among age and age^2, word reading, color naming, and incongruent color naming.

Structural model

Our structural modeling strategy was straightforward. First, we attempted to explain the more complex factors (e.g., naming colors in the incongruent condition) in terms of less complex factors (e.g., color naming and word reading, in this order). Second, we tried to explain age effects on the more complex factors in terms of age effects on the less complex factors, and we probed whether there was need for direct links between age or age^2 and any of the more complex factors.

For the first structural model, we had postulated the following links: age → word reading (note: "→" stands for "is linked to", "affects" or "is functionally related to"), word reading → color naming, and color naming → incongruent color naming. The fit indexes in Table 4 show that this model (SM1) describes the data reasonably well; it accounted for 77.1% of the variance in the latent factor corresponding to the incongruent color-naming condition. For subsequent models, we tested whether additional direct links, for example, from age^2 to word reading, or from age or age^2 to color naming or to color naming with incongruent stimuli, would improve the model fit. Table 4 shows that addition of a link between age^2 and word reading (SM2) did not improve the model fit, thereby indicating that word reading speed is a linear function of age. By contrast, the addition of two other links, one between age and color naming (SM3) and the other between age^2 and color naming (SM4), did improve the model fit. Consistent with our previous work (Graf et al., 1995), this outcome indicates that age has a direct influence on color naming speed (i.e., an influence that is not functionally linked with word-reading speed), and in addition, this influence has a small but significant non-linear component. In total, SM4 accounted for 78.6% of the variance in the factor associated with performance in the incongruent condition.

The most critical question is whether performance in the incongruent stimulus condition shows a significant direct influence due to age or age^2, that is, an any influence that is not directly linked with color naming or word reading. Do age effects in the incongruent condition provide positive evidence that requires postulating a distinct type of processing that is not required for color naming or for word reading? For the sake of completeness we also examined the possibility that age effects in incongruent color naming can be explained directly in terms of word reading speed (SM5), and the results in Table 4 show that adding such a link did not improve the model fit. By contrast, the addition of direct links between age and incongruent color naming (SM6) and between age^2 and incongruent color naming (SM7) did improve the model fit. While significant, however, the increase in model fit from SM4 to SM7 was relatively small; it amounted only to a .024 increase in the proportion of variance explained in the incongruent condition. The fit indices and the Chi-square statistics in Table 4 indicate that SM7 made a very good fit with the data. An examination of the individual links in SM7 revealed that the direct link from age^2 to color naming was not significant, and thus in the interest of parsimony, it was deleted from the final model (SM8) without any loss of fit. The final model is shown in Figure 2.

The final model shows that age effects in the incongruent condition are associated primarily and strongly with age effects in color naming and in word reading,[3] but it also shows a direct link between age and performance in the incongruent condition. What is critical is that the latter, direct link is weak; it is possible that this is an artifact due to a few odd subjects in the extreme age groups.

CONCLUSION

The goal of the present investigation was to examine whether the pattern of age changes in Stroop interference provides positive proof for a separate cognitive factor or processing resource, one that is not indexed by color naming or by word reading. The outcome of our investigation and the statistical analyses provide no such positive evidence; they show that age effects in the incongruous condition of the CWST are directly related to, are a direct function of, performance in the color naming and word reading condition. Consistent with this outcome, any explanation, such as global slowing, that can deal with age changes in color naming could easily be tweaked for explaining age effects observed in the incongruous condition (for example, by making the simple assumption that the incongruous condition differs by being more difficult than the color naming condition). The age effects we observed in Stroop interference do not provide positive evidence for a separate cogni-

[3] The data summarized in Table 2 and Figure 1 may give the conflicting impression that performance in the incongruous condition is critically different in some manner. This impression persists, and is even strengthened, when we look at Stroop interference expressed as a proportion relative to color naming, as was suggested by two reviewers of this article. Across the overlapping age groups listed in Table 2, the proportional scores are as follows, in terms of means and standard deviations (the latter are shown in brackets): 0.45 (0.18), 0.46 (0.20), 0.46 (0.18), 0.45 (0.17), 0.47 (0.18), 0.52 (0.21), 0.52 (0.19), 0.50 (0.16), 0.49 (0.20), 0.52 (0.22), 0.62 (0.21), 0.67 (0.16). The means are ordered from youngest to oldest group. When we used these scores in multiple regression analyses, that is, in order to quantify the relationship among age, reading, color naming, and naming colors in the incongruous condition, age and age^2 explained less then 7% of the variance in incongruous condition; this figure is comparable to the outcome of our structural model which showed only about a 2% direct effect for the same variables (in contrast, age and age^2 explained 18.1% of the variance in the widely used difference scores). We did not work with proportional scores, however, and we recommend against using proportions in this case, because meaningful, interpretable proportions can only be computed with ratio scale data, and this criterion is not likely to be met by performance on the CWST.

tive factor or processing resource; they do not force us to postulate that the incongruous conditions engages a separate, qualitatively different type of processing (one that also declines with age) from that required for color naming and word reading.[4] Invoking Occam's principle of parsimony, the outcome of our investigation suggests instead that age effects in the incongruent condition are due to the same factors that produce age effects in color naming.

REFERENCES

Anastasi, A. (1988). *Psychological testing*. New York: Macmillan.

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

Baddeley, A. (1986). *Working memory*. Oxford: Oxford University Press.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation, Vol. 8* (pp. 47-89). New York: Academic Press.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 51*, 35-51.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

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

[4] The possibility that the incongruous condition of the CWST engages a separate, different kind of processing from that required for color naming or word reading is not ruled out (i.e., it is simply not ruled in) by the fact that our investigation revealed no direct, strong links between age and performance in the incongruous condition. As was noted by a reviewer of an earlier version of this manuscript, the incongruous condition might engage qualitatively different kinds of processing from color naming or word reading, but our investigation would have picked up this difference in processing only if it also were changing with age. If color naming involves a distinct type of processing that does not change with age (e.g., like direct priming), our investigation would have missed it. Of course, any claim along these lines would not illuminate age changes in Stroop interference.

Cerella, J., Poon, L. W., & Williams, D. M. (1980). A quantitative theory of mental processing time and age. In L. W. Poon (Ed.), *Aging in the 80s: Selected contemporary issues in the psychology aging* (pp. 332-340). Washington, DC: American Psychological Association.

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

Cohen, P., Cohen, J., Teresi, J., Marchi, M., & Velez, C. N. (1990). Problems in the measurement of latent variables in structural equations causal models. *Applied Psychological Measurement, 14*, 183-196.

Cohn, N. B., Dustman, R. E., & Bradford, D. C. (1984). Age-related decrements in Stroop Color Test performance. *Journal of Clinical Psychology, 40*, 1244-1250.

Comalli, P. E. (1965). Cognitive functioning in a group of 80–90 year-old men. *Journal of Gerontology, 20*, 14-17.

Craik, F. I. M. (1986). A functional account of age differences in memory. In F. Klix & H. Hagendorf (Eds.), *Human memory and cognitive capabilities: Mechanisms and performances* (pp. 409-422). New York: Elsevier Science.

Cudeck, R. (1989). Analysis of correlation matrices using covariance structures. *Multivariate Behavioral Research, 18*, 147-167.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*, 450-466.

Dulaney, C. L., & Rogers, W. A. (1994). Mechanisms underlying reduction in Stroop interference with practice for young and old adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 470-484.

Graf, P., & Masson, M. E. J. (1993). *Implicit memory: New directions in cognition, development, and neuropsychology*. Hillsdale, NJ: Lawrence Erlbaum.

Graf, P., & Uttl, B. (1995). Component processes of memory: Changes across the adult lifespan. *Swiss Journal of Psychology, 54*, 113-130.

Graf, P., Uttl, B., Mori, M., Birt, A., & Shapka, J. (1997). Qualitative changes in picture descriptions and memory across lifespan. (manuscript in preparation).

Graf, P., Uttl, B., & Tuokko, H. (1995). Color- and picture-word Stroop tests: Performance changes in old age. *Journal of Clinical and Experimental Neuropsychology, 17*, 390-415.

Hartman, M., & Sweeny, S. F. (1994, February). *The Stroop test: One source of age differences identified*. Presented at the meeting of the International Neuropsychological Society, Cincinnati, OH.

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General, 108*, 356-388.

Houx, P. J., Jolles, J., & Vreeling, F. W. (1993). Stroop interference: Aging effects assessed with the Stroop color-word test. *Experimental Aging Research, 19*, 209-224.

Jensen, A. R. (1965). Scoring the Stroop test. *Acta Psychologica, 24*, 398-408.

Jöreskog, K. G., & Sörbom, D. (1984). *LISREL VI. Analysis of linear structural relationships by maximum likelihood, instrumental variables, and least squares methods*. Mooresville, IN: Scientific Software.

Jorm, A. F. (1986). Controlled and automatic information processing in senile dementia: A review. *Psychological Medicine, 16*, 77-88.

Kaplan, R. M., & Saccuzzo, D. P. (1989). *Psychological testing: Principles, applications, and issues*. Belmont, CA: Brooks/Cole Publishing.

Kausler, D. H. (1991). *Experimental psychology, cognition, and human aging*. New York: Springer-Verlag.

Kausler, D. H. (Ed.) (1994). *Learning and memory in normal aging*. San Diego, CA: Academic Press.

MacLeod, C. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Review, 109*, 163-203.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57*, 519-530.

Mardia, K. V., & Foster, K. (1983). Omnibus tests of multinormality based on skewness and kurtosis. *Communication in Statistics, 12*, 562-582

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.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. *Psychological Bulletin, 107*, 247-255.

Pauker, J. D. (1988). Constructing overlapping cell tables to maximize the clinical usefulness of normative test data: Rationale and an example from neuropsychology. *Journal of Clinical Psychology, 44*, 11-17.

Rush, M. C., Panek, P. E., & Russell, J. E. (1990). Analysis of individual variability among older adults on the Stroop Color Word Interference Test. *International Journal of Aging and Human Development, 30*, 225-236.

Salthouse, T. A. (1980). Age and memory: Strategies for localizing the loss. In L. W. Poon, J., J. L. Fozard, L. S. Cermak, D. Arenberg, & L. W. Thompson (Eds.), *New directions in memory and aging: Proceedings of the George A. Thalland Memorial Conference* (pp. 47-65). Hillsdale, NJ: Lawrence Erlbaum.

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

Salthouse, T. A. (1988). Resource-reduction interpretations of cognitive aging. *Developmental Reviews, 8*, 238-272.

Salthouse, T. A. (1991). Mediation of adult age differences in cognition by reductions in working memory and speed of processing. *Psychological Science, 2*, 179-183.

Salthouse, T. A. (1993). Speed and knowledge as determinants of adult age differences in verbal tasks. *Journals of Gerontology: Psychological Sciences, 48*, 29-36.

Salthouse, T. A., Kausler, D. H., & Saults, J. S. (1988). Utilization of path-analytic procedures to investigate the role of processing resources in cognitive aging. *Psychology and Aging, 3*, 29-37.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review, 103*, 403-428.

Spreen, O., & Strauss, E. (1991). *A compendium of neuropsychological test: Administration, norms, and commentary*. New York: Oxford University press.

Statistica/W [Computer Software]. (1994). Tulsa, OK: StatSoft.

Statistica/W [Computer Software]. (1995). Tulsa, OK: StatSoft.

Steiger, J. H. (1989). *EzPATH: A supplementary module for SYSTAT and SYGRAPH*. Evanston, IL: SYSTAT.

Steiger, J. H. (1994). *Statistica: Structural equation modeling*. Tulsa OK: StatSoft.

Steiger, J. H. (1995). Structural equation modeling. In StatSoft, *STATISTICA for Windows (Addendum): New features in STATISTICA 5.0*, 283-487.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically-based tests for the number of common factors*. Paper presented at the Annual Spring Meeting of the Psychometric Society in Iowa City,IA.