Inspection Time as a Biological Marker for Functional Age

Tess A. Gregory

School of Psychology
University of Adelaide
October 2006

TABLE OF CONTENTS

ABSTRACT

Inspection Time (IT) is a speed measure that has been primarily investigated in the field of individual differences. However, Nettelbeck and Wilson (2004) proposed that IT could have promise as a biomarker for functional outcomes, particularly cognitive aging. The premise behind biomarker research is that chronological age is simply a proxy for the physiological and cognitive changes that occur in the body with advancing age. Biomarkers are measures that 'mark' the aging process and represent the biological age of an individual rather than the years since his/her birth. Speed of processing tasks offer promise as biomarkers because decline in speed of processing is one of the most robust findings in cognitive aging research. However, traditionally used tasks are problematic because they confound speed and accuracy and some are sensitive to cohort effects. Inspection time is a speed of processing measure that is free from these problems and is therefore a promising candidate for a biomarker. This dissertation presents the first empirical investigation of this proposition.

One hundred and fifty elderly participants were assessed on IT, traditionally used biomarkers (e.g. grip strength, visual acuity), a battery of cognitive tasks (e.g. fluid ability and crystallised ability) and measures of everyday functioning (e.g. activities of daily living). These individuals were assessed on three separate occasions over a period of 18-months. For the biomarkers, initial scores, 6-month change scores and 18-month change scores were generated and used to predict final scores and 18-month change scores on the functional outcomes (cognition and everyday functioning). Results revealed that slow IT at the start of the study was associated with dependence in activities of daily living and poorer fluid ability at the end of the study. There was also evidence that slow IT at the start was associated with decline in fluid reasoning over the subsequent 18-months. Moreover, consistent with the major aims of this study, decline in IT over time was associated with more cognitive problems in daily life and poor fluid ability at the end of the study. Given that initial and change scores for IT were independent, due to the methodology used to estimate them, the two measures explained unique variance in the functional outcome measures.

These findings are extremely encouraging, particularly given the relatively short time frame for this study. IT has predictive validity for everyday functioning and cognitive aging over an 18-month period, and therefore, it is concluded that IT has promise as a valid biomarker for functional age. Recommendations for further research include investigating the link between IT and mortality, examining the association between IT and a broader range of functional age measures, the replication of these findings in a different sample, and means for improving the sensitivity and specificity of the current IT estimation procedure.

DECLARATION

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

I give consent to this copy of my thesis, when deposited in the University Library, being made available in all forms of media, now or hereafter known.

_____ _____

Tess A. Gregory Date

ACKNOWLEDGEMENTS

First, I would like to thank my supervisor Professor Ted Nettelbeck from the University of Adelaide I sincerely appreciate the time you have taken to read through my draft chapters, the many useful recommendations that you have made throughout my dissertation, and for the encouragement to attend international conferences and meet with researchers in this field, which has made the PhD experience much more rewarding.

Second, I would like to thank my supervisor Dr. Carlene Wilson from CSIRO, who presented this idea to me as a potential PhD project. While difficult at times, the investigation of IT as a biomarker ended up being an excellent PhD project and I appreciate being given the chance to work on it. I appreciate the many useful comments you have provided on my drafts and the suggestions about alternative statistical procedures to use throughout.

Third, I would also like to acknowledge the CSIRO for providing additional funding for my PhD and financial assistance for travel. I would also like to thank CSIRO for the use of their nutritional databases to analyse the nutritional data and for help with computer issues on this project.

Fourth, I would like to acknowledge Sara Howard who assisted in collecting the longitudinal data for this project, and made this great dataset possible. I also appreciate the many discussions on the theoretical issues involved in this project and the emotional aspects of completing a PhD project.

Fifth, I would like to thank all of the participants who were involved in this research project. The four-hour testing sessions were time-consuming and required intense concentration and perseverance. This research would not be possible without you and I offer my sincere thanks for your time and interest in my research.

Thanks must also go to the numerous researchers who offered advice throughout this project on psychometric tests, questionnaires, statistical procedures and relevant conferences to attend. These include Dr Nick Burns, Dr Janet Bryan, Dr Kaarin Anstey, Professor Ian Deary, Professor Timothy Salthouse, and Dr Linley Denson, to name a few.

Finally, I would like to thank my husband Shanan Gregory for the support he has given me during the past four years. I appreciate the many discussions on theoretical issues surrounding biomarkers and functional age and your ideas about the causes of some of the more unusual findings in my dataset. I would like to thank you for the emotional support and encouragement you have given me, particularly in the past 6-months.

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

ADL	Activities of Daily Living
ADAS-Cog	Alzheimer's Disease Assessment Scale - Cognitive
BP	Blood Pressure
CA	Chronological Age
CCFT	Cattell Culture Fair Test
CDL	Cognition in Daily Life
CF	Concept Formation
CNS	Central Nervous System
DS	Digit Symbol
FA	Functional Age
Gc	Crystallised Ability
Gf	Fluid Ability/ Reasoning
Gs	Speed of Processing
IT	Inspection Time
PC	Pattern Comparison
RSPM	Raven's Standard Progressive Matrices
VA	Visual Acuity
VM	Visual Matching

CHAPTER ONE: FUNCTIONAL AGE AND BIOMARKER RESEARCH

Inspection time (IT) has been studied extensively for nearly 30 years, with most research being done in the fields of intelligence and individual differences. Since the initial publication (Nettelbeck & Lally, 1976) that suggested IT was related to psychometric intelligence, numerous researchers have examined the task with respect to a wide range of outcomes. Grudnik and Kranzler (2001), in a review article, included over 90 studies, with more than 4100 participants, which reported on IT. Both a diverse range of research areas (e.g. intelligence, developmental psychology, psychophysiology, learning difficulties and degenerative diseases) and participant groups (e.g. children, young and old adults, people with an intellectual disability, gifted and multicultural groups) have been examined using the IT task.

In a recent publication Nettelbeck and Wilson (2004) suggested a novel use for the IT measure. In this study the goal was to test whether average IT became shorter over successive generations (i.e. showed a cohort effect). Flynn (1987; 1999) demonstrated that many cognitive abilities show improvement over successive generations, with the largest increases in reasoning and non-verbal tasks. Such cohort effects are particularly problematic for the interpretation of cross-sectional aging research, because decline widely found with old age is confounded by improvements in more recently born cohorts. It is not clear what causes these cohort effects but Nettelbeck and Wilson were interested in whether speed of processing might play a role. A group of children (6 – 13 years) completed the IT task and a measure of vocabulary and these data were compared to data collected from an earlier cohort of children with the same ages, collected at the same school 20 years previously. A significant improvement was seen in the vocabulary task but there was no evidence that speed of processing, as measured by IT, had improved. Given that IT appeared to be stable over a 20-year period of time, at least in children, Nettelbeck and Wilson (2004, p. 85) suggested "IT may have promise as a useful biological marker for an important component of cognitive decline during old age".

Speed of processing tasks have been used as biological markers in the past but tasks like Digit Symbol from the Wechsler Scales display cohort effects. One major requirement of a biological marker is that it is stable across generations and cultures, thus ruling out non-stable speed tasks. Because IT is free from cohort effects it offers promise as a valid biological marker

of aging. That is, IT might act as a lead indicator of unfavourable changes in cognition with advancing age. Investigation into this proposition will form the basis of this dissertation.

Functional Age

The Concept of Functional Age

The term functional age (FA) was first used by McFarland (1956; 1973; McFarland & Philbrook, 1958) to denote the level or "age" at which a person was functioning. McFarland, working in the area of occupational psychology, noted that an alternative to chronological age (CA) was needed to represent functional capacities. In many situations, particularly in an occupational setting, important decisions have been made on the basis of CA although it is clearly unreliable for predicting level of performance. McFarland (1956, p. 235) suggested, "it is evident with all productive workers, the important variable to consider is not chronological, but rather, functional age or the ability to perform required duties efficiently and safely".

The crux of the problem with CA is that it is not particularly good at indicating functional capacity. Two people of the same CA could be functioning at very different levels. One 65-year-old might be bright, healthy and living independently, while another might be experiencing striking memory and health problems and require nursing home care. The main reason that CA is so poor at indicating functional capacity is that there are marked individual differences in the onset and rate of decline of various abilities across people. Morse (1993), and many others, have shown that variability between individuals in cognitive abilities increases beyond the 50s. Increased variability is also seen on sensorimotor variables such as motor skills (Spirduso & MacRae, 1990) and sensory functioning (Heron & Chown, 1967) as people age. Furthermore, there is variability between different systems in the same individual (Fozard, Metter, & Brant, 1990). The implication of this variability is that CA is not a good predictor of an individual's level of functioning, whether defined by cognition, health, or activities of daily living.

The Measurement of Functional Age

Functional age is a measure of current functional capacity rather than the years that have passed since an individual's birth. However, measuring a person's FA is not simple. The question of how to measure FA has been debated for a long time and a number of different approaches have been used. There are essentially two issues that need to be addressed in order to

achieve a satisfactory measure. First, it is necessary to be more specific about what kind of *functioning* is of interest. Second, a statistical method is required to generate a measure of FA.

Several types of functioning have been of interest to researchers. A review article by Anstey, Lord and Smith (1996) found FA studies that had considered seven types of functioning: anthropometric (e.g. height, weight), sensorimotor (e.g. grip strength, visual acuity), cognitive (e.g. fluid reasoning, attention), psychosocial (e.g. stress, subjective health), behavioural (e.g. activities per day, sleep duration), physiological/biomedical (e.g. blood pressure, pulse rate), and dentition (e.g. number of teeth, plaque index). Indices used generally varied with the focus of the study. In a health setting, it might be the physiological/ biomedical indices that are considered to be of upmost importance. In an employment setting, both cognitive and sensorimotor ages are likely to be important. Once the type of functioning has been selected, the focus moves to generating an estimate of FA.

In early FA research, there were three main statistical methods employed: *multiple-regression*, *profiles* and *data reduction*. In the *multiple regression* approach, a large number of putative age-related variables (e.g. visual acuity, grip strength, forced expiratory volume, auditory acuity, and body mass index) were regressed on CA. The predicted value from the regression equation was taken as an estimate of FA and was considered particularly useful because it was a single score and could be compared directly to CA. In the *profile* approach, a profile was developed for each task allowing for an individual to be equated with a particular CA for each variable separately. Anstey et al. (1996) pointed to similarities between this method and a WAIS-R profile, with the IQ score being analogous to the FA score generated in the multiple regression technique. This method was considered useful because different biological and psychological systems are known to decline with age at different rates (Fozard et al., 1990) but group data cannot easily be examined. In studies that have focused on the relationships between variables, *data reduction* techniques have been employed to test whether a single 'aging factor' or multiple aging factors exists and which measures define them. This method involved entering all variables, including CA, into a large factor analysis. The factor defined by CA, was taken as the 'aging' factor and those variables that loaded on this factor were examined. To clarify further the statistical methods used, two prominent FA studies will be discussed. Murray (1951) performed the earliest study using the multiple-regression method and Heron (1962; Heron & Chown, 1967) used both data reduction and profiles to investigate FA in the Liverpool Aging Project.

Functional Age Research

Murray (1951) presented the first attempt to combine a number of variables into what he called a 'physiologic age' score. He selected five physiological variables (auditory acuity, visual accommodation, systolic blood pressure, sensitivity of the dark-adapted eye, and grip strength), all of which had previously been shown to be age related, and measured them in a group of 38 men aged 21 - 84 years. These five variables were entered into a regression equation with chronological age (CA) as the dependent variable. In addition, quadratic terms were included for each variable to account for non-linear age relations, which Murray suggested were obvious in many published graphs. The regression equation generated an estimate of physiological age, which could be compared with CA, for each subject. Rather than suggesting potential uses for this physiological score, Murray advocated using this method with a large number of physiological, anatomical and psychological variables, to generate a 'biologic age' score. If done adequately, he suggested this biological score could be used for life insurance applicants and patients in clinical medicine. That is, he suggested the biological age score could predict mortality or morbidity.

An alternative method was described by Heron and Chown (1967) in the Liverpool Age Project. A large number of tasks that were thought to be age-related were measured in a sizeable sample (N = 540) stratified by age (20 – 79 years). The measures in this study consisted of physical, physiological, sensory, cognitive and personality variables. Although development of norms was the primary goal, examination of the relationship between variables was a secondary aim (Heron & Chown, 1967).

Initially, factor analysis was used to examine the relationship between the variables. Separate factor analyses were performed for each gender by entering all measures, including CA and socio-economic status. The factor structure accounted for 65% of the variance for the males and 67% of the variance for the females and was very similar for both. Seven factors were extracted and interpreted as Age, Socio-Economic/Education, Temperament, Physique, Cardiovascular, Personal Rigidity and an unnamed factor. The only factor that CA displayed a significant loading on was the first factor "Age", which incorporated physical (e.g. sitting height, grip strength), physiological (e.g. forced expiratory volume) and sensory (e.g. visual and auditory acuity) measures. The cognitive variables loaded on both the "Age" and "Socio-economic/ Education" factors, suggesting that the factor analysis separated the individual differences and age-related variance in the cognitive tasks. However, there were gender differences in the

loadings on these two factors. For males, the "Age" factor had salient loadings from some cognitive variables (Matrices, Perceptual mazes, Digit coding and Trails) but for females the cognitive tasks were more related to the "Socio-Economic/ Education" factor. The personality variables loaded on "Temperament" and "Personal rigidity". This suggested that significant gender differences may exist in cognitive tasks and, if possible, analyses should be done for each gender separately.

In order to develop norms (or profiles), the relationship between each variable and age was examined in depth and validated against previous research. All of the physical measures were significantly related to age but some displayed non-linear age relations (e.g. weight, bicep circumference, and hearing), and hence could not be represented by a correlation. All *physiological* measures, except for pulse rate, correlated with age. The most significant correlation was forced expiratory volume (r = - 0.70). Most psychological measures were age-related, with memory and reasoning having the highest correlations, while vocabulary displayed no age-related decline. The personality measures showed some changes with age. However, these changes were not apparent in the factor analysis because they were not related to the other age-related measures (Heron & Chown, 1967).

The sample was split into age cohorts (by decade) for males and females to create profiles for FA. For each measure, the 25^{th}, 50^{th} (median) and 75^{th} percentile scores were calculated (see Heron & Chown, 1967, Appendix I). This allowed a profile to be constructed for any individual person. For example, Mr X aged 60 years might have the strength of a 65-year-old, the vision of a 70-year-old but the memory of a 30-year-old. Heron and Chown (1967) suggested this demonstrates a major advantage of profile scores over a single FA score. By amalgamating all measures into a single score a lot of information is lost and the true picture may be distorted. Furthermore, if the FA measure is used for personnel selection, as suggested by Dirken (1972), then different employers are likely to be interested in different aspects of functioning. Heron and Chown suggested that, for manual work, physical and physiological measures are likely to be of upmost importance, whereas in professional work cognitive measures are likely to be more vital. Profiles scores would allow for this distinction to be made but a single FA score would not.

These two studies illustrate the methods that were commonly used at the time. Many researchers (Bell, 1972; Damon, 1972; Dirken, 1972; Furukawa et al., 1975; Heikkinen, Kiiskinen, Käyhty, Rimpelä, & Vuori, 1974; Hollingsworth, Hashizume, & Jablon, 1965; Webster & Logie, 1976) generated estimates of FA using the multiple-regression model, with

some studies reporting R-values above 0.90. Other researchers used the profile approach (e.g. Borkan & Norris, 1980a, 1980b) or data reduction (e.g. Clark, 1960) but these methods were used much more rarely. In addition to the large number of publications on FA, there has been a lot of criticism of FA research, in particular the statistical methodologies employed.

Criticisms of Functional Age

Costa and McCrae (1980) published a detailed critique of the conceptual and empirical problems with FA studies. Although these criticisms focused on the multiple regression technique, Costa and McCrae suggested that many would apply to the factor analytical studies too. Four major criticisms were presented and will be discussed in detail below.

The first criticism of the FA methodology was that, *multiple regression is not an appropriate method*. Multiple regression assumes that the variables are ratio scales, normally distributed, and that the age associations are linear. These assumptions are often not met, particularly the linearity assumption. Murray (1951) noted that quadratic terms needed to be included but very few researchers have actually included non-linear terms in their regressions. Another problem with this method is that *regression to the mean* occurs (Bulpitt, 1995; Costa & McCrae, 1980; Hochschild, 1990). Imagine that a group of people (20 – 80 years old) are assessed on a number of physiological variables and an estimate of FA is generated. For each age (e.g. 50-year-olds) some people will have a FA older than their CA (i.e. FA > 50) and other people will have a FA younger than their CA (i.e. FA < 50). However, these FA scores should be spread about a mean equal to their CA (i.e. mean FA = 50 years). That is, within each age group, people should be equally likely to have a FA older or younger than their CA. When multiple regression is used to generate a FA estimate, this pattern is seen only in people in the middle of the age range. Older people are biased toward having lower FA scores, which indicates the mean FA for older people is significantly lower than their mean CA. Conversely, younger people are biased toward having higher FA scores. Although some researchers (e.g. Dirken, 1972) made adjustments accordingly, many ignored this problem.

The second criticism was that, *chronological age is an inappropriate criterion*. There is a conceptual problem with using CA as the dependent variable if the goal is to replace CA. If CA is not adequate at differentiating between people (due to increased variability with age) then it is not valid to select those variables that correlated maximally with CA (Brown & Forbes, 1976; Costa & McCrae, 1980; Hochschild, 1990). “If the regression succeeded perfectly, the resultant statistical age would correlate 1.0 with chronological age, and hence would be a perfect, and

perfectly useless, alternative to chronological age" (Costa & McCrae, 1980, p. 32). Furthermore, it does not make sense conceptually to consider CA as a dependent variable. CA is not dependent on anything but time and to model it as dependent on some physiological variable (e.g. blood pressure) is inappropriate. Hochschild (1990) points out that this implies that the predicted-CA from the regression does not make sense and thus equating it with FA does not make any sense either.

The third criticism was that the *interpretation of the results is implausible*. It is assumed that variation in the tasks is due to differences in the rate of aging between people. Costa and McCrae (1980; 1988) argued that this is implausible because individual differences, cohort effects, short-term variation in the measurement and error variance are bound to play a role and are not acknowledged. The assumption that these effects are much smaller than the effect of aging is questionable.

Finally, Costa and McCrae (1980) suggested the greatest shortcoming of FA research is that *validating evidence has not been offered*. Although many researchers produced estimates or profiles of FA, very few made any suggestion about how they could be utilised, let alone tested for their validity. Dean and Morgan (1988) acknowledged that validation studies have been rare but presented some research on predicting mortality, which appeared promising. Furthermore, Anstey et al. (1996), in a review article, found a number of studies that had validated their FA estimates (e.g. Borkan & Norris, 1980b; Furukawa et al., 1975; Webster & Logie, 1976). These studies generally considered two groups of people, who conceptually were expected to exhibit different rates of aging and looked for mean differences in the FA estimate. For example, Webster and Logie (1976) found healthy participants had significantly lower FA than unhealthy participants. Thus, at the time that Costa and McCrae published their paper this was a major problem. However, more recent validation has been provided such that this criticism is not as relevant now.

Considering these major statistical and conceptual criticisms of FA research, Costa and McCrae (1980; 1988) suggested that FA research had no utility and should be abandoned. However, there have been other researchers who were more positive and suggested that the concept needed re-working rather than abandoning. Anstey et al. (1996, p. 246) suggested, "the construct of functional age may be salvaged, but perhaps requires refinement and application to specific functional outcomes".

Biomarker Research

During the 1980s a number of conferences were held on the issue of FA and guidelines were developed for future research in the area (Baker & Sprott, 1988; Reff & Schneider, 1982; Regelson, 1983). This led to a major shift in the terminology used and the way that the issue was conceptualised. Ingram (1991) suggested that this shift was due to a focus on *experimental* biological research. That is, rather than simply measuring FA, people wanted to predict FA in the future, in order to test the effectiveness of intervention programs.

With respect to terminology, Ingram (1983; 1988) argued that the confusion of terminology in FA research has caused much of the controversy associated with the measurement of "biological" age. Anstey et al. (1996) noted the terms physiological, biological and functional age had been used synonymously in FA research and the term used did not necessarily relate to the type of variables operationalised in the study. The term *biomarker of aging* was introduced, with Baker and Sprott (1988, p. 223) defining it as, "a biological parameter of an organism that either alone or in some multivariate composite will, in the absence of disease, better predict functional capability at some late age than will chronological age". To link this back to the earlier FA research, systolic blood pressure or visual acuity could be termed *biomarkers* if they were more effective (than CA) in predicting functional capacity in the future. In addition, a number of papers were published that presented explicit criteria for validating biomarkers (Arking, 1991; Baker & Sprott, 1988; Reff & Schneider, 1982). Although some of the criteria have been criticised (e.g. McClearn, 1997) they are undoubtedly more useful that the original criterion of a correlation with CA.

Criteria for Validating a Biomarker

In 1982, Reff and Schneider published four essential characteristics of a biomarker. A biomarker should (1) be non-lethal in rodents and cause minimal trauma in humans, (2) provide highly reproducible results and reflect physiologic age, (3) display significant alterations during a relatively short time period, and (4) be crucial to the effective maintenance of health and prevention of disease in man. Baker and Sprott (1988) extended these criteria to a total of six requirements and they will be discussed in depth in the following section. Arking (1991) published seven requirements for validating a biomarker of aging. These raised a couple of novel points, which will also be discussed. In addition, a paper by Birren and Fisher (1992) identifies a number of variables that a biomarker should be related to, such as length of life, cognition, and

life-style factors. These requirements are very specific and are a nice adjunct to the theoretical requirements of Baker and Sprott (1988) and will also be presented in the following section.

Baker and Sprott (1988) published six requirements against which a biomarker could be validated, which span biological aging in humans and animal species. First, *the rate of change of a biomarker must, at least in mathematical terms, reflect some measurable parameter, which can be predicted at a later chronological age*. The rate of change in the biomarker (e.g. grip strength) must reflect some measurable parameter (e.g. general muscular strength or frailty), which can be predicted at a later date. Thus, if the rate of change in grip strength is measured then it should be able to predict the general muscular strength of the individual in the future. There is a clear shift, even in the first requirement, to *predictability*. The biomarker must show predictive validity rather than just concurrent validity. Regardless of the relationship between the biomarker and outcomes currently, if the biomarker cannot make predictions for the future then it is not valid.

Second, *the biomarker should reflect some basic biological process of aging and certainly not the predisposition toward a disease state or some inborn error in metabolism.* There are two main issues in this point. The first is that the biomarker must reflect a "biological" process of aging. So, regardless of the nature of the outcome measure (e.g. everyday functioning or cognitive decline), the actual biomarker must be *biological* in nature. This is similar to one of the original criteria from Reff and Schneider (1982), which suggested the biomarker should reflect *physiological* age.

The second point concerns the distinction between normal and diseased aging. *Normal aging* is defined by universal changes that everyone experiences as they get older, such as hearing loss, wrinkled skin, and a reduction in muscle strength. On the other hand, *diseased aging* reflects the fact that certain diseases are more common as people age (e.g. coronary heart disease, diabetes) but they do not represent universal changes because some people do not suffer from them. A biomarker should reflect normal aging or universal changes that all people experience with increasing age, not the predisposition toward age-associated diseases.

Third, *the biomarker should have high reproducibility in cross-species comparisons of functional or physiological age versus chronological age, particularly within the same classes and certainly within the same families of species.* All animal species experience the phenomenon of aging. Therefore, if a biomarker (e.g. grip strength) is marking the aging process in humans then it should do the same in other species, at least to the degree that they are in the same class of

species. Although, an effective biomarker in humans might not generalise to an insect or rodent population it should generalise to non-human primates because they are in the same family of species as us. It is therefore very important to investigate biomarkers in non-human primates rather than the usual laboratory rodents. Ingram, Nakamura, Smucny, Roth, and Lane (2001) have reported that the National Institute of Aging, National Center for Research Resources, and the Wisconsin Regional Primate Research Center are working together to examine biomarkers of aging in a population of rhesus monkeys. These results would be made available to researchers around the world and offer much promise for the cross-species validation of biomarkers of aging.

Another similar point is that a biomarker should have high reproducibility in humans across generations and cultures. If a biomarker is a good indicator of the aging process in one age cohort (or culture) then it should be the case for other cohorts. On the other hand, if the biomarker appears to be useful exclusively for people born in a particular era or location then it is clearly specific rather than universally valid. As mentioned in the preamble, some speed of processing tasks are compromised as biomarkers because they exhibit cohort effects. However, inspection time has been shown to have high reproducibility across generations and thus offers promise as a biomarker, at least on this level.

Fourth, *biomarkers should change independently with the passage of time and reflect physiologic (functional) age.* This statement can be taken two ways, both of which are valid when considering biomarkers. First, an individual biomarker should change independently with the passage of time. This implies that the rate of change of the biomarker should be non-constant or independent of the passage of time (i.e. CA). Early research, that regressed variables on CA made an assumption that the rate of change was linear and constant over time. This statement makes the opposite claim. That is, the rate of change can follow any mathematical curve and furthermore the rate can change over time. Second, it is clear that different systems within an individual decline at different rates. Therefore, different biomarkers should change independently of one another over time. For example, while muscle strength might be declining linearly over time, the visual system might show accelerated decline with time. Finally, the statement about the biomarker reflecting physiologic age has already been addressed in criterion 2.

Fifth, *assessment of biomarkers should be non-lethal in animal systems and should cause minimal trauma in humans. The availability of non-lethal testing in animal model systems would permit longitudinal analyses.* This is taken directly from Reff and Schneider (1982). Although,

number of dendritic connections in the brain may be a reliable biomarker for aging it is not practical to measure this. Measurement of a biomarker must be non-lethal in animals and cause minimal trauma in humans, particularly because multiple measurements (i.e. longitudinal analyses) are necessary to observe the *rate* of aging. Also, stress can produce biased results both in animals and human participants (McClearn, 1997).

Finally, *the biomarker should be reproducible and measurable during a relatively short time interval compared to the life span of the animal.* This is very similar to one of Reff and Schneider's (1982) points but they also emphasised that the biomarker should show significant change over a relatively short time interval. As mentioned in the first criterion, it is the rate of change in the biomarker that is of upmost importance. In order for the biomarker to be of predictive value, the rate of change must be accurately measured during a relatively short time period (i.e. in a way that is highly reliable or reproducible). If it takes most of the lifespan of the animal to measure the rate of change in the biomarker then it will not have much predictive validity. Baker and Sprott (1988) suggested that the most useful biomarkers would be measurable early in life and predict outcomes later in life.

These requirements are extremely useful because they provide a clear basis on which to validate proposed biomarkers. Baker and Sprott (1988, p. 230) suggested that "the basic criteria of a valid biomarker are that it will directly relate to functional age better than chronological age and will in some manner predict longevity". However, setting out the specific requirements allows one to define hypotheses against which to test the validity of biomarkers. To summarise, a biomarker must be have predictive validity, be biological in nature and reflect normal aging, show high reproducibility in cross-species comparisons and across generations and cultures in humans, change independently with the passage of time, be non-lethal to animals and minimally traumatic in humans, and exhibit reliable change over a relatively short period.

Arking (1991) published an alternate set of characteristics for a biomarker. These were based on Reff and Schneider (1982) and Baker and Sprott (1988), and so there is considerable overlap. However, two points not emphasised by the other researchers are worth discussing. The first point pertains to the issue of normal vs. diseased aging. The second point concerns the ability of a biomarker to predict lifespan or longevity.

One of the criteria listed by Arking (1991) is that *a biomarker should be crucial to maintenance of health.* In other words, people who show marked decline in the biomarker should be at risk of health problems. In so far as some diseases are more prevalent as people get older

this point is valid. If people are aging rapidly, they are approaching a point where health-related problems are more likely to occur. However, as mentioned earlier, although some diseases are more prevalent with age, not all people suffer them and they cannot be considered a part of normal aging. McClearn (1997) noted, however, that there is a wealth of literature on the concept of successful aging and therefore it is not valid for all biomarkers to be related to health. Some people live to old age without any serious health problems, yet they clearly age. Thus, a biomarker may be able to mark the aging process without being related to health outcomes. This point illustrates the difficulty in aging research to distinguish changes associated with normal aging from disease processes.

Arking (1991) has indicated that *a biomarker should serve as a predictor of lifespan and/or a retrospective marker of aging.* That is, a biomarker must be able to predict mortality or longevity. Although Baker and Sprott (1988) briefly mentioned this point, it is important and warrants further discussion. Many researchers have suggested that a major requirement of a biomarker is that it must be able to predict mortality (Birren & Fisher, 1992; Brown & Forbes, 1976; Bulpitt, 1995; Hochschild, 1990; Ingram, 1991). Baker and Sprott (1988) noted that there are three variables with respect to a biomarker: initial level, onset of decline and rate of decline. They also suggested that the rate of decline can ultimately be the most critical of the three variables. That is, the level of the biomarker at the testing session may not be as important as the rate of change of the biomarker over time. Both Baker and Sprott (1988) and Arking (1991) emphasised the rate of decline in the biomarker as being most important. Theoretically, the rate of decline in the biomarker should reflect the rate of aging (Arking, 1991) and it follows that the rate of decline in the biomarker should predict mortality. People who are declining quickly are theorised to be aging more rapidly and therefore approaching death more quickly. Thus, many researchers have emphasised that *the rate of decline in the biomarker should predict mortality.*

As mentioned earlier, these requirements are theoretical and would be strengthened with the addition of more specific and testable guidelines. These clear, testable hypotheses have been provided by Birren and Fisher (1992, p. 12 - 13). The first requirement is *a biomarker should be related to length of life*. This requirement is discussed above and appears to be one of the most central requirements across sets of criteria for biomarkers.

The second requirement is that *adjacent phylogenetic species should show changes in the biomarker with age*. *Adjacent phylogenetic species* are those that have evolved from the same node in a phylogenetic tree. With respect to humans, other primate species are considered

adjacent phylogenetically. This requirement is reflected in the criteria of both Reff and Schneider (1982) and Baker and Sprott (1988), although in a more general, cross-species way.

The third requirement is related to gender differences. That is, *since females have a longer lifespan than males, greater change in the biomarker should be seen in older males than in older females*. Many researchers report gender differences in level of sensorimotor (e.g. grip strength), cognitive (e.g. reasoning) and biological variables (e.g. blood pressure, see Anstey et al., 1996). However, this requirement refers to "rate of change" in the biomarker rather than level and for that reason is novel. Indirectly, it suggests that the biomarker should be related to mortality because females live longer than males.

The fourth requirement is that the *biomarker should correlate with physiological and anatomical indicators of aging (e.g. lung vital capacity, skin elasticity, bone mass, muscular strength, maximum heart rate, hearing threshold, glucose tolerance, measures of brain excitability, and brain metabolism).* This suggests that the biomarker should be related to physiological indicators of normal aging. For example, it is known that lung capacity declines as people age and if a biomarker is marking the aging process then it should correlate with lung capacity. However, there is a conceptual problem with this argument. It assumes that those physiological variables that decline with age are indicators of the aging process. This essentially is the idea behind FA research, which has been shown to be problematic. A correlation with CA does not establish that these variables (e.g. lung capacity) are indicators of physiological aging. Therefore, it may not be necessary for a biomarker to be correlated with them.

The fifth requirement is that the *biomarker should correlate with behavioural processes (e.g. attention, perception, memory, problem solving and reasoning*). The biomarker should be related to behavioural (or cognitive) indicators of normal aging. Since fluid ability declines with age, a valid biomarker should be related to the amount of decline in this outcome measure. The conceptual problems related to the fourth requirement are applicable here too.

The sixth requirement is that the *biomarker should be reduced but not eliminated by exercise, proper diet, not smoking and moderate use of alcohol*. This item refers to the effect of modifiable life-style factors on the rate of aging. Some life-style factors are known to increase (e.g. smoking) or decrease (e.g. exercise) the risk of mortality. Thus, a biomarker should be affected by such life-style factors. This requirement is worded in a positive way, suggesting that interventions to improve lifestyle should slow down the rate of aging.

The final requirement is that decline in a *biomarker should be exacerbated by the presence of age-associated diseases such as coronary artery disease, cerebrovascular disease, diabetes, and Alzheimer's disease*. These diseases are indicative of diseased aging rather than normal aging. Thus, this requirement suggests that these diseases produce accelerated aging, which should be reflected in the level or rate of change of the biomarker.

To summarise, the first two requirements are very similar to those of Reff and Schneider (1982) and Baker and Sprott (1988), whereas the subsequent five requirements are very specific. A biomarker should predict mortality, show changes with age in non-human primates, show gender differences in rate of change, relate to indicators of normal aging (physiological and cognitive), be reduced by positive life-style factors and be exacerbated by age-associated disease. Considering all three sets of validation criteria, it is possible to provide clearly testable hypotheses for any purported biomarker of aging.

Models for Validating Biomarkers

In addition to the criteria for validating biomarkers, two papers have presented models for validating biomarkers. Ingram (1991) presented three models, which he called the *ex post facto*, *ipso facto*, and *ad hoc* models. Each model is used in a different situation to evaluate a purported biomarker. Birren and Fisher (1992) discussed the evaluation of biomarkers at four distinct levels of importance, which they referred to as *general, superordinate, coordinate* and *subordinate*. These models are usefully added to the criteria outlined above, and will be discussed below.

Ingram (1991) suggested four main criteria for validating a biomarker. The biomarker should be able to predict: (1) *future performance (long term) including the rate of decline in test performance*, (2) *ability to withstand a specific stress or toxin*, (3) *chronic disease onset*, or (4) *life span*. These are not unlike criteria presented above. However, Ingram went a step further by detailing three models for validating biomarkers, which he called the *ex post facto, ipso facto*, and *ad hoc* models.

The *ex post facto* model is a non-experimental model, so the group is observed rather than applying any experimental manipulation. The biomarker is not validated based on its correlation with CA. Thus, Ingram (1991) proposed that the biomarker should not correlate > 0.7 with CA (levels widely accepted as indicating strong correlation) and, indeed, a correlation with age is not even necessary. Rather, the biomarker is evaluated in term of its predictive validity for future performance. Both future test performance and rate of decline on a test are considered valid

outcome measures. For example, if cognitive functioning is of concern, one could see if a biomarker (e.g. visual acuity) is predictive of cognitive level (Raven's Standard Progressive Matrices) in the future or cognitive decline between initial and final measurement.

The *ipso facto* model involves measuring the biomarker, splitting the group into an experimental and control group, applying some intervention and finally observing differences between groups. In this model, it is necessary for the biomarker to correlate with CA. Because change over time between the two groups is observed, it is necessary for the biomarker to change with age. If the intervention is effective, then the biomarker should change more/less in the experimental group than in the control group.

The *ad hoc* model can be applied when outcomes other than lifespan are of principal importance (Ingram, 1991). In some cases, a biomarker may not predict lifespan but might affect health or quality of life. In a nursing home situation, improving quality of life may be more important than longevity. The design is the same as the ipso facto model (i.e. experimental vs. control group) but the outcome is of a different nature. In this case the biomarker does not have to be indicative of aging. Rather, it has to predict outcomes (e.g. quality of life) that are considered to be part of healthy aging. Ingram (1991) suggested the outcome measure cannot be validated empirically, rather the content validity of the test should be evaluated by the gerontological community.

These three models provide a clear guide to validating any biomarker of aging; it is simply necessary to consider what functional outcomes one is interested in, such as cognition, everyday activities, health, quality of life, or mortality. In the initial stages, the *ex-post facto* model could be used to investigate the utility of the biomarker. If the results are encouraging, one might conduct further experiments using one of the two experimental models.

Birren and Fisher (1992) presented a model that detailed four levels at which a biomarker can be evaluated. The first level, *general*, involves assessing whether a biomarker is predictive of longevity or mortality. That is, the most general question with regards to biomarkers is whether they can differentiate between those people who will die and those who survive within a specified future time frame. This point was made in all three criteria for validation (Arking, 1991; Baker & Sprott, 1988; Reff & Schneider, 1982) and by many other researchers. It certainly appears to be one of the most important issues with respect to biomarkers of FA and is considered the most important level of evaluation in this model.

The second level, *superordinate*, is concerned with functioning in everyday life. Birren and Fisher (1992) suggested that functioning in everyday life should be operationalised by instrumental activities of daily living (IADL). Therefore, if a biomarker is effective at the superordinate level, it must be related to IADL.

The third level, *coordinate*, refers to the relationship between the biomarker and cognitive functioning. Cognitive functioning is known to decline with age, thus a biomarker should be able to predict future decline in test performance on cognitive tasks. This point was made explicitly by Birren and Fisher (1992) in their validation criteria. Furthermore, Ingram (1991) suggested that both future test performance and decline in test performance are suitable outcome measures for biomarker validation in the *ex post facto* model.

The final level, *subordinate*, concerns the physiological changes that occur with age. A biomarker should be related to psychophysiological and physiological processes (e.g. visual acuity, grip strength) that accompany normal aging. One criterion common to both Baker and Sprott (1988) and Arking (1991) is that biomarkers are assumed to be biological or physiological in nature. Thus, biomarkers could be thought of at the same level as physiological processes.

Birren and Fisher (1992) suggested that validation of a biomarker at the general level is most important, followed by the superordinate, coordinate and subordinate levels. In practice, it may be useful to work through the levels from least important to most important. That is, show the biomarker is related to physiological processes first, then cognition, then everyday functioning and finally mortality. It is quite likely that some biomarkers may be valid at the subordinate and coordinate levels but fail at the superordinate or general level; thus, offering further support for the notion of working through from least important to most important.

Next Steps in Biomarker Research

Hochschild (1990) noted that one of the fundamental problems in biomarker research is that different authors use different validation criteria and thus generate somewhat disparate findings. Reviewing both the major validation criteria and models of biomarkers has here generated a unified set of concepts that can be used for future validation studies. The degree of overlap among suggested criteria is quite significant, leading to a good balance of theoretical and more specific testable hypotheses. Furthermore, the models provide a clearly defined method for future validation studies. In order to test a purported biomarker, the following steps are tentatively recommended.

First, the biomarker should be considered theoretically to see if it meets the requirements for a valid biomarker. The biomarker must be shown to (1) be biological in nature, (2) reflect normal aging, (3) have high reliability, (4) show stability across generations and cultures in humans, (5) change independently with the passage of time, (6) be non-lethal to animals and minimally traumatic in humans, and (7) exhibit reliable change over a relatively short period. Second, a literature review should examine the biomarker empirically with respect to the seven specific hypotheses of Birren and Fisher (1992). That is, the biomarker should relate to length of life, show changes with age in non-human primates, show gender differences in rate of change, relate to indicators of normal aging (physiological and cognitive), be reduced by positive life-style factors and be exacerbated by age-associated disease. Third, if the results are encouraging at this point, the *ex post facto* model (Ingram, 1991) should be utilised to evaluate the biomarker. That is, a group of participants should be observed initially before any experimental manipulations are made. The biomarker can be examined at all or some of the levels of importance as defined by Birren and Fisher (1992). That is, the biomarker can be measured along with tests of physiological processes, cognition, and everyday living. If the study is prospective, longitudinal change and the predictive validity of the biomarker for mortality can also be examined. Finally, if these results are positive then the *ipso facto* model could be used to evaluate interventions, such as exercise programs or dietary change. This explicit method for evaluating biomarkers should provide a clear means to validate any purported biological marker of aging.

CHAPTER TWO: INTELLIGENCE AND SPEED OF PROCESSING

Intelligence

Theories of Psychometric Intelligence

Although speed of processing is the main focus of this chapter it is useful to view it within a larger framework of intelligence theory because speed of processing is just one of a number of cognitive abilities. In the study of psychometric intelligence, there currently exists one prominent theory of the structure of cognitive abilities. *Gf-Gc* theory (Horn, 1988, 1989, 1990; Horn & Cattell, 1967) is a structural model that was derived from primary mental abilities theory (Thurstone, 1938, 1947) and incorporates aspects of Guilford's theory (Horn & Noll, 1994). In order to give some insight into the development of this model, a brief summary of early models of intelligence will be presented, followed by a detailed description of Gf-Gc theory.

Early Models of Intelligence

Spearman (1904; 1927) in a review of intelligence research, suggested that one basic function, '*g*', was responsible for performance on a range of intellectual tests. His theory, derived from his seminal work in the development of factor analysis, proposed that there was one common factor (*g*) and all relationships between the individual tasks were due to their relationship with this factor. This theory was soon challenged by subsequent research that found that this model did not account for individual differences in test performance adequately. Thus, Burt (1909; 1911) found that two group factors (verbal and numerical), in addition to the general factor, were required to explain the variance in test performance. A further eight group factors (memory span, manipulative ability, scholastic ability, spatial, perceptual speed, mechanical reasoning and visualisation) were added in the 1920s (Horn & Noll, 1994). This model and others provided support for the notion that a single factor model did not provide a sufficient explanation for human intelligence and, instead, group factors were also necessary to explain the inter-correlations between test scores.

Thurstone (1938; 1947) proposed that at least nine common factors were required to account for individual differences in performance on a battery of tests relevant to academic achievement, and he termed them the *primary mental abilities* (PMA). These abilities were Inductive Reasoning, Deductive Reasoning, Practical Problem Reasoning, Verbal Comprehension, Associative Short-Term Memory, Spatial Relations, Perceptual Speed,

Numerical Facility and Word Fluency. Subsequent studies extended Thurstone's theory, uncovering between 28 and 40 primary abilities, which led to a somewhat cumbersome theory. The primary mental abilities needed to be combined into a more manageable theory with a smaller number of group factors.

Gf-Gc Theory

Gf-Gc theory has been described as "a second-order system for the PMA factors", (Horn & Noll, 1994, p. 172). That is, the PMA factors are considered first-order factors and postulated to load on a smaller number of second-order factors. In this sense, Gf-Gc theory is a hierarchical model of intelligence. The development of Gf-Gc theory is based on the early work of Horn and Cattell (1967) and their subsequent follow up studies. Horn and Noll (1994) have also acknowledged the contribution of more recent large-scale factor analytic studies by Carroll (1989), Gustafsson (1984), Undheim and Gustafsson (1987) and Woodcock (1990).

This theory postulates nine second-order ability factors: *Fluid reasoning (Gf)*, *Crystallised ability* (*Gc*), *Quantitative knowledge (Gq)*, *Short-term memory (Gsm*), *Long-term memory (Glr*), *Visual Processing (Gv)*, *Auditory Processing (Ga)*, *Processing Speed (Gs)*, and *Correct Decision Speed (CDS)*. Each of these factors is marked by a number of Thurstone's primary mental abilities and represents a different aspect of intelligence. No third-order general factor is included to represent Spearman's *g*. Horn and his followers have always maintained that individual differences in test performance are adequately explained by just the first and second-order factors. Horn and Noll (1994) described Gf-Gc theory as a theory of several intelligences rather than a theory of intelligence. However, it should be noted that others, most notably Carroll (1993), argue for the existence of a higher order *g* factor, that exists at the third stratum and represents general intelligence.

Woodcock (1990) and McGrew (1997) analysed the major intelligence batteries and located each subtest within Gf-Gc theory. Furthermore, the Woodcock-Johnson Psycho-Educational Battery (WJ-R, Woodcock & Johnson, 1989), a prominent intelligence battery, was developed based upon Gf-Gc theory. Each of the ability factors (except CDS) is defined by two subtests in the battery. Therefore, a clear model of intelligence and the actual tests that measure each construct are available. In the following sections each of the nine abilities in Gf-Gc theory will be described, including examples of tests that mark that ability.

Fluid reasoning (Gf) is defined as the ability to solve novel problems and deal with information not previously seen. Because these tests do not depend on previous knowledge they

are sometimes thought of as being "culture fair". Horn (1990) suggested that Gf involved many mental operations such as identifying relations, drawing inferences, concept formation, concept recognition, identifying conjunctions, and recognising disjunctions. Tests that measure *Gf* include Analysis-Synthesis and Concept Formation (WJ-R), Raven's Standard Progressive Matrices (de Lemos, 1995), and sections of the Cattell Culture Fair Test (Cattell & Cattell, 1959).

Crystallised ability (*Gc*) is defined as the degree of cultural knowledge an individual has acquired. More specifically, Gc incorporates general knowledge, vocabulary, verbal analogies, and problem definitions (Horn, 1990). Tests that measure Gc include Oral Vocabulary and Picture Vocabulary (WJ-R), and a number of subtests from the verbal scale of the Wechsler Adult Intelligence Scale – Revised (WAIS-R, Wechsler, 1981) including Information, Similarities, Vocabulary and Comprehension.

Quantitative knowledge (*Gq*) is defined as the knowledge and application of mathematical concepts and skills. This is the most recent addition to Gf-Gc theory, stemming from studies done in the 1980s. Horn (1990) suggested that test variance associated with quantitative skills is distinct from knowledge associated with Gc. Gq involves both basic calculations and more complex applied problems and can be assessed by Calculation and Applied Problems (WJ-R), and Quantitative, Equation Building and Number Series from the fourth edition of the Stanford-Binet Intelligence Scale (SB-IV, Thorndike, Hagen, & Sattler, 1986).

Short-term memory (Gsm) requires the encoding and recall of information over a relatively short period of time (i.e. a couple of minutes). Both verbal and non-verbal stimuli are used in the assessment of *Gsm*. Tests that measure Gsm include Memory for Words and Memory for Sentences (WJ-R), Digit span (WAIS-R), and Memory for Objects (SB-IV).

Long-term memory (Glr) is defined by the ability to retrieve information stored minutes, hours, weeks and years earlier. Some tests (e.g. from WJ-R) use the same stimuli from the Gsm tests but retrieval is performed after a longer interval of time. Other tests require the retrieval of expressions, ideas or words from long-term memory. Tests of *Glr* include Memory for Names and Visual-Auditory Learning (WJ-R), and Rebus Learning from the Kaufman Adolescent and Adult Intelligence Test (Kaufman & Kaufman, 1993).

Visual Processing (Gv) is defined as the ability to visualise objects in space as they are being manipulated. Gv involves visual scanning, Gestalt closure, mind's-eye rotations, spatial orientation, flexibility of closure, and length estimation (Horn, 1990). Tests of Gv include

Visual Closure and Picture Recognition (WJ-R), Block Design and Object Assembly (WAIS-R), and Pattern Analysis and Copying (SB-IV).

Auditory Processing (Ga) is defined by the ability to perceive sounds and the relationship between sounds under different conditions (e.g. distractibility). *Ga* involves perception of sound patterns, awareness of order and rhythm, and the comprehension of groups of sounds such as chords (Horn, 1990). Sound Blending, Incomplete Words, and Word Attack (WJ-R) all provide strong measures of Ga (McGrew, 1997).

Processing Speed (Gs) is defined as the speed of scanning and responding to simple tasks, that everyone could get correct, given enough time. This ability is thought to play a role in all other cognitive ability factors and has been referred to as perceptual speed by other researchers. Tests of Gs include Cross Out and Visual Matching (WJ-R) and Digit Symbol (WAIS-III: Wechsler, 1997).

Correct Decision Speed (*CDS*) is defined as the speed at performing more complex cognitive tasks. There is evidence that CDS and Gs do not correlate highly, confirming that they represent separate constructs (Horn, 1988). Often, the tests used to measure other abilities (e.g. Gf, Gc, and Gv) are utilised to measure CDS. That is, a Gf test could be administered (e.g. Raven's Progressive Matrices) and the speed at providing correct answers, rather than number correct, would be a measure of CDS.

Age trends in Gf-Gc factors

In general, the Gf-Gc ability factors can be divided into those that are "vulnerable" to age-related decline particularly during old age and those that are "maintained" throughout the lifespan. Vulnerable abilities show decline from early adulthood and include Gf, Gsm, Gv, Gs and CDS. Conversely, maintained abilities show stability and sometimes improvement with age and include Gc, Glr and Gq. The evidence with respect to Ga is less clear (Horn, 1990; Schaie, 1994). The following section will discuss the age trends with respect to Gf, Gc and *Gs*.

Fluid intelligence (Gf). The term "fluid" is used because these abilities are changing or fluid throughout the lifespan. Cross-sectional research suggests that Gf peaks as early as the late 20s and shows accelerated linear decline thereafter (Horn, 1990; Schaie, 1994). However, longitudinal studies have suggested that Gf might actually be maintained until much later. Schaie (1994) found that, longitudinally, Gf peaked at about 50 years of age and declined thereafter. This suggests that cross-sectional studies might be overestimating the decline of *Gf* with age. A possible explanation for this phenomenon is the Flynn effect (Flynn, 1987, 1999);

i.e. improvement in scores on cognitive tasks over successive generations. This implies that the true decline with age, as seen in longitudinal studies, might be inflated in cross-sectional research by cohort effects because more recent cohorts perform more effectively. Schaie (1994) provided support for this proposition, demonstrating positive cohort effects for inductive reasoning, a first-order *Gf* factor. On the other hand, longitudinal investigation might underestimate decline, because of selective survival or dropout, which means that both methods are probably necessary at this stage.

Crystallised ability (Gc). The term "crystallised" is used for abilities that are preserved or crystallised into old age. In fact elderly people often perform slightly better on crystallised tests of ability that younger adults (Horn & Cattell, 1967). Schaie (1994) found that, compared with age 25, at age 88 there was virtually no difference in verbal abilities. However, there is evidence that when verbal abilities do begin to decline, they decline quite sharply (Schaie, 1994). Given that crystallized abilities are so stable, the point of decline has been postulated to indicate impending death (e.g. Cooney, Schaie, & Willis, 1988).

Processing Speed (Gs). This ability has received a lot of attention in the aging literature because it displays the largest decline of any ability with age. In a meta-analysis, Verhaeghen and Salthouse (1997) found that speed was more strongly related to age (r = -.52) than was any other ability (including Gf, Gv, and Gsm). Kail and Salthouse (1994) examined the two WJ-R speed tests (Cross Out and Visual Matching) cross-sectionally and found that, with performance standardised, the decline from age 25 to age 75 was equivalent to nearly 2 standard deviations. Schaie (1994) found that although a number of abilities showed early decline cross-sectionally, Gs was the only ability to show longitudinal decline from age 25 onwards. In fact, Schaie (1989) suggested that, in contrast to the general finding that cross-sectional studies overestimate the extent of age related change in cognitive abilities, the cross-sectional analysis had actually underestimated the amount of decline in Gs with age. In a review of age and processing speed, Salthouse (2000b, p. 38) concluded that, "speed variables are among the biological and behavioural variables with the strongest relations to age".

Speed of Processing

Processing-Speed Theory

Large declines in speed of processing with age have lead to the suggestion that slowing speed of processing might be a mechanism to explain decline in a range of cognitive abilities

with age (Birren, 1965, 1974; Salthouse, 1985, 1993, 1996). That is, it may be that many cognitive abilities decline with age largely because people are slower at the individual components of the tasks. Salthouse (1996) proposed and provided empirical support for three major hypotheses concerning this processing-speed theory. Firstly, decline in a range of different speed measures should share a large amount of common variance. Second, processing speed should act as a mediator between age and cognition. That is, statistical control of speed measures should substantially reduce the age-cognition correlation. Third, two mechanisms (limited time and simultaneity) are primarily responsible for the relations between speed and cognition.

The first hypothesis concerns the relationship between measures of processing speed. There are a number of different types of processing speed measures, with Salthouse (2000a) describing six types: *decision speed* (similar to CDS in Gf-Gc theory), *perceptual speed, psychomotor speed, reaction time, psychophysiological speed* and *time course of internal response* (e.g. Event Related Potentials). However, most research in this area concerns perceptual speed, psychomotor speed and reaction time. There exists a lot of evidence that the age-related variance on these speed measures is to a large degree shared or common. For example, Salthouse (1993) administered 11 paper and pencil speed measures to 305 adults. For each pair of tests, the proportion of shared age-related variance was estimated, with a median value of .842. That is, on average the tests shared 71% age-related variance, which is consistent with the hypothesis of a general speed factor.

The second hypothesis, that speed mediates the age-cognition relationship, has also received a lot of empirical support. Many studies have found that the age-related variance in a range of cognitive tasks can be explained by simple measures of processing speed (Babcock, 1994; Bryan & Luszcz, 1996; Lindenberger, Mayr, & Kliegl, 1993; Nettelbeck & Rabbitt, 1992; Salthouse, 1991, 1993, 1996; Salthouse & Babcock, 1991; Salthouse, Hambrick, & McGuthry, 1998; Verhaeghen & Salthouse, 1997). Speed has been shown to mediate the age effects in Gf (e.g. Babcock, 1994), Gv (e.g. Nettelbeck & Rabbitt, 1992), and Gsm (e.g. Salthouse, 1993). In a comprehensive review of the topic, Salthouse (1996, p. 420) concluded “an average of 75% or more of the age-related variance in a wide range of memory and cognitive variables is shared with measures of processing speed”.

Salthouse (1996) proposed two mechanisms to explain the relationship between speed and cognition: *Limited time* and *simultaneity mechanisms*. The *limited time mechanism* suggests that cognitive operations are performed too slowly to complete the overall task in the time required.

It is suggested that this mechanism is likely to be relevant when time limits are imposed or there are other restrictions on time available for processing. Kersten and Salthouse (1993) had a group of young (*M* = 20.5 years) and older (*M* = 67.9 years) adults complete an associate memory task. They manipulated the amount of time available to view the stimuli and found that older people needed considerably longer than younger people in order to achieve the same level of accuracy. They concluded that older people complete less processing in a set period of time than younger people, hence lending support to the *limited time mechanism*. However, Salthouse (1996) and others have also shown that the correlation between speed and cognition is apparent in situations where no time limit is imposed. Thus, this mechanism alone cannot account for the speed-cognition relationship.

The *simultaneity mechanism* suggests that slow processing speed reduces the amount of information that can be used in higher order processing. That is, the information from earlier processing may be lost or be no longer valid by the time later processing is done. Salthouse (1996) draws on the concept of working memory (WM) to evaluate the simultaneity mechanism because these tasks require storage of information for later processing. Verhaeghen and Salthouse (1997) performed a meta-analysis on the relationship between age and a number of ability factors including WM. They generated a structural equation model where the relationship between WM and age was mediated by processing speed and found that speed mediated 92.5% of the age-related variance in WM. This study and those like it offer support for the *simultaneity mechanism* in explaining the relationship between age and cognition.

The processing speed theory has generated massive interest in the gerontological literature. One reason is that speed measures are extremely simple to assess and often take a minimal amount of time and resources. Perceptual speed tasks, such as Digit Symbol and Visual Matching, are generally performed with paper and pencil and take less than 5 minutes to complete. This suggests that a 5-minute test can provide valuable information about the degree of age-related decline in cognitive performance. To summarise, processing speed is a simple, non-invasive test that shows considerable age-related decline and is predictive of changes in a wide range of cognitive outcomes.

Speed of Processing as a Biomarker

Given the strong relationship between speed of processing and cognitive decline, some researchers have suggested that speed of processing might provide a valid biomarker for functional age. Birren (1965, p. 289) suggested that, “age-related changes in perceptual speed

may be a *primary marker of central nervous system aging*, reflecting an adaptive capacity to resist the cumulative effects of disease". In one of the earliest biomarker publications by Reff and Schneider, Salthouse (1982) wrote a chapter on speed of processing measures as biomarkers. Speed of processing has been linked to *mortality* (e.g. Bosworth & Schaie, 1999), *everyday functioning* (e.g. Fleischmann, 1994), *cognition*, and *physiological functioning* (e.g. Lindenberger & Baltes, 1997). It is also related to *life-style factors* such as exercise (e.g. Bashore, 1989) and to *age-associated diseases* such as Alzheimer's disease (e.g. Deary, Hunter, Langan, & Goodwin, 1991). The prospects for speed of processing as a biomarker appear quite positive. However, there are a number of problems with the currently used speed of processing measures for use as biomarkers. The following section will detail these problems as they pertain to the criteria for biomarkers (see Chapter 1).

Traditional Speed Measures

Most frequently, studies that have examined speed as a biomarker have focused on perceptual speed and Reaction Time (RT) tasks. Both types of tasks are valid for use in cognitive aging research but somewhat problematic for use as biomarkers. The problems pertain to two particular criteria for biomarkers. First, there is some doubt over the degree to which these tasks reflect a basic biological process of aging. Second, the reproducibility of these tasks over generations is questionable.

For speed of processing measures to be valid biomarkers they *must reflect a basic biological process*. That is, the level of explanation needs to be reduced from cognitive to biological. With all speed of processing measures, there is a belief that they are tapping some aspect of speed or efficiency of the central nervous system (CNS). Madden (2001, p. 288) stated, "speed is often viewed not only as a behavioural measure but also as a fundamental property of the central nervous system". This implies that superior performance on speed of processing tasks is indicative of a CNS where neural impulses are transmitted quickly and efficiently. However, the real issue here is how well do perceptual speed and RT measure this neural efficiency?

There are two main concerns with using perceptual speed and RT as proxies for neural efficiency. First, both measures contain a large psychomotor element. With RT, much of the measure is determined by the speed at which people physically operate the response button. As for perceptual speed, the measure is largely determined by the speed at which people write their answers. Clearly, motor capacities are likely to deteriorate as people age, thus perceptual speed and RT confound true CNS slowing with peripheral changes in physical response. Second, these

measures inevitably confound accuracy and speed of responding. In effect, there are individual differences in a tendency to "trade" accuracy against speed that is difficult to control. Furthermore, these individual differences are known to be affected by age such that older people emphasise accuracy more than younger people (Welford, 1977). Thus, older people may be slower not only because their CNS is slowing but also because they are more concerned with accurate responding.

The other area of concern for these measures is their ability to show *high reproducibility across generations*. Essentially, the point here is that if a biomarker is a good indicator of the aging process in one age cohort then it should be the case for other cohorts. Perceptual speed tasks, such as Digit Symbol, are known to be sensitive to generational effects; that is, average performance on this test has been shown to improve across successive generations (see Wicherts et al., 2004). Thus, although there is considerable evidence for a real decline in perceptual speed with age, this may be exaggerated in cross-sectional studies that have used the Digit Symbol test because the better average performances of participants within later born cohorts will also have been improved by the Flynn effect. In other words, younger cohorts tend to perform better on this test and this would result in inferences of larger effects than really exist.

These findings indicate that perceptual speed and RT tasks are not ideal for use as biomarkers because they incorporate psychomotor speed, confound accuracy and speed of response, and perceptual speed tests, at least, show cohort effects. If possible, a speed measure that was free from these particular problems would be more suitable. That is, a measure of speed that is stable across generations, free from psychomotor confounding and does not lead to an speed-accuracy trade off, would be more suitable as a biomarker, at least as defined by generally accepted criteria.

Alternative Speed Measures

As mentioned earlier, Salthouse (2000a) suggested that there are six types of speed measures: *decision speed*, *perceptual speed*, *psychomotor speed*, *reaction time*, *psychophysical speed* and *time course of internal responses* (e.g. Event Related Potentials). *Decision speed* is considered analogous to Horn's CDS and is derived from the speed of answering items in moderately complex cognitive tasks. Given that this is cognitively complex (more so than perceptual speed) it is not likely to be measuring neural efficiency very directly. Furthermore, the level measures (e.g. Gf) from which decision speed is derived show cohort effects, and it is possible that decision speed does too. The problems with *perceptual speed* and *RT* have already

been discussed and the problems with *psychomotor speed* have been mentioned, which leaves *psychophysical speed* and *time course of internal processes*.

Psychophysical speed involves briefly presenting a target (usually visual or auditory) on which a simple decision is made. It is possible to calculate how long the target needs to be presented in order for the participant to make correct decisions with high accuracy. In this way, *psychophysical speed* is an inferred speed measure and hence not confounded by speed of physical movement to a response key. Furthermore, the nature of the task means that the participant does not have to trade accuracy for speed. Response speed is not important so the participant can focus solely on accuracy. Thus, a *psychophysical speed* measure that can be shown to be stable across generations would be very promising as a valid biomarker.

Time course of internal processes (Salthouse, 2000a) refers to components of event related potentials (ERP). In order to measure ERPs, a participant is fitted with electrodes that monitor the electrical responses of the brain to images presented on a screen. The same image (e.g. checkerboard pattern) is flashed repeatedly and the electrical responses are measured and then averaged to generate a waveform. This waveforms shows when the neuron or group of neurons fired (i.e. latency) and how large the response was (i.e. amplitude). In the current context, it would be the latency of particular parts of the waveform that would be operationalised. Conceptually, it could certainly be argued that this is a direct way to measure neural efficiency. However, the problem with this type of speed measure is the complex and time-consuming nature of its measurement. Sophisticated equipment is required, a trained expert is needed to measure the ERP, application of the electrodes and analysis of the waveform are time consuming, and the latency variables have questionable reliability. Although conceptually this type of speed might seem ideal, in a practical sense it is not.

To summarise, of the six types of speed measures proposed by Salthouse (2000a), the most promising for a biomarker of functional age is *psychophysical speed*. It is simple to measure, highly reliable, free from psychomotor speed, and does not lead to the confounding of accuracy and speed. If a measure of *psychophysical speed* can be found that is free from generational effects, this offers the best prospects for a speed of processing measure as biomarker of functional age.

Inspection Time

Inspection time (IT) is a construct from psychophysiology that stemmed from Vickers' accumulator model (see Vickers, Nettelbeck, & Willson, 1972). This model postulates that

people make a series of observations of sensory information on which to make a decision. When enough information is accumulated to satisfy some criterion a decision will be made. Vickers et al. suggested that in some cases a single observation might provide adequate information and they were interested in how long such an observation might take. The term *Inspection Time* was introduced and defined as the "time required by a subject to make a single observation or inspection of the sensory input on which a decision of relative magnitude is based" (Vickers & Smith, 1986, p. 609).

In order to estimate IT, a task was needed where just a single observation was adequate to make a decision. Vickers et al. developed a discrimination task, where two lines of markedly different lengths were presented side-by-side and the participant had to indicate the shorter (or longer) line. The discrimination was so simple that given enough time anyone would be able to make a correct decision. However, the IT task involved presenting the stimulus for a brief period immediately followed by a backward mask. A backward mask is simply an image that completely covers the stimulus, the purpose of which is to restrict further accumulation of information from stored visual traces (Vickers et al., 1972). By presenting the task at a series of different durations (or stimulus onset asynchronies) it is possible to estimate an individual's IT.

Before any experiments were done, Vickers et al. (1972) hypothesised that IT would be about 100 ms and would probably not vary much between people. The first experiment used the method of constant stimuli to estimate IT and an accuracy level of 95% was set. That is, they desired to know how long people needed to see the stimulus in order to make a correct decision 95% of the time. Ten psychology students completed the experiment and the average IT was 105 ms, quite close to expectations. However, these estimates ranged from 74 to 144 ms suggesting there was significant variation in IT between people. Vickers et al. therefore concluded that IT might provide a useful index of individual differences in speed of perception.

It is proposed that IT might be a suitable psychophysical task for use as a biomarker for a number of reasons. First, IT has been studied extensively for over 30 years and is related to performance on range of different cognitive tasks including fluid reasoning, visualisation, perceptual speed and omnibus IQ measures. Second, IT is easy to measure and does not suffer the problems of other speed of processing tasks such as perceptual speed and reaction time tasks. Third, IT has recently been shown to be stable across generations (Nettelbeck & Wilson, 2004). In summary, IT is a measure of *psychophysical speed* that has been studied extensively, is easy to measure, and is free from generational effects. Thus, it is proposed that IT may be a useful

biological marker of aging and may act as a lead indicator for decline in cognitive abilities with advancing age.

In Chapter 1 it was suggested that a biomarker must be assessed both theoretically and empirically to see whether it meets the validation criteria. In Chapter 3, the validity of IT will be considered. First, the evidence will be evaluated as to whether IT (1) is biological in nature, (2) reflects normal aging, (3) has high reliability, (4) shows stability across generations in humans, (5) changes independently with the passage of time, (6) is non-lethal to animals and minimally traumatic in humans, and (7) exhibits reliable change over a relatively short period. Second, the literature will be reviewed to see whether IT is (1) related to length of life, (2) shows changes with age in non-human primates, (3) shows gender differences in rate of change, (4) relates to indicators of normal aging (physiological and cognitive), (5) is reduced by positive life-style factors, and (6) is exacerbated by age-associated diseases. Finally, these results will be summarised and a decision will be made about applying the *ex post facto model* to further examine IT as a biomarker.

CHAPTER THREE: VALIDATION OF INSPECTION TIME AS A BIOMARKER

Theoretical Validation

There are seven theoretical requirements that Inspection Time (IT) must meet to be considered a valid biomarker (see p. 17). This section will consider each of the requirements and present supporting evidence from previous speed of processing and inspection time research.

Biological in Nature

The first step is to show that IT is *biological in nature.* There is widespread acceptance that speed of processing tasks tap some kind of speed or efficiency of the central nervous system (CNS), although no detailed account of how yet exists. For example, Osmond and Jackson (2002) described IT as a measure of neural efficiency; although IT is technically a low level psychological construct it can also be thought of as indicative of CNS efficiency, which is *biological in nature*. The question that was posed with respect to the traditionally used speed tasks was, how well do they actually measure CNS efficiency or speed? The answer may be not very well because they (a) confound CNS slowing and peripheral motor slowing and (b) permit a trade-off between speed and accuracy, which clearly implicates higher level cognitive monitoring. Because the IT measure derives from a method that circumvents these two problems, IT may theoretically be argued to be a better measure of CNS efficiency. In order to illustrate how the IT measure avoids the problems discussed above, the task will be briefly described.

Any task that requires a speeded response will inevitably confound CNS speed with psychomotor speed. Due to the nature of the IT task, the speed of processing measure is *inferred* rather than measured directly. To describe what is involved briefly, a target is presented at various stimulus durations and, depending upon which items are completed correctly, an estimate of processing speed is calculated, which is totally independent of response speed. In fact, people who are confused about responding or have some physical disability can still be assessed on IT, provided that they can articulate or indicate their responses in some manner. Given, that the estimate is not based on a speeded response, it is not confounded by psychomotor speed. In addition, because speed is not important, all of the focus is on accuracy and therefore the participant has no opportunity to trade-off accuracy for speed. A respondent can take as long as s/he likes to answer each item, and the next stimulus is not presented until they respond to the previous one. Therefore, the IT measure avoids the problems inherent in both perceptual speed

and RT estimates and, on this basis, could be argued to be a more pure measure of decision speed and hence CNS efficiency.

Although it is currently not possible to confirm that IT is measuring CNS functioning, various researchers have been able to eliminate alternate explanation of what IT is measuring. For example, Nettelbeck and Wilson (1985, Study 1) showed that the target and mask in the IT task are integrated centrally rather than binocularly. This suggests that IT measures the speed of some central mechanism rather than the speed of the peripheral visual system. Of course, visual acuity might affect speed if uncorrected but the discrimination is occurring at the level of the CNS and relies on much more than simply visual acuity.

Other researchers have stated that IT is overly affected by lapses in attention and may simply provide a measure of attentional processes. However, Nettelbeck and Wilson (1985, Study 2) showed that differences between 7-year old and 11-year old children on the IT task were not explained by attentional differences, with both groups exhibiting virtually error free performance on random unmasked trials.

Another issue with the IT task is the use of apparent movement as a strategy. When the mask is presented immediately after the stimulus figure some people report that one leg of the figure appear to "grow" more rapidly and this helps them to make a decision. Some researchers have argued that the IT estimate was dependent on strategy use and that this accounted for the differences between people and consequently for the IT – IQ relationship. However, Grudnick and Kranzler (2001) showed that the correlation between IT and IQ was larger for the non-strategy users, therefore indicating that the IT measure is not dependent on strategy use.

To summarise, IT is a speed of processing task that can be thought of as measuring CNS efficiency, which is biological in nature. As a measure of CNS efficiency, IT is theoretically superior to perceptual speed and reaction time tasks because it does not rely on psychomotor speed and does not confound speed and accuracy of responding. Although, it is impossible to prove that IT is measuring CNS efficiency, it has been shown that IT involves central mechanisms, is not measuring simple visual acuity or attention and is not dependent on strategy use.

Reflect Normal Aging

The second criterion is that IT must be shown to *reflect some element of "normal" aging.* Assuming IT reflects efficiency of the CNS this leads to the question, is slowing of the CNS an element of normal aging or is it only associated with diseased aging? Schaie (1989) examined

cross-sectional and longitudinal age changes in perceptual speed using data from the Seattle Longitudinal Study. He found large declines in speed (longitudinally) starting from the youngest age group, people aged 25 to 32 years, suggesting slowing of the CNS begins in early adult years. People in this age group are certainly aging but a minority would have any sort of chronic disease. To suggest that this effect is due to a minority of people with some sort of diseased aging is unlikely. Furthermore, many researchers have established that CNS slowing is a common experience with aging, with Birren and Fisher (1992, p. 31) stating, "slowness of behaviour with age has become the most robust phenomenon seen in research on aging". This suggests that slowing of the CNS (as measured by speed of processing) does reflect an element of normal aging.

Highly Reliable

The third criterion is that IT must *have high reliability*. Grudnik and Kranzler (2001) performed a meta-analysis on the relationship between IT and psychometric intelligence and, in the process, also examined test-retest reliability. As mentioned in Chapter 1, it is in this context that the majority of IT research has been done. They evaluated 90 studies that had examined this relationship and were able to get an estimate of reliability from these studies. The average test-retest reliability for the visual IT task was 0.83, with adult IT estimates being slightly more reliable than IT estimates from children. This confirms that IT has *high reliability*, as a result minimising the effects of error variance in longitudinal age changes in IT.

Stable across Generations

Fourth, IT must *be stable across generations*. Nettelbeck and Wilson (2004) tested the stability of IT across twenty years in a sample of school children. In 1981, Wilson assessed IT and vocabulary in a group of children (6 – 13 years) at a suburban school in Adelaide. Twenty years later, another group of children (matched for age and other demographic variables) was assessed at the same school. Australian census data confirmed that these children matched the earlier sample on SES. Although, the children displayed significant improvements in vocabulary, consistent with the Flynn effect, the mean IT scores were remarkably similar in both groups. This suggests that IT is *stable across generations*, at least in children.

This findings needs replication but, if IT is stable across generations, this suggests that IT is a measure of processing speed that is not confounded with the putative environmental variables (not yet identified) that influence rising IQ-type abilities across generations. Cohort effects in

cognitive abilities are generally thought to reflect improving environmental circumstances and performance on these tasks tend to be effected by environmental factors such as education. On the other hand, biological variables (e.g. grip strength, visual acuity) do not tend to improve over generations and are generally not affected by environmental factors. That IT is free from cohort effects and estimated from a task with very low knowledge requirements suggests that it is a variable representing lower level psychological processes than those involved in other speed tasks like perceptual speed (that does suffer cohort effects) and RT (that is confounded by both motor and speed-accuracy differences).

Change Independently with Passage of Time

Another of the requirements was that IT should *change independently with the passage of time*. This implies that the change over time on IT should be non-constant. This can be examined cross-sectionally or longitudinally and, given that IT appears to be free from cohort effects, the cross-sectional results might be a good representation of true longitudinal age trends. Nettelbeck and Rabbitt (1992) examined the relationship between age and mental speed in a group of 104 people aged 54 to 85. IT was correlated with age (r = .37), so that older people needed to see the stimulus for a longer period of time. This result confirms that IT changes with the passage of time in the elderly but in order to show the change is non-constant, the rate of change in one time period (e.g. 60 – 69 years) should be different to the rate of change in another time period (e.g. 70 – 79 years). There is insufficient evidence in the literature to answer this question with respect to elderly people.

However, there is a study by Nettelbeck and Wilson (1985) that found some evidence of a differential relationship between age and IT in a young sample. A group of primary school children and a small sample of university students completed the IT task. Mean IT scores for each age indicated that IT declined quite markedly from 6 to 13 but then levelled off such that the difference between the 13 year olds and the university sample was marginal. This suggests that the rate of change between 6 and 13 years is different to the rate of change from 13 to early adulthood, offering some evidence that the change in IT over time is non-constant. This implies that IT might indeed *change independently with the passage of time*. Nonetheless, it terms of utility of IT as a biomarker it would be necessary to establish this pattern in an elderly sample, ideally using longitudinal data.

Minimally Traumatic to Measure in Humans

The sixth requirement is that IT needs to *be non-lethal to animals and minimally traumatic in humans*. Speed of processing, as operationalised by RT, has been successfully measured in animals and is certainly non-lethal (see p. 40 for further discussion). As for IT, there is one report of a researcher successfully training a mouse to make the line discrimination but an estimate was not successfully generated (Welsh, 2003). Given the nature of the task is seems plausible that non-human primates could be trained to perform it. If they could do the task, it would certainly be non-lethal and measurable multiple times.

In humans, IT takes about 15 minutes at most to complete, including an initial practice session. The task is not considered to be traumatic at all. The nature of the task means the participant can have a rest whenever they like, without affecting the estimate. The practice period makes it clear what is required and the relatively short time required means that it is not tiring. Furthermore, recent investigations in our laboratory are investigating Bayesian algorithms for even quicker estimation of IT.

Exhibit Reliable Change over Short Period of Time

Finally, IT must *exhibit reliable change over a relatively short period*. There are two parts to this statement. The first suggests that the *rate of change* of IT, rather that just the estimate, must be highly reliable. When dealing with the IT estimate the best way to establish the reliability of the initial score is to administer a re-test soon after the initial test to get a measure of test-retest reliability. As for the *rate of change* of IT, the reliability of this estimate can be calculated once IT has been assessed on two occasions separated by a reasonable period of time. Of course this needs to be measured in a period of the lifespan where significant change is expected to occur. Using the test-retest reliability and the correlation between IT at Time 1 and 2, it is possible to estimate the reliability of the *rate of change* score. However, this information is not available from the literature and empirical work must be done to obtain it.

The second part requires that this change must be measurable over a *relatively short period of time*. When considering biological changes as people age, a *relatively short period of time* could be argued to be 1 to 5 years. If IT is marking the aging process, then a group of elderly people (e.g. 70+ years) should show reliable change in IT over a 1 to 5 year period. Nettelbeck, Rabbitt, Wilson and Batt (1996) examined longitudinal changes in IT in a group of 76 people aged 55 to 86 years old. On average, IT was stable over an 18-month period and the

slight reduction in IT was explained as a practice effect. This may suggest that 18-months is not a sufficiently long time frame to monitor change in IT. However, there is another possible interpretation of these results. Perhaps some people were exhibiting longer IT scores, some were exhibiting shorter IT scores and the remainder were stable. In this case the average IT would appear quite stable and uninformative but examination of these three groups could actually be very informative. For example, the group whose IT estimates were getting longer might be showing signs of accelerated aging. Due to the focus of the Nettelbeck et al. (1996) study, this proposition was not examined. However, it would be a highly informative path of investigation to follow.

To summarise, it has been argued that IT *is biological in nature, reflects some element of normal aging, has high reliability*, is *stable across generations,* and is *non-lethal to animals and minimally traumatic to humans*. Whether IT *changes independently with the passage of time* and *exhibits reliable change over a relatively short period* remains to be seen. At this point, the prospects for IT as a marker task are positive. IT appears to meet most of the theoretical requirements but does it meet the empirical ones? The following section will address whether IT meets the specific empirical criteria from Birren and Fisher (1992).

Empirical Validation

Birren and Fisher (1992) specified seven criteria for validating biomarkers that are specific and testable. In this section, each one will be considered in turn and the empirical evidence will be presented. In some cases, there is literature on speed of processing but not specifically on IT. For example, although speed of processing has been linked to mortality, there is no research on the link between IT and mortality. In this case, research from other types of speed of processing measures will be presented.

Speed of Processing and Mortality

The first requirement is that *a biomarker should be related to length of life; that is, faster processing speed should be associated with a longer life.* It is important to make the distinction between 'initial level' and 'rate of change' in the biomarker. In studies that investigate initial level, a group of people are assessed, and then some time in the future their survival status is examined. If the biomarker were predictive of mortality, it would be expected that the survivors would show superior performance at initial level compared with decedents. In studies that investigate rate of change, a group of people are assessed multiple times, rate of change is

calculated, and then in the future their survival status is investigated. In this case, the decedents would be expected to show larger rates of change in the biomarker than the survivors. This approach focuses on examination of patterns of change, which may be much more informative than the static view provided by studies using initial level. Furthermore, both Baker and Sprott (1988) and Arking (1991) made it clear that the rate of change in the biomarker is of utmost importance (see p. 12). There is evidence in the literature, that both initial level and rate of change in speed of processing are predictive of mortality.

Initial level. Anstey, Luszcz, Giles and Andrews (2001) assessed a large group of elderly adults on measures of cognition (memory, verbal ability, and processing speed) and sensory functioning (visual acuity, auditory acuity, and grip strength) in 1992. Six years later the survival status of the participants was investigated and survivors showed superior initial performance on all cognitive and sensory variables. However, there were substantial age effects, in that older people were more likely to die, so these were controlled and the results re-examined. In order of significance, a verbal ability test (Similarities), perceptual speed (Digit Symbol), the two memory tests (Symbol Recall and Picture Recall) and a dementia-screening test (Mini-Mental State Exam) all predicted mortality. None of the sensory variables showed significant differences between survivors and decedents once age was controlled.

Bosworth and Schaie (1999) reported findings from the Seattle Longitudinal Study on a group of 605 decedents (M = 73 years), and a group of 613 survivors (M = 72 years) matched for age and education. Testing was performed every seven years and scores at last testing session were compared for survivors and decedents. The test battery included measures of crystallised ability, visualisation, fluid reasoning, perceptual speed, and behavioural rigidity (motor-cognitive flexibility, attitudinal flexibility and psychomotor speed). Survivors had significantly higher initial levels of crystallised ability and visualisation and faster perceptual and psychomotor speed. There were gender effects in psychomotor speed, in that male survivors had a higher initial level than decedents, but this pattern was not apparent for females.

Singer, Verhaeghen, Ghisletta, Lindenberger and Baltes (2003) examined the utility of initial level of cognitive performance (perceptual speed, episodic memory, fluency, and knowledge) and sensory measures (visual and auditory acuity) in predicting mortality. Participants (n = 516), who ranged in age from 70 to 103, were initially tested from 1990 – 1993. Although, there were four testing sessions in total, the cognitive battery was measured on only three occasions: T_1 (1990 – 1993), T_3 (1995 – 1996) and T_4 (1997 - 1998). After the final testing

phase, mortality was examined for all participants who had completed at least one testing session. They found that people with higher initial levels of cognitive and sensory performance were more likely to survive. That is, all cognitive and sensory variables were predictive of mortality.

These three studies show that cognitive variables, including speed of processing measures, are effective at differentiating between survivors and decedents. In some cases, these variables predicted mortality up to six years later. Thus, initial level of speed of processing is related to length of life.

Rate of change. Bosworth and Siegler (2002, p. 300) performed a review of terminal change in cognitive functioning, where terminal change was defined as a "general association between mortality and *change* in cognitive measures". Nine studies were found that met their inclusion criteria, and five of these included measures of speed of processing. Of these five studies, two studies found evidence of an association between speed of processing and the other three found no association. In addition, the study by Singer et al. (2003), published since this review, is also relevant to this argument. These studies will be briefly discussed, followed by a summary of findings.

Mortensen and Kleven (1993) examined a group of 689 people, born in 1914, at 50, 60, and 70 years of age. Subsequently, in 1991, the participants were followed-up and mortality data were available for 141 participants. In this study, cognition was assessed by the WAIS, and significant differences in rate of change, from 60 to 70 years, were apparent between survivors and decedents on three performance subtests: Digit Symbol, Object Assembly and Picture Arrangement.

Bosworth and Schaie (1999) examined rate of change longitudinally, in addition to initial level. In this study, individuals were assessed longitudinally every 7 years. Bosworth and Schaie found that decedents declined more than survivors on two tasks only, verbal meaning and psychomotor speed, and that these declines could be seen over 7-year and 14-year periods.

Singer et al. (2003) presented longitudinal findings from the Berlin Aging Study. The initial testing was completed on a group of 516 participants, ranging in age from 70 to 103 years. The second testing session was completed almost four years later, the final testing session was about six years later, and 132 people completed all test sessions. With respect to rate of change, only perceptual speed and knowledge were predictive of survival.

Three studies found no association between speed of processing and mortality. Van der Wal and Sandman (1992) examined the ability of electroencephalogram waveforms and three

cognitive tasks (Digit Symbol, Digit Span, and Vocabulary from the WAIS-R) to predict terminal decline in a small group (7 survivors, 7 decedents). In the group that died, the stability of the waveform declined in the year prior to their death, and was thus predictive of mortality. However, there was no difference in rate of change on the cognitive tasks between the two groups. It is important to note that this sample was extremely small and consequently so was the statistical power, so that even if there were a significant difference in rate of change between the two groups, it would have been extremely difficult to detect.

Anstey et al. (2001) assessed rate of change, in addition to initial level of performance, as a predictor of mortality. Two-years after initial assessment, the group was reassessed, and rate of change was used to predict mortality over the following four years. Mortality data were collected on a sample of 1947 people. In this study, the highest quintile of the change distribution (i.e. those people who declined the most) was taken to represent significant decline. After adjusting for health and demographics, significant decline on Similarities (WAIS-R), visual acuity and hearing were predictive of mortality. Two possible explanations of the null finding for rate of change in speed of processing are the adjustments for health and demographics and the statistical method (lowest quintile) used to measure rate of decline.

Hassing et al. (2002) examined cognitive decline on measures of inductive reasoning, perceptual speed, spatial ability, and memory in a sample of 466 people aged 80 – 98 years. After the initial assessment, the group was reassessed at 2-years and then at 4-years, followed by an examination of mortality status. In this study, participants were tested for dementia at each stage and those with the disorder were excluded from further analyses. There were no significant differences in rate of change between survivors and decedents in any of the cognitive variables. This null finding for all cognitive variables is unusual, particularly because the sample was substantial in size and age. Anstey et al. (2001) and other studies found the only predictor of mortality was decline in verbal abilities. If this measure had been included in the Hassing et al. (2002) study, it would have been interesting to see whether it would have been a significant predictor. If not, the null result might effect exclusion of those classified as showing dementia.

In summary, there is insufficient evidence to establish that rate of change in speed of processing predicts mortality. Nonetheless, the Seattle Longitudinal Study and the Berlin Aging Study, both large-scale studies, found that rate of change in processing speed predicted mortality. Mortensen and Kleven (1993), in a smaller sample, also found rate of change in perceptual speed (60 –70 years) predicted mortality. Although, some studies have found no effect, tentative

explanations of these findings have been presented. There has been no research investigating the relationship between Inspection Time and mortality. The above findings are certainly sufficiently encouraging to suggest that this issue is worth investigating.

Speed of Processing in Animal Research

The second of Birren and Fisher's (1992) requirements is that *adjacent phylogenetic species should show changes in the biomarker with age*. Species that are considered adjacent phylogenetically to humans are those that have evolved from the same node in a phylogenetic tree. Specifically, these are non-human primate species. Most research with animals is performed in laboratory rodents but the degree to which these findings generalise to human behaviour is sometimes questionable. In non-human primate species, the generalisability of findings would arguably be higher, but testing these animals is more problematic both ethically and practically. The first question that needs to be asked is whether it is possible to measure speed of processing in non-human primates? If so, do non-human primate species show declines in speed of processing with age?

It is clear that primates are unable to complete perceptual speed tasks that require paper and pencil responding. An alternative is to use the Reaction Time (RT) paradigm where responses are made via a button or lever to more basic stimuli such as lights or tones. Unlike human subjects, primates cannot be told to respond as quickly as possible, so the RT task has to be modified so the animal is forced to respond as quickly as possible. Without going into the details of these modifications, it is clear from the literature that RT can and has been measured in primates.

A number of studies have utilised the Cambridge Neuropsychological Test Automated Battery (CANTAB; CeNeS, Cambridge, UK) for testing non-human primates. This test battery requires minimal verbal explanation, is presented on a computer screen and responses are made via a touch screen. Weed et al. (1999) confirmed that it has been used extensively with human subjects and suggested that, due to the nature of the battery, it can be used effectively with non-human primates. The CANTAB is made up of a number of subtests including spatial memory, recognition memory, self-ordered spatial search, RT, and a bimanual motor skill task. However, most of the research on primates with this battery has focused on the effects of drug administration such as scopolamine (Taffe, Weed, & Gold, 1999) and ketamine (Taffe, Davis, Gutierrez, & Gold, 2002), rather than declines in these measures with age. Weed et al. (1999) established norms for the CANTAB tests for the rhesus monkey but, given the focus of most

research in the area, they are based on young (3 – 4 year old) monkeys and therefore not age normed. Thus, although this battery offers the basis for studying decline with age in RT in primates, it has not been utilised for that purpose to date. Burbacher and Grant (2000) described some alternative methods for studying neurological behaviour or cognition in non-human primates. They suggested that both simple and choice RT can be assessed in these animals but, again, the focus in on toxicology and teratology rather than aging.

Voytko and Tinkler (2004) in a recent paper have reviewed the literature on cognitive functioning and aging in non-human primates. Most research has focused on rhesus monkeys (*Macaca mulatta*), which have a lifespan of 35 – 40 years. Voytko and Tinkler concluded that reliable age-related cognitive decline was apparent from about 20-years of age. Specifically, older monkeys take longer to learn new information, have reduced cognitive flexibility, impaired recent memory, and there is some evidence of attentional deficits. They found just one study that had investigated declines with age in speed of processing in monkeys.

Baxter and Voytko (1996) trained a group of adult (10 – 15 years) and aged (28 – 33 years) rhesus monkeys to perform simple RT. In this task, the monkey had to start by touching a home button then, when a target light illuminated, the home button had to be released and the response button pushed. With each subsequent trial, the target was illuminated for a reduced amount of time until the monkey could not longer reach it in time. The fastest reaction time was the shortest presentation where the monkey could reach the response key in time. In this way the monkey was encouraged to respond as quickly as possible to receive a food pellet in reward. Baxter and Voytko (1996) found that older monkeys had comparable RTs to young monkeys.

It is important to be cautious about drawing conclusions from just one study. Although this study suggests that RT is stable across age in rhesus monkeys, more research is needed in order to confirm this. It is clear from the literature that the means for answering this question are indeed available but more research needs to be done before we can be confident about drawing conclusions.

Speed of Processing and Gender

The third requirement is related to gender differences. S*ince females have a longer life span than males, greater changes in speed should be seen in older males than in older females.* This proposition can be measured indirectly by means of cross-sectional studies and directly through longitudinal investigations. Longitudinal data would be most informative because it provides information about the true rate of change with time rather than inferring this from

differences between age groups, which may be confounded by cohort effects. There are a few studies that have investigated gender differences in longitudinal rate of decline in speed of processing. Four studies (Aartsen, Martin, & Zimprich, 2004; Anstey, Hofer, & Luszcz, 2003; Finkel, Reynolds, McArdle, Gatz, & Pedersen, 2003; Singer et al., 2003) found no difference in rate of change between genders. Although these studies generally had large samples there was not a statistically significant difference between the genders. However, one study by Mortensen and Kleven (1993) did find evidence that males decline more on speed tasks than do females.

Mortensen and Kleven (1993), in the study described above, examined performance on the WAIS in a group of 68 females and 73 males at ages 50, 60, and 70 years. There was significant decline in a number of subtests over the 20-year period and males declined significantly more than females on four subtests: Digit Symbol, Object Assembly, Picture Arrangement and Information. Mortensen and Kleven (1993) claimed the difference in rate of decline was unrelated to the superior performance of the males on the WAIS at age 50.

At this stage, it is unclear whether there are reliable gender differences in rate of decline on speed of processing tasks. The one study that found an effect was conducted over a period of 20-years, which suggests that if gender differences do exist they might only be apparent over a long period of time. In the other four studies, the rate of decline over about six years was assessed and this might not be a sufficiently long time frame to see differential rates of decline between genders. There is no research regarding gender differences in rate of decline in IT.

Speed of Processing and Physiological Aging

Birren and Fisher's (1992) fourth requirement is that the *biomarker should correlate with physiological and anatomical indicators of aging (e.g. lung vital capacity, skin elasticity, bone mass, muscular strength, maximum heart rate, hearing threshold, glucose tolerance, measures of brain excitability, and brain metabolism).* There is a lot of evidence that perceptual speed and RT tasks correlate with physiological and anatomical indicators of aging. However, there is little evidence of an association between IT and these measures. This is not because an association does not exist; rather IT has not been considered in the context of biomarkers before and therefore this question has not been examined.

Speed of processing tasks have been shown to correlate with sensory variables including *visual acuity* (Anstey, 1999; Anstey, Luszcz, & Sanchez, 2001; Anstey & Smith, 1999; Anstey, Stankov, & Lord, 1993; Baltes & Lindenberger, 1997; Lindenberger & Baltes, 1994, 1997; Salthouse et al., 1998), *lens accommodation* (Clark, 1960), *visual contrast sensitivity* (Anstey,

Lord, & Williams, 1997; Anstey et al., 1993), *auditory acuity* (Anstey, Luszcz, & Sanchez, 2001; Anstey & Smith, 1999; Baltes & Lindenberger, 1997; Clark, 1960; Lindenberger & Baltes, 1994, 1997), *vibration sense* (Anstey & Smith, 1999; Anstey et al., 1993), and *proprioception* (Anstey et al., 1993).

Perceptual speed and RT have also been linked to measures of motor functioning such as *balance-gait* (Lindenberger & Baltes, 1997), *sway* (Anstey et al., 1993), *muscle strength* (Anstey, Lord et al., 1997; Anstey et al., 1993), and *grip strength* (Anstey & Smith, 1999; Clark, 1960; Salthouse et al., 1998). Furthermore, there is a link between speed measures and what Anstey et al. (1996) would term physiological/ biomedical variables. These include *lung function* (Anstey & Smith, 1999; Cerhan et al., 1998) and *blood pressure* (Aleman, Muller, de Haan, & van der Schouw, 2005; Blumenthal, Madden, Pierce, Siegel, & Appelbaum, 1993; Clark, 1960; M. F. Elias, Robbins, Elias, & Streeten, 1998; Salthouse et al., 1998; Swan, Carmelli, & Larue, 1998). To illustrate these relationships more clearly, one study will be described in detail that included a number of different cognitive and sensorimotor variables.

Lindenberger and Baltes (1997) examined the relationship between cognition and sensorimotor variables within the Berlin Aging Study. As described earlier, this sample was substantial in size, with 516 participants aged from 70 – 103, stratified by age and gender. Overall, there was a considerable association between the sensorimotor (balance-gait, vision and hearing) and cognitive variables. Moreover, they found that the sensorimotor variables were more highly related to perceptual speed than to any other ability construct including fluency, reasoning, memory and knowledge. Lindenberger and Baltes (1997, p. 428) remarked, “the magnitude of the relationship between perceptual speed and sensory-sensorimotor functioning was especially impressive: The two constructs shared 72% of their variance”. Thus, this specific study and the others outlined above, confirm that speed of processing correlates with physiological and anatomical indicators of aging.

Speed of Processing and Cognition

The fifth requirement is that the *biomarker should correlate with behavioural processes (e.g. attention, perception, memory, problem solving and reasoning*). The relationship between IT and cognitive abilities has been studied extensively. Early studies focused on the relationship between IT and omnibus IQ tests with three reviews (Grudnik & Kranzler, 2001; Kranzler & Jensen, 1989; Nettelbeck, 1987) estimating the correlation at about -0.5. However, later research has aimed at locating IT within the group factors of *Gf-Gc* theory. The following section will

review the relationship between IT and fluid reasoning, crystallised ability, visualisation, short-term memory and speed of processing.

Fluid reasoning (Gf). There is evidence of a moderate correlation between IT and measures of *Gf*. Mackintosh and Bennett (2002) demonstrated a small association between IT and Raven's Standard Matrices (r = -.29). Burns and Nettelbeck (2003) found a moderate correlation between IT and sections of the Cattell Culture Fair Test (median correlation = -.42). Osmon and Jackson (2002) demonstrated a large correlation between IT and Analysis-synthesis and Concept formation from the WJ-R (mean correlation = -.63). These studies suggest that people with shorter ITs perform better on measures of fluid reasoning. This is consistent with research showing an association between *Gf* and other speed measures such as perceptual speed and RT.

Crystallised ability (Gc). The relationship between IT and measures of *Gc* is considerably smaller. Many of the early studies on IT examined the relationship between IT and the WAIS. It was consistently found that IT correlated more highly with the performance section than with the verbal section of the battery (see Kranzler & Jensen, 1989 for review). Burns, Nettelbeck and Cooper (1999) found virtually zero correlation between IT and Picture Vocabulary (r = -.05). Osmon and Jackson (2002) measured both picture and oral vocabulary and found a median correlation with IT of .05. This relationship is consistent with that of other speed measures. In general, people who are quick do not have an advantage on measures of crystallised ability such as general knowledge and vocabulary.

Visualisation ability (Gv). An association between IT and measures of *Gv* has been established. As mentioned above, IT correlates with the performance scale of the WAIS. McGrew (1997) showed that most of the subtests from this scale provide measures of *Gv* with the exception of Digit Symbol, which measures *Gs*. Burns and Nettelbeck (2003) found that IT correlated with Block Design (r = -.40), Object Assembly (r = -.32), and Picture Arrangement (r = -.27), all from the performance scale of the WAIS-R. Furthermore, IT has been shown to correlate with the *Gv* tests from the WJ-R (Burns & Nettelbeck, 2003; Osmon & Jackson, 2002) and mental rotation (Mackintosh & Bennett, 2002). Therefore, IT is related to performance on visualisation tasks, which is to be expected given the nature of the IT tasks.

Short-term memory (Gsm). IT has not been studied with respect to memory very extensively. However, two studies by Nettelbeck (Nettelbeck & Rabbitt, 1992; Nettelbeck et al., 1996) demonstrated an association between IT and short-term memory tests, cumulative learning

and free recall. In addition, Burns and Nettelbeck (2003) examined the relationship between IT and the two WJ-R short-term memory tests. IT was significantly correlated with memory for words (r = -.33) and memory for sentences (r = -.28).

Speed of processing (Gs). IT is a measure of psychophysical speed and therefore it would be expected to correlate with measures of *Gs* (i.e. perceptual speed). There is substantial evidence that IT does correlate with perceptual speed tests. IT has consistently been shown to correlate with Digit Symbol from the Wechsler batteries (Burns & Nettelbeck, 2003; Crawford, Deary, Allan, & Gustafsson, 1998; Deary, 1993; Nettelbeck, Edwards, & Vreugdenhil, 1986; Nettelbeck & Lally, 1976; Nettelbeck & Rabbitt, 1992). Burns et al. (1999) also found that IT correlated significantly with the two *Gs* tests from the WJ-R: Cross out (r = -.42) and Visual matching (r = -.38). Thus, people who perform well on the IT task are likely to perform well on other speed measures including perceptual speed. These studies have established that IT is correlated with the behavioural processes of *Gf*, *Gv*, *Gsm* and *Gs*. That, is superior performance on the IT task is related to better performance on a range of cognitive tasks.

The location of IT within Gf-Gc theory. In addition to these findings it would be desirable to know, in a factor analytic sense, which of the Horn and Cattell model factors IT loads on. A number of studies have attempted to answer this question with Burns and Nettelbeck (2003) providing the most comprehensive investigation thus far into this issue. Burns and Nettelbeck (2003) performed a confirmatory factor analysis on IT and tests from the Woodcock-Johnson Psycho-Educational Battery- Revised (Woodcock & Johnson, 1989), the Wechsler Adult Intelligence Battery - Revised (Wechsler, 1981), and the Cattell Culture Fair Test (Cattell & Cattell, 1959). A second-order general factor and five group factors emerged, which were interpreted as *Gs*, *Gv*, *Gf*, *Gc*, and *Gsm*. IT loaded unambiguously on *Gs*, which was primarily marked by Digit Symbol. This suggests that, although IT correlates with all of these behavioural processes, it is primarily a measure of general speed of processing. Furthermore, the link between IT and general intelligence is likely mediated by speed of processing.

At the level of *Gf-Gc* theory, IT is clearly measuring speed of processing. However, it would also be desirable to show that IT is distinct from other types of speed measures such as perceptual speed, reaction time and psychomotor speed. It was argued that there were problems with perceptual speed and reaction time tasks for use as biomarkers so it would be useful to demonstrate that IT is distinct from these measures. O'Connor and Burns (2003) performed an exploratory factor analysis to locate IT within a model of speed of processing abilities. Their

test battery included simple RT measures to more difficult CDS tasks. Factor analysis revealed five first-order group factors and a second-order general speed of processing factor. The group factors were CDS, perceptual speed, visualisation speed, decision time (also called RT) and movement time (also called psychomotor speed). IT did not load with CDS, perceptual speed, reaction time or psychomotor speed. Rather it loaded on the fifth factor, which was tentatively labelled visualisation speed. This factor cannot be interpreted as representing psychophysical speed because IT was the only measure of this type in the battery. However, this study did confirm that IT is distinct from perceptual speed, reaction time and psychomotor speed.

Speed of Processing and Life-Style Factors

The sixth requirement is that the *biomarker should be reduced by not smoking, limited use of alcohol, proper diet and exercise.* There is evidence that all four of these lifestyle factors are related to speed of processing as measured by perceptual speed and RT. In addition, there is literature on the relationship between nicotine and IT, which is relevant to the discussion of the effect of smoking on speed of processing.

Smoking. There are two seemingly contradictory research findings on the relationship between cigarette smoking and IT. On one hand, there is a wealth of literature on the relationship between cigarette smoking and cognition, which suggests that current smokers have poorer performance on a range of cognitive tasks, including speed of processing when compared with non-smokers (Cerhan et al., 1998; Hill, 1989; Kalmijn, van Boxtel, Verschuren, Jolles, & Launer, 2002; Whalley, Fox, Deary, & Starr, 2005). However, there is also evidence from psychopharmacological studies that acute nicotine intake can enhance performance on IT (see Stough, Thompson, Bates, & Nathan, 2001 for review). Stough et al. suggested that the effect of nicotine on IT is mediated by the cholinergic system, with nicotine causing the release of acetylcholine, which leads to faster IT performance. These two findings will be discussed in more detail in an attempt to draw some conclusions on the question of whether *speed of processing is slowed by smoking*.

With respect to speed of processing as a biomarker, the pertinent question is whether the lifestyle choice to smoke cigarettes impacts on the rate of functional aging. A study by Hill (1989) is particularly relevant because the participating groups were matched on contextual variables including age, gender and education. Seventy-six elderly adults (aged 64 – 83 years) who were classified as non-smokers, current smokers or ex-smokers completed a range of cognitive measures including problem solving, psychomotor speed, memory, attention span and

visuospatial reasoning. The non-smokers performed significantly better than current smokers on all of the speeded tasks (problem solving and psychomotor speed), with no differences in the non-speeded tasks. This finding and others appear to confirm that cigarette smoking does have a deleterious effect on speed of processing in the long-term.

Studies on the acute effect of nicotine on IT, which suggest nicotine can enhance IT performance, do not necessarily contradict the above findings. These studies have been extremely short term, with within-subject designs. They have suggested that nicotine administration can produce improvements on speed of processing tasks in the minutes or hours immediately after administration within that individual. They do not suggest that prolonged cigarette smoking improves or maintains speed of processing. That is, these studies focus on the acute effects of cigarette smoking rather than the chronic effects. However, it is possible that some of the previously discussed studies may have been affected by this nicotine effect. Let us assume that a group of age-matched participants were to be assessed on a range of speed measures. The smokers all had a cigarette before entering the testing room and the non-smokers did not. It is plausible that the smokers may display quicker performance on the speed tests due to recent nicotine intake but the deleterious effects of long-term cigarette smoking should counteract this. If acute nicotine intake has a larger positive effect on speed performance than the negative long-term smoking effect then smokers should perform at a superior level. On the other hand, if acute nicotine intake has a smaller effect then the non-smokers should still come out ahead. Evidence suggests that the latter is correct. Although, the smokers in a sense had an immediate advantage, many studies have still found that the non-smokers display superior performance, which suggests that the deleterious effects of long-term smoking may be underestimated. To conclude, smoking cigarettes does appear to have a negative impact on speed of processing performance but this may be underestimated in some studies due to the short-term effect of nicotine intake by the smokers in the group.

Alcohol consumption. There is a wealth of literature of the impact of alcohol consumption on cognition. Although alcoholics are clearly impaired on a range of cognitive tasks, there appears to be some protective effect for low to moderate drinkers. A number of studies have shown that people who consume up to two alcoholic drinks per day display superior performance on cognitive tasks than abstainers and excessive drinkers (P. K. Elias, Elias, D'Agostino, Silbershatz, & Wolf, 1999; Hendrie, Gao, Hall, Hui, & Unverzagt, 1996; Kalmijn et al., 2002; Schinka, Belander, Mortimer, & Borenstein Graves, 2003). However, many of these

studies have utilised gross measures of cognitive functioning, such as IQ scores or dementia screening tests such as the Mini-Mental State Exam, which do not allow for the assessment of the effect of alcohol consumption on speed of processing. Of the four studies that included speed of processing measures, all confirmed a relationship with alcohol consumption, although the nature of this relationship is inconclusive.

Two studies (Cerhan et al., 1998; Kalmijn et al., 2002) reported a u-shaped relationship, with abstainers and heavy drinkers performing poorly and moderate drinkers showing superior performance. On the other hand, Aleman et al. (2005) reported a linear trend where abstainers performed quickest, suggesting that any alcohol consumption may be detrimental. Richards, Hardy, and Wadsworth (2005) demonstrated that women who drank alcohol had a more rapid decline in speed of processing over a 10 years period to their mid 50's. Furthermore, they found that, if alcohol consumption increased during this period, then the decline was further exacerbated. Therefore, the evidence is quite compelling that alcohol consumption is related to speed of processing but the nature of this relationship (i.e. u-shaped or linear) is not clear.

Nutrition. Many studies have demonstrated a link between nutrition and cognition (see Bryan, 2003 for review). However, as in the studies on alcohol consumption, many have used general measures of cognitive functioning or impairment. As Calvaresi and Bryan (2001) pointed out, these tests may not be able to discriminate between those aspects of cognition that are vulnerable to inadequate nutrition and those that are resistant. Birren and Fischer (1992) suggested that a biomarker should be sensitive to diet and this implies that speed of processing should be vulnerable to inadequate nutrition. Is there any evidence of this?

Berr, Richard, Roussel and Bonithon-Kopp (1998) examined the effect of antioxidants on cognitive performance in a group of 1,389 people aged between 59 and 71 years. Damage by free radicals have been implicated in the aging process and antioxidants are thought to protect against this damage. Thus, one might expect people with higher levels of antioxidants to have experienced less cognitive decline than those with insufficient levels. The results showed that participants with low levels of plasma carotenoids (an antioxidant marker) had increased risk of poor performance on Trail Making B, Digit Symbol and Auditory Verbal Learning. Low levels of selenium (another antioxidant marker) were associated with poorer performance on Trail Making B. Digit Symbol is clearly a measure of perceptual speed and Trail Making B has been described as a measure of attention or cognitive flexibility; but it undoubtedly measures speed to

some degree. Therefore, people with inadequate levels of antioxidants do indeed display impairments in speed of processing.

Two recent studies (Bryan, Calvaresi, & Hughes, 2002; Lindeman et al., 2000) have established a link between folate and speed of processing. This is particularly important because folate deficiencies are thought to be increasingly prevalent with advancing age (Bryan et al., 2002). Lindeman et al. (2000) examined the effect of vitamins B_{12}, C and folate on cognition in a group of 883 participants all aged over 65. People with low serum folate levels performed more poorly on a number of tests, including a version of the Trail Making test, which they interpreted as measuring psychomotor speed. Bryan et al. (2002) examined the effect of vitamins B_{12}, B_6 and folate on cognition in a sample of 211 women ranging in age from 20 – 92 years. They found that people with low levels of folate in their diet were impaired on a measure of speed of processing but vitamin B_{12} and B_6 had no effect. Thus, there is some evidence that folate is linked to speed of processing performance.

These three studies provide initial evidence that speed of processing is improved by proper diet, at least with respect to antioxidants and folate. In order to answer this question comprehensively it would be necessary to establish clearly what constitutes a "proper diet" for elderly people and then investigate the degree to which these key nutrients are related to speed of processing. However, at this early stage it is at least possible that antioxidants and folate (both implicated in aging) do have an impact on speed of processing.

Exercise. Spirduso (1975) demonstrated that active individuals had quicker RTs than their sedentary counterparts and this led to a great deal of research on the link between exercise and speed of processing. Chodzko-Zajko and Moore (1994) reviewed the literature and found 15 studies that had linked physical fitness and cognitive processing speed. Furthermore, they found a number of studies that showed that short-term exercise programs can actually improve processing speed. For our purposes these studies are particularly interesting because they imply a causal relationship between exercise and speed of processing.

Dustman et al. (1984) assigned a group of 43 sedentary adults to aerobic exercise, strength and flexibility training or a control group. The exercise groups completed a one-hour session, three times a week for four months. The aerobic group showed significant increases in Critical Flicker Fusion, Digit Symbol, simple RT and Stroop. However, the strength and flexibility group did not show significant change in any of the variables. One explanation for this is the *cerebral circulation hypothesis* (Spirduso, 1980), which posits that regular exercise

enhances oxygen transportation in the brain. Although the people completing strength and flexibility training were indeed exercising, they were not increasing their heart and breathing rate as much as the aerobic group. Therefore, the beneficial effect on the CNS was not as great and thus the effect on speed of processing was not apparent. This study, and many others have established the strong link between exercise and speed of processing performance. That is, people who regularly exercise have faster speed of processing than sedentary adults.

Speed of Processing and Disease

The final requirement is that a *biomarker should be exacerbated by the presence of age-associated diseases such as coronary artery disease, cerebrovascular disease, diabetes, and Alzheimer's disease*. There is a large quantity of literature demonstrating a link between speed of processing and Alzheimer's disease (AD). Furthermore, a direct link between AD and Inspection Time (IT) has been established. As for coronary artery disease, it is just one of a number of diseases that comprise cerebrovascular disease. Both heart disease and other manifestations of cardiovascular disease (e.g. hypertension) are associated with poorer cognitive performance and, specifically, slower speed of processing.

Alzheimer's Disease. Dementia is a major health issue in the elderly and it demonstrates increased incidence with advancing age. Corsini (1999, p. 262) defined dementia as "*a lasting deterioration of memory, judgement, and emotions generating erratic behaviour*". The leading cause of dementia in the elderly is AD, which is thought to define over 65% of dementia cases (Kolb & Whishaw, 1996). In the initial stages of dementia, people commonly complain of memory difficulties but changes in a range of cognitive abilities are observable. A number of studies have found a relationship between AD and speed of processing and these studies can be split into four different types.

First, studies have shown that AD patients have significantly slower speed than age-matched controls on measures of *perceptual speed* (Berg et al., 1984; Devanand, Folz, Gorlyn, Moeller, & Stern, 1997; Larrabee, Largen, & Levin, 1985), *RT* (Pate, Margolin, Friedrich, & Bentley, 1994; Saito et al., 2001) and *IT* (Deary et al., 1991).

Second, the severity of dementia diagnosis is significantly related to scores on Digit Symbol (Larrabee et al., 1985) and IT (Deary et al., 1991). For example, Deary et al. (1991) examined AD patients, Korsakoff's patients and controls who were matched on age and premorbid IQ. Participants were assessed on two speed measures, Digit Symbol and IT, and the CAMDEX-Cog, a clinical dementia-screening test. In the AD group, scores on the CAMDEX-

Cog correlated highly with both IT (r = -.81) and Digit Symbol (r = .80). This is despite the fact that this correlation was based on a total sample size of 13.

Third, the rate of decline in speed of processing tests has been shown to be significantly larger in AD patients than normal elderly adults. Botwinick, Storandt and Berg (1986) followed a group of 18 subjects diagnosed with mild senile dementia of the AD type (aged 64 – 80) and an age-matched control group for 4 years. The AD group showed significant declines in all measures, with the largest seen in sections of the Wechsler Memory Scale, Digit Symbol, WAIS (total score) and Trail Making. The control group showed little decline over the four years. Botwinick et al. (1986) concluded that memory and speed of processing declines may be predictive of dementia.

Fourth, speed of processing has been effectively used to predict dementia diagnosis in the future. This is particularly important because a biomarker must be shown to have predictive rather than just concurrent validity. Devanand et al. (1997) administered a number of neuropsychological tests to a group of 62 individuals with "questionable dementia". At least one year later the group was re-assessed and initial scores on the neuropsychological tests were used to predict dementia diagnosis. Low scores on a number of memory measures, category naming for animals, Digit Symbol (a perceptual speed test), Picture Arrangement and Block design were all predictive of final dementia diagnosis.

These studies lead to the unequivocal conclusion that speed of processing is impaired in people with dementia of the Alzheimer's type. Both level and rate of decline are effected by AD and speed of processing can effectively be used to predict dementia status in the future. Perhaps most importantly, some part of these results has been confirmed with IT. However, the predictive validity of IT for dementia diagnosis has not been studied. It is desirable to examine whether IT can effectively predict dementia diagnosis in the future.

Cardiovascular Disease. One of the earliest manifestations of cardiovascular disease is hypertension, which affects around one in four Australian adults (Australian Institute of Health and Welfare, 2004). A number of studies have linked hypertension to cognitive performance, in particular speed of processing. First, studies have shown that hypertensives perform more poorly than age-matched normotensives on perceptual speed tasks (Blumenthal et al., 1993; Cerhan et al., 1998; Miller, Shapiro, King, Ginchereau, & Hosutt, 1984). Second, regression analyses have demonstrated that blood pressure is a significant predictor of perceptual speed (Blumenthal et al., 1993). Third, hypertensives demonstrate more longitudinal decline on speed of processing than

do normotensives (M. F. Elias et al., 1998; Haan, Shemanski, Jagust, Manolio, & Kuller, 1999). Fourth, medicated hypertensives have shown differential performance to un-medicated hypertensives on speed of processing. Miller et al. (1984) assessed hypertensives and normotensives on Digit Symbol and followed them up 15-months later. The medicated hypertensives showed substantial improvements while the normals and un-medicated hypertensives were relatively stable. Finally, there is some evidence that people with low blood pressure also perform more poorly on speed of processing tasks. Swan et al. (1998) assessed blood pressure at middle age (mean age = 45) and again at old age (mean age = 75) in a group of 717 male survivors from the Western Collaborative Group Study. They found that people whose blood pressure had reduced from middle to old age showed particularly poor performance on the Digit Symbol task. In fact, they performed more poorly than people who had sustained high blood pressure.

In addition to hypertension, later manifestations of cardiovascular disease, including atherosclerosis, congestive heart failure and stroke, have been linked to speed of processing. Cerhan et al. (1998) examined correlates of cognitive performance in a large group (N = 13,913) of middle aged people. They found that high carotid artery intima-media thickness (a marker of atherosclerosis) was associated with poor performance on the Digit Symbol task. Verhaeghan, Borchelt and Smith (2003) examined the impact of a number of somatic diseases on cognition in a group of 516 people (aged 70 to 103). They found that after demographics were entered (age, sex, SES and dementia diagnosis), both stroke and congestive heart failure were significant predictors of perceptual speed performance. That is, people who had experienced either stroke or congestive heart failure were impaired on speed of processing performance. These studies confirm a link between cardiovascular disease and speed of processing. It appears that people who suffer from cardiovascular disease perform more poorly on speed tasks and show more longitudinal decline than normals although the direction of causation is not clear.

Plan for Experimental Investigation

The aim of this chapter was to consider whether IT met the theoretical and empirical requirements for a biomarker and subsequently to decide whether an experimental investigation into IT as a biomarker is warranted. In the section on theoretical validation, it was argued that IT was *biological in nature* and *reflected some element of normal aging, had high reliability*, was *stable across generations*, was *non-lethal to animals and minimally traumatic to humans*. There

was insufficient evidence to decide whether IT meets the two other requirements of *changing independently with the passage of time* and *exhibiting reliable change over a relatively short time period*. In the section on empirical validation, an extensive literature review linked speed of processing to *mortality*, *physiological aging*, *cognitive aging*, *life-style factors* and *disease*. Where IT had been studied, it conformed to all expectations (e.g. related to cognitive aging and age-associated disease). There were only two criteria that speed of processing had not fully met. First, it is not yet clear whether *non-human primates show declines in speed of processing with age*, due to a lack of research in this area. Second, the results were not supportive of the notion that *older males show more decline on speed of processing than older females*. There was some evidence for this proposition but it appeared to be apparent only over a long period of time. However, overall the prospects for IT as a biomarker are extremely encouraging and there are enough unanswered questions to warrant an investigation of IT as a biomarker of aging. Based on the plan presented at the end of Chapter 1, the next step is to use the ex-post facto model to examine the predictive validity of IT. The plan for an experimental investigation of IT will be presented below.

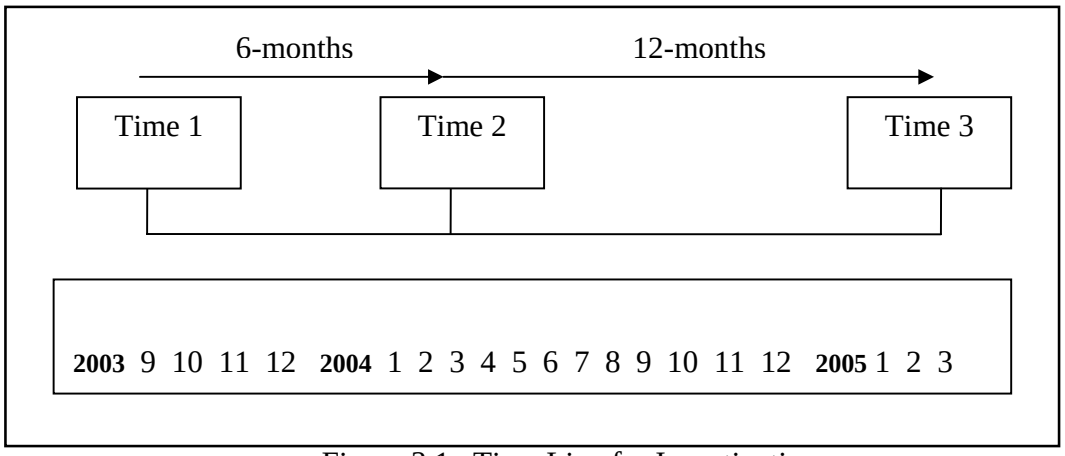

Figure 3.1. Time Line for Investigation

Because this is the first investigation into IT as a biomarker, the *ex post facto* model (Ingram, 1991) was used. That is, elderly adults were recruited and asked to complete the IT task on a number of occasions, in order to examine longitudinal age trends. This investigation was purely observational without any experimental manipulation. Figure 3.1 shows the time line for the investigation. An estimate of initial score on IT was taken at Time 1. An estimate of IT change over a 6-month period was derived from the difference between Time 1 and 2. Finally, an

estimate of IT change over 18-months was derived from the difference between Times 1 and 3. This allowed initial and change scores on IT to be used as predictors of functional age at Time 3, in order to investigate whether the biomarker has predictive validity. Furthermore, functional age was assessed at Times 1 and 3, so that change over 18-months on functional age could be estimated. This also allowed the examination of whether initial scores on IT predicted change over 18-months on functional age.

In addition to establishing predictive validity, we have seen in this chapter that a biomarker should be related to factors such as mortality, life-style and health. Therefore, an attempt was made to clarify the relationship between IT and a number of these factors defined by Birren and Fisher (1992). It is beyond the scope of this investigation to examine the relationship between IT and mortality or show that IT changes with age in adjacent phylogenetic species. However, it is possible to examine many of the other issues, such as whether males show more decline on IT than females, and whether IT is related to physiological aging markers.

Finally, one of the most important issues that must be addressed is how functional age was measured. Based on the work of Birren and Fisher (1992), functional age was measured at two levels: everyday functioning (quality of life and activities of daily living) and cognition (fluid reasoning). That is, the validity of IT as a biomarker will be primarily decided based upon its ability to predict everyday functioning and cognition in the future. IT must have predictive validity but it must *also* be more useful than chronological age and ideally other previously used physiological markers (e.g. grip strength). To evaluate this idea, we included a number of previously used biological marker tests in the test battery. For comparison sake, several measures of perceptual speed were also included in the study, to test whether IT is indeed more useful in terms of predictive validity.

Concurrent Validity

In Chapter 4, data from the first testing phase will be examined to establish concurrent validity. First, the relationship between IT and a number of age-related factors will be examined. The measures include demographics (e.g. gender, education), life-style factors (e.g. exercise, nutrition) and health (e.g. cardiovascular disease, self-reported health). The relationship between IT and physiological markers of the aging processing such as grip strength and visual acuity will also be examined. Finally, although IT needs to be able to predict functional age outcomes in the future, it is interesting to examine the concurrent relationship between IT and these measures.

Therefore, the ability of IT to explain variance in the functional age outcomes will be examined. IT will be compared to age and physiological markers for this purpose.

Reliability and Six-month Change

In Chapter 5, the results of the second testing phase will be discussed. In this testing phase, IT and the physiological measures will be assessed to allow for estimation of 6-month change scores. Chapter 5 will also present a general discussion of the issues surrounding change scores, followed by details of the methodology chosen. The reliability of the 6-month change scores will also be examined in order to see whether they are valid to use as predictive variables.

Assessment of Functional Age

Chapter 6 will examine the functional age outcomes at the initial testing session and the degree to which they have changed over 18-months. This will allow us to see whether initial and 18-month change scores for the functional age measures are suitable outcome measures for assessing the predictive validity of IT.

Predictive Validity

In Chapter 7, initial scores on IT and change scores over 6-months and 18-months will be used as predictors of the functional age outcomes. As above, IT will be compared to age and physiological markers for its ability to predict these outcomes. Multiple regression will be used to examine the predictive validity of IT.

As for the outcomes, the two major functional age measures will be everyday functioning and cognition. However, it would be ideal if we could use IT to predict mortality or longevity. As mentioned earlier this cannot be ascertained within the time frame of the current investigation. However, there is one way this could be measured indirectly. Crystallised ability is a cognitive ability that is largely maintained throughout the lifespan. However, some researchers have suggested that when it does begin to decline, this may be predictive of impending death or mortality (see Cooney et al., 1988). If IT can predict decline in crystallised ability this might indirectly suggest it is related to mortality. Therefore, this issue will also be examined.

Test Battery

Given the design of the study, the following factors will be included in the test battery: inspection time, *demographics* (age, gender, and education), *life-style factors* (exercise, smoking, alcohol, and nutrition), *health* (stroke, coronary heart disease, hypertension), *physiological*

markers of aging (grip strength, systolic blood pressure, diastolic blood pressure, height, weight, visual acuity,), *cognitive measures* (fluid reasoning, crystallised ability, perceptual speed) and *everyday functioning* (quality of life, activities of daily living). In addition, it is necessary to screen out people with dementia because this condition impacts on the aging process and represents diseased aging.

Hypotheses

In Chapter 4, there are essentially three issues to be considered: IT and age-associated factors, IT and physiological markers of aging, and IT and functional age outcomes. The age-associated factors include demographics, life-style factors and health. The functional age outcomes are quality of life, activities of daily living and cognition. The following hypotheses pertain to these issues.

1. *Age*. There will be a positive correlation between IT and age.
2. *Gender*. Males will show more cross-sectional change on IT than females. That is, the correlation between age and IT will be larger for males than females.
3. *Education*. Education will not have a significant effect on IT scores. To the degree that IT is biological it should not be affected to a large degree by environmental factors such as education.

4. *Smoking*. Current smokers will have the longest IT scores, followed by ex-smokers and then non-smokers. In order to differentiate more adequately between current smokers and ex-smokers, it is hypothesised that years of smoking will be positively related to IT.
5. *Alcohol*. There will be a relationship between alcohol consumption and IT performance. It is expected that a quadratic function will be found, such that people who drink a lot or abstain will have longer IT scores than those who drink a little each day.
6. *Exercise.* Regular exercisers will have shorter IT scores than sedentary adults. This effect will be most pronounced for cardiovascular exercisers.
7. *Nutrition*. Nutritional intake will be related to IT performance. People with low intake of important micronutrients (e.g. folate) and antioxidants will perform significantly worse on the IT task than people with an adequate or high intake.

8. *Stroke*. A history of stroke will be associated with poorer performance on the IT task.
9. *Coronary Heart Disease*. A history of coronary heart disease will be associated with poorer performance on the IT task.
10. *Hypertension*. Hypertension will be associated with poorer performance on the IT task.

11. *Physiological aging indicators*. IT will be correlated with physiological indicators of normal aging including grip strength, blood pressure, weight, height, and visual acuity.

12. *Quality of life.* IT will explain more of the variance in quality of life than will chronological age.
13. *Everyday functioning.* IT will explain more of the variance in everyday functioning than will chronological age.
14. *Cognition.* IT will explain more of the variance in fluid reasoning than will chronological age.

CHAPTER FOUR: STUDY 1 – TESTING CONCURRENT VALIDITY

This chapter presents the findings from the first phase of data collection. Given that this study uses a longitudinal research design, this chapter will present an extensive description of the sample, materials and procedures used throughout the entire program of research. In the results section, the concurrent relationships between IT and all of the constructs of interest (e.g. cognition, health) will be presented. Based on these findings, the prospects for IT as a biomarker will be clarified and any necessary changes to the test battery or procedures will be made.

Method

Participants

The participants (N = 150) were all living in their own homes and were recruited through radio, television, and print media. All participants were required to be fluent in English, living in metropolitan Adelaide, and were screened for dementia. Their ages ranged from 70 to 91 years; 99 females (*M* = 77.7 years, *SD* = 4.8) and 51 males (*M* = 77.4 years, *SD* = 3.6). Table 4.1 shows the number of male and female participants in each age group. First, it shows that almost twice as many females participated as males. Second, the majority of people were in the age group 75 – 79 years, which is somewhat surprising because there are clearly more people in the population aged 70 – 74.

Table 4.1. Sample distribution by Age and Gender

Age	Males	Females	Total
70 – 74	12	30	42
75 – 79	26	34	60
80 - 84	11	27	38
85+	2	8	10
Total	51	99	150

Population data from the Australian Institute of Health and Welfare (AIHW, 2004) allows for examination of the representativeness of this sample with respect to age and gender. In the group 65 – 74 years, males made up 48.6% of the population, so that gender distribution is almost equal. Our sample started from age 70 and in the age group 70 – 74 years males made up only 28.6% of the sample and were thus under-represented. In the group aged 75 to 84, males made up 42.7% of the population and 37.8% of the sample, and therefore the sample was considered a reasonable representation of the population. Finally, in people aged 85 and over, males make up

31.6% of the population and only 20% of the sample. Thus, again males were under-represented in this age group. This trend for males to be under-represented may indicate a gender difference in tendency to volunteer for this type of research project. Alternatively, it is important to consider the possibility that the sample is biased toward healthy participants. That is, older males may not be as healthy as older females and therefore less likely to volunteer.

With respect to education, participants reported both their highest level of education (e.g. high school) and years of formal education. 12.8% attended primary school only, 38.3% attended high school, and the remaining 48.1% completed higher education of some type. Of this group, 18.8% completed an apprenticeship or certificate, 14.1% completed a diploma, 10.7% completed a bachelor degree and 5.4% completed a higher degree. Therefore, almost half of the participants completed some form of higher education after high school. This level of education is well above that of the wider population in this age group. Census data from the Australian Bureau of Statistics (ABS, 2001) indicated that 15% of adults aged over 75 years old have completed some type of higher education (certificate = 8.8%, diploma = 3.1%, bachelor = 2.8% and higher degree = 0.8%). Therefore, the current sample was highly educated compared with similar aged people in the wider population.

Information about the time spent in formal education was available for 93% of the sample and the average was 11.7 years. Anstey et al. (2003) examined years of formal education in a representative sample of 1,823 Australian adults aged over 70 years. Those participants completed an average of 9.3 years of formal education, which provides further evidence that our participants were more highly educated than the average for the wider population. There were no significant differences in level or years of education between males and females.

One hundred and one of the participants were born in Australia. Of those who immigrated here, 36 were born in the United Kingdom, 7 were born in other areas of Europe with the remainder from Sri Lanka, New Zealand, and Egypt. On average, people who did immigrate arrived over 40 years ago (M = 42.8 years, range = 15 – 79 years).

Table 4.2 presents details of the marital status of the sample divided by gender. The vast majority (89%) of the sample were either married or widowed but there was a significant difference between males and females. Most of the males (71%) were still married with just under a quarter being widowed. However, only 33% of the females were still married with over half being widowed. This is probably a reflection of the fact that, on average, females tend to outlive males by 5.1 years in Australia (AIHW, 2004).

Table 4.2. Sample distribution by Marital Status

	Males	Females	Total
Married	36	33	69
Widowed	11	53	64
Separated/ Divorced	3	7	10
Never married	1	6	7

Materials and Apparatus

Dementia

The *Alzheimer's Disease Assessment Scale – Cognitive* (*ADAS-Cog*: Mohs, Rosen, & Davis, 1983) was used to screen out people with dementia. This test was chosen over the frequently used Mini-Mental State Exam (*MMSE*: Folstein, Folstein, & McHugh, 1975) because it is more sensitive to severity of dementia (Rosen, Mohs, & Davis, 1984). The ADAS-Cog includes measures of recall and recognition memory, copying shapes, orientation (space and time) and naming objects. The scale is scored from 0 to 70, with higher scores indicating more cognitive impairment. Although universally agreed cut-off scores for this test are not available, a study by Weyer, Erzigkeit, Kanowski, Ihl and Hadler (1997) found that a score of 22 on the ADAS-Cog was equivalent to the cut-off score of 24 on the MMSE, which is used to indicate mild dementia. Therefore, a cut-off score of 22 was used for inclusion into this study. None of the participants was excluded based on their ADAS-Cog score, with scores ranging from 1 to 18.

Table 4.3 shows the distribution of ADAS-Cog scores for the sample split by gender and age group. First, the mean score for each of the age groups was similar for the males and females. The largest difference in ADAS-Cog scores between males and females was in the group aged between 80 and 84 and this difference approached significance (t (36) = 2.07, p = .053, d[1] = 0.72). This may suggest that as people get older, the difference between males and females becomes more apparent. This is consistent with the argument from Birren and Fisher (1992) that men show more decline later in life than females because they have a shorter lifespan. The results from the final group appeared somewhat inconsistent with this assertion but they were based on a very small sample (n = 10) and therefore should be given less weight.

[1] Effect size refers to Cohen's *d*, which is calculated by the mean difference divided by an estimate of the pooled standard deviation.

Table 4.3. ADAS-Cog score by Age and Gender

	Males		Females		Total	
Age	Mean	SD	Mean	SD	Mean	SD
70 – 74	3.87	(1.78)	4.34	(3.33)	4.21	(2.96)
75 – 79	5.20	(2.62)	4.12	(2.37)	4.59	(2.52)
80 – 84	7.08	(3.06)	5.01	(3.82)	5.61	(3.00)
85+	7.65	(4.74)	7.84	(3.26)	7.80	(3.28)
Total	5.39	(2.80)	4.73	(3.01)	4.95	(2.95)

Note. ADAS-Cog = Alzheimer's Disease Assessment Scale - Cognitive

Second, there is a clear pattern for older people to show more cognitive impairment on the ADAS-Cog task. An analysis of variance showed that the age groups had significantly different ADAS-Cog scores (F (3, 146) = 5.39, $p < .01$, partial $\eta^2 = .10$[2]). When examined in more detail it was found that the significant difference was between the people aged over 85 and the people aged 70 – 74 and 75 – 79, which suggests that people aged over 85 are quantifiably different to those in these two relatively younger groups in terms of cognitive impairment.

Third, these mean values can be compared to other studies on healthy elderly adults. Graham, Cully, Snow, Massman and Doody (2004) examined the ADAS-Cog in a normative sample of older adults aged between 55 and 89 and found that the mean ADAS-Cog score was 5.0. Given that our sample is, on average, older and has a near identical mean value, it suggests that our sample may have less cognitive impairment than the general population. There are three plausible explanations for this finding; (1) our sample had high pre-morbid intelligence, (2) our sample is declining at a slower rate or (3) a combination of both (1) and (2). Given that this sample is known to be more highly educated than the general population, this provides evidence that they probably did have high pre-morbid intelligence and this may account for the low cognitive impairment.

Inspection Time

A small cross (5 x 5 mm) was presented, in the centre of the screen, for 520 ms immediately before the target figure, to act as a warning cue (see Figure 4.1). The target figure consisted of two vertical lines, 10mm and 21mm, connected at the top by a horizontal line of 17mm. The shorter line was on the left or right with equal probability. The target figure was

[2] Partial η^2 is an effect size estimate calculated by SS effect divided by (SS effect + SS error).

presented for a short period followed by a flash mask (Evans & Nettelbeck, 1993) for a period of 375 ms. This mask consisted of two vertical lines, 24mm in length, shaped like lightening bolts. Participants were required to indicate which line was shorter, the left or right. Responses were made via the keyboard[3] or verbally (if preferred) and the next item did not appear until a response was made. If the participant chose to respond verbally then the experimenter pressed the answer on the keyboard for them. In order to avoid bias, the experimenter sat in a position from where the screen was not visible, thus, the correct alternative could not be observed. Nonetheless, the majority of people chose to press the answer on the keyboard themselves.

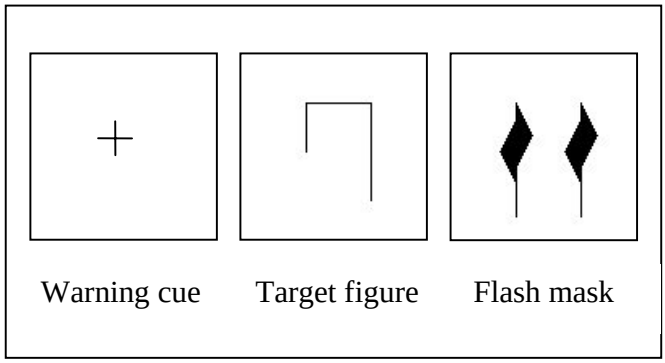

Figure 4.1. Stimuli for Inspection Time task

Life-style

In the first test session, a short questionnaire on *smoking, alcohol consumption* and *exercise* was administered (see Appendix A). The *smoking* questions classified people as current smokers, ex-smokers or non-smokers, gauged for how many years the ex-smokers and current smokers had smoked and asked how many cigarettes they smoke or did smoke per day. The *alcohol* section classified people as drinkers or abstainers and then estimated the amount of alcohol consumed per day. The *exercise* items addressed whether the participant did exercise, how often and what type of exercise they undertook.

At the same time *nutritional intake* was assessed with a diet diary, which is presented in Appendix B. On the first page, detailed information was provided about how to fill in the diet

[3] For the keyboard responses, the participants pressed the left or right button on the mouse pad built into the keyboard. They put their index finger from their left hand over the left button and the index finger from their right hand over the right button. A regular mouse was not used because of a tendency for people to push the left button since this is a learned response in general use of the mouse.

diary and participants were urged to complete the diet diary as soon as possible after they consumed any food or drink. The next four pages present a sample of a completed food diary for them to view and refer to when completing their own food diary. This was followed by a blank food diary, with room to record their food consumption for three days. The participants were required to fill in the food diary on three specific days, and these dates were entered into the top section of each page. Of the three days, two were weekdays and one was on the weekend because people tend to eat differently on the weekend.

This information was entered into a dietary database, which was supplied by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) - Human Nutrition Division, to generate various indices of nutritional intake. The major nutritional indices of interest were folate and antioxidants, as these had been linked to speed of processing in previous research (see p. 48). However, the relationship between IT and nutrition had not been investigated in the literature, therefore a range of nutritional indices were calculated. The nutritional indices generated from the diet diary were minerals (calcium, magnesium and iron), essential vitamins (vitamin A, thiamine, riboflavin, niacin, folate, vitamin B_6 and vitamin B_{12}), fatty acids (omega 3, omega 6 and VLC-omega3) and antioxidants (vitamin C, vitamin E, and β-carotene). Selection of these indices was done in conjunction with researchers from the CSIRO with the goal of representing a wide range of micronutrients.

Health

A health questionnaire provided a list of a number of diseases that are prevalent in elderly adults[4]. The participants recorded whether they suffered from any of these diseases and at what age they had first been diagnosed.

Physiological Aging

Six indices of physiological aging were obtained: grip strength, systolic blood pressure, diastolic blood pressure, height, weight and visual acuity. *Grip strength* (in kg) was measured with a dynameter. Blood pressure (BP: mmHg) was assessed using an automatic blood pressure monitor (Omron T5) to generate a measure of *systolic* and *diastolic blood pressure*. *Height* without shoes was measured with a standard tape measure and *weight* (clothed but with shoes removed) was measured with a set of digital scales. *Visual acuity* was measured binocularly

[4] This list was derived from Table 4.3 in Fry (1985).

with a Snellen chart and the participant wore their corrective glasses if applicable. Visual acuity was defined as the natural logarithm of the minimum size of letters of the alphabet (i.e. 60, 36, 24, 18, 9, 6 or 5) that the participant could read at a distance of 4 meters.

Cognitive Aging

Fluid ability.

A computerised form of *Raven's Standard Progressive Matrices* (*RSPM*: de Lemos, 1995) was used. The original 60-item test was split into odd and even questions, thereby creating two versions (Form A – odd questions, Form B – even questions) each with 30-items and the same level of difficulty. In the original test, the timed version suggests a 20-minute time limit and each of the new 30-item versions therefore imposed a 10-minute time limit. Instructions were given verbatim from the manual, with responses made via the keyboard after selecting the answer from the six or nine alternatives. Number correct for each section (*Sets A – E*) and the total score were recorded. In the first testing phase, Form A was used.

A computerised version of Scale 2, Form A from the *Cattell Culture Fair Test* (*CCFT*: Cattell & Cattell, 1959) provided a second measure of fluid ability. For each item, the picture was presented on the screen and participants selected a response from the alternatives by pressing the number on the keyboard that corresponded to their answer. All instructions were given verbatim from the test manual including the practice items before each section of the task. However, participants had to proceed through the items in order and were instructed to guess if they did not know the answer. Time limits as in the test manual were applied. Number correct was recorded for each section (*Series, Classification, Matrices, and Conditions*) and for the total score. The total score ranged from 0 to 36.

Concept Formation (*CF*: Woodcock & Johnson, 1989) was presented on the computer. Participants were presented with a group of shapes inside a box and another group outside the box. They had to deduce the rule to explain why some pictures were inside a box and others were outside. They responded verbally as in the original task. The instructions were given from the test manual and number correct ranged from 0 to 35.

Crystallised ability.

The *Information* task provided a measure of general knowledge and was adapted from the Information sub-test from the Wechsler batteries. In a previous project in the Adelaide laboratory, a general knowledge test had been developed that could be used across the whole life span and was in multiple-choice format for computer administration. Thus, a version of the

Information task with 50-items had been generated by adapting the Information sub-tests from the WAIS-III and the Wechsler Intelligence Scale for Children – III as a Process Instrument (Kaplan, Fein, Kramer, Delis, & Morris, 1999). This scale was reduced to a 40-item task for this study because the items at the beginning were too easy for elderly participants. The items became more difficult as the test progressed and a stopping rule was applied if the participant achieved fewer than two items correct in any set of six consecutive items. This test is presented in Appendix C.

In *Spot-the-Word* (Version A: Baddley, Emslie, & Nimmo-Smith, 1992), 60 pairs of words were presented, in which one word from each pair was real and one was nonsense. Participants were required to circle the real word, thereby providing a measure of vocabulary with a maximum score of 60. No time limit was applied.

In the *Similarities* task, from the WAIS-III, participants were required to describe the similarity between two objects or concepts (e.g. orange – apple). The test was administered according to the instructions in the manual (Wechsler, 1997) and scored in terms of number correct, with a maximum score of 35.

Perceptual speed.

In the *Digit Symbol* task (*DS*; Wechsler, 1997), a code (e.g. 6 = O, 7 = X) was presented at the top of the page followed by a series of random numbers with empty boxes below. The participant was required to draw the correct code in each empty box as quickly as possible and the number correct in 120 seconds was recorded.

In *Visual Matching* (*VM*; Woodcock & Johnson, 1989), participants were presented with six groups of numbers, in a straight row, and had to circle the two groups that were exactly the same. At the start, each group is made up of just one number but this increases to three by the end of the task, thereby increasing difficulty. The number correct in two minutes was recorded.

Pattern Comparison (*PC*[5]) is a 30-item, paper and pencil test. Each item consists of two line patterns presented side-by-side and participants were required to decide, as quickly as possible, whether the patterns were the same or different. Participants were given 40 seconds to complete as many items as possible and number correct was recorded.

---

[5] Pattern Comparison was generously provided by Professor T. A. Salthouse.

Everyday Functioning

Two aspects of everyday functioning were assessed: *quality* and *independence* of everyday life. *Quality of everyday life* was assessed via the Life Satisfaction Scale (Salamon & Conte, 1988). This self-report scale was chosen because it was designed specifically for elderly people and can be split into eight aspects of life satisfaction: daily activities, meaning, goals, mood, self-concept, health, finances and social contacts. Each section generated a score between 1 and 25 and the total score therefore had a maximum value of 200.

A self-report Activities of Daily Living scale was used to assess *independence of everyday life*. This scale was constructed by the author of this thesis by adapting questions from: the Instrumental Activities of Daily Living Scale (Lawton & Brody, 1988) and the Bristol Activities of Daily Living Scale (Bucks, Ashworth, Wilcock, & Siegfried, 1996). Therefore, the scale included questions on both basic and instrumental activities of daily living (see Appendix D).

Procedure

The first testing session was completed between September and December 2003. All participants had the choice of completing the test session in their own home or at the university. Of the 150 participants, 81% chose to complete the testing session at home. Those people who completed the testing session at the university (n = 29) were significantly younger than the rest of the sample (t (148) = 2.61, $p < .05$, d = 0.54) but did not differ on any other demographic variable.

First, a package of questionnaires was sent to the participant's home, with instructions that these be completed within a 2-week period. This package included questionnaires on general demographics, Life-Style, Diet, Health, Quality of Life and Activities of Daily Living. Second, a convenient time was organised for the testing session. On arrival, the author of this thesis talked with the participant for a short time to establish rapport and ensure that the participant felt comfortable. What the study would involve was explained and answers to the questionnaires were checked. Finally, the cognitive and physiological procedures were completed.

The ADAS-Cog was presented first because people with dementia needed to be excluded. The author sat opposite the participant and explained that the first task was a relatively simple test of cognitive functioning. Participants were not informed that it was a dementia-screening task. The scale was administered according to instructions in the test booklet and it took approximately 20 minutes.

Blood pressure was measured on the left arm while participants were seated. They were instructed to sit with feet flat on the ground and their arm extended on the table. Participants were encouraged to try to relax and not talk while their BP was being assessed. After each reading the participant was allowed to view their BP if they so desired. The average of three measurements was used to generate estimates of systolic and diastolic BP.

For the Spot-the-word task, participants were given written instructions, which were also read out. Practice questions were provided first to familiarise participants with the task and then the actual test was given. Participants were instructed to guess if they did not know the real word. This task took about 5 minutes to complete.

In the Inspection Time task, participants were seated in front of the laptop computer in a comfortable position. The task requirements were presented on the screen with verbal description. The response keys were explained but if the participant was uncomfortable or confused about how to respond the author operated the computer as determined by the participant's verbal answers. When this was necessary the author sat so that the computer screen was not visible. For the first practice phase, the stimulus onset asynchrony (SOA; time that the stimulus is presented on the screen before being replaced by the backward mask) was set at 830 ms. The participant continued until s/he achieved 10 correct answers in a row. Next, participants had to get 10 out of 10 at an SOA = 420 ms and finally 9 out of 10 at an SOA = 320 ms, to progress to the estimation phase. Estimation began at SOA = 320 ms and followed an adaptive staircase procedure (Wetherill & Levitt, 1965). According to this algorithm, a single error causes the SOA to increase exposure duration of the target by 17ms (one refresh rate of the monitor screen), whereas three successive correct answers are required for the SOA to decrease by 17ms. The average SOA over eight reversals of the staircase gives the inspection time estimate, which represents a probability of 79% of making a correct response. This task took approximately 15 minutes to complete.

Digit Symbol was administered according to instructions in the WAIS-III manual (Wechsler, 1997). The task was explained and, as according to instructions, a short practice phase preceded the actual test. It took about 3 minutes to complete.

Raven's Standard Progressive Matrices was the last task in this block and instructions were read out and printed on the screen. Once familiar with the task requirements, three items were practiced, the final instructions given, and the task began. If participants were

uncomfortable with the computer they verbalised their answers and the experimenter pressed the response keys. This task took approximately 20 minutes to complete.

At this stage participants were asked if they would like to stop for a 20-minute break. Most people were happy to have a break and stopped for tea or coffee. If participants wanted to continue or have a later break, this was permitted.

The second block of procedures began with the measurement of height and weight. To measure height, shoes were removed and the participant stood against a door or wall, as straight as possible. A tape measure was lined up next to them and a piece of cardboard was extended from their head to the tape measure to get a clear reading. For weight, the scales were positioned on a flat surface (not carpet) in the participant's home (or in the laboratory at the university). If the participant was unstable on his/her feet, the scales were positioned near a structure (cupboard or chair) to aid balance. Shoes remained off and participants were asked to take any heavy objects (e.g. keys or wallet) out of their pockets before weight was measured.

The Similarities task was presented verbally and took about 10 minutes to complete. A stopping rule was applied if a score of zero was given for five successive items, as instructed in the manual.

Next, the Pattern Comparison task was presented. Participants were shown written instructions that were also read out. They attempted three practice questions to familiarise themselves with the task then proceeded immediately with the task.

Concept Formation was presented next and participants simply responded verbally to this task. In each section a number of practice questions were given to familiarise the participant with the task. This was followed by the actual task and a stopping rule was applied if the participant did not get a specified number correct in each section. This task took up to 20 minutes and many participants had a lot of trouble with the concept of this task.

Visual Matching, a relatively simple task, was administered following the instructions in the manual. However, two minutes rather than three minutes were allowed because, during piloting of the task, some people complained that their eyes were blurring after two minutes.

At this stage, measurement of the participant's grip strength and visual acuity broke up the cognitive regime. This also gave the participant a chance to walk around, get a drink or go to the toilet if desired. Grip Strength was measured on each hand in turn. Participants stood up with their hand extended down to their side and were instructed to squeeze the dynameter as hard as possible then release it. Three attempts on each hand were made and from these six estimates

(three left, three right) an average grip strength measure was generated. That is, the estimated grip strength was the average of the six individual estimates.

For Visual Acuity, a Snellen chart was held at head height by the researcher, while the participant stood four metres away. The chart was positioned in a well-lit area, while avoiding glare from lights or the sun. The participant was instructed to wear distance spectacles (if applicable) and attempt to read the chart. The last line from which the participant made at least 50% correct identification was used to calculate visual acuity. These two tasks took about five minutes to administer in total.

The next test was the Cattell Culture Fair Test. The participant was instructed to sit down in a comfortable position with the laptop directly in front of him/her. All instructions were presented on the screen and read out simultaneously. The response keys were explained and if the participant was unclear about what was required, the experimenter could operate the computer, based on verbal answers. For each section (Series, Classification, Matrices and Conditions) the task was explained and practice items completed. Like the paper and pencil test, time limits were applied for each section. This task took approximately 25 minutes to complete.

Finally, the *Information* task was administered. Again, the instructions were presented on the screen and read out aloud and the response keys were explained. Participants were allowed to respond verbally if they wished. This task took approximately 10 minutes to complete, with variation depending on whether the stopping rule was employed.

For a sub-sample of participants (n = 26), the IT task was completed on a second occasion at the end of the session in order to estimate test-retest reliability. The participants were asked whether they were willing to try this task again and told that they were not obliged to do so. The participant was reminded of the task and requirements but did not do another set of practice questions. The estimation phase of the task was exactly the same as described above. If the participant became confused right at the start of the re-test, because they had not done the test since the start of the session, then the test was restarted.

Results

Table 4.4 presents a summary of the characteristics of the sample in terms of age, gender, education and disease history. The education level of males and females was similar with a tendency for people aged over 80 years to have less education than the other two groups. There was also a tendency for people aged 75 – 79 to have more education than people aged 70 to 74,

which was surprising. In terms of health, the prevalence of stoke and diabetes was low but heart disease and hypertension were more common. In both males and females, the prevalence of heart disease increased with age, and about 30% of the sample over 80 years had experience heart disease. Similarly, the incidence of hypertension increased with age, and a higher percentage of females reported having hypertension. This gender difference is interesting and might indicate a difference between males and females in reporting hypertension.

Table 4.4. Age, Gender, Education and Health characteristics of the sample

	Females			Males		
Age group	70 – 74	75 – 79	80 +	70 – 74	75 – 79	80 +
Years of Education	11.48	12.55	10.63	12.15	12.52	10.25
Stroke[1]	7%	9%	6%	8%	4%	8%
Heart Disease	17%	24%	29%	8%	35%	31%
Diabetes	13%	12%	9%	-	8%	15%
Hypertension	43%	47%	57%	17%	42%	39%
n	30	34	35	12	26	13

Note. [1] Measures of disease history were self-reported.

Descriptive Statistics

Table 4.5 presents the descriptive statistics for dementia, IT, the physiological variables and the cognitive measures in the dataset. Two points should be noted. First, there were missing data for some of the measures (IT, CCFT, CF, and BP). The reasons for this were (1) participant availability and fatigue, (2) complaints of vision problems, and (3) complaints that the BP cuff became too tight. Second, there was a clear ceiling effect on the Spot-the-word task, which suggests that on average, this sample had a high vocabulary, consistent with high pre-morbid intelligence.

There were a number of gender differences in the cognitive and physiological measures. For the cognitive measures, males performed significantly better than females on RSPM (t (148) = 4.05, $p < .001$, $d = 0.70$), CCFT (t (112) = 3.78, $p < .001$, $d = 0.75$), and Information (t (146) = 3.86, $p < .001$, $d = 0.67$). As for the physiological measures, as expected males had significantly stronger grip strength (t (148) = 14.11, $p < .001$, $d = 2.64$), were taller (t (148) = 13.71, $p < .001$, $d = 2.36$) and weighed more (t (148) = 4.59, $p < .001$, $d = 0.79$) than the females. On the IT task, males had a mean IT score of 81.73 (SD = 23.24) and the females had a mean IT score of 92.16

(*SD* = 32.47). The mean difference between males and females was statistically significant (*t* (121) = 2.14, *p* < .05, *d* = 0.35).

Table 4.5. Descriptive Statistics at Time 1

Measures	N	Mean	SD	Range
ADAS-Cog	150	4.95	2.95	1 – 18
Inspection Time (ms)	132	88.44	29.85	32 – 215
Grip strength (kg)	148	18.64	8.74	1 – 44
Systolic BP (mmHg)	137	147.98	22.70	97 – 208
Diastolic BP (mmHg)	137	79.23	11.53	56 – 119
Weight (kg)	150	70.77	12.57	43 - 105
Height (mm)	150	163.43	8.86	147 – 187
Visual Acuity (log units)	150	1.82	0.32	1.6 – 3.2
RSPM	150	16.61	4.73	7 – 26
CCFT	114	23.56	5.97	8 – 36
CF	101	21.92	7.61	0 – 35
Information	148	27.95	5.76	7 – 38
Spot-the-word	150	53.33	5.25	30 – 60
Similarities	148	22.05	4.77	8 – 32
DS	149	53.83	13.25	20 – 85
VM	147	32.46	4.95	22 – 45
PC	148	16.45	3.41	8 – 26

Note. ADAS-Cog = Alzheimer's Disease Assessment Scale – Cognitive, BP = Blood Pressure, RSPM = Raven's Standard Progressive Matrices, CCFT = Cattell Culture Fair Test, CF = Concept Formation, DS = Digit Symbol, VM = Visual Matching, PC = Pattern Comparison.

Inspection Time

There were missing data for 18 people on the IT task. This comprised people who failed to complete the task (n = 11) and people whose scores were excluded due to one or more outliers in their reversals (see Appendix E for discussion). The re-test reliability for IT was established for the sub-sample of 26 participants and was high (*r* = .826, *p* < .01). This reliability estimate is similar to the value reported in a large-scale meta-analysis by Grudnik and Kranzler (2001). However, the IT estimates for the total sample were quite variable as indicated by the standard deviation and range. Nonetheless, the mean score was comparable to that reported for a recent study using exactly the same estimation procedure with people aged over 55 (Burns, Bryan, & Nettelbeck, 2006).

Figure 4.2 shows the distribution of scores on the IT task. Although scores were normally distributed there were five scores over 175 ms that would be considered outliers. These were

dealt with in the following way. All analyses were completed on the full IT dataset (i.e. n = 132) and then with the outliers excluded (n = 127). Results are reported for the full dataset unless exclusion of the outliers led to disparate findings, in which case, this is explained and results from the smaller sample are reported.

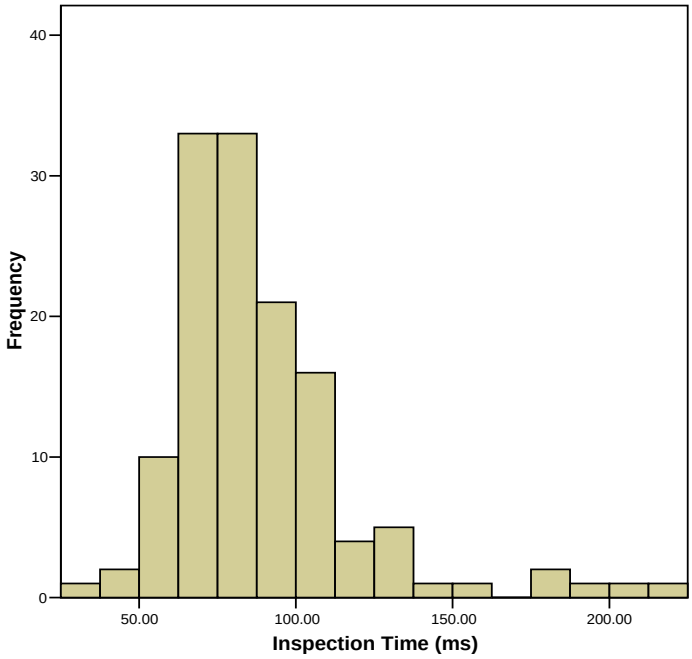

Figure 4.2. Distribution of Inspection Time estimates

Demographics

The first hypothesis (see p. 56) was that there would be a positive association between IT and chronological age. That is, older people would have longer IT scores than younger people. The data verify that there was a positive correlation between IT and age (r (130) = .210, $p < .05$), thus confirming the first hypothesis.

The second hypothesis concerned gender effects in IT scores. Specifically, the hypothesis predicted that men would show more cross-sectional change with age on IT than women. Comparing the correlations between age and IT for men and women tested this hypothesis. When the five outliers were included, the correlation between age and IT was significant in the women but not in the men. However, a visual inspection of the scatterplots showed that the IT outliers greatly influenced these correlations. Specifically, the outliers made the relationship between IT and age appear stronger in the females and weaker in the males. Therefore, results from the smaller sample (n = 127) will be reported. These results are based on 81 women and 46

men. The correlation between age and IT was non-significant in women (r (79) = .203, $p > .05$) and men (r (44) = .201, $p > .05$). The cross-sectional association between age and IT was near identical for males and females, suggesting that the correlation for the full sample represents both males and females very well. These correlations within gender were simply not significant due to a decreased sample size. To conclude, there was no evidence in this sample that males were showing more cross-sectional decline on IT than women.

The third hypothesis relating to demographic factors tested the prediction that education is not related to IT. First, the raw correlation between IT and years of education was non-significant (r (124) = -.03, $p > .05$). However, it is possible that age differences confounded the result and therefore age was partialled out and the correlation recalculated. Again, the partial correlation between IT and years of education was not reliably different from zero (r (123) = -.01, $p > .05$). Therefore, the finding met the prediction that IT is not dependent upon years of education. Moreover, consistent with theory that biological markers should not be influenced by environmental circumstances like years of education, none of the physiological tasks correlated with this variable whereas, as would be expected, years of education was significantly correlated with all of the cognitive measures (except for PC).

To summarise, IT was significantly related to age but not to education. For gender, there was no evidence that males showed more cross-sectional change on the IT task but there was evidence that males had a shorter mean IT score than females. Based on these findings, a regression analysis was run with IT as the dependent variable and age, gender and education entered as independent variables (n = 132). As expected, age was a significant predictor of IT but education and gender were not. The model accounted for just 4% of the variance in IT (R_{adj}^2 = .043, F (3,122) = 2.85, $p < .05$).

Life-Style

There were four hypotheses presented in Chapter 3 (p. 56) that related to life-style factors. These hypotheses concerned cigarette smoking, alcohol consumption, exercise and nutrition. This section deals with each of these hypotheses in turn and evaluates the observed evidence. Given that these analyses primarily involved group comparisons, it was noted that extreme IT scores could influence results. Therefore, these analyses were completed on both the full dataset (n = 132) and the smaller dataset with the outliers removed (n = 127). As expected, these outliers did lead to disparate outcomes and results in this life-style section have therefore been based on

the smaller sample of 127 people. This sub-sample was representative of the larger sample in terms of demographics (i.e. age and education).

Smoking

The hypotheses on cigarette smoking all predicted that a history of smoking would have a negative effect on IT. Specifically, the first hypothesis was that smoking status would have an effect on IT scores. It was predicted that non-smokers would have the shortest IT scores followed by ex-smokers and finally current smokers. The second hypothesis concerned the length of time during which ex-smokers and smokers had actually smoked. It predicted that years of smoking would be related to IT scores. That is, people who had smoked for a longer time would display longer IT scores than people who had smoked or did smoke for just a short period of time. It was also hypothesised that the number of cigarettes smoked per day would influence IT; heavy smoking would lengthen IT. However, the relevant question in the questionnaire was left unanswered by so many participants and there were insufficient data to analyse.

Table 4.6. Smoking Status and IT scores

Smoking Status	n	Mean	*SD*
Non-smoker	54	88.00	23.70
Ex-smoker	67	82.43	18.23
Current smoker	6	69.92	18.25
Total	127	84.21	21.00

The first hypothesis was that smoking status would have an effect of IT. Table 4.6 shows the number of people in each smoking group and their mean IT score. It is clear that a minority of people were current smokers, about half were ex-smokers, and just under half were non-smokers. In order to test the effect of smoking status on IT, an ANCOVA was used with age and pre-morbid IQ (defined by Spot-the-Word) entered as covariates, because they were both related to IT and could therefore confound the results. With respect to IT scores, current smokers were the quickest, followed by the ex-smokers and finally the non-smokers. Thus, this pattern is in the opposite direction to the hypothesis. However, the superior performance of the current smokers might be attributable to recent nicotine intake and thus would be consistent with the literature on the acute effects of cigarette smoking (see Stough et al., 2001). However, these mean differences

were not statistically significant (F (2, 122) = 2.39, $p > .05$, partial $\eta^2 = .04$) and it is therefore concluded that smoking status does not impact IT.

The second hypothesis predicted that years of smoking and IT would be positively correlated. The correlation between years of smoking and IT was examined for the sub-sample classified as either ex-smokers or current smokers (n = 73). If the non-smokers had been included then the distribution of scores would be skewed because they all reported smoking for zero years. The ex-smokers and current smokers reported smoking for between 1 and 55 years, which would correspond to most of the adult life for some. The ex-smokers reported smoking for an average of 22.9 years, while the current smokers had been smoking for an average of 42.2 years. The correlation between years of smoking and IT was non-significant (r (71) = -.121, $p > .05$). Both analyses therefore lead to the same conclusion; cigarette smoking is not related to IT performance.

Alcohol

The effect of alcohol consumption on IT was hypothesised to follow a quadratic function. That is, people at both extremes (i.e. abstainers and excessive drinkers) were expected to have slower IT scores but people who drank a little each day were predicted to have quicker IT scores. To analyse this question, participants were allocated to groups according to the extent that they reported drinking alcohol and these groups were compared on IT.

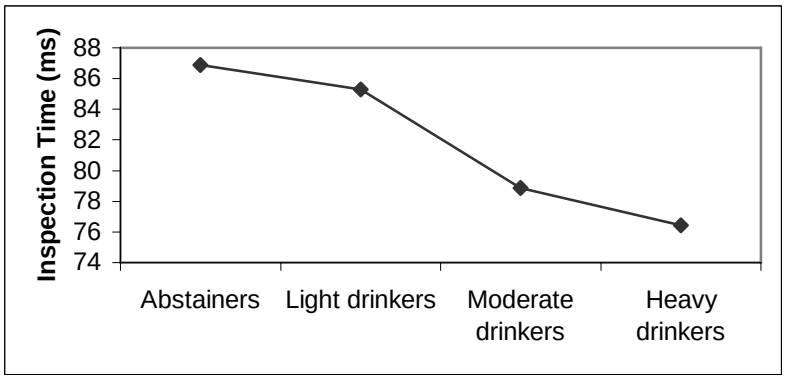

Figure 4.3. Alcohol Consumption and IT scores

The participants reported drinking between 0 and 25 standard alcoholic drinks per week and were subsequently divided into *abstainers*, *light drinkers* (1 – 6 drinks), *moderate drinkers* (7 – 13 drinks) and *heavy drinkers* (14+ drinks). Almost half (43%) of the sample were classified as abstainers, 32% were light drinkers, 19% were moderate drinkers and just 6% were classified as heavy drinkers. As previously described, the effects of alcohol consumption on IT were

examined by using an ANCOVA with age and pre-morbid IQ as covariates. Ideally, the effect of alcohol consumption on IT would have been examined separately for males and females. However, the samples were too small to make valid comparisons of IT scores for the four levels of alcohol consumption and analysis was therefore limited to the group as a whole.

Figure 4.3 shows the IT scores for the alcohol consumption groups. There appears to be a linear trend with abstainers displaying the slowest IT scores and heavy drinkers displaying the quickest IT scores. However, these mean differences were not statistically significant (F (3, 121) = 1.11, $p > .05$, partial $\eta^2 = .03$). Therefore, it is concluded that alcohol consumption within the limits estimated does not have an effect on IT performance.

Exercise

The third life-style factor of interest was exercise. First, it was hypothesised that there would be a negative relationship between time spent exercising and IT. That is, people who spent a lot of time engaged in exercise would have shorter IT scores than people who completed minimal or no exercise. Second, people who engaged in exercise that involved the cardiovascular system (e.g. walking, team sports) would have shorter IT scores than people who engaged in non-cardiovascular exercise (e.g. toning and stretches).

Table 4.7. Frequency of Exercise groups and IT scores

Exercise Frequency	n	Mean	*SD*
None	27	82.77	19.79
Brief	38	84.17	23.65
Moderate	25	88.39	19.14
Frequent	13	91.66	22.63
Extensive	7	74.38	6.17
Total	110	85.04	20.95

Unfortunately, the data for time spent exercising were incomplete, with 17 of the 127 people having missing data for this question. Therefore, this analysis was based on a group of people with full data (n = 110). The participants reported how many times per week they engaged in exercise and how long each session was. These values were used to calculate hours per week spent exercising, to give a global measure of the time spent exercising. This measure was quite variable between people and ranged from 0 to 18 hours per week. However, most people (92%) spent between 0 and 9 hours exercising per week. The sample was divided into

five groups: *non-exercisers* (0 hours), *brief exercisers* (up to 3 hours per week), *moderate exercisers* (3 – 6 hours per week), *frequent exercisers* (6 – 9 hours per week) and *extensive exercisers* (over 9 hours per week). As with alcohol consumption, it would have been desirable to examine males and females separately but the sample was too small.

Table 4.7 presents the number of people in each exercise group and it is clear that most people would be classified as *brief exercisers* spending less than 3 hours per week exercising. The second largest group were the people who reported doing no exercise. The *moderate exercisers* made up the third largest group, with two small groups classed as *frequent* and *extensive exercisers*. The hypothesis predicted that people who spent the most time exercising would have the shortest IT scores and that people who did minimal or no exercise would have the longest ITs. The data did support shortest IT for *extensive exercisers*. However, the *frequent exercisers* had the longest mean IT followed by the *moderate exercisers.* An ANCOVA with age and pre-morbid IQ as covariates was used to test the effect of exercise time on IT. Differences were not statistically significant (F (4, 103) = 1.25, $p > .05$, partial $\eta^2 = .05$). Therefore, the data do not support the prediction that frequency of exercise is related to IT performance.

Table 4.8. Type of Exercise and IT scores

Exercise Type	n	Mean	*SD*
Sedentary	28	82.89	19.44
Body toning	17	93.98	27.49
Walking	56	84.74	20.62
Sports	26	78.10	17.01
Total	127	84.21	21.00

The second issue was the role of cardiovascular exercise on IT scores. The sample reported participating in a wide range of types of exercise activities and based on this they were allocated to one of four groups. Group 1 was the *sedentary* group and consisted of people who reported no involvement in exercise and people who listed housework or gardening as their main type of exercise. Group 2 was the *body toning* group who engaged in activities including stretching, toning, weights, and hydrotherapy. Group 3 was a *walking* group consisting of people who listed walking as their main form of exercise. Group 4 was the *sports* group who engaged in a range of sports including croquet, bowls, golf, bike riding, and volleyball. For people who engaged in multiple activities, the most frequent exercise was used for classification. These four groups were ordered according to the degree of assumed cardiovascular involvement. *Body*

toning was judged to place the least demand on the cardiovascular system, followed by *walking*, and *sports*. Therefore, the hypothesis was that the *sports* group would have the shortest mean IT followed by the *walking* group, the *body toning* group and finally the *sedentary* group.

Table 4.8 shows that almost half of the exercisers reported *walking* as their main form of exercise. About 20% were *sedentary* and another 20% were involved in *sports*. The remaining 15% completed *body toning* exercises. Was there evidence of a difference in IT according to the type of exercise completed? The *sedentary* group was clearly out of place with the second shortest IT score. The other three groups were in the expected order, with the *sports* group having the shortest IT scores and the *body toning* group having the longest IT scores.

However, comparing mean IT scores using an ANCOVA, with age and pre-morbid IQ as covariates, the differences were not statistically significant (F (3, 121) = 1.90, $p > .05$, partial $\eta^2 = .05$). Therefore, the hypothesis that cardiovascular exercise has an impact on IT performance was not supported.

Table 4.9. Micro Nutritional Intake and IT scores

Nutritional Index	Range	Median	% < RDI	p-value
Calcium	195 – 5856	858.93	50.4	.724
Magnesium	97 – 786	282.03	50.4	.687
Iron	4 – 326	11.81	3.5	.653
Vitamin A	160 – 11372	890.23	36.5	.748
Thiamine	0.3 – 228	1.34	8.7	.971
Riboflavin	0.4 – 512	1.82	7.8	.668
Niacin	16 – 318	32.62	0.0	.973
Folate	49 – 3920	353.55	16.5	.199
Vitamin B_6	0.5 – 322	1.60	9.6	.720
Vitamin B_{12}	0.5 – 1135	3.41	20.0	.876
Omega 3	0.3 – 4.1	1.05	-	.455
Omega 6	1.6 – 20.5	6.90	-	.624
VLC Omega 3	0.0 – 3.6	0.23	-	.447
Vitamin C	13 – 7736	128.76	9.6	.272
Vitamin E	1 – 763	7.01	59.1	.625
β-Carotene	157 – 12383	3237.53	-	.608

Note. VLC = Very Long Chain. All minerals and vitamins are in measures of milligrams (mg) except for Vitamin A, Folate, Vitamin B_{12}, and β-Carotene, which are in micrograms (μg). The fatty acids are in grams (gm).

Nutrition

The final hypothesis with relation to lifestyle factors concerned the role of nutrient intake on IT. First, it was predicted that people with a low intake of micronutrients would have a longer IT than people who had an adequate or high intake of micronutrients. The micronutrients under investigation were the minerals (calcium, magnesium and iron), the essential vitamins (vitamin A, thiamine, riboflavin, niacin, folate, vitamin B_6 and vitamin B_{12}) and the fatty acids (omega 3, omega 6 and VLC-omega3). The second hypothesis was that people with a low intake of antioxidants would have longer IT scores than people with an adequate or high intake of antioxidants. Three important antioxidants (vitamin C, vitamin E, and β-carotene) were used as markers of antioxidant intake.

Table 4.9 presents details of the micro nutritional intake of the participants. Some of the food diaries (n = 12) were too incomplete to generate micronutrient intake measures. Data were available for 115 of the 127 participants, which represented 91% of the sample. In addition to their food consumption, about half of the participants took nutritional supplements, which were included to calculate total nutritional intake. The first column of Table 4.9 indicates the range of daily intakes of each nutrient. The second column gives the median daily intake. Because some people consumed large quantities of nutrients, the median is a better representation of central tendency than the mean in this case. The third column gives the percentage of people who were consuming less than the recommended daily intake (RDI).

There were a number of concerning trends apparent in these data. First, about 50% of people were not consuming enough calcium, magnesium or vitamin E on a daily basis. Secondly, some people appeared to be consuming excessive amounts of some micronutrients including Vitamin A, a number of B-vitamins (particularly folate and B_{12}) and Vitamin C. It was clear from the data that most of this excess consumption was due to the supplementation of the diet with nutritional pills, which is a worrying trend. Moreover, Vitamin A is fat-soluble so the body stores excess amounts of this vitamin rather than flushing it out. This means that excess intake of Vitamin A is potentially quite problematic.

Was the amount of micronutrients and antioxidants consumed related to IT? This was investigated by dividing each nutrient intake into quartiles and using an ANCOVA (with age and pre-morbid IQ as covariates) to examine whether the groups had significantly different IT scores. The last column of Table 4.9 shows the p-value for each of the ANCOVAs. None of the micronutrients was significantly related to IT performance, and none approached significance.

On the basis of these data, the hypotheses that current micronutrient and antioxidant intake have an effect on IT were rejected. This result does not, however, test the possibility that current speed of processing is influenced by past dietary choices.

Health

The hypotheses with relation to health were that the presence of age-associated diseases would be associated with longer IT scores. The diseases considered were stroke, coronary heart disease, diabetes mellitus, and hypertension. The health data were generally self-report except for BP, which was also assessed at the testing session. First, the findings from the self-reported health questionnaire with respect to stroke, coronary heart disease, diabetes and hypertension were analysed. Second, the hypertension information collected at the testing session was used to allocated people into groups and IT scores were compared. The numbers of people in each group were generally low and hence outliers had a large effect on the results. Therefore, the smaller sample (n = 127) was used for all group comparisons in this section. For each of these group comparisons, an ANCOVA was used with age and pre-morbid IQ as covariates.

Stroke. In the sample, a total of 10 people reported having a history of stoke and these are referred to here as the stroke group. The stroke group had a mean IT of 86.6 ms (*SD* = 23.2) compared with 84.0 ms (*SD* = 20.9) for the rest of the group. This difference was not statistically significant (F (1, 123) < 1.0, $p > .05$, partial $\eta^2 = .00$).

Coronary Heart Disease. In the group, 34 people reported a history of coronary heart disease. The heart disease group had a mean IT score of 89.5 ms (*SD* = 20.8) compared to the control group with a mean IT score of 82.3 ms (*SD* = 20.9). The difference was in the hypothesised direction but was not statistically significant (F (1, 123) = 2.09, $p > .05$, partial $\eta^2 = .02$).

Diabetes Mellitus. A total of 11 people reported a presence of diabetes with all but one suffering from Type II diabetes. The mean IT for the diabetes group was 84.3 ms (*SD* = 20.7) compared to 84.2 ms (*SD* = 21.1) for the control group. These groups had very similar mean IT scores and the difference was not statistically different (F (1, 123) < 1.0, $p > .05$, partial $\eta^2 = .00$).

Hypertension. In the health questionnaire, the participants were asked to indicate whether they suffered from high BP. Based on their responses, the participants were classified into a hypertensive group (n = 59; *M* = 86.1 ms, *SD* = 17.9) or a normotensive group (n = 68; M = 82.6 ms, *SD* = 23.4). There was no significant difference in IT scores between the hypertensive and

normotensive group (F (1, 123) = 1.06, $p > .05$, partial $\eta^2 = .01$). Based on self-report data, the presence of stroke, coronary heart disease, diabetes and hypertension are not related to IT.

Table 4.10. Blood Pressure and IT scores

Blood Pressure Classification	n	Mean	*SD*
Normal	11	81.78	11.57
Pre-hypertension	31	86.20	23.69
Stage 1 Hypertension	42	81.81	24.29
Stage 2 Hypertension	31	85.20	17.21
Total	115	83.90	21.30

At the testing session measures of systolic and diastolic BP were obtained and, based on these, the participants were classified into one of four groups. These data were available for 115 participants. The *Normal* group consisted of people with a systolic BP less than 120 mmHg and a diastolic BP less than 80 mmHg. The *Pre-hypertensive* group included people for whom systolic BP was between 120 and 139 mmHg or diastolic BP was between 80 and 90 mmHg. The *Stage 1 Hypertensive* group included people with a systolic BP between 140 and 159 mmHg or diastolic BP between 90 and 99 mmHg. Finally, the *Stage 2 Hypertensive* group had a systolic BP above 160 mmHg or a diastolic BP over 100 mmHg.

Table 4.10 shows the number of people in each BP group. There were relatively few classified as *Normal* in this group, which is not surprising since BP is known to increase with age. The other three groups were relatively even in terms of sample size. With respect to IT, the normal group had the shortest IT but the other groups were not as expected. These differences were not statistically significant (F (3, 109) < 1.0, $p > .05$, partial $\eta^2 = .01$). Therefore, both the self-report and directly measured data led to the conclusion that hypertension does not have an effect on IT. To summarise, there was no evidence that age-associated diseases were associated with IT.

Physiological Aging

It is hypothesised that if IT is a biological marker of aging then it should be related to physiological aging indicators. In Table 4.11 the correlations between IT and the physiological and cognitive measures are presented. Before running the correlations, the effects of gender and education were partialled out because these variables confound the relationships between the

physiological and cognitive measures. The physiological measures in this study were grip strength, systolic BP, diastolic BP, height, weight and visual acuity.

From Table 4.11, it is clear that IT does not correlate significantly with any of the physiological measures. However, there are a number of significant inter-correlations between the physiological measures. First, as would be expected weight and height were significantly correlated (r (117) = .444, $p < .001$). Second, systolic and diastolic BP were significantly related (r (105) = .673, $p < .001$), as expected. Third, weight was positively correlated with grip strength, visual acuity, and diastolic BP. That is, heavier people tended to have more strength in their hands, poorer visual acuity and higher diastolic BP. The relationship between weight and visual acuity is unexpected and may be informative. Perhaps, people with poor visual acuity are limited in some of their daily activities, are more sedentary and as a result weigh more. Finally, taller people also tended to have stronger grip strength and this probably reflects a tendency for bigger (heavier and taller) people to have more strength in their hands. Another interesting finding was that none of the physiological measures was significantly related to age. This suggests that, in this age group (i.e. 70+), there was very little evidence that these physiological measures are declining with advancing age. Although there is evidence in the literature of a decline in physiological measures over the lifespan (e.g. 20 – 80 years), this pattern was not apparent in the restricted range in this study. To conclude, there were a number of relatively small correlations between the physiological measures in this study but IT was not significantly related to any of them.

Outcome Measures

There were three main outcome measures: everyday functioning, quality of life and cognition. Although the ultimate test of the biomarker will be the predictive validity, this section examines the concurrent relationship between IT and the outcomes. Each outcome measure was considered in turn, in order to see how effective IT was at explaining the total variance. In each case IT was compared with chronological age and the other physiological variables that have often been used as biomarkers in the past.

These analyses were run with and without the five IT outliers and this did lead to disparate findings. However, it was decided to leave in the outliers for the assessment of the outcome measures for the following reasons. First, large outliers have less of an impact on correlation and regression analyses than on statistical methods that compare mean differences

Table 4.11. Correlation matrix for physiological and cognitive measures at Time 1

1. RSPM	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.	16.
2. CCFT	**.642**															
3. CF	**.431**	**.397**														
4. Info	**.401**	**.437**	**.235**													
5. STW	**.315**	**.389**	**.489**	**.426**												
6. Similar	**.412**	**.375**	**.404**	**.318**	**.384**											
7. DS	**.421**	**.591**	**.452**	**.211**	**.246**	**.286**										
8. VM	**.356**	**.455**	**.415**	**.186**	**.364**	**.200**	**.658**									
9. PC	**.351**	**.485**	**.346**	**.247**	**.294**	**.258**	**.635**	**.575**								
10. GS	**.187**	.149	-.053	.173	.129	.049	.162	.005	.145							
11. SBP	.154	-.010	.004	-.004	.066	.095	-.054	.008	.011	.010						
12. DBP	**.235**	.200	.049	-.126	.068	.040	.044	.058	.050	.128	**.673**					
13. Weight	**.214**	.193	.180	-.045	.088	.085	.119	.114	.215	**.217**	.131	**.213**				
14. Height	-.049	.123	.096	-.164	-.050	-.015	.095	-.002	.067	**.197**	-.092	.040	**.444**			
15. VA	**.219**	.207	.207	.047	.075	.112	.174	.108	**.261**	.116	.047	.096	**.198**	.158		
16. IT	.170	**.306**	**.320**	.124	**.241**	.110	**.334**	**.278**	**.359**	-.007	.005	-.004	.017	.114	.049	
17. Age	-.159	**-.263**	-.185	-.079	.036	-.040	**-.317**	-.117	**-.253**	-.094	.024	-.157	-.157	-.030	-.124	**-.208**

Note. RSPM = Raven's Standard Progressive Matrices, CCFT = Cattell Culture Fair Test, CF = Concept Formation, Info = Information, STW = Spot-the-Word, Similar = Similarities, DS = Digit Symbol, VM = Visual Matching, PC = Pattern Comparison, GS = Grip Strength, SBP = Systolic Blood Pressure, DBP = Diastolic Blood Pressure, VA = Visual Acuity, IT = Inspection Time.
Visual Acuity and Inspection Time have been reflected to make correlations positive. Gender and education have been partialled out.
Significant correlations ($p < .05$) are shown in **bold**

such as t-tests or ANOVA. Second, the five people with high IT scores also had scores on all three of the perceptual speed tasks that were consistent with them actually experiencing some degree of decline in speed of processing. This implies that the high IT scores for these five individuals may actually be informative about central nervous system slowing, rather than representing some problem with the IT task. Third, most of the people with high IT scores (n = 4) also showed impairment on Raven's Standard Progressive Matrices and the Cattell Culture Fair Test, suggesting that they may be experiencing impairment on more global cognitive measures. All of these points led to the conclusion that the five IT outliers were indeed informative about the aging process. Furthermore, in a reasonably healthy sample of elderly people, it is the people with the most impairment who are likely to be the most informative about the processes associated with physiological, cognitive and everyday aging. Therefore, the five IT outliers were included in all of the analyses presented below.

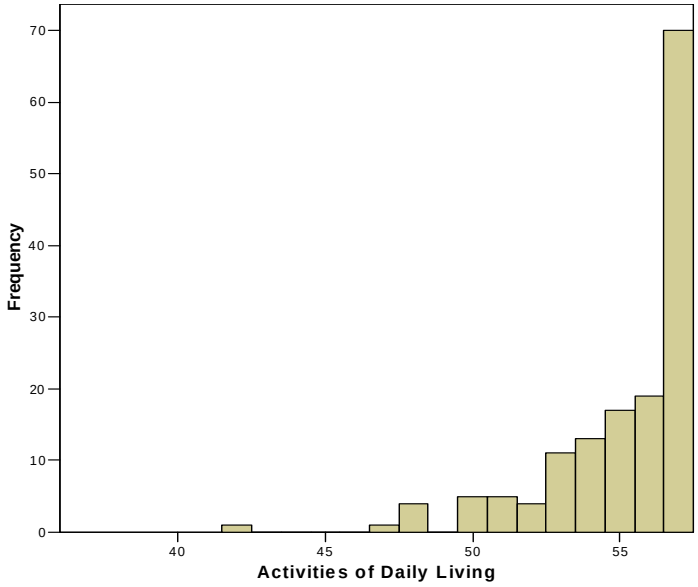

Figure 4.4. Activities of Daily Living at Time 1

Everyday Functioning and Quality of Life

The Activities of Daily Living (ADL) scale was used to provide an index of everyday functioning. This scale produced scores between 0 and 57 with higher scores representing more independence in everyday life. Figure 4.4 shows the distribution of scores on the ADL scale. It is clear that there was a marked ceiling effect on this scale, with approximately half of the sample getting the maximum score for independence. The remaining scores were between 42 and 56, suggesting that, on the whole, the sample was very independent. Furthermore, it suggested that

the scale was inadequate in discriminating between levels of independence of everyday functioning in this high functioning sample.

Quality of Life was assessed using the Life Satisfaction Scale, which has been designed for use in the elderly. For this scale, there was a much wider distribution of scores, no indication of ceiling effects and the scores were normally distributed. The mean score for quality of life was 146.7 (*SD* = 14.7, range = 101 – 186).

The question with respect to these outcome measures is whether IT can explain a significant amount of the variance in them. If IT can explain some of the variance, it is important to see whether it does so better than chronological age and the other physiological variables. It is also interesting to see whether IT explains a similar amount to the perceptual speed tasks. There were missing data for some of these variables, so the analysis was done on the sample of people who completed all the 13 relevant measures (n = 117).

Table 4.12. Predictors of Everyday Functioning and Quality of Life

	% Total variance	
Predictor	Everyday Functioning	Quality of Life
Age	10.23**	3.21
Inspection Time	7.91**	1.62
Grip Strength	9.39**	0.56
Systolic BP	0.21	1.55
Diastolic BP	0.15	3.33
Weight	5.62*	3.81*
Height	1.02	3.93*
Visual Acuity	1.23	0.00
Digit Symbol	14.56**	0.00
Visual Matching	6.42**	1.74
Pattern Comparison	7.25**	0.17

Note. * $p < .05$, ** $p < .01$

The methodology was as follows. Correlations were run between the outcome measures and the eleven predictors after partialling out the effects of gender and education because these are known to impact on cognitive and physiological measures. These correlations were then converted to represent percentage of total variance explained (i.e. r^2 x 100) and entered into the table. This allowed for the comparison of how much variance each of the predictors could explain once gender and education were accounted for. This method was also used for the fluid ability outcomes. To re-iterate, the hypothesis was that IT would be a better predictor of the outcomes than age and the physiological measures. The variance accounted for by the perceptual

speed tasks was calculated for interest sake but due to the problems with these measures we would not advocate using them as biomarkers of aging.

From Table 4.12 it is clear that there are four significant predictors of everyday functioning: age, grip strength, IT, and weight. First, age is the best predictor, so younger people are functioning more independently than older people in their everyday lives. Second, grip strength is a predictor, which suggests that people who have maintained the strength in their hands are more independent in their everyday functioning. Third, IT is related to everyday functioning so people with shorter IT scores are more independent. Finally, weight is a predictor, which suggests that people who weigh less are more independent in their everyday lives. The other physiological variables (visual acuity, BP, and height) do not explain a significant amount of the variance in everyday functioning. The perceptual speed tasks all explained a significant amount of the variance in everyday functioning and the magnitude of this effect was about the same as or better than IT. This analysis has shown that IT is not as effective as age at explaining the variance in everyday functioning. However, it remains to be seen whether IT and age explain the same variance in the outcome and this issue is considered below.

Table 4.13. Hierarchical Regression for Everyday Functioning

Predictor	β	*t*	R^2	R^2 change
Step 1				
Gender	.241	2.61*		
Education	.057	0.62	.063	
Step 2				
Age	-.314	-3.53**	.159	.096*
Step 3				
IT	-.205	-2.96*	.198	.039*
Step 4				
Grip strength	.536	3.95**		
Weight	-.354	-3.95**	.357	.158**

Note. * $p < .05$, ** $p < .01$

An alternative method to consider the concurrent validity of IT is to examine whether IT remains a significant predictor of everyday functioning once age has been entered into a regression equation. If IT remains a significant predictor, then it would suggest that IT and age are explaining different aspects of everyday functioning and that IT may indeed be a useful

biomarker. This was achieved using hierarchical regression with everyday functioning as the dependent variable and the results are presented in Table 4.13.

In the first step, gender and education were entered as independent variables since these variables are known to impact everyday functioning and physiological measures but also for consistency with Table 4.12. This model accounted for 6% of the variance in everyday functioning. In the second step, age was entered resulting in a significant improvement in R^2 and a model that explained 16% of the variance. The third step was to enter IT and the crucial question was whether the inclusion of IT at this stage would cause a significant R^2 change. IT was a significant predictor of everyday functioning and the R^2 change was indeed significant. Finally, the physiological variables of grip strength and weight were entered and both were significant predictors with the final model accounting for 36% of the variance (R^2 = .357, F (6,106) = 9.80, $p < .001$). The results of this analysis suggest that all three of the biomarkers (IT, grip strength and weight) contribute unique information and independently predict the outcome of everyday functioning. This suggests that IT may indeed be a useful biomarker for everyday functioning and reinforces the idea that a range of biomarkers are more useful that any single one.[6]

The second outcome variable was quality of life and it is clear from Table 4.12 that just weight and height were significantly related to this measure. In both cases, the relationship only just reached significance and was negative. That is, shorter and lighter people had a higher quality of life. Because gender was partialled out this was not a gender effect. Nonetheless, IT was not related to quality of life and therefore the hypothesis was not confirmed.

Cognition

There were varying degrees of missing data for the three fluid ability tasks. Consequently, the analyses were done on the people who completed IT, the physiological measures, the perceptual speed tasks and the individual fluid tasks. This corresponded to 119 for Raven's Standard Progressive Matricis (RSPM), 92 people for Cattell Culture Fair Test (CCFT) and 84 people for Concept Formation (CF).

---

[6] This analysis was re-run with the perceptual speed tasks entered as a final block after grip strength and weight. None of the perceptual speed tests were significant predictors of everyday functioning. That is, the variance in everyday functioning that was related to perceptual speed could be explained by the other predictor variables.

Table 4.14. Predictors of Fluid Ability

	% Total variance		
Predictor	RSPM	CCFT	CF
Age	3.55*	4.55*	9.87*
Inspection Time	7.97**	28.51**	10.42**
Grip Strength	4.49*	2.94	0.69
Systolic BP	2.57	0.76	0.14
Diastolic BP	6.05**	6.97**	1.43
Weight	1.05	0.39	0.12
Height	0.04	2.93	2.69
Visual Acuity	9.01**	12.41**	11.24**
Digit Symbol	20.77**	34.46**	22.27**
Visual Matching	20.03**	25.99**	18.47**
Pattern Comparison	10.13**	16.83**	12.23**

Note. * $p < .05$, ** $p < .01$

How effective was IT at explaining the total variance in the fluid ability task? First, IT was better than chronological age in all cases. Second, visual acuity consistently explained a significant amount of the variance and, in two cases, was a better predictor than IT. Third, in two of the three outcomes diastolic BP was a significant predictor of fluid reasoning. However, from the raw correlations in Table 4.11 it is clear that higher diastolic BP scores were associated with higher fluid reasoning scores. This is unusual, because past research has suggested the opposite. To conclude, IT explained a significant amount of total variance in the fluid reasoning measures and was always a better predictor than chronological age.

One question that arises from Table 4.14 is the degree to which IT and visual acuity are explaining unique or shared variance in the fluid ability tests. It is clear that visual acuity would have an impact on IT performance but if IT is a useful biomarker then it needs to explain some additional variance in the fluid ability tasks since visual acuity is considerably quicker and easier to assess. This was investigated using hierarchical regression analyses, which are presented in Appendix F, for each of the fluid ability tests separately. At Step 1, gender, education and age were entered as independent variables. In Model 1, visual acuity was entered at Step 2 and IT was entered last. In Model 2, this order was reversed, with IT entered at Step 2 and visual acuity entered last. This allowed for the examination of the unique and shared variance that IT and

visual acuity contributed to the fluid ability tests once gender, education and age were accounted for[7]. These results are presented in Table 4.15.

Table 4.15. Unique and Shared Variance between IT and VA on Fluid Ability

Type of Variance	RSPM	CCFT	CF
Shared	2.1%	4.0%	2.3%
IT unique	2.9%	14.8%	3.0%
VA unique	3.4%	2.4%	3.2%
Total IT-VA	8.4%	21.2%	8.5%

Note: IT = Inspection Time, VA = Visual Acuity, RSPM = Raven's Standard Progressive Matrices, CCFT = Cattell Culture Fair Test, CF = Concept Formation.

In all three fluid ability tasks, there was some variance that was shared between IT and visual acuity (2 – 4%) but there was also some variance unique to IT. This was most pronounced in CCFT where IT explained 14.8% unique variance in addition to that explained by gender, education, age and visual acuity. In the other two tasks, the unique variance explained by IT was small. However, the hierarchical regression for RSPM showed that IT was a significant predictor even in Model 1 where it was entered last. For CF, IT was non-significant when entered last and this must be attributed to the smaller sample size, given that the β-value was nearly identical to that from RSPM analysis. Therefore, it is concluded that IT does explain additional variance in the fluid ability tasks after visual acuity has been entered but, although the unique variance was considerable for CCFT, it was small in the RSPM and CF.

Another interesting finding was that the perceptual speed tasks explained much more variance in RSPM and CF than did IT. One of the reasons that perceptual speed tasks are problematic for use as biomarkers is that they suffer from cohort effects and are confounded by psychomotor speed. Furthermore, some of these tasks involve other cognitive skills such as attention and memory. Therefore, it is not surprising that they can explain more of the variance in the fluid ability tasks, given that they (1) account for some of the differences between cohorts, (2) explain some degree of physical slowing and (3) they share a number of components (i.e. attention) with reasoning tasks. However, this does not imply that the perceptual speed measures are better biomarkers than IT. As discussed by Deary (2001), perceptual speed tasks are not elementary speed measures, when compared with IT, and at an explanatory level are not very

[7] Details of these calculations are presented in Appendix F.

useful. Rather, perceptual speed tasks are relatively complex and bound to share variance with higher order cognitive abilities factors. Furthermore perceptual speed tasks fail to meet a range of requirements for biomarkers. Thus, although it is interesting to see how much variance they can explain, they are not considered analogous to the biomarker variables in this study.

Discussion

In this first cross-sectional study, the main aim was to explore the relationship between IT in a group of healthy elderly adults and age-associated factors, markers of physiological aging and various outcome measures. Several factors are thought to accelerate the aging process (e.g. smoking) or help maintain healthy aging (e.g. exercise) and these factors are referred to as age-associated factors. The age-associated factors investigated here were chronological age, gender, education, smoking, alcohol consumption, exercise, nutrition, and disease (stroke, coronary heart disease, diabetes, and hypertension). The markers of physiological aging were grip strength, BP, weight, height and visual acuity. In terms of outcomes, we examined whether IT was more useful than chronological age and the physiological measures in explaining variance in everyday living, quality of life, and cognition. The discussion will be split into three sections: age-associated factors, markers of physiological and cognitive aging and outcome measures.

IT and Age-Associated Factors

Chronological age. A number of demographic variables were hypothesised to have an effect on the aging process and therefore IT, the first being chronological age. The first hypothesis was that older people would have longer IT scores than younger people. There was a significant correlation between chronological age and IT so that, cross-sectionally at least, there is a trend for IT to become longer as age increases.

Gender. Second, it was hypothesised that gender should have an effect on the rate of decline in IT with age. The theory from Birren and Fisher (1992) was interpreted to suggest that, because females live longer than males on average, older males would decline on the IT task more than older females. At this stage, the data are purely cross-sectional but there was no evidence to support this hypothesis, the correlations being near identical for males and females. There are at least two possible explanations for this finding. First, because males do not on average live as long as females, the current sample might represent higher functioning males, who are by definition therefore not declining as much as the average older male. There was some evidence that the current male group was more homogeneous because their IT scores were less

variable than scores for females (see p. 70). Second, the cross-sectional nature of the investigation might not be a good representation of the true longitudinal decline with age. The one study that did find evidence supporting the gender effect (Mortensen & Kleven, 1993) used longitudinal data over a 20-year period.

Education. The third hypothesis with relation to demographics concerned the role of education on IT. In general, biological variables (e.g. grip strength) do not correlate with education, although cognitive variables often do. To validate IT as a biomarker is it important to show that it is unrelated to education. The results supported this (null) hypothesis, with the correlation between IT and years of education being near zero. However, it should be noted that this sample was highly educated compared with the general population and thus this outcome has been found in a sample that is skewed towards higher levels of education with the restriction in the range attenuating the correlation.

Smoking. There were a number of hypotheses about the impact of life-style factors on IT, which proposed that some life-style choices (e.g. smoking) should accelerate age-related decline while others (e.g. exercise) should slow down or stabilise age-related decline. Smoking is thought to accelerate the aging processing through an increased risk of various diseases. Therefore, it was hypothesised that current smokers and ex-smokers should have longer IT scores than non-smokers. However, there is another line of research suggesting that the acute effects of nicotine can enhance speed of processing performance and it was not therefore entirely clear what, if any, effects would emerge concerning IT and cigarette smoking. The results showed that the current smokers tended to register faster IT estimates but the mean differences were not significant. Possible explanations for this null result are, first, insufficient power in this dataset. Only six participants were classified as current smokers. Almost all of the previous research on smoking and speed of processing compared current smokers and non-smokers but there was insufficient power for this comparison in this dataset. Second, it may not be valid to group all ex-smokers together, in effect assuming that they are a homogeneous group with respect to their smoking behaviour. Differences in the number of cigarettes consumed per day, the number of years for which they smoked and the period of time since they stopped smoking may confound the results. Whalley et al. (2005) have suggested that the effects of smoking on cognition may depend on prolonged exposure to cigarettes or exposure during later life. If this is the case then future research may benefit from examining in much more detail the smoking behaviour of the

ex-smokers. Nonetheless, the current results found no association between cigarette smoking and IT scores.

Alcohol. The effect of alcohol on aging and cognition is not straightforward. Although it is clear that excessive alcohol consumption is detrimental to cognitive performance and health, there is evidence that a moderate intake of alcohol is more beneficial than complete abstinence (e.g. Kalmijn et al., 2002). Therefore, it was hypothesised that people who abstained or drank a lot would have longer IT scores than those who drank a little per day. There was an apparent linear trend in the current data, with abstainers having the slowest IT scores and heavy drinker showing the quickest IT but this difference was not statistically significant and the effect size was very small. One limitation of this study with respect to this issue was the lack of heavy drinkers. However, there may not be many heavy drinkers amongst the elderly and a relationship between IT and alcohol consumption may need to be established in a middle-aged group. To conclude, there was no evidence of an association between IT and alcohol consumption in this sample.

Exercise. The third life-style factor to be investigated was exercise, and it was hypothesised that people who engaged in exercise should have shorter IT scores than sedentary people. This was investigated by comparing IT scores based on (1) time spent exercising and (2) type of exercise completed. Firstly, the hypothesis that time spent exercising has an effect on IT scores was not supported. Although extensive exercisers did have the shortest IT scores, this was followed by sedentary adults, brief, moderate, and finally frequent exercisers. Thus, the pattern was not as expected, leading to the conclusion that time spent exercising does not impact on IT scores.

The second hypothesis was that people who engaged in exercise that involved the cardiovascular system (e.g. bike riding, tennis) would have shorter IT scores than people who did exercise such as body toning and stretches. The sample was split into four groups: body toning, walking, light sports and active sports, according to the assumed degree of cardiovascular involvement in the exercises. There was a trend in the hypothesised direction among those people who reported engaging in exercise but the sedentary group also had relatively short IT scores. Furthermore, these differences were small and not statistically significant.

One possibility is that current exercise behaviour may not be indicative of exercise behaviour in the past. There is some suggestion that early life exercise is particularly important to later life cognition (e.g. Dik, Deeg, Visser, & Jonker, 2003) and it is quite plausible that a change in exercise behaviour over time may occur. For example, it is possible that some people

who are currently sedentary (i.e. due to injury or health ailment) may have been quite active for most of their lives and this may influence results.

Another possibility is that differences between individuals in the amount of "incidental exercise" undertaken might be important and not detected by the exercise questionnaire. While some individuals report that they do not engage in any formal exercise, they may still work around the garden, take the stairs and so on. If participants do not report incidental exercise in the exercise questionnaire, then the measures of exercise behaviour that are subsequently derived may not be complete. Nonetheless, this study has shown that current level of exercise is not related to IT performance.

Nutrition. The final hypothesis with respect to life-style factors was that nutritional intake should be related to IT scores. Although folate and antioxidants have explicitly been shown to effect speed of processing, few studies have used speed as an outcome measure. Therefore, this current study used a broad range of nutrients and antioxidants to see whether IT performance was dependent on nutritional intake. However, not one of the 16 nutrients examined had a significant effect on IT scores.

Possible explanations of this null effect are, first, the method of assessing nutritional intake from food diaries may not have been sufficiently accurate. Although this method gives an indication of the nutrients that are being consumed, it does not take into account the absorption of those nutrients. For a number of reasons (e.g. disease processes), elderly people do not absorb a number of key nutrients as well as younger people (see Russell, 2001). Therefore, it would be necessary to measure blood levels of nutrients because this indicates how much has actually been absorbed rather than just consumed. Second, the intake of nutrients throughout the lifespan will impact on the aging process but the current diet may not be indicative of dietary habits in the long term. For example, there are a number of circumstances (such as dentition problems, reduced appetite, and eating alone) that might explain why the diets of elderly people are qualitatively different from those of younger people (see Cobiac & Syrette, 1995). Nonetheless, it is possible that IT is simply not related to nutritional intake. Of the three other speed tasks, only one suggested a significant association, which was between PC and iron intake. Therefore, in this study, similar perceptual speed tasks to those used in other nutritional studies failed to provide convincing evidence for association between speed and nutrition, contrary to earlier findings. This could mean that the methodology, rather than the IT measure, was responsible for the null effect.

Health. A number of diseases are more common in the elderly and are known to accelerate the aging process. Therefore it was hypothesised that people with the presence of age-associated diseases should be impaired on the IT task. The role of stroke, coronary heart disease, diabetes and hypertension on IT were all studied. For all of the self-reported diseases, the disease group showed slower means IT but these differences failed to reach significance.

The main problem for this investigation was that very few people had a history of these disorders. Therefore, an essentially healthy group was compared to a small sample of people with a particular disease history. For example, just 34 people reported a history of coronary heart disease and although they displayed slower IT scores, their mean was not significantly different to the majority group. In order to do this type of analysis it is essential to have a larger group. Alternatively, if the participants were to complete health questions that generated continuous variables (rather than categories) then the power issue would be improved.

IT and Physiological Markers

Six measures were included to represent physiological aging because decline in a range of physiological measures as people age is inevitable and these measures might therefore be able to provide an "index" of physiological aging. For example, as people get older their strength declines, they often lose weight and height, experience more sensory problems (e.g. impaired visual acuity) and tend to experience higher BP. There were a number of significant correlations between the physiological measures, which could be taken to suggest that they are providing an index of physiological aging. However, none of these physiological measures showed a significant correlation with age. Therefore, the inter-correlations may be telling us nothing about the aging process and simply that these indices are related to one another independent of age.

The main reason for including a range of physiological markers was a statement by Birren and Fisher (1992) that a valid biomarker should be related to other physiological measures. Our analyses showed that IT was not significantly related to any of the physiological measures. This was a surprising finding and requires some contemplation. One important consideration is whether the physiological measures are themselves valid biomarkers because it does not seem logical to require that IT be related to other biomarkers if they have not been previously validated. Detailed validation of these physiological variables with respect to formal criteria did not appear in the literature reviewed. However, a number of these physiological measures have been shown to predict cognition, everyday functioning, or mortality. Grip strength, blood pressure and visual acuity have all been linked to cognitive performance in later life (Anstey,

Lord et al., 1997; Baltes & Lindenberger, 1997; Swan et al., 1998). All six of the physiological measures have been linked to everyday functioning (Davis, Ross, Nevitt, & Wasnich, 1999; Guo, Viitanen, & Winblad, 1997; Judge, Schechtman, & Cress, 1996; Marsiske, Klumb, & Baltes, 1997; Tully & Snowdon, 1995). Grip strength and visual acuity, at least, have been linked to mortality (Anstey, Luszcz, Giles et al., 2001). Based on these results it seems that these physiological measures are probably valid biomarkers, and perhaps it is valid to claim that IT should be related to them. One possible explanation for the lack of correlation between IT and the other physiological variables is discussed below.

There is some evidence from the literature that difference systems (e.g. sensory, muscular) begin to decline at different points in time and at different rates. Therefore, we should certainly expect systolic and diastolic BP to correlate highly because they are marking the same system (i.e. cardiovascular health). However, we should not necessarily expect IT to show a correlation with these variables because it is theorised to measure a different system (i.e. speed of the central nervous system), which may be declining at a different rate. IT does show a significant correlation with the other speed measures, which, despite their problems, would also be theorised to measure CNS slowing. To conclude, various systems are expected to start declining at different times and rates. This might explain the lack of correlation between IT and the other physiological measures, because IT is purported to mark a different system to the other measures. Nonetheless, it a still a little concerning that IT appears to have no relationship with these physiological measures.

IT and Outcome Measures

The most important test of a biomarker is its ability to predict outcomes in the future and this will be investigated in Chapter 7. However, at this stage, the concurrent relationships between IT and the outcome measures were investigated, in the expectation that this would give some hint as to the efficacy of IT as a biomarker.

Everyday Living. An activities of daily living scale was used to measure everyday living. Four measures (age, grip strength, IT and weight) were effective in explaining the variance in this outcome measures. That is, younger people, those with strong grip, shorter IT estimates and weighing less are more independent in their daily lives. All of these relationships are sensible and in the expected direction. However, the pertinent question was whether IT was more informative than chronological age, in predicting performance on the activities of daily living task. A regression analysis showed that even after the variance associated with gender,

education, and age had been accounted for, IT was still a significant predictor of activities of daily living. That is, IT made a significant improvement to the regression model. In addition, grip strength and weight were also significant predictors suggesting that a range of biomarkers can be more informative than any single one. Thus, the evidence if positive for IT and suggests that it can indeed provide information about performance in everyday activities. However, there were a number of problems with the activities of daily living scale, which should be addressed.

First, the everyday living scale was not ideal because there was a clear ceiling effect, with approximately half the sample getting the highest score. In a less able sample, it is likely that this scale would be useful but in our high functioning sample it simply does not differentiate adequately between people. Second, the scale was not normally distributed and could not be converted to a normal distribution by any transformation attempted. Therefore, the scale was not an ideal variable to use as an outcome measure. Ideally, we needed an everyday living questionnaire that is normally distributed in the population. That is, most people should score around the mean value, with some people showing more independence than most at their age and other people needing more help in their everyday functioning than most. In subsequent phases of this study, an alternative scale will be sought.

Quality of Life. The second outcome measure was quality of life and the scale used was quite effective in discriminating between people. Scores were normally distributed, with an acceptable range. Weight and height explained a significant amount of variance in this outcome measure but only just reached significance. None of the other predictor variables was significant. This suggests that quality of life is largely determined by variables other than age, IT, and the physiological measures. Although it might be possible to mark the aging process by these types of variables, they were not effective at all in predicting quality of life. Given these results are very clear, the decision was taken not to continue with this outcome measure but, instead, to focus on independence of everyday living and attempt to measure that variable more effectively.

Cognition. Fluid reasoning was adopted as an outcome measure for cognition. There were a number of measures that explained a significant amount of variance in fluid ability but the measures that were consistently useful (i.e. explained significant variance in all three fluid ability tests) were IT, visual acuity, and age. Overall, the most effective measure was clearly IT. Furthermore, hierarchical regression showed that the inclusion of IT produced a significant improvement to the regression model, even when gender, education and age were already entered. In all three fluid ability tasks, visual acuity accounted for a substantial proportion of the

variance and the question was posed as to whether IT was explaining any additional variance after visual acuity. Hierarchical regression confirmed that, for RSPM and CCFT, IT explained a significant proportion of unique variance not related to visual acuity. These analyses confirmed that IT is an important predictor of fluid ability and can explain variance additional to gender, education, age and visual acuity. This is an encouraging result but, as discussed above, although this establishes concurrent validity, the major test of IT as a marker of cognitive decline depends on predictive validity

General Conclusions

The findings for IT are encouraging and certainly justify further investigation. IT is related to chronological age, independently from education. IT is not related to lifestyle or disease but this may in part be due to power limitations when the sample was split into groups and to the generally healthy nature of the sample. IT has explained a significant amount of variance in everyday living and cognition. These findings therefore lead to the conclusion that IT does decline with advancing age, is independent of education, and it useful in explaining variance in everyday living and cognition. Therefore, it is proposed that these concurrent results warrant an investigation of the predictive validity of IT as a biomarker of aging. However, due to the findings of this study, changes to the test battery will be made, and these will now be briefly outlined.

First, there was little evidence of a link between IT and life-style in this sample so life-style will not be included further. In some cases, this result may have been due to the methods used (e.g. diet diaries vs. blood samples). In other cases, it was probably due to the nature of the healthy elderly sample (e.g. few current smokers or heavy drinkers). Regardless of the cause, the problems cannot be rectified in this study and this link will not be investigated further. Second, both the original and a new questionnaire will be administered to assess everyday living. It is hoped that the new questionnaire will more adequately discriminate between the everyday activities of this sample. Finally, the outcome variable of quality of life will no longer be examined because no evidence for the efficacy of IT (or any of the other variables) in predicting it has been found.

CHAPTER FIVE: STUDY 2 - RELIABILITY AND STABILITY OF THE BIOMARKERS

The aim of this chapter is to consider the stability and reliability of the biomarkers over a period of 6-months. For each of the biomarkers, there are two variables of major concern: the initial value and the 6-month change score. The reason that they are important is that these variables will ultimately be used as predictors for a range of functional outcomes at the end of the study. This chapter will deal with the following questions.

1) How reliable are the initial values?
2) How reliable are the change scores?
3) How stable are the constructs over a 6-month period?
4) Are there individual differences in stability?
5) Are there gender differences in the stability of the biomarkers?

In order to answer these questions, there are a number of statistical issues that must first be resolved. These are; to determine methods for estimating reliability; how best to calculate change scores; and problems inherent in the use of change scores. Answers to these questions will permit considerations of whether the initial score and 6-months change scores have potential as predictors of functional age.

Method

Participants

Of the original 150 participants, 137 completed the second testing phase, which represents an attrition rate of just 9%. There were 87 females (mean age = 78.0, *SD* = 4.7) and 50 males (mean age = 77.7, *SD* = 3.6) with ages between 71 and 92 years. Of the 13 people who discontinued, all but one were females. This suggests that there was a gender difference in the attrition rate. There were a number of reasons given for discontinuation, including health problems (with themselves or spouse), moving interstate and disinterest.

We were interested in whether the 13 people who did not continue participation were substantively different to the 137 people who chose to continue. Therefore, we compared these two groups on a number of variables of interest, including age, IT, the physiological measures and the cognitive abilities measures. There were two points that needed to be considered before

this comparison could be made. First, 12 of the 13 people who discontinued were female and it was therefore necessary to co-vary for gender in each of the group comparisons because many of the measures show gender effects. Thus, a difference between the two groups in grip strength is likely to be caused, at least partially, by the over-representation of females in the drop-out group. Second, a group comparison of 13 people vs. 137 is not ideal because the variance in each group is likely to be different (i.e. violating the assumption of homogeneity of variance) and statistical power will be low. Nonetheless, an analysis of covariance was performed for each of the variables of interest with gender entered as a covariate.

There were differences between the two groups on three of the variables and in all cases the people who did not continue performed less well. The "drop-out group" were significantly slower at Time 1 on the Digit Symbol task (F (1, 146) = 4.90, $p < .05$, partial $\eta^2 = 0.03$), completed significantly fewer correct items on the Concept Formation task (F (1, 98) = 4.13, $p < .05$, partial $\eta^2 = 0.04$) and had a poorer vocabulary as indexed by the Spot-the-Word task (F (1, 147) = 6.31, $p < .05$, partial $\eta^2 = 0.04$). Contrary to expectations, there were no problems with the assumption of homogeneity of variances but there was evidence of low statistical power.
Thus, there is some evidence, despite low statistical power, that the drop-out group were slower, had poorer reasoning abilities and had a smaller vocabulary. As a result, the sample that completed this round of testing (n = 137) were more homogeneous than the full sample and therefore more restricted in range.

Materials and Apparatus

The materials and apparatus used in this testing phase were exactly the same as described in Chapter 4. However, not all tests were administered in this second test phase. The tests that were administered were Inspection Time, the physiological measures (*grip strength, visual acuity, systolic BP, diastolic BP, height* and *weight*) and the perceptual speed tests (*Digit Symbol, Visual Matching* and *Pattern Comparison*). These tests were administered so that the change scores over 6-months could be calculated.

Procedure

Testing sessions were completed between March and June 2004. As with the first session, participants had the choice of completing the session in their home or at the university and, in almost all cases, they chose to complete at the same place as the first session. Each participant was contacted and a suitable time was organised with him or her. An attempt was

made to test the participants in a similar order to the first occasion although this was not always possible (i.e. people who completed during September 2003 were encouraged to complete the second testing phase in March rather than June 2004). The average time between the first and second test session was 5.5 months (SD = 0.8, range = 3.3 – 8.5 months). No questionnaires data were collected on this occasion. The tests were administered in the same order and manner as in the first testing session (see p. 67). The order was blood pressure, Inspection Time, Digit Symbol, weight, height, Pattern Comparison, grip strength, visual acuity and Visual Matching. The tests were administered in a single block, without a break, for a total assessment time of about 2 hours.

Table 5.1. Descriptive Statistics for Biomarkers at Time 1 and 2

Biomarker	n_{12}	$Mean_1$	SD_1	$Mean_2$	SD_2
Inspection Time (ms)	113	86.36	(27.97)	83.63	(23.56)
Grip Strength (kg)	136	18.99	(8.85)	19.11	(9.18)
Systolic BP (mmHg)	121	146.35	(25.99)	141.33	(21.16)
Diastolic BP (mmHg)	121	78.95	(11.52)	76.20	(10.79)
Weight (kg)	137	71.12	(12.81)	70.42	(12.75)
Height (mm)	137	163.86	(8.83)	163.37	(8.75)
Visual Acuity (log units)	137	1.80	(0.29)	1.79	(0.30)

Results

Table 5.1 presents the descriptive statistics for the biomarkers on the first and second occasion. The second column shows the number of people who completed the task on both occasions (e.g. 113 people had scores on IT at Time 1 and 2). The variables with the most missing data were IT and the two BP estimates. The reasons for these missing data have already been discussed (see p. 70) and were essentially the same in this round of data collection. As with the first testing phase, a small number of IT scores (n = 6) were excluded due to problems with their IT scores (see Appendix E). The mean values were very stable over the 6-month period for all of the biomarkers, with none of the differences reaching statistical significance. Furthermore, the standard deviation estimates were also very similar. This does not imply that there was no decline in the biomarkers over a 6-month period but rather that the average change was small.

Question 1: How Reliable are the Initial Values?

The first question addressed in this chapter was: how reliable are the initial values of the biomarkers, assessed at the first testing phase? Given that the biomarkers were measured at

multiple time points, it was highly desirable to have marker tests that showed high reliability because this is known to decrease when change scores are calculated. In order to answer this question, it was necessary to use a range of different methods. First, in the IT task a sub-sample (n = 26) completed the task a second time at the end of their first testing session and this allowed for an estimate of test-retest reliability. Second, for grip strength, systolic BP and diastolic BP there were multiple measurements, which allowed for the use of Cronbach's alpha to calculate a coefficient of reliability. Third, the lower limits for the reliability of weight, height and visual acuity were calculated from a formula in Rudinger and Rietz (2001), based on the correlation between the biomarkers at Times 1 and 2. A brief account of the theoretical basis of these methods is presented below.

The first method used was test-retest reliability. Test-retest reliability provides a measure of how consistent the test is when people complete the same test on two occasions over a short period of time. This was calculated by getting a small group to complete IT a second time, at the end of their first test session. Scores from the first and second estimate were correlated leading to an estimate of the reliability of the measure.

$$\alpha = \frac{N * \bar{r}}{1 + (N - 1) * \bar{r}} \qquad (1)$$

The second method used was Cronbach's alpha (α). Cronbach's alpha provides a measure of how well a group of tests measure a latent factor (e.g. grip strength); internal consistency. For grip strength, systolic BP and diastolic BP, multiple estimates were taken at the first session and the average of these was used to estimate the biomarker. The strength of relationship between these individual measures provides an indication of the reliability of the total score. The formula for Cronbach's alpha is given in Equation 1 and it is based on just two variables: the number of estimates (N) and the mean correlation of all estimates ($\bar{r}$). For example, there were six estimates of grip strength (N = 6) and the average correlation between them was very high ($\bar{r}$ = .939). When these values are entered into Equation 1, the reliability coefficient (α) is equal to .989 suggesting that the grip strength score is very reliable.

$$r_{12} = \rho_{12} * (rel_1 * rel_2)^{1/2} \qquad (2)$$

$$r_{12} = \rho_{12} * rel_1 \quad (3)$$

For the variables weight, height and visual acuity only one estimate was taken at the first test session and reliability could not therefore be estimated by normal means. However, it was possible to derive lower bounds for the reliability estimates using an equation from Rudinger and Rietz (2001). The formula is presented above as Equation 2. r_{12} denotes the correlation between scores on the measure of interest (e.g. weight) at Times 1 and 2, ρ_{12} denotes the stability of the construct over 6-months, rel_1 denotes the reliability of the score at Time 1 and rel_2 denoting the reliability of the score at Time 2. Essentially, it means that the correlation between two scores over a period of time is a product of the stability of that construct but also the reliability of the individual estimates since measurement error contributes to the correlation. If we assume that $rel_1 = rel_2$, that is the reliability at Time 1 and 2 are equal then the formula simplifies to Equation 3. Furthermore, we know that ρ_{12} and rel_1 must lie between 0 and 1. Based on these two pieces of information, it is possible to prove mathematically that ρ_{12} and rel_1 must lie between r_{12} and 1. That is, the reliability of the measure at Time 1 must be greater than or equal to the correlation between the measures at Time 1 and 2. For example, the correlation between weight at Time 1 and 2 is .989. Therefore, it follows that the reliability of weight at Time 1 must lie between .989 and 1, which would mean that it is highly reliable. This method was used to estimate the lower bounds of the reliability for weight, height and visual acuity.

Table 5.2. Reliability Estimates for Biomarkers

Biomarker	Method	Initial value	Change score
Inspection Time	Test-retest	.826	.461
Grip Strength	Cronbach's alpha	.989	.667
Systolic BP	Cronbach's alpha	.918	.740
Diastolic BP	Cronbach's alpha	.880	.580
Weight	Lower limits	≥.989	-
Height	Lower limits	≥.980	-
Visual Acuity	Lower limits	≥.686	-

Table 5.2 shows the reliability estimates for each of the biomarkers. For the initial value, the most reliable measures were grip strength, weight, height and systolic BP, all of which would be deemed highly reliable with scores above 0.90. Diastolic BP and IT were somewhat less reliable, with scores between 0.80 and 0.90, but would still be considered to have adequate reliability. Finally, the reliability of visual acuity is unknown except to say that it is greater than

.686. If the reliability was at the lower end (i.e. close to .686) then it would be considered quite low but it is possible that the measure is actually reliable and simply unstable over time. It is impossible to confirm either conclusion at this time. With the exception of visual acuity, it is concluded that the initial value of the biomarkers are reliable.

Question 2: How Reliable are the Change Scores?

A major problem with change scores is that they tend to be unreliable. The reason for this is that error variance associated with both of the individual estimates (i.e. Times 1 & 2) is incorporated into the change score. Therefore, the change score is almost always less reliable than the original estimates. Cohen and Cohen (1983, p. 414) provided a formula for calculating the reliability of change scores and this is presented in Equation 4[8]. The following notation is used: $rel_{(cc)}$ denotes the reliability of the change score, $rel_{(12)}$ denotes the reliability of the measure (e.g. IT) for both occasions (operationally define as a test-retest coefficient or an alpha coefficient, on the first occasion), and r_{12} denotes the correlation between Times 1 and 2 scores on the measure of interest. Consider the IT measure as an example. From Table 5.2, we know that $rel_{(12)} = .826$ and from Table 5.3 we can see that $r_{12} = .677$. When these values are substituted into Equation 4, the reliability of the change score is .461, which is considerably lower than the reliability of the individual estimates. So, although the individual estimates of IT are moderately reliable, the reliability of the change score is lower because it incorporates error variance from both individual estimates.

$$rel_{(cc)} = \frac{rel_{(12)} - r_{12}}{1 - r_{12}} \qquad (4)$$

The final column of Table 5.2 shows the reliability of the change scores. Given that reliabilities of the initial estimates were not available for weight, height and visual acuity, the reliability for the change scores for these measures could not be calculated. The reliabilities of the change scores were highest for systolic BP and grip strength, with reliability coefficients of 0.67 and 0.74, respectively. As for IT and diastolic BP, the reliabilities of the change scores were 0.46 and 0.58, respectively, which is somewhat lower. As a rule of thumb, the reliability of a test

[8] This formula is discussed in detail in Rogosa, Brandt and Zimowski (1982) and details of its derivation are provided.

should be at least 0.7 if it is going to be used to assess individual differences (Kline, 1998). Therefore, it is questionable how useful the change scores for grip strength, IT and diastolic BP are going to be when predicting outcome measures. However, this point will be expanded upon in the discussion section. To summarise, the change scores for systolic BP were reliable, for grip strength, IT and diastolic BP they were far less reliable and for weight, height and visual acuity they were unknown.

Question 3: How Stable are these Constructs over a 6-month period?

When discussing the reliability of the initial values (see p. 103), we noted that the correlation between two scores over a period of time is the product of the stability of that construct and the reliability of the individual estimates. Having estimated the reliability of the individual tests and the correlation over time, this formula can be used to estimate the stability of the constructs over 6-months. For those biomarkers without reliability estimates (weight, height and visual acuity), we showed previously that the stability of these biomarkers (ρ_{12}) was at least as large as the correlation over time. Therefore, the correlation was used as a lower limit for the stability of these biomarkers, as in Table 5.3.

Table 5.3. Correlation and Stability of the Biomarkers over 6-months

Biomarker	r_{12}	ρ_{12}
Inspection Time	.677**	.820
Grip Strength	.967**	.978
Systolic BP	.685**	.746
Diastolic BP	.714**	.811
Weight	.989**	≥.989
Height	.980**	≥.980
Visual Acuity	.686**	≥.686

Note. r_{12} = Correlation over 6-months, ρ_{12} = Stability over 6-months
** $p < 0.01$

Table 5.3 shows the correlations between scores at Times 1 and 2 and the stability coefficients of the constructs over 6-months. The most stable biomarkers were weight, height, and grip strength. This means that, as a group, the participants showed very little change over 6-months on these measures. IT and diastolic BP were relatively stable but with sufficient instability to suggest that these change scores have promise as predictors of functional outcomes. Systolic BP was less stable and, because it was also reliable, the changes scores should prove useful for predicting other outcomes. The stability of visual acuity is unknown at this point

except to say that it is greater than .686. To summarise, at a group level the biomarkers were relatively stable over the 6-months. However, the more pertinent issue with respect to prediction is whether there are individual differences between people in the stability of the biomarkers over time and this issue will be dealt with in the following section.

Question 4: Are there Individual Differences in Stability of the Biomarkers?

Before this question can be answered it is necessary to provide a short discussion of some of the issues surrounding change scores. There are many issues with the calculation and use of change scores, many of which have not been fully resolved by methodologists themselves (Cohen & Cohen, 1983). To some degree, the decision about which method to use depends upon the subsequent use of the change scores. In a very comprehensive paper on the measurement of change scores, Rogosa and Willett (1983) stated that the major question should always be – How much has person *j* changed on the variable of interest? In this case, it is quite valid to use a standard difference score (i.e. difference = Time 2 – Time 1) to answer this question. However, the question that is more pertinent to the current project is - Which people have changed more than expected compared to the rest of the group? When attempting to answer this question, there are some problems with the standard difference score and Cohen and Cohen (1983), amongst others, have advocated for the use of residual change procedures to deal with these problems. In the following paragraphs, the two major issues with the difference score will be discussed, followed by an explanation of the residual change procedures.

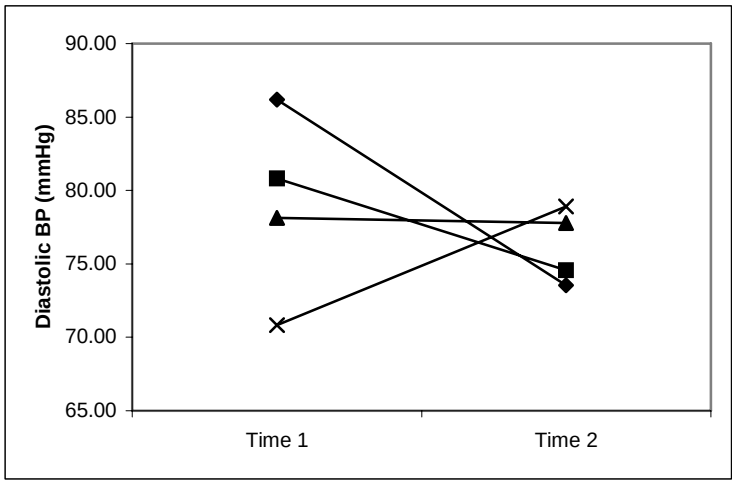

Figure 5.1. Six-month difference scores on Diastolic BP by quartile

There are two basic problems with using difference scores when attempting to compare the amount that individuals have changed. The first problem is referred to as *regression to the*

mean and will be described with the use of Figure 5.1 using diastolic BP as the example. On the first occasion, there was a wide spread of diastolic BP measures, which were allocated into quartiles (e.g. the top line represents the mean for people with the highest 25% of diastolic BP measures at Time 1). On the second occasion, the diastolic BP measures were centred around the overall mean value and thus each quartile group tended towards that mean value. As a result, people with higher measures at Time 1 tended to show decline over time (i.e. approach the mean value from above) and people with lower measures tended to show an increase over time (i.e. approach the mean from below). Thus, these trends produce a situation where the change scores also reflect a statistical artefact of regression, which is clearly undesirable. The question that we would really like to answer is – if everyone started at the same point, how much would person j have changed?

The second related problem is that change scores are statistically dependent on the initial score (see Figure 5.1). That is, the change over a 6-month period on any of the biomarkers is dependent upon the score that a participant achieved at the first testing session. This is a serious problem in the current context because the goal is to calculate change over 6-months and use this to predict the outcomes at the end of the study. If change scores are statistically dependent upon initial scores then any relationship between change and the outcome may be spurious because it might actually reflect a relationship between the initial value and the outcome measure.

In order to circumvent these problems, a number of researchers have encouraged the use of residual change procedures (e.g. Cohen & Cohen, 1983). Essentially, the variance associated with the initial score is partialled out of the change score by using regression techniques. This treatment results in a change score that is statistically independent of the Time 1 score and thus avoids the problems of regression to the mean and spurious results due to dependence on the initial value. This method will therefore be used to calculate the change scores for this research project. However, it must be acknowledged that there are still methodological problems with this approach (see Rogosa et al., 1982 for discussion). Using this method does not necessarily correct all the problems with standard difference scores; but theoretically it is a better method to use in order to answer the questions in which we are interested for this project.

As a result of calculating change scores using with the residual change method, the reliability of the change score needs to be re-calculated using a slightly different formula (presented in Equation 5). This formula was taken from Rogosa et al. (1982) but derives from much earlier papers (e.g. Linn & Slinde, 1977; E. F. O'Connor, 1972). Using this new formula,

the reliability of the change score for IT is 0.532, for grip strength is 0.672, for systolic BP is 0.773 and for diastolic BP is 0.630. Although these reliability estimates are very similar to those presented in Table 5.2, the most important point is that the reliability for IT is improved using this method.

$$rel_{(cc)} = \frac{rel_{(12)} - [(r_{12})^2 * (2 - rel_{(12)})]}{(1 - r_{12}^2)} \quad (5)$$

For replication purposes, the method that has been used to calculate the change scores is provided here. One small adjustment was made to the method following Rogosa et al., in order to take into account one further variable, the time between the first and second test phase. People completed the second testing session on average approximately 6-months from the first one. However, there were individual differences in this delay with some people completing the second session after just 3.3 months and others completing it after 8.5 months. These individual differences introduced unwanted error variance to the change scores and were therefore statistically controlled.

To simplify this explanation, it is necessary to define notation. Let *a* denote the scores at Time 1 (i.e. initial scores), *b* denote the scores at Time 2, and *c* denote the change scores, and *d* denote the time in months between the first and second testing session. First, the change scores were calculated as the difference between *a* and *b (c = b – a).* Second, a regression analysis was run with *c* as the dependent variable and *a* and *d* as the independent variables. Third, the unstandardised residuals were saved because these represent the change scores with all the variance associated with *a* and *d* removed. These new change scores (c_1) by definition correlate zero with the initial scores and remove the error variance due to differences in the time between testing sessions. That is, this new variable represents the change over 6-month, independent of the initial score.

Due to the residual method of calculating these score, the change scores have a mean score of zero and the scale is not easily interpretable. This is not a problem when considering correlations or regression analyses because the mean value and scale are not important. However, it does make interpretation of individual differences in the degree of change difficult. To accommodate this issue for the purpose of illustrating relative changes, the following method was used. First, the change scores (c_1) were sorted from largest increase to largest decrease and split into quartiles. Second, for each of the quartiles the mean score at Time 1 and 2 was

calculated. Third, the difference between mean scores at Times 1 and 2 was calculated to see the magnitude and direction of change for each quartile. This change was plotted on a graph by using zero as the starting point so that all quartiles could be compared. This indicates in a descriptive sense, the characteristics of those people with different degrees of change. For all of the measures, Quartile 1 represents the largest decline in the construct, even though this may actually represent an increase in some of the scales (i.e. IT and Visual Acuity).

Figure 5.2 illustrates the outcomes generated by using this method for IT. The group with the largest increase on IT show a mean increase of 18 ms. For analyses to follow that address the central question of whether change scores can predict subsequent outcomes, the new change scores (c_1) have been used because these are independent from both the initial scores and variation in time between the two testing occasions. The versions shown in Figures 5.2 to 5.11 are only included here because they are informative about the extent of differences in stability of the biomarkers over approximately six months.

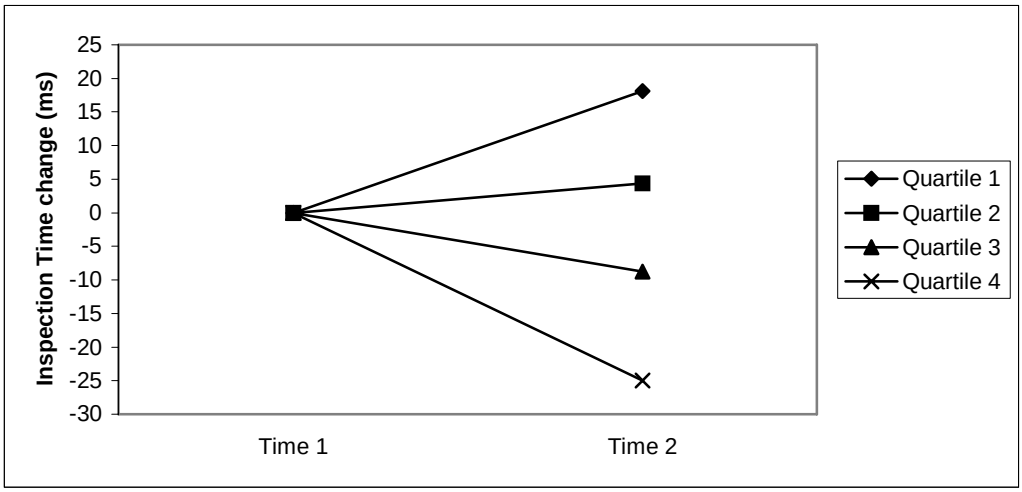

Figure 5.2. IT change over 6-months by quartile

Inspection Time. In the IT task, higher scores are indicative of slower processing speed. Therefore, the top line in Figure 5.2 summarise the performance of the people whose processing speed slowed the most over the 6-month period. These people showed a mean increase in IT of 18 ms, which is a relatively large decline, considering that the time interval was just 6-months. On the other hand, those in the fourth quartile showed a large degree of improvement on the IT task with a reduction of 25 ms. It is possible that this improvement was due to practice effects. If this were the case, then it would follow that the decline seen in the first quartile was an underestimation of the true decline, because these participants should also have been affected by

practice. To conclude, Figure 5.2 clearly illustrates that there have been substantial individual differences in the 6-month change scores in IT, which suggests that the variable has potential for predicting functional outcomes.

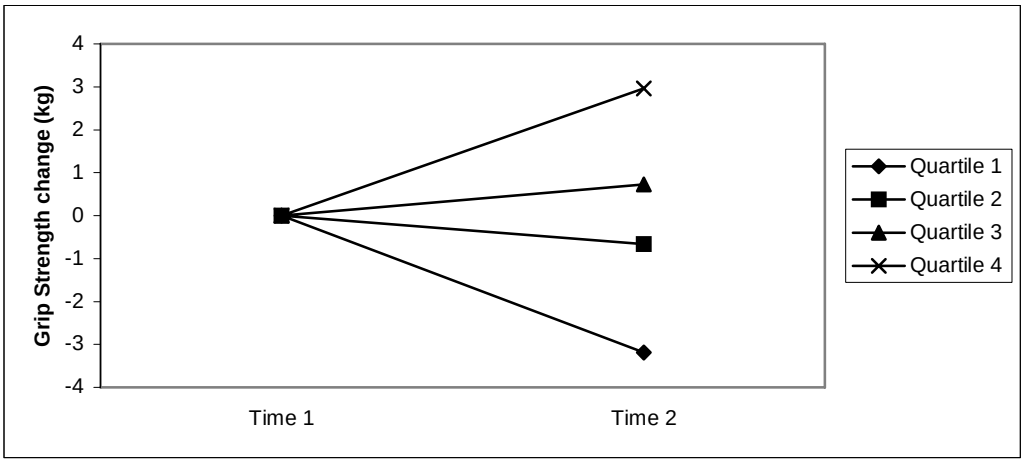

Figure 5.3. Grip Strength change over 6-months by quartile

Grip Strength. One aspect of normal aging is the loss of muscle mass and an increase in physical frailty. The measurement of grip strength is one way of assessing the decline in physical strength and muscle mass with age. If a person was to show substantial decline in grip strength over a relatively short period then this might be indicative of accelerated aging. Figure 5.3 shows changes (c_1) in grip strength over the 6-month period, rescaled as described above for IT. There was a group who showed more decline than the rest of the sample. Quartile 1 showed a mean decline of 3 kg over 6-months. Alternatively, Quartile 4 showed an improvement over the 6-months of about the same magnitude. Therefore, there were individual differences in change in grip strength over 6-months but overall the measure was very stable (stability = .978). For most people, there was little change over the 6-months, which may indicate that the change score is very stable across people and may not therefore be very useful for predicting future outcomes. On the other hand, this marked stability might suggest that those people who do decline are very unusual and therefore informative. Furthermore, the reliability of this change scores was relatively high (.667), which further suggests that the 6-month change score on grip strength could prove useful as a predictor.

Systolic Blood Pressure. This measure provides an indication of the pressure in the blood vessels when blood is being pumped through them. Blood pressure tends to increase with age and it therefore follows that a large increase over a short period of time may be indicative of accelerated aging or health problems. From Figure 5.4 it is apparent that those participants in the

fourth quartile have registered a marked increase in systolic blood pressure over the 6-month period, compared with the other three groups. This increase is equivalent to about 11 mmHg and may equate to a clinically significant increase in blood pressure, particularly given that the mean systolic BP was reasonably high at Time 1 for the group as a whole (see Table 5.1). Furthermore, the reliability of this change scores was high (.773), giving further weight to the importance of this change. Overall, there were large and reliable individual differences in systolic blood pressure, suggesting that this change score might be a useful predictor of future functional outcomes.

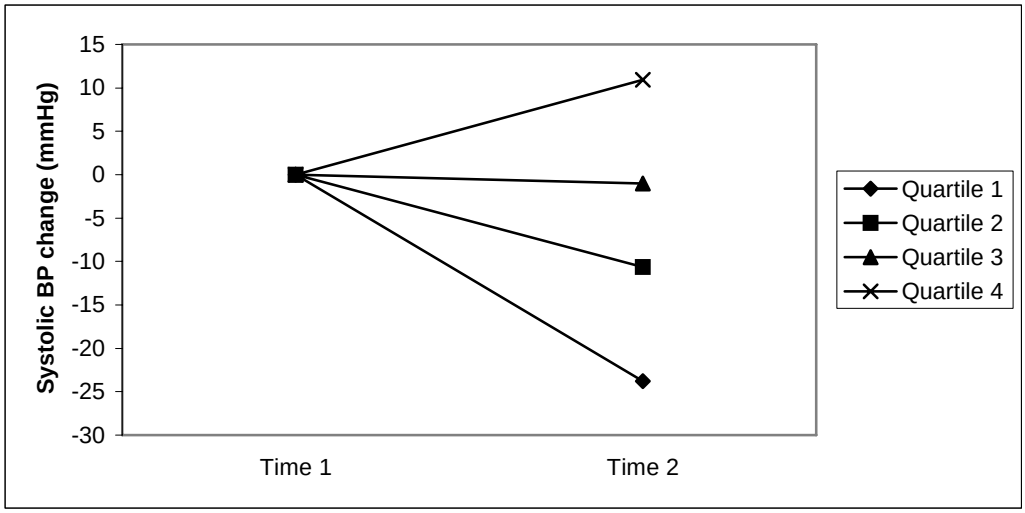

Figure 5.4. Systolic BP change over 6-months by quartiles

Diastolic Blood Pressure. This biomarker indicates pressure in the blood vessels when the heart is resting. For this reason, high values on this measure are often considered a more serious problem than high values on systolic blood pressure. The group with the largest increase in diastolic blood pressure have an average increase of 7 mmHg, which is a reasonably small change. Examination of Figure 5.5 shows that the largest change was actually experience by a group of people in the first quartile whose diastolic BP dropped over the 6-months by an average of 11 mmHg. Although high BP scores are generally thought to be concerning in a health sense, it is possible that a large drop in BP may be informative from an aging point of view. Furthermore, as noted in Chapter 4, low diastolic blood pressure has been linked to low fluid ability. To conclude, there were individual differences in diastolic BP and it may be necessary to consider those individuals who show substantial decline in diastolic BP, in addition to considering those cases registering an to increase in this measure.

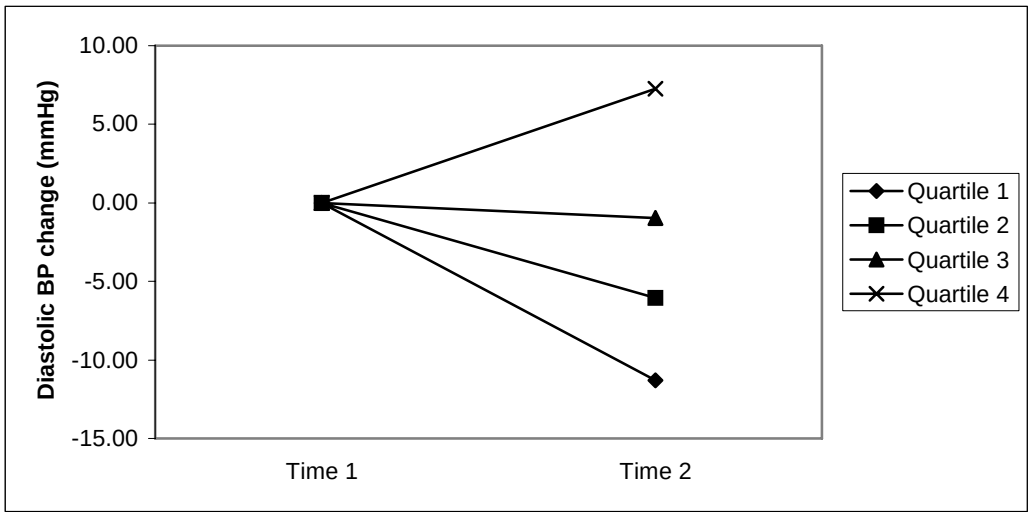

Figure 5.5. Diastolic BP change over 6-months by quartile.

Weight. In term of biomarker theory, it is thought that older people shrink in height and lose weight as they become more aged and frail. In addition, weight loss can also be indicative of a number of age-associated diseases. Therefore, individuals of most concern should be those people who have lost the most weight in the 6-month period. Figure 5.6 illustrates the largest weight change over 6-months for Quartile 1 who registered average weight of about 3 kg. On the other hand, there was a group who gained a little weight (mean = 1.5 kg).

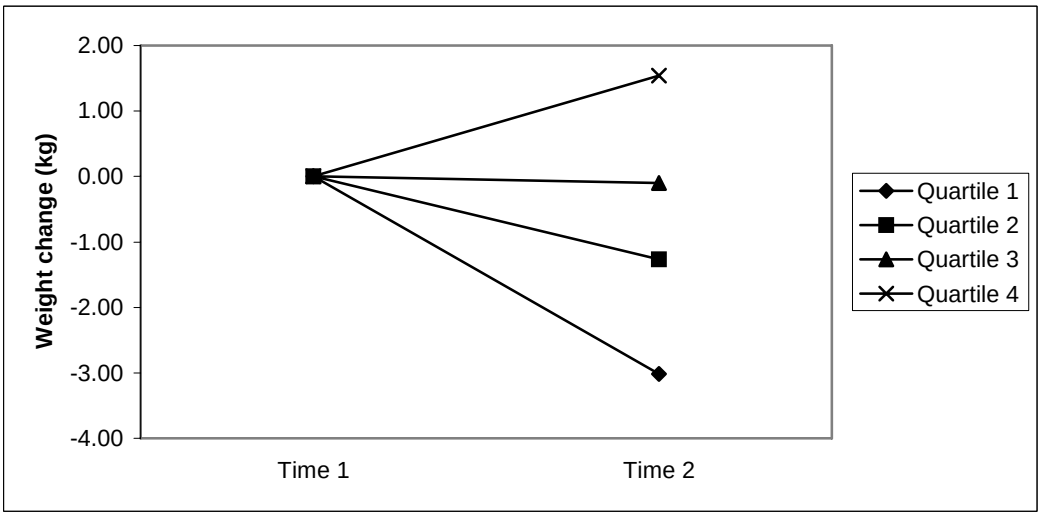

Figure 5.6. Weight change over 6-months by quartile

It is impossible to know to what degree such differences were due to daily fluctuation in weight. However, if these daily fluctuations are independent of the mean weight then the weight loss seen in this figure is still informative. Overall, weight was very stable over 6-months

(stability ≥ .989). This, together with very little change, suggests that the change score may not be very informative. To summarise, there were small individual differences in weight over 6-months but there are also daily fluctuations, the magnitude of which was unknown.

Height. Initially, separate figures were generated for height for males and females because they are known to have significantly different heights. However, the main issue of concern here was not the initial level of height but change over time. Given that the patterns of change were similar in both males and females it was deemed reasonable to combine them. The biggest concern with Figure 5.7 is that some participants appear to have grown taller over the 6-month period. Quartile 4 has grown an average of 2 cm over 6-months, which suggests that some individuals within that group have grown even more than 2 cm. There are a number of possible explanations for this finding.

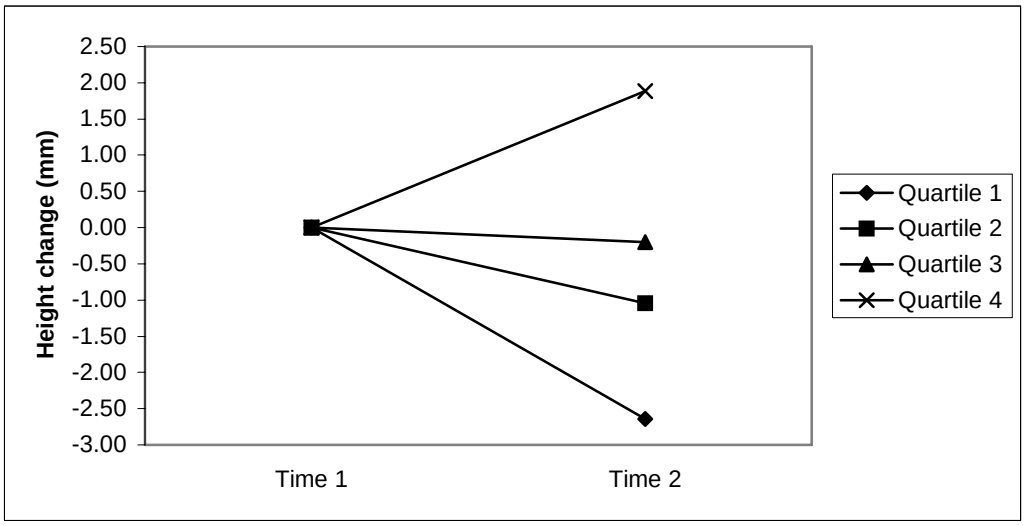

Figure 5.7. Height change over 6-months by quartile

First, it is possible that these participants had some chiropractic manipulation that straightened their spine and made them taller. Second, there could have been a significant change in their posture from one occasion to the next, causing them to stand up straighter and appear taller. Third, there may have been measurement error and unreliability when measuring height. If only one or two people had exhibited increased height over 6-months, then this might indicate that the first or second explanation was plausible. However, because a large number of people displayed this pattern, the third explanation seems more likely. On page 103, it was shown that the reliability of the initial estimate of height was very high (≥.980). However, this does not establish that the change score is reliable. In fact, the reliability of the change score can still take any value between 0 and 1. Even though height was stable over 6-months (≥ .980) it is still

possible that the change scores were unreliable. In a sense, the "signal" or true change is so small that it is very difficult to rank people in terms of their change and the "error" is likely to be substantial when compared with such a small effect. To summarise, there were small individual differences in the stability of height over a 6-month period are serious questions about the reliability of these change score arise, meaning that this measure may be questionable as a predictor for functional outcomes.

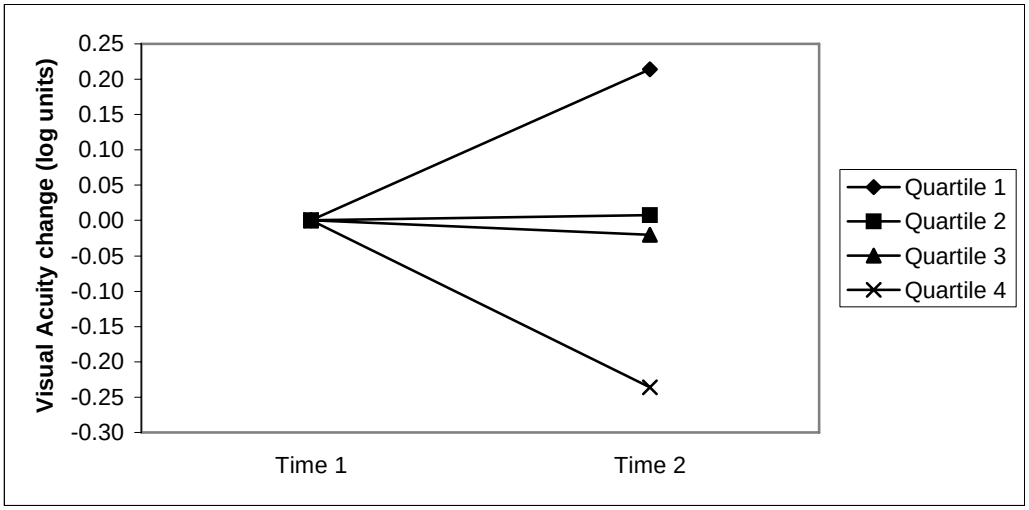

Figure 5.8. Visual Acuity change over 6-months by quartile

Visual Acuity. Both visual and auditory acuity are known to decline with advancing age. Therefore, people who show a large decline in visual acuity over the 6-months might be showing accelerated aging. The visual acuity test was scored so that low scores indicate superior vision. Examination of Figure 5.8 shows the group whose vision declined the most over 6-months (Quartile 1) appeared to show a relatively small change but it is difficult to interpret the degree of change from the scale. On a Snellen chart, this degree of decline is equivalent to reading the second to bottom line on the first occasion but then only being able to read the third to bottom line at the second session. As for Quartile 4, which showed improvement, their change is equivalent to reading the second to bottom line on the first occasion and then reading the bottom line on the last occasion.

There are two other points that cast doubt over the utility of the visual acuity change measure, at least when measured over this short time frame. First, there were a large number of people who showed no change at all (n = 65; indicated by Quartiles 2 and 3), which suggests that the construct was highly stable over 6-months in most people. Second, it was impossible to estimate the reliability of visual acuity except to say that it was at least .686. Therefore, it was

not clear from this data whether visual acuity was a reliable measure to use as a biomarker. However, a study by Anstey, Smith and Lord (1997) found that the test-retest reliability for visual acuity was .82 over a period of 3-months. Using this estimate in Equation 4, the reliability of the change score for visual acuity would be .423, which is low. This indicates that initial scores on visual acuity are reliable but that 6-months may be too short a time period to take a useful or reliable estimate of change in vision. To conclude, visual acuity shows marked stability over 6-months and there is some doubt over the reliability of the change score, which suggests that it might not be the best candidate for predicting functional outcomes.

For interest sake, the decline over 6-months on the perceptual speed tasks will also be examined. Although there are problems with the perceptual speed tasks, they are the closest measures that we have to IT. Therefore, it should be informative to see whether they also display a reasonable decline over 6-months. Furthermore, by calculating change scores for these tasks, we can ultimately compare the predictive validity of IT to that of other speed tasks.

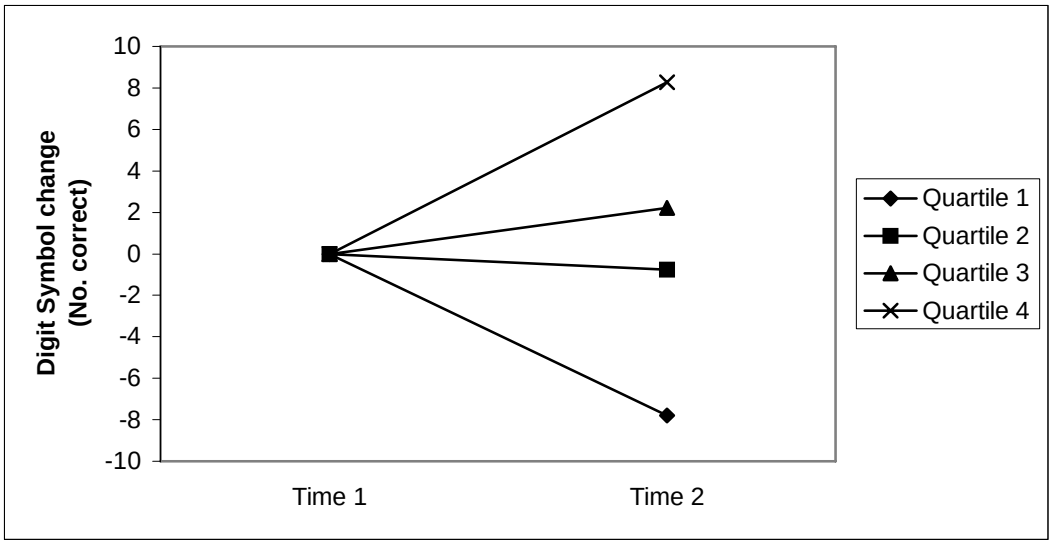

Figure 5.9. Digit Symbol change over 6-months by quartile

Digit Symbol. This was the first of three perceptual speed tasks that were administered to the participants. Although designed to measure speed of processing, it is known to involve other cognitive abilities such as attention and memory and also to confound speed and accuracy. One hundred and thirty six participants completed DS on both occasions and the mean scores were very similar from Time 1 (M = 54.57, SD = 13.3) to Time 2 (M = 55.04, SD = 14.4). The correlation between the Time 1 and 2 scores was also high (r (134) = .889), suggesting that DS has marked stability over a period of 6-months. Figure 5.9 shows the individual differences in

the change score for DS. The group with the largest decline (Quartile 1) completed an average of 8 less correct items at Time 2. Decline of this magnitude may be important and given that the task is so stable, those people who do decline may be experiencing accelerated aging. However, due to the nature of the task, it is impossible to tell whether these people are slowing down, experiencing attention problems or something else. Given that IT is less stable (stability = .677) and less complex, it should provide a more useful index of central nervous system slowing.

Visual Matching. This task was quite stable over the 6-month period, as indexed by the high correlation between Time 1 and 2 score (r (132) = .791) together with the near identical mean scores at Time 1 (M = 32.72, SD = 4.6) and Time 2 (M = 32.85, SD = 5.68). Given that the correlation between Time 1 and 2 scores for VM is smaller than for DS, it may be less stable and therefore more informative. Figure 5.10 shows the individual differences in the VM task. The group that showed the largest decline completed an average of 4 items less at Time 2. To conclude, there are moderate changes in the VM task over 6-months for some individuals and thus the change score might be quite useful in a predictive sense.

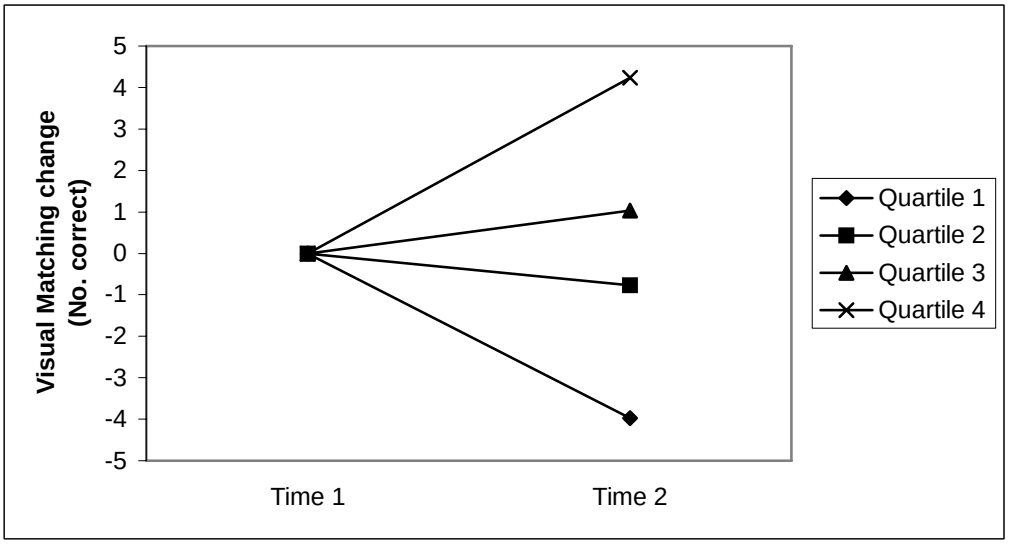

Figure 5.10. Visual Matching change over 6-months by quartile

Pattern Comparison. In this task, participants have just 40 seconds to complete as many items as possible, and this task was the least variable of all three of the perceptual speed measures. An examination of the mean scores showed near identical performance at Time 1 (M = 16.52, SD = 3.5) and Time 2 (M = 16.56, SD = 4.3) on PC. However, the correlation from Time 1 to 2 was actually the lowest of all three perceptual speed tasks (r (132) = .709), which indicates that, for some people at least, it may be less stable across time than the other perceptual

speed tests. Figure 5.11 shows the change over 6-months of the PC task in the quartile groups. The group with the largest decline completed 4 less items correct at Time 2. Given that the task was so stable, a change of this magnitude may be quite important. Thus, there were relatively large changes occurring over a 6-month period in some individuals and these changes may be indicative of accelerated aging.

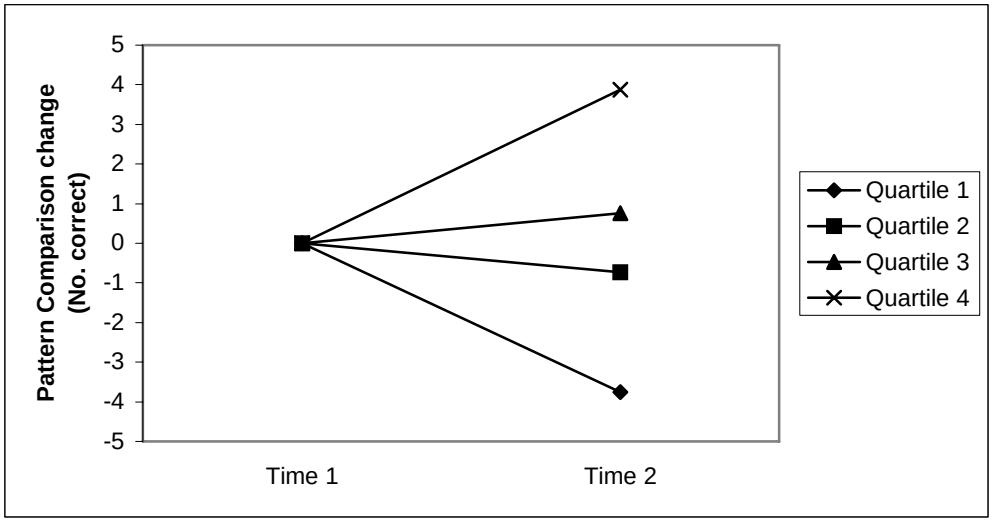

Figure 5.11. Pattern Comparison change over 6-months by quartile

Question 5: Are there Gender Differences in the Stability of the Biomarkers?

One issue that was considered in Chapter 4 was the existence of gender differences in the decline of the biomarkers over time. At that time it could only be examined in a cross-sectional manner but it is now possible to test this idea longitudinally. For each of the biomarkers, the mean change score (c_1) was compared for males and females. There were significant differences in the change scores for just one variable: diastolic BP. For diastolic BP, both groups showed a decrease over the 6-months. However, the males showed a significantly larger decrease than the females (t (118) = 2.41, $p < .05$, d = 0.43) and this effect size was moderate. The theory by Birren and Fisher (1992) posits that males have a lower life expectancy, so that they should show larger declines in biomarkers than similar aged females. This theory would predict that males would show more of an increase in diastolic BP. Therefore, these findings do not offer any support for this theory.

Discussion

The aim of this chapter was to explore the reliability and stability of a number of potential "biomarkers" over a 6-month period. The initial score and the 6-month change scores will be used as predictors for a range of functional outcomes at the end of this study. Therefore, it is necessary to show that there exist reliable individual differences in both the initial level and the change scores between people. For each of the seven biomarkers, the results of these investigations will be discussed and a decision will be made as to whether the initial score and change score are likely to be useful predictors of the functional outcomes.

Inspection Time. The reliability of the initial score was moderately high and this value can be used as a valid predictor of the functional outcomes. The reliability of the change score was lower, at .532. Kline (1998) proposed that any variable that is used to investigate individual differences should have a reliability coefficient of at least 0.7. Based on this assertion, we would have to conclude that the change score for IT is not a valid measure to use as a predictor of the outcomes. However, using this rule-of-thumb to make such an important decision is actually highly questionable; and, the issues surrounding the reliability of change scores are actually much more complicated than this rule would suggest.

Reviewing Figure 5.2, we can see that about half of the sample showed very little change over the 6-month period on the IT task. When individual show almost the same change (e.g. zero) it is very difficult to distinguish between them in terms of the change score (Rogosa & Willett, 1983). To think about it in statistical terms, we have a very small "signal" or effect, embedded in "noise" or error variance. Therefore, it is not surprising that it is difficult to rank people based on their change score when the change is small (i.e. low inter-individual variability). In this situation, the reliability coefficient will always be low. Thus, the reliability coefficient for IT is low but this is due to the large number of people whose IT scores are very stable over the 6-month period, not those people who are showing decline or improvement. Given that we are primarily interested in those people who show a large decline in IT over the 6-months, this issue of low reliability may not be as "catastrophic" as we had previously surmised. Furthermore, consistent with this line of argument, Rogosa et al. (1982, p. 730) stated that, "the difference score can be an accurate and useful measure of individual change even in situations where the reliability is low".

It appears that the change score for IT is indeed a valid and important measure to consider as a predictor of functional outcomes given that there are substantial individual differences in the

amount of change over 6-months in IT. Although the reliability of the change score is moderate, it is still valid to consider whether decline over this period is indicative of accelerated aging and therefore predictive of functional outcomes. Because IT is relatively stable over the 6-month period, people who do show slowing are quite unusual and may be qualitatively different. To conclude, the initial IT score and the 6-month change score can be tested for utility as predictors of functional outcomes.

Grip Strength. The initial value of grip strength was very reliable and could be used as a predictor of functional outcomes. The reliability of the change score was also quite high and it therefore scores can be used. Grip strength is not only one of the most reliable biomarkers but also among the most stable over the 6-month period. The mean scores at Time 1 and 2 were almost identical and the stability coefficient was close to unity. This indicates that the individuals who did show decline were highly unusual and therefore may be informative. From Figure 5.3 it is apparent that there was a sub-group showing a mean decline of about 3 kg and it is quite likely that this group was losing some muscle mass and therefore strength. That is, individuals whose grip strength declined over the 6-month period may indeed be showing signs of accelerated aging and, given that this change is measured relatively reliably, this may be very informative and potentially predictive of the outcome measures.

Systolic Blood Pressure. The initial estimate for systolic blood pressure was very reliable and using the average of three measurements helped increase this. The change score was a little less reliable but was still above 0.7 and therefore would be deemed to have an adequate reliability for use as a predictor of the functional outcomes. The stability of systolic blood pressure was low compared to the other biomarkers. This indicates that there is quite a bit of change in systolic blood pressure over time and it is unclear how important or informative the observed changes actually are. There were individual differences in the change over time, with one group showing in increase in mean systolic blood pressure of about 11 mmHg at Time 2. Given that blood pressure is known to increase with age, this increase over 6-months may be indicative of accelerated aging. Furthermore, a change of this magnitude may represent a clinically significant change, putting the participants in this group at a higher risk for heart attack and stroke. To summarise, the initial estimate and change score for systolic blood pressure were reliable and could be used as valid predictors of functional age. The magnitude of change is quite large and important in clinical terms and, given that the change score is reliable these differences might be quite informative.

Diastolic Blood Pressure. The reliability of the initial score for diastolic blood pressure was quite high and certainly acceptable. The change score was less reliable, but given the previous discussion of this problem, we know that a change score with low reliability can still be quite informative. This is fortunate because there were considerable individual differences in the change in diastolic blood pressure over the 6-month period, which could be quite informative. Diastolic blood pressure was slightly more stable that systolic blood pressure but this might be because it was less variable. Finally, there was evidence that men showed more of a decrease over time in diastolic blood pressure than women. The decrease in diastolic blood pressure over time was surprising but, given that there was a significant gender difference, it offers more evidence that a decline over time might be just as informative as an increase over time. To conclude, there were moderate increases and decreases in diastolic blood pressure over time in sub-samples of the group and both the initial score and change score are therefore expected to be informative about the aging process.

Weight. It is clear that the initial score for weight was highly reliable, even though it was not possible to estimate the reliability coefficient exactly. Given that the measurement of weight is so simple and quick, a second estimate could have been done easily; and on reflection should have been done. Future studies should take multiple estimates of all biomarkers, in order to test the reliability of the biomarkers under investigation.

Using the same proof, as mentioned above, it was possible to prove that the stability of weight was very high. Given that weight was so stable over the 6-month period, any decline in weight would probably be small and therefore quite possibly unreliable. Given that the effect or signal is so small, even a small error variance would be enough to produce an unreliable estimate. To summarise, the initial value was very reliable and could be sensibly used as a predictor of the functional outcomes. There were small individual differences in the stability of weight over 6-months and, given that the measure was so stable over time, is it very likely that the reliability coefficient was low (see Rogosa & Willett, 1983 for discussion).

Height. The initial value was very reliable for height and the construct was very stable over the 6-months. The reliability of the change score could not be estimated, and so it was impossible to know whether the individual differences in the 6-month change scores reflected real effects. However, there was some evidence that this change may have been unreliable. There were a group of people who appeared to have ‘grown’ taller (M = 2 cm) in the 6-month period, an outcome that is difficult to explain in an elderly sample. Given that so many people were in this

group, it points to measurement error in the height estimate and suggests that the change estimate may be unreliable. Therefore, the initial values can certainly be used as a predictor but the change score may be unreliable and one should be careful about drawing conclusions if using it as a predictor of functional outcomes.

Visual Acuity. Visual acuity should have been assessed on two different occasions in order to get an estimate of reliability. Since this was not done, it was impossible to estimate the reliability of the initial score, the stability of the construct over 6-months or the reliability of the change score. All that could be ascertained about this measure was that the reliability of the initial score and the stability of the construct over 6-months were greater than .686. Given that this value is less than 0.7 it is impossible to say whether the initial score was sufficiently reliable to be used as a predictor but the chances are that it was reliable enough. The examination of the individual differences in visual acuity showed that about half of the group showed absolute stability on visual acuity, with one quarter displaying decline and the other quarter displaying improvement. This suggests that the measure may be quite stable, which in turn implies that the change scores are likely to be unreliable. Therefore, due to the single measurement of visual acuity taken at the first testing session, it is impossible to estimate the reliability of the initial estimate or the change score. However, there is some evidence that the change score may not be very reliable.

Perceptual Speed. Although the perceptual speed tasks are not here regarded as valid biomarkers, it is worth considering the decline over 6-months in these tasks. For DS, the scores were very stable and the decline was small, even in the group with the most decline. However, VM and PC showed less stability and, therefore, it is expected that the change scores for VM and PC might be more useful in terms of predicting functional outcomes than DS. However, all three of these measures were more stable than IT, which suggests that in some ways IT might be more sensitive to changes over time than these perceptual speed measures.

This chapter has examined in depth the reliability and stability of all seven of the biomarkers under investigation. In terms of the initial value, it was confirmed that all of the biomarkers, with the possible exception of visual acuity, had adequate reliability. It is therefore reasonable to use these measures as predictors of the functional outcomes in the following chapter. As for the change scores, although some did have low reliability, we demonstrated that this was largely due to the majority of people with very stable scores from Time 1 to 2 and not the people showing decline over time. Furthermore, we showed that change scores with low

reliability can still be useful and informative about the true change and should not be disregarded. However, the change score for height appears to be somewhat problematic since it seems to involve some measurement or random error that we cannot explain and thus may not be very useful as a predictor. Overall, it appears that the initial values and change scores are quite valid to examine as potential predictors of functional age. The next step is therefore to consider the validity of the functional outcomes measures in order to ultimately answer the question – can the initial values and the change scores for IT predict functional age? An in depth examination of the functional outcomes will be presented in Chapter 6.

CHAPTER SIX: STUDY 3 - THE ASSESSMENT OF FUNCTIONAL AGE

The aim of this chapter was to examine the reliability and stability of the functional outcomes over a period of 18-months. As was the case in the previous chapter, it was important to consider the stability of the functional outcomes and whether the measures were reliable, before examining the predictive validity of the biomarkers in Chapter 7. Two aspects of the functional outcomes were considered; (i) the final score and (ii) the 18-month change score. Therefore, this chapter describes how various aspects of these scores were examined including normality of the distributions, the stability of the constructs over 18-months and gender effects in the change scores.

Before considering issues of reliability and stability of the functional outcomes the constructs and tasks that were used to measure functional age are defined. The validity of each biomarker clearly depends upon the choice of functional outcomes but, as will be clarified by what follows, this is not a straightforward matter. In this project, the decision was made to focus on two aspects of functional age; everyday functioning and cognition. It was assumed, moreover, that normally distributed continuous variables representing each of these constructs would be achieved.

The concept of *Everyday Functioning* was assessed by two questionnaires; (1) Activities of Daily Living (ADL; a composite of basic and instrumental activities) and (2) Cognition in Daily Life (CDL; see p. 125). Because the ADL scale was administered at Times 1 and 3, both the final score at Time 3 and the 18-month change score were available for consideration. However, CDL was only assessed in the final testing session, so that it was only possible to examine the final score for this scale.

Conventional psychometric intelligence or *cognition* was assessed by measures of fluid reasoning and crystallised ability. Fluid reasoning was measured by three tests; (1) Raven's Standard Progressive Matrices (RSPM), (2) Cattell Culture Fair Test (CCFT), and (3) Concept Formation (CF). Given that previous research has shown fluid reasoning shows marked decline over time for the age range considered here, a primary aim was to test whether the biomarkers predicted the final scores at Time 3 and/or 18-month decline on these measures. Crystallised ability was also measured by three tasks; (1) Information, (2) Spot-the-Word, and (3) Similarities. Crystallised ability is generally maintained into old age but it has widely been accepted that, when it does begin to decline, this indicates terminal decline that accompanies the final years of

life. Therefore, a primary concern was whether the biomarkers could predict the 18-month change scores for crystallised ability, rather than the final scores at Time 3.

Method

Participants

The final test session was completed by a total of 127 people, reflecting a total attrition rate of 15%. Their ages ranged from 72 to 93 years; 82 females (M = 79.0 years, SD = 4.5) and 45 males (M = 78.5 years, SD = 3.4). Within this group, there were three people who did not complete the second test phase but did complete test phase 3. These two did not have 6-month change scores on the biomarkers. Overall, 17 females and 6 males did not continue in the study over a period of 18-months. Most people gave reasons for discontinuing similar to those described in Chapter 5 and two women died between the second and third testing phases.

Once more, whether the 23 people who did not continue were significantly different to the 127 people who continued until the end was investigated. The two groups were compared on age, IT, the physiological measures and the cognitive abilities measures from the initial testing session. Given that proportionally more females did not complete the study, it was necessary to co-vary for gender in the group comparison, using the ANCOVA procedure. There were several significant group differences between the "drop-out group" and those who continued.

First, the "drop-out group" were significantly older at Time 1 than the group who continued (F (1, 147) = 5.76, p < .05, partial η^2 = 0.04). Second, there was a significant difference in IT (F (1, 129) = 4.60, p < .05, partial η^2 = .03) between the two groups, with the drop-out group having markedly longer mean initial IT (M = 102.4 ms, SD = 36.7) than those who continued (M = 86.0 ms, SD = 27.9). Third, the drop-out group had substantially poorer visual acuity at the start of the study (F (1, 147) = 9.53, p < .01, partial η^2 = .06); but the two groups did not differ on any of the other physiological measures. Finally, the two groups differed on every single one of the cognitive abilities tasks, including the dementia test. In all cases, the drop-out group displayed poorer cognitive performance at Time 1, with effect sizes (partial η^2) between .03 and .11.

There are a number of possible interpretations of the differences between these two groups. First, it was obvious from observations that a number of the participants who were having problems with the cognitive tasks were visibly uncomfortable and chose not to 'endure' the testing on subsequent occasions. Another possibility is that people who suspected at the time

of initial recruitment that they were having cognitive problems found this confirmed when attempting the tasks at the first session and therefore felt that they had no reason to continue. Regardless of the reason for this effect, the result was that the final group who completed all three test sessions was even more homogenous and restricted in range than the original sample. As a result, it became even less likely that people who demonstrated accelerated aging could be identified, simply because many of the less able people dropped out before the end.

Materials and Apparatus

The majority of the materials and apparatus used in this testing phase were exactly the same as has been described in Chapter 4. However, some questionnaires were added, as described below. In addition, some of the cognitive tasks had alternate forms, which were used in the final testing session to try to reduce practice effects. For Raven's Standard Progressive Matrices, Form B was used and for Spot-the-Word an alternative word list (Version B) was used.

There was also a problem with one of the perceptual speed tasks, Pattern Comparison. In the first two phases of data collection, participants were given 40 seconds to complete as many items as possible. However, in the final testing phase participants were accidentally instructed to complete as many items as possible in 30 seconds. To permit direct comparison, scores for the final session were adjusted to represent a 40 second time period. This is problematic, however, for a number of reasons including the improvement that is often seen as participants become familiar with the task. Therefore, results for Pattern Comparison should be viewed with caution; and they have always been compared to performance on the other two perceptual speed tasks.

Cognition in Daily Life (CDL). It was already clear (see Figure 4.4, p. 84) that the original Activities of Daily Living questionnaire was not ideal because scores had such a marked ceiling effect (i.e. most people reported being very independent). Therefore, in the final testing phase a second questionnaire was added. This was the Subjective Scale to Investigate Cognition in Schizophrenia (*SSTICS*: Stip, Caron, Renaud, Pampoulova, & Lecomte, 2003), a 21-item scale that asks questions about the degree of memory, attention and language difficulties that the participant is experiencing in everyday life; (for example – "Do you have trouble focusing you attention on the same thing for more than 20 minutes?"). Although the SSTICS was designed for testing schizophrenics, consideration of the questions suggested that they would be applicable to any population experiencing some degree of cognitive impairment or decline. Hereafter, therefore, this scale will be referred to as the CDL scale. For each item, the participant indicated on a Likert scale (0 = never, 1 = rarely, 2 = sometimes, 3 = often, 4 = very often) how often the

problem was occurring. Therefore, the total score could take a value between 0 and 104, with higher scores indicating more difficulties with cognitive aspects of everyday life.

Activities of Daily Living (ADL). As already mentioned, the original ADL scale was problematic because many people reported that they were completely independent. On some of the items (e.g. eating, drinking and medication) there was absolutely no variation, with all participants registering the highest score. On other items, there was very little variation, with only a small number of participants scoring slightly below the maximum. Therefore, for the final testing phase a shorter version of the original questionnaire was administered, which included only those items that had previously differentiated between the participants. In Appendix D these items are identified by an asterisk. These questions tended to be the more complex instrumental activities of daily living, rather than the basic activities of daily living. The shorter version included nine questions on food preparation, transfers (e.g. in and out of chairs), mobility, mode of transport, shopping, telephone, housekeeping, laundry, and ability to handle finances. It therefore generated a score between 9 and 34, with higher scores indicating more independence in everyday life. Given that all of these items were administered at Time 1, it was possible to calculate a comparable score (i.e. 9 to 34) from the original questionnaire, to permit calculation of a change score.

Procedure

Testing sessions were completed between March and August 2005. An attempt was made to test the participants as close to 18-months from their initial session as possible and this was achieved quite well. The average time between the initial and final testing session was 18.96 months (SD = 0.8, range = 17.8 – 21.8 months).

The procedure was very similar to the initial testing session. Participants were contacted and a suitable time was organised for the session, either at their home or at the university. About two weeks before the session, a package of questionnaires was mailed to each participant's home for them to complete. The questionnaires were on cognition in daily life and activities of daily living. Once at the testing session, the experimenter examined the questionnaires for completeness, in order to reduce the problem of missing data, encountered at the initial session.

At the final test session, all of the tasks were re-assessed using exactly the same procedures as has been described in Chapter 4. However, the participants were given the opportunity to complete the testing over two different days if they preferred. Some participants were unwilling to complete another 4-hour testing session, and in an attempt to retain them in the

study we offered this alternative. If they chose this option, two 2-hour sessions were organised within a 2-week period and as close together as possible. The participant completed as much as possible on the first occasion, with no break and then finished the remainder of the tasks on the second occasion. At the end of the testing session, those people who completed the IT retest at Time 1 were asked to complete another IT retest. Of the 26 people completing on the first occasion, a total of 20 people completed the re-test a second time. Full data collection took an average of 4-hours with each participant to complete.

Results

Two main aspects of the functional age outcomes were of interest; the *final score* and the *18-month change score.* The results section had therefore been divided into two sections, to deal with each in turn. For the *final score* the distribution of each measure was examined for normality, because linear regression analysis assumes that the residuals are normally distributed, which is largely determined by whether the dependent variable is normally distributed[9]. In addition, each measure was examined for ceiling and floor effects and dispersion (standard deviation and range), in order to decide whether they were valid outcome measures. For those tasks that were administered twice, the correlation over time as a lower bound for test-retest reliability was also examined. For the *18-month change scores,* the extent of missing data was considered, together with the stability of each of the constructs and the extent of individual differences in the change scores. In addition, we considered whether each of the change scores were normally distributed and whether there were gender differences in the amount of change over 18-months. It was anticipated that these analyses should provide a clearer picture of which measures were likely to be the most useful functional age outcomes.

The Final Score

Table 6.1 shows descriptive statistics for the measures assessed at the final testing session. The first column shows the number of people who completed each tasks and it is clear that there was very little missing data. The next three columns show the mean, standard deviation and range. The final two columns show skewness and kurtosis information. The

[9] Linear regression analyses are planned for Chapter 7, to evaluate the utility of the biomarkers to predict the functional outcomes.

values in the table are the actual statistic for skewness (or kurtosis) divided by the standard error, thereby permitting direct comparisons of these values across tasks. As a rule-of-thumb, statisticians often suggest that these values should be between –2 and 2 for normally distributed variables.

Table 6.1. Descriptive Statistics for Functional Outcomes at Time 3

Outcome	n_3	$Mean_3$	SD_3	Range	Skewness	Kurtosis
ADL	126	32.46	(2.62)	19 - 34	-11.41	16.99
CDL	125	44.31	(9.72)	25 - 70	0.21	-0.90
RSPM	125	14.02	(4.38)	3 - 23	-0.68	-1.16
CCFT	122	23.81	(5.73)	12 - 36	0.21	-1.20
CF	122	21.43	(9.19)	1 – 35	-1.32	-2.53
Similarities	127	22.97	(4.89)	11 – 32	-1.56	-1.85
Information	123	28.71	(5.07)	16 – 37	-1.93	-1.80
Spot-the-word	127	53.53	(3.42)	44 – 60	-2.26	-0.60

Note. ADL = Activities of Daily Living, CDL = Cognition in Daily Life, RSPM = Raven's Standard Progressive Matrices, CCFT = Cattell Culture Fair Test, CF = Concept Formation.

Were the Final scores normally distributed?

Everyday Functioning. The first everyday functioning measure was ADL as assessed by the reduced version and it is quite clear from Table 6.1 that this measure was not normally distributed. More than half of the participants (59%) registered the top score for total independence and this led to a distribution that was highly negatively skewed, with a very high kurtosis value. Although, 41% of the participants scored below the maximum, most still received high scores, so that the range and standard deviation were quite small (i.e. low dispersion). The one positive attribute of this task was that the test-retest correlation over 18-months was high (r = .825), suggesting that the task was reliable (see Table 6.2). Overall, this scale was not a good functional outcome measure, because it was non-normally distributed, had a clear ceiling effect and had low variability.

The other measure used to assess everyday functioning was the CDL questionnaire. An examination of this scale showed that it was normally distributed, confirmed by the low skewness and kurtosis scores and the Shapiro-Wilk Normality Test ($S\text{-}W$ (125) = .985, $p > .05$). There were no problems with ceiling or floor effects and the range and standard deviation were reasonably high, indicating that this scale differentiated between the participants quite well. Because this task was only measured at the final testing phase it was not possible to examine the reliability but high internal consistency and test-retest reliability have been established in the

literature[10]. The CDL measure therefore appears to be a suitable outcome measure for *everyday functioning* because it is normally distributed, reliable and showed a wide dispersion of scores.

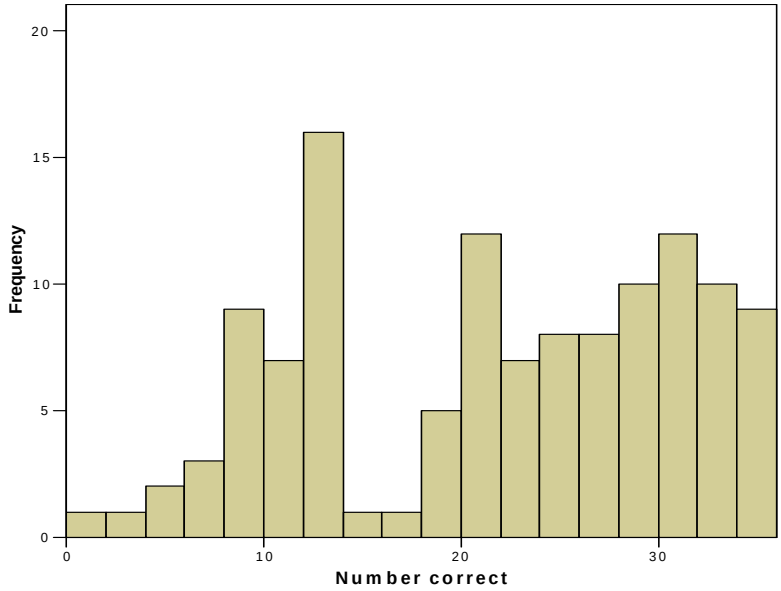

Figure 6.1. Performance on Concept Formation at Time 3

Cognition. This construct was divided into fluid ability and crystallised ability because these two broad abilities have been well established as showing differential decline with age. *Fluid ability* was assessed by three tasks: RSPM, CCFT and CF. Scores were normally distributed for RSPM (*S-W* (125) = .985, $p > .05$) and for CCFT (*S-W* (122) = .984, $p > .05$). Examination of Table 6.1 shows that there were no problems with ceiling or floor effects for these two tasks and that the range and standard deviation were good, suggesting that the tasks differentiated between individuals on fluid ability. Furthermore, the correlations in Table 6.2 indicated that both tests had adequate test-retest reliability. However, the distribution for CF was clearly non-normal, as indicated by the normality test (*S-W* (125) = .942, $p < .01$) and the high kurtosis value. In fact, this distribution appeared to be bi-modal (see Figure 6.1); that is, there appeared to be one group who scored between 0 and 15 and another qualitatively different group who scored between 15 and 35. This outcome may lead to a violation of the assumptions in the subsequent linear regression analyses planned. Thus, RSPM and CCFT were more suitable outcome measures for fluid ability than CF.

---

[10] Stip et al. (2003) found that the CDL measure had good internal consistency (Cronbach's alpha = .858) and test-retest reliability (r = .82).

For *crystallised ability*, none of the three measures met the normality requirements of the formal normality test. However, there is some indication that the Shapiro-Wilk Normality Test is oversensitive in cases of large samples and that, in this case, the skewness and kurtosis values should be examined. An examination of these values showed that all three tasks were slightly negatively skewed (i.e. some indication of ceiling effects) but the only one that exceeded the rule-of-thumb was Spot-the-Word. None of the tests had serious kurtosis problems, with all values between –2 and 2. In terms of dispersion, the Spot-the-Word task also has the smallest standard deviation but all three tasks were able to differentiate between the crystallised ability of the sample quite well. Furthermore, Table 6.2 shows that all three tasks were highly reliable. To conclude, there were some problems with ceiling effects in the crystallised ability tasks; and the Spot-the-Word task was the most problematic of all three tasks.

The 18-month Change Score

Four issues were examined with respect to the change scores. First, the stability of each of the measures was tested by comparing the means at Times 1 and 3 and the considering the correlation over time. Second, individual differences in the change over time were examined, to try to identify a group of people who were showing signs of accelerated aging. Third, the distributions of the change scores were examined, to determine whether they were approximately normally distributed. Finally, Birren and Fisher's (1992) suggestion that, because males have a shorter life span, they should decline more in a range of age-related tasks than females, was tested by comparing male and female change over the 18-month period.

How Stable were the Functional Outcomes?

Table 6.2 shows descriptive statistics for the functional outcomes completed at both Time 1 and 3. Three points should be noted. First, considerable data were missing for CCFT and CF. These missing data were generated at Time 1 and fewer than 100 participants completed these tasks on two occasions. One possible problem with these missing data is that the people who completed CCFT and/or CF on both occasion might be qualitatively different from the people who did not. Therefore, the results for CCFT and CF may not be directly comparable to the results from RSPM, although they are all fluid ability tasks.

Second, most of the tasks were stable on average over the 18-month period (i.e. similar mean values at Times 1 and 3). For each of the seven measures, a repeated measures ANOVA tested whether there was a significant mean difference between scores at Times 1 and 3. A

significant decline was found in mean performance on RSPM (Wilks' Lambda = .561, F (1, 124) = 96.94, $p < .01$, partial $\eta^2 = .44$). Similarly, there was a significant decline in ADL (Wilks' Lambda = .919, F (1, 125) = 10.99, $p < .01$, partial $\eta^2 = .08$).

Third, the correlations between Time 1 and 3 scores were high, suggesting that the individual differences at Time 1 were, on average, quite stable. Correlations of this magnitude suggest that between 35 and 65% of the variance in the Time 3 score was due to the Time 1 score. Therefore, other between-subject factors (e.g. aging) have certainly contributed to the Time 3 score. On the whole, however, we could conclude that the measures were very stable over the 18-months. We next consider which measures were the most stable and which showed the most change.

Table 6.2. Stability of Functional Outcomes over 18-months

Outcome	n_{13}	$Mean_1$	SD_1	$Mean_3$	SD_3	r_{13}
ADL	126	32.90	(1.89)	32.46	(2.62)	.825**
RSPM	125	17.13	(4.60)	14.02	(4.38)	.691**
CCFT	95	23.99	(5.58)	24.17	(5.75)	.763**
CF	85	22.85	(7.13)	22.55	(8.83)	.657**
Information	120	28.87	(4.80)	28.68	(5.11)	.785**
Spot-the-word	127	53.99	(4.43)	53.53	(3.42)	.690**
Similarities	125	22.55	(4.61)	22.93	(4.91)	.758**

Note. ADL = Activities of Daily Living, RSPM = Raven's Standard Progressive Matrices, CCFT = Cattell Culture Fair Test, CF = Concept Formation.
** $p < .01$

Two constructs used to measure functional age: *everyday functioning* and *cognition*[11]. For ADL, there was a significant mean decline over 18-months. Although the mean difference was small, it represented a significant decline because the standard deviation for the task was so small. The correlation from Time 1 to 3 was the highest of all the measures (i.e. 69% shared variance), which suggests that most the sample were showing a small decline over the 18-month period in everyday functioning.

For *cognition*, it is useful to consider the fluid ability and crystallised ability tasks separately. One of the *fluid ability* tasks, RSPM, showed marked decline over the 18-month

[11] *Everyday functioning* is defined by ADL and CDL. However, there were no change scores available for CDL so this section simply discusses ADL as the sole measure of *everyday functioning*.

period, while the other two tasks were relatively stable. One possible explanation for this discrepancy is the effect of the missing data for CCFT and CF. It is certainly possible that the people who failed to complete these tasks at Time 1 tended to have a lower fluid ability and/or a higher rate of decline in fluid ability over time. If this was the case then we would be less likely to detect a significant decline in CCFT and CF, due to the higher homogeneity of these samples and also because of reduced power.

This possibility was examined by splitting the sample into three groups depending on which fluid ability tasks they completed at Time 1. The three groups were (1) those participants who completed all three fluid tasks at Time 1 (n = 83), (2) those who completed two fluid tasks (n = 49), and (3) those who completed just one fluid task (n = 18). These three groups were compared on their initial RSPM score and the 18-month change in RSPM, using ANCOVA with age and gender as covariates. If the hypothesis were correct, then we would expect the people who completed just one fluid task (i.e. missed CCFT and CF) to have lower RSPM scores at Time 1 and/or show more decline over the 18-month period than the other groups.

For initial level, the group who completed just one fluid task did have marginally lower RSPM than the other two groups ($M_1 = 16.43$, $SD_1 = 4.94$; $M_2 = 17.29$, $SD_2 = 4.20$; $M_3 = 15.61$, $SD_3 = 5.08$), but the difference was not statistically significant (F (2, 145) = 0.96, $p > .05$, partial $\eta^2 = .01$). Similarly, the group who completed one fluid task experienced more decline than the other two groups over 18-months in RSPM ($M_1 = 0.09$, $SD_1 = 3.27$; $M_2 = 0.29$, $SD_2 = 3.18$; $M_3 = -1.19$, $SD_3 = 2.46$) but again this difference was not statistically significant (F (2, 120) = 1.21, $p < .05$, partial $\eta^2 = .02$). However, the observed power for each of these tests was low (< .30) and therefore the probability of a Type II error was high. To conclude, there was no convincing evidence that the samples who completed CCFT and CF at Time 1 and 3 were more homogeneous because they did not include those participants with the lowest fluid ability and highest rate of change. However, it remains possible because comparisons lacked sufficient power to prove this adequately.

Finally, for *crystallised ability*, all three tasks showed marked stability and this was consistent with expectations. There was no evidence of mean decline in any of the tasks and the 18-month correlations were all high ($r \geq .690$). If a subgroup does show decline in these tasks then it is quite possible that they are showing accelerated aging, given that these tasks were, on average, so stable. To conclude, the most stable construct was crystallised ability, followed by everyday functioning and fluid ability.

Were there Individual Differences in the Stability of the Functional Outcomes?

There is evidence that some of the constructs were relatively stable over 18-months while others declined. However, we are primarily concerned with individual differences and would like to ascertain whether a sub-sample of people displayed signs of accelerated aging. For each of the seven functional age measures, a change score was calculated following the methods described in the previous chapter (see p. 108). The distribution for each change score (c_1) was then split into quartiles and the mean Time 1 and Time 3 score calculated for each quartile. The difference between these mean scores was calculated to see the magnitude and direction of change for each quartile. This change was plotted on a graph by using zero as the starting point so that all quartiles could be compared and the extent of individual differences between people in their change over 18-months could be examined.

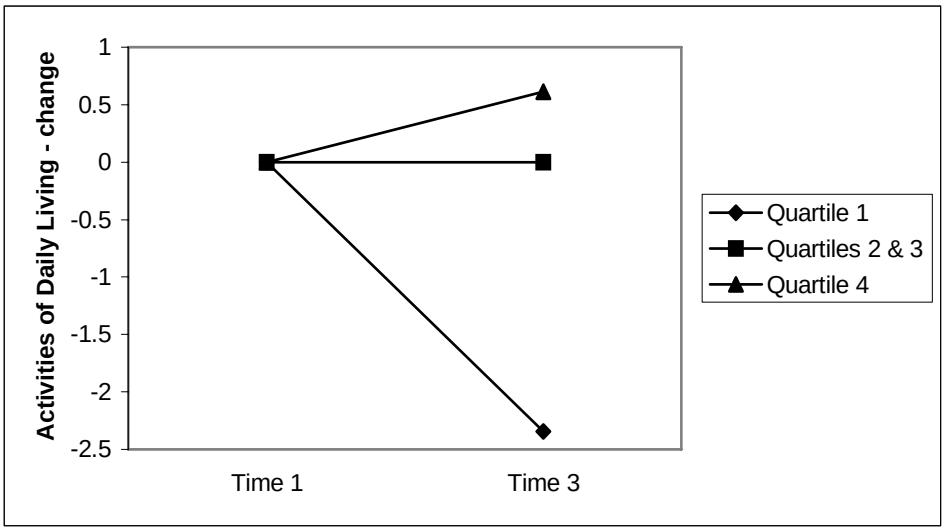

Figure 6.2. Activities of Daily Living change over 18-months by quartile

Everyday functioning. There was high stability in this ADL measure, with 59% of the sample showing absolute stability at the highest level of independence (i.e. 34/34). Therefore, the middle line in Figure 6.2 actually represents two-quartiles of the sample or one half. One group (Quartile 1) declined an average of 2.3 points over the 18-months. Although, a change of this magnitude appears small, it actually represents a decline of about 1 standard deviation. Furthermore, this is the average decline for that quartile, so that some individuals declined more than this. The maximum decline over the 18-month period was an individual who had an initial score of 31 and a final score of 24, which represents a decline of over 3 standard deviations.

Therefore, there were large individual differences in the change over 18-months in ADL, with some people showing sizable declines while others exhibited stability.

Raven's Standard Progressive Matrices. This task was one of the functional outcomes where the participants showed a significant mean decline over the 18-month period. This is very clear from an examination of Figure 6.3. Participants in the fourth quartile showed a small improvement in the task but all other groups declined. Quartile 1 declined considerably over the over 18-months (mean decline = 7 points), which equates to about a 1.5 standard deviations decline from the start of the study. This suggests that RSPM is very sensitive to declines in cognition with age and may therefore be a useful outcome measure.

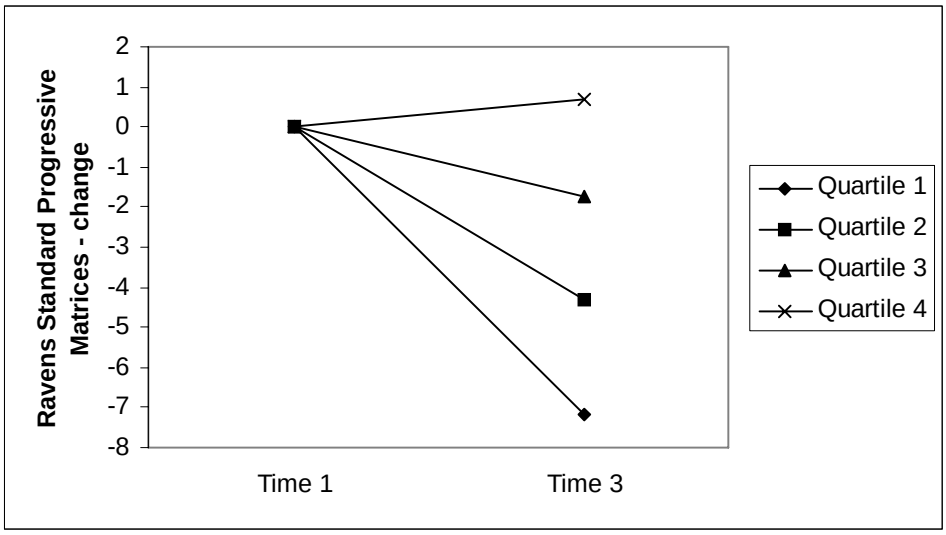

Figure 6.3. Raven's Standard Progressive Matrices change over 18-months by quartile

Cattell Culture Fair Test. Performance on this task was quite stable, on average, but it is apparent from Figure 6.4 that there was a group (Quartile 1) who declined over the 18-months by an average of 5 correct items. A decline of this magnitude is equal to just under one standard deviation. Another group (Quartile 4) displayed a large improvement in the task over 18-months. This improvement could be explained as a practice effect, and if so, may indicate that the decline experienced in Quartile 1 was actually an underestimation of their true decline because they should also have been affected by practice[12].

[12] Comparison between results for RSPM (where parallel forms were used) and CCFT (where they were not) support this suggestion.

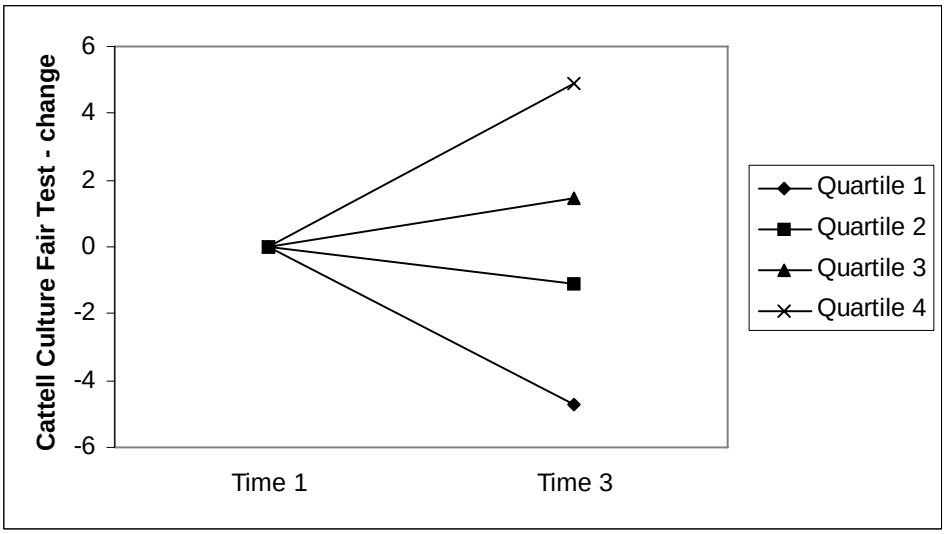

Figure 6.4. Cattell Culture Fair Test change over 18-months by quartile

As reported earlier, the sample of people who completed the CCFT (n = 95) on both occasions was quite likely more homogeneous than the full sample. Yet, despite this, some individuals still exhibited sizable declines in this task over 18-months, which may be an underestimate of their true decline due to practice effects. This suggests that the CCFT may indeed be as sensitive to age-related decline as RSPM, and prove to be a useful functional outcome.

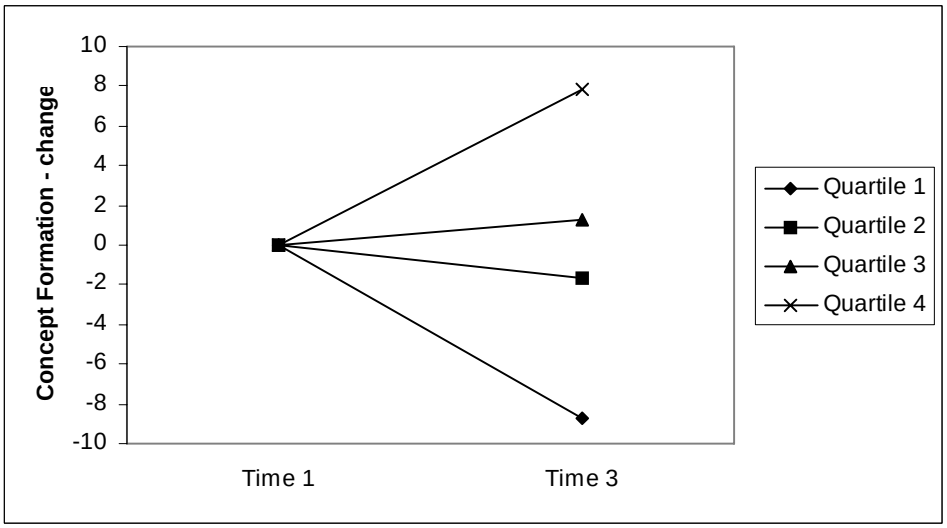

Figure 6.5. Concept Formation change over 18-months by quartile

Concept Formation. Just 82 people completed CF at Times 1 and 3, the smallest sample for any task. On average, the sample showed a small and non-significant decline in this task over

the 18-months. However, examination of Figure 6.5 shows a group with a mean decline of about 9 items correct. Further examination showed that this quartile achieved 21 items correct at Time 1 and 12 items correct, which represents a large decline of 1.3 standard deviations. Furthermore, Figure 6.1 shows that the distribution of scores for Concept Formation was bi-modal, with one mode at about 14 and the other at about 22. Thus, this group of people could be thought of as declining from the top distribution into the bottom distribution, which could indeed be a very important change. Furthermore, large practice effects were apparent in this task. with Quartile 4 displaying an improvement of about 1 standard deviation over 18-months. Therefore, as for CCFT, the true decline for CF by Quartile 1 may have been underestimated due to the effects of practice on all groups.

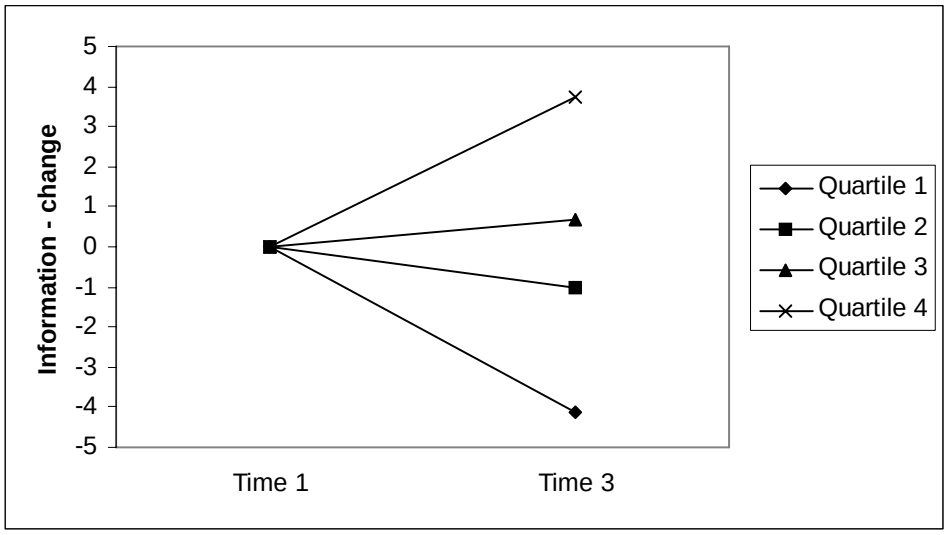

Figure 6.6. Information change over 18-months by quartile

Information. This is a test of crystallised ability, which is generally considered to be relatively stable throughout the lifespan. Information, like the other crystallised ability measures, did not show a significant mean decline over the 18-months. However, the *terminal drop hypothesis* suggests that when people begin to experience a drop in crystallised ability, this can be indicative of impending death. There was evidence of a decline in crystallised ability for one of the groups, where the mean score dropped by 4 items correct over 18-months. A change of this magnitude represents less than a standard deviation but, because tasks of this nature tend to be so stable, a decline of this magnitude may be important.

Spot-the-word. This task assesses vocabulary, which is one of the variables that tend to increase or remain stable into late life. Overall, the participants had an extremely high vocabulary and the task showed evidence of ceiling effects. Nonetheless, examination of Figure

6.7 shows that one group registered a drop in vocabulary of 4 words, which is just less than one standard deviation. This indicates that some individuals were experiencing a decline in vocabulary with age. However, there is another possible explanation. For this task, the participants completed parallel versions of the task at Times 1 and 3. Each version had a different list of words, knowledge for which was deemed to be equivalent difficulty. It is possible therefore that the words from the second list were marginally more obscure, which would account for the apparent decline over the 18-month period. Baddley, Emslie and Nimmo-Smith (1993) administered Form A and B to 50 adult participants aged 20 – 85 years (M = 38 years). They found a high correlation between the parallel forms (r = .776) and found that performance was slightly better on Form A (M = 53.0 items correct) than on Form B (M = 52.3 items correct). This pattern is consistent with the findings from this current study, and may suggest, that some of the decline from Time 1 to 3 was due to marginally more obscure words in Form B. Although it is not possible to be sure of the cause of the decline in some individuals, there certainly did appear to be a decline in vocabulary for some individuals.

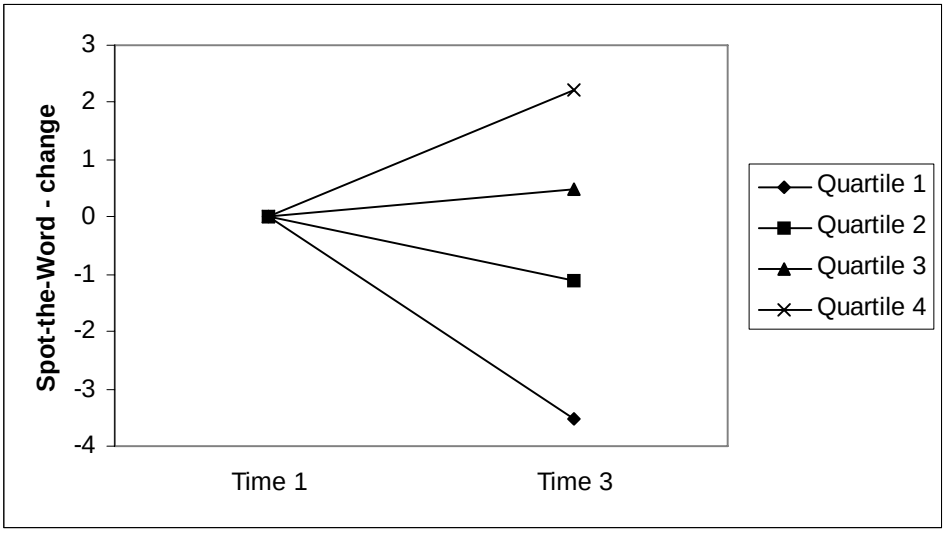

Figure 6.7. Spot-the-Word change over 18-months by quartile

Similarities. The third crystallised ability task was Similarities from the WAIS-III (see Figure 6.8). The same version of the task was administered at both times, so that any decline over the 18-months cannot be attributed to a different set of items. Although the mean score was quite stable, one group showed a moderate decline in the task (mean change = 4 points), which is equivalent to a decline of 0.8 of a standard deviation. Moreover, there were some individuals who declined up to 7 points on this task over 18-months. Thus, once again there were individuals

who declined in crystallised abilities and this decline was of about the same magnitude as for the other crystallised ability tasks.

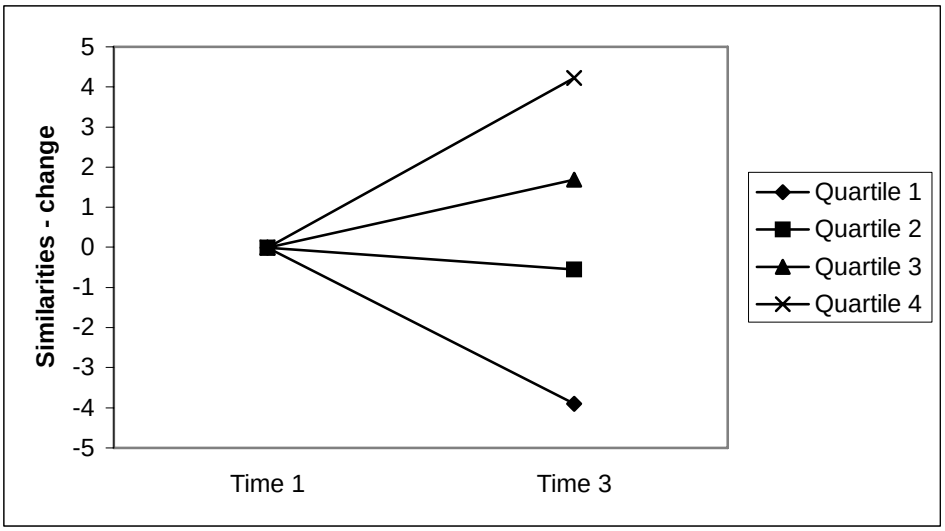

Figure 6.8. Similarities change over 18-months by quartile

Were the 18-month Change scores Normally Distributed?

As with the final scores, the change scores should also be approximately normally distributed if they are going to be suitable to use in linear regression analyses in Chapter 7. Thus, the skewness, kurtosis and Shapiro-Wilk normality test for each of the variables were examined and these values are presented in Table 6.3[13]. For ADL it is very clear that the distribution was not normally distributed. The major problem is one of kurtosis, with the distribution being markedly leptokurtic. Because so many people showed no change, there was a huge peak at zero. In addition the distribution was negatively skewed and the normality test confirmed these problems. However, for fluid ability, it was pleasing to see that all three of the measures had normally distributed change scores, despite the fact that the final score for Concept Formation was non-normally distributed. For crystallised ability, the Information and Similarities tests would be considered approximately normally distributed based on the skewness and kurtosis values, although the Spot-the-Word test was significant negatively skewed and leptokurtic and thus not ideal as an outcome measure. To conclude, the ADL scale was very problematic and there were problems with one of the crystallised tasks, but all fluid tasks were suitable. Based on these findings it was decided not to exclude any of the change scores but to take these issues of

[13] The skewness and kurtosis values are the statistic divided by the standard error.

non-normality into account when interpreting the findings in Chapter 7. Specifically, the ADL scale will be analysed using logistic rather than linear regression. The scale will be converted to a dichotomous variable representing decline or stability in ADL and logistic regression will test whether the biomarkers can correctly identify which group the participants belong to. As for Spot-the-Word, this will be analysed using linear regression and the assumptions of the test will be examined. If the assumptions are violated then logistic regression will be used. I

Table 6.3. Normality statistics for 18-month Change scores on Functional Outcomes

Outcome	Skewness	Kurtosis	S-W statistic
ADL	-5.19	11.95	.781**
RSPM	-0.89	-0.37	.992
CCFT	-0.26	-1.53	.982
CF	-0.86	0.45	.983
Information	0.99	0.60	.990
Spot-the-word	-2.49	2.32	.971**
Similarities	-1.91	-1.14	.969**

Note. ADL = Activities of Daily Living, RSPM = Raven's Standard Progressive Matrices, CCFT = Cattell Culture Fair Test, CF = Concept Formation.
** p < .01

Were there Gender Differences in the Stability of the Functional Outcomes?

Functional age outcomes were examined for gender differences in the mean change over 18-months. Females showed higher mean decline in ADL than did the males (t (123) = 2.02, p < .05, d[14] = 0.31). For CF, the females showed a decline in mean performance while the men showed an improvement and this difference was statistically significant (t (83) = 2.44, p < .05, d = 0.56). Therefore, within this sample, females displayed greater decline in everyday functioning and fluid reasoning with age.

Discussion

The aim of this chapter was to investigate the functional outcomes measures extensively, in order to decide whether they were suitable statistically, before considering whether the

[14] Cohen's d is calculated by the mean difference divided by an estimate of the pooled standard deviation.

biomarkers were reliable predictors of the outcomes. For each outcome the *final score* and the *18-month change score* have been examined separately, to test adequacy. Taken together, the individual analyses for the *final score* (i.e. normality tests, test-retest reliability) and the *18-month change score* (i.e. repeated measures ANOVA, correlations) have suggested an answer to the question of whether the operationalisations for these constructs have provided suitable outcome measures for functional age.

Everyday functioning. It is clear that the composite ADL scale, assembled from basic and instrumental activities of daily sources, was not sufficiently sensitive to putative individual differences in level of everyday functioning in this sample. Participants in this study were functioning very independently in their day-to-day lives and the ADL scale was not able to differentiate between people at the high end of ability. As a result, final scores revealed a large ceiling effect, resulting in a negatively skewed and leptokurtic distribution. Furthermore, the change score had an even higher kurtosis value, given that 68% of the participants displayed no change over 18-months. These findings do not necessarily represent a general problem with the ADL scale because results also showed that the correlation was high over the 18-month period (i.e. high test-retest reliability). Rather, these results suggest that the scale is not suitable for use in a sample that as a whole was as independent as ours.

Due to these problems with the ADL scale, a second everyday functioning scale was therefore introduced for the final testing phase (see p. 125). The Cognition in Daily Life (CDL) scale asked questions that more specifically related to the cognitive aspects of everyday functioning and results confirmed that the scale was approximately normally distributed, with no signs of ceiling or floor effects. This outcome confirms that the CDL questionnaire is more suitable for use with high functioning samples and if further waves of data collection were possible then it is recommended that this scale should be re-administered. However, it should be noted that this scale does not assess exactly the same abilities as the ADL scale and should not be thought of as a replacement for that scale. Ideally, a scale that provides a global measure of the independence of everyday functioning, rather than just cognitive aspects, and that differentiates between the highest functioning participants, is still required.

Fluid Reasoning. Two of the fluid reasoning tasks – RSPM and CCFT – proved to be very good outcome measures. For both, the final scores and the change scores were normally distributed, the final score was highly reliable, and there were large individual differences in 18-month stability. However, the other fluid reasoning task – CF – has some problems, at least

when the final score is used as an outcome measure. The final score was bi-modally distributed, which may violate the assumptions of linear regression analysis in the next chapter. Nonetheless, the change score for CF was normally distributed and thus presents no problem for use as an outcome measure.

As a personal comment, having now administered CF to a large number of elderly people, this bi-modal distribution is not surprising to the author; and it may be useful for future researchers to summarise at this point these personal observations about this instrument. There are three main types of questions or rules that must be mastered in order to complete this task. The third rule (the "or" rule) posed significant problems to a lot of the participants, and if they could not solve this rule, they failed to score any more items correct. Given that instructions for administrating this task included a stopping criterion, those participants who were not able to master the "or" rule were limited to scores around 13. Thus, people who did not understand the "or" rule were represented by the lower distribution and those who mastered this rule were represented by the higher distribution. This is not an ideal characteristic for a psychometric task and, in hindsight, it was not a suitable fluid ability task to include.

The other point worth mentioning concerns the missing data in the CCFT and CF. It is unfortunate that these data were not collected but this was primarily the result of such a large test battery, which some participants found onerous. Analysis has not established statistically that the group of people with missing data for these tasks had lower fluid ability or a higher rate of decline over 18-months but the results were at least in that direction. Given that these missing data may not have been random, the effects for RSPM may have been larger than for CCFT or CF and this should be taken into account when evaluating the regression results in the next chapter.

Crystallised Ability. As with the fluid ability tasks, two of the crystallised tasks – Information and Similarities - proved to be adequate outcome measures. The final scores for these two measures were approximately normally distributed, although slightly negatively skewed, and were highly reliable. Furthermore, the change scores were normally distributed and there was evidence of declines in some individuals of up to one standard deviation from the initial session. However, there were problems with the third crystallised ability task – Spot-the-Word. The distribution of the final scores was negatively skewed, because there was a ceiling effect, and there was quite a small range and standard deviation. Furthermore, change scores were also negatively skewed and leptokurtic and therefore non-normally distributed. Given that

the mean vocabulary score was so high at Time 1, most people had to show stability or decline, because there was little room for improvement and this led to the negatively skewed distribution for change scores. Furthermore, given that vocabulary was so stable, there were many individuals who showed very little change, and, consequently the distribution was very peaked (i.e. leptokurtic).

At the start of this chapter it was acknowledged that crystallised ability tends to be very stable, even into old age. Given this, it is therefore encouraging for the aims of this thesis that declines in crystallised ability over the relatively short period of 18-months were detected. For all three tasks, a mean decline of up to one standard deviation in some individuals was observed. This is a substantial decline, given that it occurred over a period of just 18-months, and it may be particularly important because the *terminal drop hypothesis* (Riegel & Riegel, 1972) posits that death is preceded by a decline in crystallised cognition over about a 5-year period. This would suggest that those people who are currently experiencing declines in crystallised ability might be at a higher risk of mortality in the next five years. Ideally, we would follow those participants for the next five or more years and examine mortality data but this commitment was beyond the scope of this investigation. Nonetheless, it appears that we have been able to observe some participants during a period of their lifespan where they are actually experiencing detectable cognitive decline.

Gender effects. One final point to be discussed is the gender effect for ADL and Concept Formation. Females displayed larger 18-month declines in these measures than males. This finding was contrary to the hypothesis of Birren and Fisher (1992) that predicted that males should show more decline, because they have shorter average lifespans. However, it was consistent with the results from Chapter 5 that found females showed significantly more decline in grip strength and a significantly higher increase in diastolic blood pressure over 6-months. It is worth speculating why females in this sample did show larger declines than males.

Males have a shorter life expectancy than females. So, males who are still alive into their 70s and 80s could be thought of as examples of ‘successful’ aging, simply because they are still here. By contrast, a group of females alive into their 70s and 80s could be thought of as including some females who are aging successfully and others that have simply not reached their life expectancy. It is those people who are not showing successful aging who would be expected to show declines in some of the measures. However, the male group included few, if any, people

who were not aging well therefore it might be expected that the female group would show more signs of decline.

In the literature, there is one study that found evidence of larger declines among males than females. Mortensen and Kleven (1993) found that males declined more from 50 to 70 years than females in Digit Symbol, Object Assembly, Picture Arrangement and Information. However, this period of the lifespan (50 – 70 years) is prior to the current average life expectancy for males or females. Therefore, it may be valid to expect males to show more decline during the years when males are approaching their average life expectancy. But, during a later period of life (i.e. 70+), it may not be valid to expect that males should show more decline, because the males who are still alive are in a sense a more homogeneous and highly functioning group. To conclude, there is evidence in this study that females declined more in physiological, cognitive and everyday living measures, which may provide some evidence of a flaw with the underlying theory of Birren and Fisher (1992).

Having considered each of the seven biomarkers and each of the eight functional age measures we are now in a position to evaluate the predictive validity of IT. Chapter 7 will present extensive regression analyses to determine whether IT has predictive validity. That is, the main issue is whether IT is able to predict performance on the functional outcomes at the end of the study and/or decline in the functional age outcomes over the 18-month period of the study. This is the ultimate test of whether IT could be used a valid biological marker for functional aging.

CHAPTER SEVEN: STUDY 4 - PREDICTIVE VALIDITY

The aim of this chapter was to determine whether Inspection Time (IT) is a valid biological marker for functional age. In Chapter 4, correlations between IT and everyday functioning, fluid ability and a measure of crystallised ability were established cross-sectionally. However, a biomarker must be able to *predict* performance on functional aging outcome measures. That is, level or change in a biomarker must be predictive of functional outcomes in the future. Chapter 7 will extend the work presented in Chapter 4 by examining the predictive validity of the biomarkers. Figure 7.1 illustrates four different ways in which the predictive validity of IT and the other markers can be tested and therefore sets the scene for this chapter.

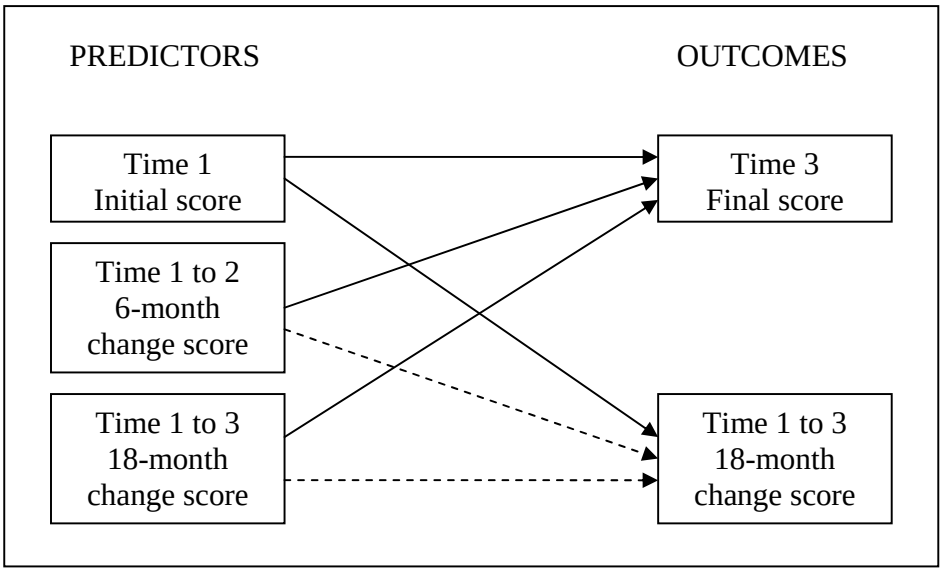

Figure 7.1. Models for Assessing Predictive Validity of Biomarkers

The solid lines in Figure 7.1 illustrate the four valid methods of assessing the predictive validity of the biomarkers; the dotted paths are not valid, as will be clarified below. The first method is to test whether initial scores on the biomarker can predict final scores on the functional outcome measure. For example, do people with quick ITs at Time 1 tend to have high fluid ability scores at Time 3? The second method involves testing whether 6-month change scores on the biomarker can predict final scores on the functional outcome measure. For example, does someone whose IT becomes disproportionately slower over 6-months than the average of the sample tend to have lower scores on tests of fluid ability at Time 3? The third method is to examine whether individual differences in 18-month change scores on the biomarker can predict

final scores on the functional outcome measure. However, this method does not test predictive validity in the true sense because the predictor variable can only be calculated at the same time as the outcome measure and not before. Therefore, for each of the biomarkers, there are three measures (initial score, 6-month change and 18-month change) that can be used to predict the final score on the functional outcome measure. It is important to note that initial scores from biomarkers are statistically independent from both the 6-month and 18-month change scores, due to the residual change method that has been used to calculate them. However, 18-month change scores are linearly dependent on 6-month change scores and hence inclusion of both in a multiple regression model may violate the multi-collinearity assumption. Therefore, it would be valid to consider the amount of variance that initial scores and a single set of change scores (not both) could explain in final scores. In situations where 6-month and 18-month change scores were significant predictors, it would be more appropriate to consider 6-month change scores because these confirm true predictive validity for the final outcomes.

The fourth method of establishing predictive validity is to establish decline over 18-months in the functional outcome measure and test whether the biomarker can predict this. As shown in Figure 7.1, it is appropriate to consider initial scores as predictors of functional decline (e.g. are slow ITs at Time 1 associated with decline in fluid reasoning over 18-months?); but invalid to consider 6-month or 18-month change scores as predictors, because the time frame for the predictor and outcome measures would overlap.

Method

Details of the sample, materials and procedures have been provided in the previous three chapters. Therefore, this method section will focus on the statistical analyses that were used to test the predictive validity of the biomarkers in this chapter using the four methods detailed above.

Statistical analyses described in this section were completed for each of the eight functional age outcome measures (ADL, CDL, RSPM, CCFT, CF, Information, Spot-the-Word and Similarities). First, each of the biomarkers was evaluated individually to test whether they explained significant variance in the functional outcome measure. Using Method 1 as an example, initial scores from each biomarker were considered to test whether they were significant predictors of final scores on each of the functional outcomes. Some of the biomarkers and the functional age outcome measures had significant gender effects, which resulted in the biomarkers

appearing to be significant predictors when they were actually acting as proxies for gender. To circumvent this possibility, hierarchical regression analyses were performed that entered gender at Step 1 and the biomarker at Step 2. These analyses allowed for the calculation of the R^2 change at Step 2; and where this statistic was statistically significant it was concluded that the biomarker was a significant predictor of the functional outcome measure. This model was run for each of the biomarkers, for each of the four methods and for each of the eight outcome measures. In addition, when initial scores provided the predictor variable (i.e. Methods 1 and 4) the predictive validity of age was examined.

Once initial analyses were completed it was possible to combine the results from all of the individual analyses, by running a second set of multiple regression analyses to determine the most important predictors of final scores and 18-month change scores for each of the functional age outcome measures. Gender and age were entered first, if they had been significant predictors in the initial analyses. Moreover, if gender-related biomarkers were entered into the model (i.e. grip strength, height and weight) it was necessary to include gender as a predictor regardless of whether it was related to the outcome measure. Second, the significant IT measures were entered (initial scores and/or change scores). If the original analyses found that 6-month and 18-month change scores were significant then just 6-month change scores were entered. Third, other significant biomarkers were entered into the model. This allowed for consideration of the most powerful predictors for each functional outcome but also whether the biomarkers were still significant once age-related variance had been considered.

Three further points should be addressed; the predictive validity of the perceptual speed tasks; the increased probability of Type I errors with multiple testing; and the interpretation of the direction of effects in the regression analyses. As already described in Chapter 2, many widely used perceptual speed tasks are problematic for use as biomarkers because they tend to reflect cohort effects and almost all involve a potential trade-off between speed and accuracy. Arguably, also, those requiring both speed and accuracy involve higher order cognitive processes, so that decline in speed of responding might be attributable to a range of sources. Nonetheless, despite these problems, perceptual speed tasks can be regarded as sharing processes with IT, to the extent that all load together on a common actor, so examination of the predictive validity of the perceptual speed tasks may help to shed light on the relationship between IT and the outcome measures. Therefore, the perceptual speed tasks were evaluated in the same way as the biomarkers but were not included in secondary multiple regression analyses.

The second point involves the increased likelihood of Type I errors, given so many individual analyses. Given that we are evaluating seven biomarkers and three perceptual speed tasks for each functional outcome, there is an increased likelihood that significant results could occur due to chance. However, when considering the predictive validity of change scores, consistent results for both 6-month change scores and 18-month change scores would indicate that chance effects are unlikely. Thus, if, for example, the 6-month and 18-month change scores for weight were significant predictors of fluid ability, then it is less likely that both effects have occurred by chance. On the other hand, inconsistencies between the two change scores might indicate that significant results are simply due to chance. When considering initial scores as predictors, we can examine the effect size (R^2 in regression analyses) and in the case of cognitive outcomes, compare the results across different measures. To summarise, it is acknowledged that the method of statistical analysis used increases the likelihood of Type 1 errors; but this will be reduced by comparing results across predictors (6-month and 18-month change scores) and across measures (e.g. RSPM and CCFT), rather than adjusting the significance level for each analysis.

The third point involves the interpretation of the direction of effects in the regression analyses. In most regression analyses it is straightforward to identify the direction of the effects by examining the sign of the β-value in linear regression or the Wald statistic in logistic regression. However, it is more difficult to establish the direction of these effects when the predictor or the outcome measure is a difference score. A negative difference score indicates that the individual got a lower score on the second occasion. So, in most cases a *positive β-value* (or Wald Statistic) indicates that people who decline over time on the biomarker tend to score lower on the functional outcome at Time 3; or, when considering Method 4 (see Figure 7.1), that people with lower initial scores on the biomarker tend to decline over the subsequent 18-months on the functional outcome measures. However, interpretation is further complicated because a few of the measures (Blood Pressure, IT and Visual Acuity) are scored such that high scores represent poorer performance. As a consequence, a negative change score indicates that the individual actually improved the second time (e.g. reduced BP, with improved speed or better vision). Thus, for these measures *negative β-values* indicate that people who deteriorate over time on the biomarker tend to get lower scores on the functional outcome at Time 3. Or conversely, that people with high initial scores (indicating poor performance) tend to decline over the subsequent

Table 7.1. Hypothesised Direction of Effects for Predictive Validity

Biomarker	Direction of Predictor	Direction of Outcome
Inspection Time	Longer IT at Time 1 Increase in IT over 18-months	*Everyday Functioning* 1) Higher dependence in ADL at Time 3 2) Decline in ADL over 18-months 3) Higher scores on CDL scale at Time 3 *Cognition* 1) Lower scores on fluid ability measures 2) Decline in fluid ability over 18-months 3) Decline in crystallised ability over 18-months
Age	Older chronological age	
Grip strength	Lower scores on grip strength at Time 1 Decrease in grip strength over 18-months	
Blood pressure	Higher scores on BP at Time 1 Increase in BP over 18-months	
Weight	Lower body weight at Time 1 Decrease in weight over 18-months	
Height	Shorter height at Time 1 Decrease in height over 18-months	
Visual Acuity	Higher scores on Visual Acuity at Time 1 Increase in Visual Acuity scores over 18-months	
Perceptual Speed	Lower scores at Time 1 Decrease in perceptual speed over 18-months	

Note. BP = Blood Pressure

18-months on the functional outcome measures. The point is that care must be taken when interpreting the direction of the effects.

To assist interpretation of results, Table 7.1 sets out the hypotheses for this chapter and indicates the direction that is expected for each of the effects. For instance, it is hypothesised that higher scores on IT at Time 1 will predict higher dependence in ADL, decline in ADL over 18-months, higher scores on CDL, lower scores on the fluid ability tasks, decline in fluid ability over 18-months, and decline in crystallised ability over 18-months. In terms of declining performance on the biomarkers, it is hypothesised that increases in IT over 6-months and 18-month would predict higher dependence in ADL, higher scores on CDL, and lower scores on the fluid ability tasks. The expected direction of the effects for chronological age, each of the biomarkers and perceptual speed can also been seen from Table 7.1.

Results

This results section is divided into two main sections; everyday functioning and cognition. Within each of the main sections are sub-sections dealing with the more specific functional age outcome measures (e.g. ADL). Each sub-section begins with an overview of the measures and the statistical methods are described. For each sub-section, initial analyses are presented and interpreted, followed by secondary analyses, where applicable, to determine which measures were the best predictors of final and 18-month change scores for the functional outcome measures.

Everyday Functioning

The measures of everyday functioning were Activities of Daily Living (ADL) and Cognition in Daily Life (CDL). A detailed analysis of the ADL scale in Chapter 6 found that it was non-normally distributed and potentially problematic for use in linear regression. An attempt to use linear regression to predict ADL scores confirmed that the assumptions underpinning linear regression were indeed violated (i.e. normality and size of residuals). Therefore, the ADL scale was transformed into a variable with two levels and logistic regression was used. The *dependent (in at least one area)* group consisted of all individuals who scored below the top score (n = 52) and the *fully independent* group consisted of all individuals who scored the top score (n = 74). Thus, the aim was to identify measures that were able to classify participants correctly as either *dependent (in at least one area)* or *fully independent* on the ADL scale.

Similarly, 18-month change scores for ADL were non-normally distributed and violated the assumptions of linear regression. ADL change scores were therefore transformed into a variable with two levels. The *stable* group consisted of 96 people who showed stability or slight improvement (Mean change = 0.2 points, *SD* = 0.6) in their ADL score over 18-months. The *decline* group consisted of 30 people who showed decline in their ADL score (Mean change = -2.5, *SD* = 1.6). A set of logistic regression analyses were run, to identify measures that predicted whether people would belong to the *stable* or *decline* group on ADL.

The CDL scale was normally distributed at Time 3, and linear regression was used to evaluate predictive validity using Methods 1 to 3. However, this scale was not measured at the start of the study, so that there were no 18-month change scores to predict.

Activities of Daily Living

Table 7.2 shows results of the three sets of logistic regressions run to test whether initial scores, 6-month change scores and 18-month change scores for the biomarkers predicted whether people were classified as *dependent (in at least one area)* or *fully independent* on ADL. The second column indicates the sample size for each of the regression analyses. Although 127 people completed the final testing phase, most analyses were based on a slightly smaller sample, due to missing data. The χ^2 step statistic indicates whether the model was significantly improved by the addition of the biomarker. The Wald statistic[15] indicates the size and direction of the effect of the biomarker in the final model and can be thought of as analogous to the β-value in linear regression. Finally, the p-value indicates the probability that the Wald statistic was a change outcome – i.e. statistically significant. In some cases, the χ^2 step can be significant where the Wald statistic is not but the degree of discrepancy is generally low.

For initial scores, there was evidence that age, IT and grip strength were significant predictors of membership into the *fully independent* group. That is, younger people, individuals with shorter ITs (i.e. quicker processors) and individuals with stronger grip strength at Time 1 were more likely to be members of the *fully independent* group at Time 3. In addition, high scores on all three perceptual speed tasks were predictive of membership into the *fully independent* group.

[15] The Wald statistic presented throughout this chapter is equal to B/SE. In some statistical packages (e.g. SPSS) the Wald statistic is the square of B/SE, but this value does not indicate the sign and thus the direction of the effect.

Table 7.2. Predictors of Activities of Daily Living at Time 3

	Initial value				6-month change				18-month change			
Predictor variable	n	χ^2 step	Wald statistic	p-value	n	χ^2 step	Wald statistic	p-value	n	χ^2 step	Wald statistic	p-value
Age	126	10.56**	-3.10**	.002								
Inspection Time	111	4.34*	-1.88*	.049	101	2.78	-1.54	.103	102	2.93	-1.67	.094
Grip Strength	126	5.77*	2.30*	.021	122	9.86**	-2.90**	.004	124	3.96*	-1.93	.053
Systolic BP	114	0.00	0.00	.982	107	0.47	-0.68	.495	109	0.23	-0.50	.629
Diastolic BP	114	0.50	0.71	.483	107	0.73	0.83	.401	109	0.63	0.78	.431
Weight	126	0.22	-0.47	.638	123	0.04	0.19	.837	126	3.94*	1.92	.055
Height	126	0.63	0.78	.430	123	0.44	0.66	.508	126	0.48	-0.69	.490
Visual Acuity	126	0.01	-0.11	.915	123	1.10	-1.04	.298	125	0.18	0.42	.677
Digit Symbol	125	24.72**	4.44**	.000	122	0.06	0.25	.806	125	0.33	0.58	.568
Visual Matching	123	18.34**	3.85**	.000	120	8.25**	2.60**	.009	123	6.30*	2.40*	.016
Pattern Comparison	124	11.69**	3.16**	.001	120	3.48*	1.83	.069	122	4.46*	2.04*	.040

Note: Gender effects were removed before the examination of each biomarker. BP = Blood Pressure.
* $p < .05$, ** $p < .01$

For 6-month change scores, grip strength was the only significant biomarker but the direction of the effect in this case was unexpected and counter-intuitive; i.e. the negative sign of the Wald statistic suggested that people whose grip strength declined over 6-months were more likely to be members of the *fully independent* group at Time 3. The effect for IT approached significance (p = .103), indicating, as expected, that people whose IT increased over 6-months (i.e. whose speed decreased) were more likely to belong to the *dependent (in at least one area)* group. Finally, 6-month change scores were significant for Visual Matching (VM) and approached significance (p = .069) for Pattern Comparison (PC). That is, decline in VM and PC over 6-months was associated with membership into the *dependent (in at least one area)* group at Time 3. Somewhat surprisingly, change scores for Digit Symbol (DS) were not significant, suggesting that although initial scores were almost always significant, change score were not.

Results for 18-month change scores largely mirrored the results for 6-month change scores. For grip strength, 18-month change scores approached significance (p = .053) as did change scores for IT (p = .094) and weight (p = .055). Furthermore, change scores for VM and PC were significant while DS was not.

To investigate these findings further, a secondary logistic regression analysis was run with five predictors (gender, age, IT, grip strength and grip strength - 6-month change) to see to what degree this model could classify people as *dependent (in at least one area)* or *fully independent* group. One hundred and eight people completed all the measures and the five-predictor model was significantly better than the constant only model, χ^2 (5) = 27.92, $p < .001$. This model explained 31% of the variance in ADL (Nagelkerke R^2 = .308) but it did not classify individuals with very high accuracy, with just 54% of *dependent (in at least one area)* people predicted correctly and 80% of the *fully independent* group classified correctly, for an overall success rate of 69%. There were three significant predictors in this model: age, IT and grip strength (6-month change), which indicates that the most important predictors of membership into the *dependent (in at least one area)* group for ADL at Time 3 were old age, slower IT scores at Time 1 and decline in grip strength over the first 6-months of the study. Furthermore, it indicates that IT and grip strength (6-month change) were still significant predictors after the age-related variance had been accounted for. So, it can be concluded that IT is a significant predictor of ADL, at least as classified in the current study.

The final method of examining the predictive validity of the biomarkers tested whether initial scores could predict stability or decline in ADL over 18-months. These results are

presented in Table 7.3. Initial scores for IT approached significance (p = .052) and were in the hypothesised direction. Two of the perceptual speed tasks were significant predictors and the other approached significance, indicating that people with slower perceptual speed were more likely to belong to the *decline* group for ADL. Subsequent analyses were not performed on this outcome measure because none of the biomarkers was a significant predictor in the initial analyses.

Table 7.3. Predictors of change in Activities of Daily Living over 18-months

Predictor variable	n	χ^2 step	Wald statistic	p-value
Age	126	2.68	1.64	.104
Inspection Time	111	3.82	1.88	.052
Grip Strength	126	1.57	-1.24	.216
Systolic BP	114	0.23	-0.50	.631
Diastolic BP	114	0.45	-0.65	.509
Weight	126	2.76	1.71	.098
Height	126	0.08	0.27	.777
Visual Acuity	126	1.22	1.12	.262
Digit Symbol	125	17.75**	-3.71**	.000
Visual Matching	123	7.67**	-2.63**	.008
Pattern Comparison	124	4.07*	-1.94	.051

Note: Gender effects were removed before the examination of each biomarker.
BP = Blood Pressure.
* p < .05, ** p < .01

Cognition in Daily Life

Table 7.4 shows the results of three sets of linear regression designed to assess whether the biomarkers could predict the final scores on Cognition in Daily Life (CDL). It is clear that there were few significant predictors. For initial level, the three perceptual speed tasks were significant, indicating that low scores on perceptual speed at Time 1 were associated with high scores on CDL at Time 3 (i.e. more problems in cognitive aspects of daily living). However, to the extent that the perceptual speed tasks also tap higher order cognitive performance as argued above, it is not surprising that they correlated with the CDL questionnaire.

As for change scores, decline over 6-month on systolic and diastolic blood pressure (BP) approached significance, indicating that increases in BP over 6-months were associated with more cognitive problems in daily life at the end of the study. This finding was consistent with previous research suggesting that higher BP is associated with poorer cognitive abilities. However, initial BP scores did not predict CDL and nor did 18-month change scores, which casts some doubt over the finding.

Table 7.4. Predictors of Cognition in Daily Life at Time 3

	Initial value				6-month change				18-month change			
Predictor variable	n	R^2 change	β-value	p-value	n	R^2 change	β -value	p-value	n	R^2 change	β -value	p-value
Age	125	.003	.059	.516								
Inspection Time	111	.002	.046	.637	101	.000	.020	.846	102	.059*	.245*	.014
Grip Strength	125	.004	-.102	.488	121	.014	.124	.191	123	.012	.110	.234
Systolic BP	113	.008	-.090	.345	107	.036	.189	.053	108	.007	.085	.383
Diastolic BP	113	.004	-.066	.487	107	.027	.164	.095	108	.001	-.030	.756
Weight	125	.009	-.104	.289	122	.000	.010	.913	125	.020	-.140	.120
Height	125	.000	.031	.824	123	.003	-.052	.571	125	.011	-.104	.250
Visual Acuity	125	.003	.051	.574	123	.003	.052	.569	124	.000	-.014	.881
Digit Symbol	124	.072**	-.268**	.003	121	.000	-.018	.847	124	.015	-.122	.180
Visual Matching	122	.089**	-.299**	.001	119	.007	-.085	.358	122	.015	-.124	.178
Pattern Comparison	123	.039*	-.198*	.029	119	.050*	-.224*	.015	121	.010	-.098	.286

Note: Gender effects were removed before the examination of each biomarker. BP = Blood Pressure.
* $p < .05$, ** $p < .0$

Six-month change scores for PC were significant and in the expected direction, indicating that decline over 6-months in PC was associated with more cognitive problems at the end of the study. However, if decline in PC over 6-months was predictive of CDL, then it is logical that 18-month change scores should also be predictive; failure to observe this raises questions about the finding. Moreover, declines in DS and VM were not predictive of CDL, which is unusual.

Finally, 18-month change scores for IT predicted CDL. The positive β-value indicated that increases in IT over 18-months (i.e. greater slowing) were associated with more cognitive problems in daily life at the end of the study. Given this finding, subsequent hierarchical analyses were not conducted on CDL. The CDL questionnaire did not yield an 18-month change score and it was therefore not possible to examine the predictive validity of initial scores on the biomarkers for decline in CDL. Thus, the only significant predictor of CDL, apart from the perceptual speed tasks, was change over 18-months on IT.

Cognition

The cognition results will be divided into separate sections for fluid and crystallised ability. The first set of analyses examined the predictive validity of initial scores, 6-month change scores and 18-month change scores on the biomarkers for fluid and crystallised ability at Time 3. The six cognitive measures were entered into a factor analysis, which generated a two-factor solution. The two factors were unequivocally interpreted as fluid and crystallised ability (see description of procedures below) and used as the outcome measures for the regression analyses. In addition, the predictive validity of the biomarkers for each of the six individual cognitive measures was also examined and these are presented in Appendix G. The second set of analyses examined the predictive validity of the biomarkers at Time 1 for the change in the cognitive measures over 18-months. These analyses were completed on the six cognitive measures rather than factors scores for a number of reasons, including missing data in some of the fluid ability tasks at Time 1 and differences in stability between RSPM and the other two fluid ability measures, as shown in Chapter 6.

Factor scores

The six cognitive measures were highly correlated and the correlation matrix (n = 118) was considered suitable for factor analysis according to the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO = .739) and Bartlett's Test of Sphericity (χ^2 (15) = 217.4, $p < .001$). A two-factor solution was extracted using Principal Axis Factoring and rotated using the direct

oblimin method. The two-factor solution presented in Table 7.5 explained 49% of the variance. All factor loadings less than .10 were omitted for ease of interpretation. The two factors were interpreted as fluid and crystallised ability and all variables had their salient loading on the expected factor. However, Similarities had almost identical loadings on the Gf and Gc factors.

Table 7.5. Pattern Matrix for Cognitive Measures at Time 3

Cognitive Measures	Gf	Gc
RSPM	.693	
CCFT	.918	
CF	.566	
Information		.727
Spot-the-Word		.551
Similarities	.365	.375

Note: RSPM = Raven's Standard Progressive Matrices, CCFT = Cattell Culture Fair Test, CF = Concept Formation, Gf = Fluid ability, Gc = Crystallised ability

Fluid Ability

The factor scores for fluid ability were normally distributed, despite the fact that Concept Formation had a bi-modal distribution, and thus linear regression method were used. Table 7.6 shows the linear regression results for each of the biomarkers on the factor score for fluid ability. First, all three IT estimates (initial score, 6-month and 18-month change) were significant predictors of fluid reasoning. That is, slow IT and slowing in IT over time were related to poor fluid ability at the end of the study. Second, age was a significant predictor and it will therefore be necessary to test whether IT scores are still significant predictors after age is entered into a regression model. Third, decline in height over 6-months was associated with poor fluid ability performance, suggesting that shrinking is indicative of accelerated aging and predicts poorer fluid reasoning abilities. However, 18-month change scores were not significant predictors of fluid ability, so that the finding can be questioned. Finally, slow performance on the perceptual speed tasks at Time 1 was associated with poorer fluid ability at the end of the study. There was also evidence that decline in VM and PC, but not DS, was associated with poor fluid ability at the end of the study.

Table 7.6. Predictors of Fluid Reasoning at Time 3

	Initial value				6-month change				18-month change			
Predictor variable	n	R^2 change	β-value	p-value	n	R^2 change	β -value	p-value	n	R^2 change	β -value	p-value
Age	118	.078**	-.280**	.002								
Inspection Time	108	.088**	-.302**	.002	100	.081**	-.285**	.004	102	.099**	-.317**	.001
Grip Strength	118	.023	.245	.100	115	.004	-.067	.491	117	.011	.105	.268
Systolic BP	107	.001	-.030	.758	100	.000	-.005	.963	103	.004	-.067	.502
Diastolic BP	107	.004	.060	.537	100	.000	-.019	.856	103	.001	.037	.716
Weight	118	.030	.184	.061	116	.000	.021	.822	118	.001	.035	.708
Height	118	.005	.105	.447	116	.037*	.193*	.039	118	.007	.083	.376
Visual Acuity	118	.004	-.066	.483	116	.004	-.061	.515	118	.002	-.039	.677
Digit Symbol	117	.336**	.580**	.000	115	.009	.095	.317	117	.009	.095	.307
Visual Matching	115	.272**	.522**	.000	113	.074	.271**	.004	115	.042*	.206*	.027
Pattern Comparison	116	.312**	.558**	.000	113	.087**	.295**	.002	115	.031	.176	.060

Note: Gender effects were removed before the examination of each biomarker. BP = Blood Pressure.
* $p < .05$, ** $p < .01$

Table 7.7. Predictors of change in three measures of Fluid Reasoning over 18-months

	Standard Matrices				Cattell Culture Fair				Concept Formation			
Predictor variable	n	R^2 change	β-value	p-value	n	R^2 change	β -value	p-value	n	R^2 change	β -value	p-value
Age	125	.001	-.036	.696	95	.014	-.118	.259	85	.082**	-.287**	.006
Inspection Time	111	.031	-.179	.067	88	.003	.059	.595	78	.000	-.001	.994
Grip Strength	125	.000	-.023	.872	95	.067*	.412*	.012	85	.010	.161	.354
Systolic BP	113	.021	-.144	.130	84	.001	-.024	.828	77	.010	-.100	.374
Diastolic BP	113	.006	-.075	.434	84	.004	.059	.592	77	.012	.110	.329
Weight	125	.001	-.025	.794	95	.020	.145	.178	85	.031	.182	.097
Height	125	.002	.064	.637	95	.007	.127	.424	85	.014	-.177	.266
Visual Acuity	125	.001	.032	.725	95	.011	-.105	.314	85	.010	-.100	.352
Digit Symbol	124	.070**	.264**	.003	94	.049*	.221*	.033	84	.030	.174	.101
Visual Matching	122	.063**	.251**	.005	93	.018	.136	.198	82	.010	.098	.365
Pattern Comparison	123	.039*	.197*	.030	94	.122**	.351**	.001	84	.033	.182	.087

Note: Gender effects were removed before the examination of each biomarker. BP = Blood Pressure.
* $p < .05$, ** $p < .01$

A secondary regression analysis was run to explore how much of the variance in fluid reasoning could be explained by the biomarkers. Based on the significant findings from the original analyses, five measures were entered into the model (gender, age, IT, IT - 6-month change and height – 6-month change). One hundred individuals completed all six measures and the analysis was run on this sample. The regression model explained a significant proportion of the variance in fluid reasoning at Time 3 (R^2 adj = .239, F (5, 94) = 7.23, $p < .001$). All predictors (except gender) were significant and the most important predictor was initial IT score with a β-value of -.291 (t (94) = -3.23, $p < .01$). Six-month change scores for IT were the second largest predictor (β = -.258, t (94) = -2.93, $p < .01$) followed by age (β = -.234, t (94) = -2.66, $p < .01$) and 6-month change in height (β = .185, t (94) = 2.10, $p < .05$). Therefore, initial scores for IT, and more importantly, change scores for IT were important predictors of performance on fluid reasoning, even after the age-related variance was accounted for.

Table 7.7 shows the results for the 18-month decline in the three fluid ability measures. First, initial IT scores approached significance as predictors for RSPM ($p = .067$) but not for CCFT or CF. Therefore, there is minimal evidence that slow IT performance is predictive of larger declines in fluid reasoning over 18-months. Second, age was a significant predictor of decline in CF but not RSPM or CCFT. This discrepancy is not altogether surprising, given that CF seems to involve a different type of fluid reasoning than the other two tasks. Third, there was some evidence that weak grip strength was associated with decline in CCFT but not the other two tasks. Finally, evidence suggested that slower perceptual speed was associated with decline in fluid reasoning measured by RSPM and CCFT but not CF. Overall the results were mixed and suggested that an 18-month period may be too short to allow firm conclusions to be drawn. To summarise, there was little evidence that IT predicts decline in fluid reasoning but there is also little evidence that any of the biomarkers (or age) could perform any more successfully.

Crystallised Ability

The distribution for the crystallised ability factor score was approximately normal and hence linear regression techniques were used to assess the predictive validity of the biomarkers. It is worth noting that the major aim associated with the use of crystallised ability as an outcome measure was to determine whether initial scores on the biomarkers could predict decline in crystallised ability over 18-months (see Table 7.9). However, whether initial scores, 6-month and 18-month change scores on the biomarkers were related to the final scores on the crystallised

Table 7.8. Predictors of Crystallised Ability at Time 3

	Initial value				6-month change				18-month change			
Predictor variable	n	R^2 change	β -value	p-value	n	R^2 change	β -value	p-value	n	R^2 change	β -value	p-value
Age	118	.022	-.147	.107								
Inspection Time	108	.020	-.145	.130	100	.013	-.115	.231	102	.025	-.160	.103
Grip Strength	118	.033*	.292	.045	115	.009	-.095	.317	117	.002	.051	.587
Systolic BP	107	.000	-.016	.871	100	.001	-.030	.761	103	.001	-.033	.736
Diastolic BP	107	.005	.072	.451	100	.001	-.038	.701	103	.015	.125	.204
Weight	118	.009	.101	.296	116	.001	.025	.789	118	.004	.060	.513
Height	118	.006	.117	.389	116	.051*	.226*	.013	118	.000	.017	.853
Visual Acuity	118	.016	-.130	.159	116	.006	-.075	.418	118	.003	-.055	.546
Digit Symbol	117	.155**	.394**	.000	115	.004	.060	.519	117	.003	.052	.574
Visual Matching	115	.144**	.380**	.000	113	.016	.127	.172	115	.022	.148	.110
Pattern Comparison	116	.122**	.350**	.000	113	.045*	.213*	.021	115	.040*	.201*	.028

Note: Gender effects were removed before the examination of each biomarker. BP = Blood Pressure.
* $p < .05$, ** $p < .01$

Table 7.9. Predictors of change in three measures of Crystallised Ability over 18-months

	Information				Spot-the-Word				Similarities			
Predictor variable	n	R^2 change	β-value	p-value	n	R^2 change	β -value	p-value	n	R^2 change	β -value	p-value
Age	120	.000	.018	.848	127	.005	-.068	.444	125	.000	-.016	.861
Inspection Time	108	.002	.049	.615	112	.023	.152	.113	111	.005	.074	.444
Grip Strength	120	.043*	.338*	.021	127	.012	-.177	.218	125	.000	-.003	.985
Systolic BP	109	.007	.082	.397	115	.002	-.047	.619	113	.053*	-.230*	.013
Diastolic BP	109	.004	.067	.489	115	.000	.003	.973	113	.022	-.150	.109
Weight	120	.021	.154	.111	127	.007	-.087	.364	125	.003	.056	.558
Height	120	.028	.248	.067	127	.000	.017	.898	125	.000	-.033	.807
Visual Acuity	120	.008	-.092	.319	127	.000	-.017	.850	125	.004	-.065	.470
Digit Symbol	119	.006	.076	.407	126	.000	-.008	.930	124	.028	.167	.063
Visual Matching	118	.005	.069	.452	124	.012	.109	.228	122	.002	.042	.648
Pattern Comparison	119	.003	-.050	.583	125	.001	.033	.712	123	.021	.146	.107

Note: Gender effects were removed before the examination of each biomarker. BP = Blood Pressure.
* $p < .05$, ** $p < .01$

ability factor score were also examined and these results are presented in Table 7.8. There was some evidence that weak grip strength and slow perceptual speed at Time 1 were associated with poor crystallised ability at the end. There was also evidence that declines in height and PC were associated with poor crystallised ability at Time 3 but once again the findings for height were questionable. There appeared to be a consistent link between PC and crystallised ability but there was no evidence that IT was related to crystallised ability.

A hierarchical analysis was run with three predictor variables (gender, grip strength and height - 6-month change) based on the significant findings from the original regression analyses. The model (n = 116) explained a significant proportion of the variance in crystallised ability at Time 3 (R^2 adj = .104, F (3, 112) = 5.44, $p < .01$). In the final model, height – 6-month change, was the only significant predictor of G*c* (β = .216, t (112) = 2.44, $p < .05$). Therefore, there was some evidence that decline in height was predictive of crystallised ability but the evidence remains doubtful.

The major question of interest was whether any of the biomarkers could predict decline in crystallised ability over 18-months and these results are presented in Table 7.9. There were few significant effects indicting that the biomarkers were not useful predictors of decline in Gc over 18-months. There was some evidence that weak grip strength was associated with high decline in Gc as measured by Information but not the other two Gc tests. There was also evidence that high systolic BP was associated with decline in Gc as measured by Similarities. This effect was mirrored for diastolic BP but did not reach significance (p = .109). There was no evidence that slow performance on IT or perceptual speed was associated with decline in Gc.

Discussion

This chapter presents the major analyses of this research project, designed to answer the question of whether IT is a valid biological marker for functional age. That is, whether IT is a lead indicator for abnormalities in functional age outcomes. Before this question is addressed, the predictive validity of age and the six traditionally used biomarkers (grip strength, systolic blood pressure, diastolic blood pressure, weight, height and visual acuity) will be discussed in light of past research and the findings from Chapter 5. If the findings from the other biomarkers are inconsistent with previous research, this might be due to the sample or the time frame of the study, which may have implications for the IT effects, and so it is worth examining them first.

Following this, the predictive validity of IT will be examined and compared to the other biomarkers and the three perceptual speed tasks.

Age

*C*hronological age has been found to predict performance on a number of the functional outcome measures including ADL and fluid ability. Thus, the common assumption that older people will perform less well than younger people on functional age outcomes has been confirmed here, to some degree, even within the relatively narrow age range for this group of elderly adults. However, there was no evidence that age was related to CDL or the crystallised ability tasks. Rather, it seems that for people aged over 70, factors other than age determine whether they will experience difficulties in cognitive aspects of daily life and start to decline in crystallised abilities. In a sense, therefore, biomarkers that can predict these functional outcomes (e.g. CDL and Gc tasks) are particularly important, because chronological age of itself revealed little or scarcely anything about them.

As for the expectation that older people would *decline* more than younger people on the functional age outcomes, there was very little support in these data. Within this elderly sample, older people showed more decline in just one outcome measure: Concept Formation. There was no evidence that older people were declining more than younger people on any of the other seven outcome measures. Once more, biomarkers that can predict decline on functional outcomes may be important because age cannot, at least within a restricted range in a homogeneous sample.

Grip Strength

Of the traditionally used tasks, grip strength was the most successful predictor, with initial scores and change scores predicting performance on a range of functional outcomes. For initial scores, there was evidence that weak grip strength at Time 1 was associated with a number of outcomes at the end of the study, including membership into the *dependent (in at least one area)* group for ADL, poor fluid ability as measured by CCFT, and poor crystallised ability as indicated by the factor score. Moreover, weak grip strength at the start was associated with decline in functional ability over the subsequent 18-months in CCFT and Information. Thus, initial grip strength scores predicted performance on measures of everyday living, fluid ability and crystallised ability (i.e. all constructs in the study).

In addition, there was evidence that weakening in grip strength over time was associated with ADL at Time 3. However, the direction of this effect was unexpected and counterintuitive ,

suggesting that decline in grip strength was predictive of membership into the *fully independent* group for ADL. One would expect that decline in grip strength would be associated with membership into the *dependent (in at least one area)* group if grip strength was to serve as a biomarker for subsequent decline. However, this study has found the opposite and it does not seem to represent a chance effect because the 18-month change scores supported it. Why might this occur? Perhaps, within the current sample, people who are beginning to have difficulties in their everyday life take steps to improve their muscle strength to counteract it or set up systems and support within their living environments that enhance their independence (e.g. grips on taps, rails in toilets etc).

Consistent with this possibility, there were a number of people in our study who attended regular exercise sessions at a gym or swimming pool. In fact, a number of the retirement villages accommodating several participants specifically catered to the exercise needs of elderly people on site. It is plausible that these people were maintaining or improving their strength in response to a perceived decline in level of independence in their daily life. Moreover, people who were experiencing decline in independence of everyday living in their own home might have moved into a retirement village and taken up exercise classes. If this were so, then it would be the individuals who were having difficulties who would be improving their grip strength. This could account for the relationship between decline in grip strength and membership into the *fully independent* group. However, this suggestion is purely speculative; and we were not able to revisit these people and to test this hypothesis.

Blood Pressure

There was just one significant effect for blood pressure (BP), indicating that a high systolic BP at Time 1 was associated with decline in Similarities over 18-months. There was additional support for this idea, with initial scores for diastolic BP approaching significance for the same effect. Given that age was unable to predict level or decline in the crystallised ability tasks, this might be important indicating that high BP scores are predictive of terminal decline. However, if this was the case then BP should also be able to predict a decline in Information and Spot-the-Word; and there was absolutely no evidence for this in these data.

Another interesting finding was that 6-month change scores for systolic and diastolic BP approached significance as predictors of CDL. That is, increases in BP over the first six months of the study were associated with more problems in CDL at the end of the study. However,

initial scores and 18-month change scores for systolic and diastolic BP were non-significant predictors of CDL and this therefore challenges the reliability of the result.

Perhaps, the most important observation concerning blood pressure was that it had such a *small* relationship with everyday functioning and cognition. It seems surprising that blood pressure was not related to cognition, given that many studies have established that hypertensives have poorer performance on a range of cognitive abilities measures, in particular fluid reasoning (see Waldstein, Manuck, Ryan, & Muldoon, 1991 for review article). However, although many studies have reported this relationship, there is some evidence in the literature that the difference between hypertensives and normotensives occurs within younger age groups but not in older groups. Waldstein (2000) reviewed the literature and found that differences between normotensives and hypertensives appeared to diminish among middle-aged and older adults (56 to 72 years). Waldstein suggested that this might be due to survival effects. That is, people with hypertension develop cardiovascular problems in midlife and therefore do not survive to participate in studies later in life. The implication of these results is that blood pressure may be a useful biomarker for mortality or cognitive decline if measured early in the lifespan but it may not be as useful within a group of healthy elderly adults.

Weight

Body weight was a significant predictor of just one functional age outcome measure (CCFT) but approached significance (i.e. $p < .10$) for an additional five outcome measures. In five of the six cases, the results suggested that thinner people performed more poorly on the functional outcome. The final effect indicated that people who lost weight over 18-months were more likely to be in the *dependent (in at least one area)* group for ADL. This indicates that low body weight and weight loss may be important predictors of poorer functional outcomes in the elderly. Given that weight is such an easy biomarker to measure and appears to be quite useful, inclusion of body weight in future studies on biomarkers would be strongly advocated.

Although body weight approached significance for a number of outcomes, it was only statistically significance for one measure. It is possible that the true effect of weight on functional outcomes was confounded by obesity. If everyone in the sample was classified within the range of recommended weight (i.e. BMI 20 to 25) then the thinnest people (within genders) might indeed be showing signs of accelerated aging. However, when some people are obese (or were obese) then people who have been declining for a period of years might still not be in the bottom 25% of the distribution for weight. Consequently, initial scores for body weight may be

confounded with obesity. However, change scores are statistically independent from initial scores and therefore should still be effective. It should be noted that body weight is highly stable over six months (r = .989) and 18-months (r = .972). Therefore, 18-months may not be long enough to establish whether decline in body weight predicts functional decline. To conclude, body weight appears to be an important biomarker but it may be necessary to consider decline over a longer period to establish predictive validity.

Height

Overall, height was not a very successful biomarker. Just two of the height effects were significant and an additional one approached significance (p < .10). Biological aging theory posits that people get shorter as they get older and large declines in height are indicative of accelerated physiological aging. There was some evidence for this proposition in these data, with higher decline in height over 6-months associated with lower scores on CCFT and Information at Time 3. However, 18-month change scores did not support this finding. In addition, there was evidence that shorter people showed more decline in Information over the subsequent 18-month period, which suggests that height might be a predictor of terminal decline. However, if this were true then it should also predict a decline in the other two crystallised ability tasks and there was no evidence for this. These three effects were in the expected direction but were questionable. In terms of evaluating height as a biomarker, the conclusion is that height is not particularly informative and this may be due to the marked stability of height (r = .975 over 18-months), the relatively short period of the change measurement and as insufficiently reliable method for measuring height. If height was assessed over a longer period of time and was measured more accurately then it might prove to be an informative biomarker for functional decline.

Visual Acuity

Many studies of biological markers have shown that visual and auditory acuity are among the most important predictors of functional age. However, there was little support for this in these data. There was one effect that approached significance (p = .058) but the direction was contrary to expectations. It suggested that deterioration in visual acuity over 18-months was associated with membership into the *independent* group for CF. However, visual acuity was not a significant predictor of any measures of everyday functioning or cognition.

One possible explanation of these null effects is that low test-retest reliability of visual acuity might have affected the results. The correlation between scores at Times 1 and 2 was the

lowest of all tasks (r = .653) and the correlation between scores at Times 1 and 3 was even lower (r = .468). This means that the rank order of people changed appreciably from one occasion to the next, which might indicate that factors other than physiological aging were affecting scores or that the construct has not been measured reliably.

One issue that should be addressed is the inconsistency between the encouraging findings on visual acuity in Chapter 4 and the null findings in this chapter. In Chapter 4, visual acuity was a significant predictor of performance on all three fluid ability tasks, with poor vision associated with poor reasoning ability. However, in this chapter there was no evidence for this association. One possibility is that some of the people with very poor vision dropped out over the course of the investigation. There is evidence for this proposition because the analyses in Chapter 6 showed that the group of 27 people who dropped out had significantly poorer visual acuity than those who remained in the study. Therefore, one reason for the null effects for visual acuity in the research reported in this chapter might be the homogeneity of vision of this sample.

Inspection Time

Inspection Time (IT) was the most successful marker task of all seven purported biomarkers and was particularly useful for predicting everyday living skills and fluid reasoning. A summary of the IT results is presented in Table 7.10. The first column shows predictive validity for IT scores at Time 1 where a tick indicates the effect was significant at the 5% level and a cross indicates that it was not. Slow IT scores at Time 1 were associated with membership into the *dependent (in at least one area)* group for ADL, and poor fluid ability at Time 3 on all three measures. There was some evidence that slow IT scores predicted decline in Standard Matrices and ADL over 18-months but these effects did not reach significance ($p < .10$). Thus, initial IT scores were able to predict performance on ADL and fluid reasoning and all effects were in the hypothesised direction. Furthermore, it should be noted that there were missing data for IT, so that these comparisons were made on a reduced sample, which resulted in reduced power.

Most importantly, change scores for IT were predictive of a wide range of outcomes and in almost all cases 6-month and 18-month change scores were consistent. Large increases in IT over time (i.e. slowing down) were associated with more problems in cognitive aspects of daily life at Time 3, poor fluid reasoning as indexed by all three tasks, and poor Similarities performance at Time 3. There was also weak evidence that slowing in IT predicted membership

into the *dependent (in at least one area)* group for ADL but this effect was only significant at the 10% level. Therefore, IT change scores were more successful than initial scores and, because the two measures were independent of one another, they were able to explain unique variance in the outcomes measures.

Table 7.10. Predictive Validity of Inspection Time for Functional Outcomes

Predictors Outcomes	Initial Score	6-month Change Score	18-month Change score
Everyday Functioning			
ADL – Time 3	✓	*x*	*x* [a]
ADL – Decline	*x* [a]	-	-
CDL – Time 3	*x*	*x*	✓
Fluid Reasoning			
RSPM – Time 3	✓	✓	✓
RSPM – Decline	*x* [a]	-	-
CCFT – Time 3	✓	✓	✓
CCFT – Decline	*x*	-	-
CF – Time 3	✓	✓	✓
CF - Decline	*x*	-	-
Crystallised Ability			
Information – Time 3	*x*	*x*	*x*
Information - Decline	*x*	-	-
Spot-the-Word – Time 3	*x*	*x*	*x*
Spot-the-Word – Decline	*x*	-	-
Similarities – Time 3	*x*	*x*	✓
Similarities - Decline	*x*	-	-

✓ = Significant at the .05 level, *x* = non-significant at .05 level, a = significant at the .10 level. ADL = Activities of Daily Living, CDL = Cognition in Daily Life, RSPM = Raven's Standard Progressive Matrices, CCFT = Cattell Culture Fair Test, CF = Concept Formation.

There are three issues that will now be discussed with regards to these findings; first, the success of 18-month change scores to predict CDL; second, the ability of 18-month change scores to predict the Similarities score, and third, the success of change scores for IT in comparison to change scores for perceptual speed.

The CDL questionnaire provided a measure of the degree of difficulty with cognitive aspects of everyday life (e.g. remembering appointments, organising daily routines, maintaining concentration, and word finding in conversations). On the whole, the biomarkers (including chronological age) were unsuccessful in differentiating between people in terms of degree of problems-solving difficulty in everyday activities. However, decline in IT over 18-months was

predictive of a higher degree of difficulty in CDL at Time 3. One argument could be that IT is a low order cognitive task and therefore change scores should be related to CDL. However, there was very little evidence that change scores for the perceptual speed tasks were predictive of CDL. Decline over 6-month on PC was a significant predictor of CDL but this was not replicated for the 18-month change scores, so that the result is less reliable. Furthermore, change scores for VM and DS were not significant predictors of CDL. Therefore, change score for IT can predict an important functional age outcome that is not predicted by chronological age, any of the biomarkers or change scores for perceptual speed.

A second important finding was that 18-month change scores for IT predicted final scores on Similarities (see Appendix G, Table G6). Although Similarities is generally considered a crystallised ability task, it can be seen from the pattern matrix in Table 7.5 that it is factorially similar to both the fluid and crystallised ability tasks. Therefore, it might be expected that IT change scores would predict final scores on Similarities, because they also predicted performance on the fluid ability tasks. Although this is a reasonable argument, it follows that change scores for the perceptual speed tasks should also predict the Similarities score. There is minimal evidence for this, with only one change score (18-month change for Visual Matching) approaching significance. One interpretation of this finding is that 18-month change scores for IT predicted the crystallised aspect of the Similarities score, in which case change scores for IT might be even more important.

The final issue requiring some consideration is a comparison of the predictive validity of change scores for IT and the perceptual speed tasks. Most literature on biomarkers emphasises the importance of change scores as predictors rather than just initial scores. For instance, Baker and Sprott (1988, p. 234) remark that, “it is the rate of decline which can ultimately be more critical than the initial value or the time of onset of that change”. The results from Chapter 4 and the current chapter showed that initial scores for the perceptual speed tasks were consistently among the most significant predictors but that was not the case for scores indicting change in perceptual speed. Change scores for Digit Symbol (DS) were not significant predictors of any of the eight outcome measures. Thus, although DS is usually the most significant predictor of the functional outcome measures, change scores for DS are not. Change scores for Visual Matching and Pattern Comparison were a little more successful but change scores for IT were the most successful of all four measures. If the success of a biomarker is primarily determined by the

predictive validity of change scores then IT was clearly the most successful biomarker, even when the perceptual speed tasks were included.

To conclude, initial scores for IT are important predictors of ADL and fluid reasoning. Change scores for IT are important predictors of ADL, CDL, Gf and Similarities and are more successful than change scores for any of the other biomarkers, including the perceptual speed tasks. Moreover, initial scores and 6-month change scores explained unique variance in the outcomes measures and in some cases both were important predictors of the functional outcomes. Thus, evidence supports the notion that IT has potential as a lead indicator for individual differences in functional age.

CHAPTER EIGHT: FINAL DISCUSSION

This chapter summarises the evidence on the predictive validity of IT as a biomarker or lead indicator for individual differences in functional aging. However, before the major findings of this investigation are reviewed there is one question that will be considered. The broader aim behind examining the predictive validity of marker test is to provide the means for early identification of "less" successful aging. In the current study, IT has successfully predicted everyday living and fluid reasoning; one question that arises is whether IT could be an effective screening tool to identify "at risk" individuals. This question will be examined in the following section. Next, assessment of the efficacy of IT as a biomarker will be discussed based on the plan proposed in Chapter 1 for evaluating a biomarker. Finally, difficulties arising during the current investigation will be discussed and recommendations made for further research into this topic.

Inspection Time: A Screening Tool?

Screening is a method used to identify the presence of, or risk factors for, a disease or disorder and it is generally relevant under circumstances where some individuals are not yet aware that they have the problem. By identifying those people at risk for developing the disease or disorder, it may be possible to intervene before they develop the full-blown problem. For example, in the field of medicine, doctors screen for high blood pressure and, if necessary, administer medication to avoid the development of chronic hypertension. Similarly, a prognostic screening test that could be used to identify those persons at risk for dependence in everyday living or cognitive problems would be useful, assuming that means of intervention were available or that forward planning was possible. Given that IT proved to have predictive validity for these two outcomes, it might provide a useful screening tool.

In order for IT to be an effective screening tool, it would need to be able to identify those people who were at risk for experiencing problems in everyday living and fluid reasoning in the future and/or abnormal decline in everyday living and fluid ability over time. For the research project reported here, rather than examine the efficacy of IT as a screening tool for all of these outcomes, a decision was made to focus on just one outcome measure. Initial and 6-month change scores for IT were most successful at predicting performance on the fluid ability tasks at Time 3. Specifically, the effect sizes (see Appendix G, Table G2, R^2 statistic) were largest for

the Cattell Culture Fair Test (CCFT). Thus, if IT was an effective screening tool then it is most likely that it would be successful for CCFT. Therefore, the efficacy of IT as a screening tool was investigated by examining whether it could detect those individuals who demonstrated poorer fluid ability at the end of the study.

In psychology, screening tests have often been used to identify those individual "at risk" for developing dementia. For example, De Jager, Hogervorst, Combrinck and Budge (2003) administered a range of neuropsychological tests to 51 controls, 29 individuals with mild cognitive impairment, and 60 people with possible or probable Alzheimer's Disease. When comparing the control group to the Alzheimer's group, the sensitivity and specificity of many of the neuropsychological tests was high (i.e. above 80%). However, the sensitivity and specificity values dropped considerably when comparing the control group to the individuals with mild cognitive impairment because these two groups are *relatively* similar compared to the difference between the control and dementia groups. This point is particularly relevant to the current study because the people in this sample with lower scores on fluid reasoning do not even reach the definition for mild cognitive impairment. Therefore, the difference between those individuals with "high" fluid ability and those with "low" fluid ability is smaller than the difference between controls and individuals with mild cognitive impairment and considerably smaller than the difference between a control group and a dementia group. Thus, the sensitivity and specificity of IT for low scores on fluid reasoning are expected to be much lower than the values found in the De Jager et al. (2003) study. Essentially, the use of IT as a screening tool to detect low scores on fluid reasoning is more demanding than using IT to predict dementia.

Method

To evaluate this idea, it was necessary to generate a dichotomous outcome measure to represent CCFT performance. The lowest quartile (n = 32; *low Gf group*) was isolated, which represented people with scores of 19 items correct or less. These were compared to the other three quartiles (n = 90; *normal-high Gf group*), with scores between 20 and 36 items correct. The aim was to test whether initial scores and/or 6-month change score for IT could identify or screen out the people who subsequently formed the *low Gf* group.

A screening test is evaluated in terms of the accuracy with which it classifies people in terms of four possible outcomes: true positive, false positive, true negative and false negative. A *true positive* occurs when the test (e.g. IT) predicts that the individual will develop the problem (e.g. *low Gf*) and they ultimately do. A *false positive* occurs when the test predicts that the

individual will develop the problem but s/he does not. A *true negative* occurs when the test predicts that an individual will not develop the problem and s/he does not. Finally, a *false negative* occurs when the test predicts that an individual will not develop the problem but sh/e does. The accuracy of a screening test is generally qualified by examining the sensitivity and specificity of the test. *Sensitivity* is defined as the probability that a person with the problem (in this case *low Gf*) will test positive. This is calculated by dividing the number of true positives by the number of people actually in the *low Gf* group. *Specificity* is defined as the probability that a person without the problem will test negative. This is calculated by dividing the number of true negatives by the total number in the *normal-high Gf* group. Finally, the *overall accuracy* of the test in terms of the probability that an individual will be classified correctly can be examined. Overall accuracy is calculated by adding the number of true positives and true negatives and then dividing by the total number of people in the sample.

In general, a screening tool is deemed accurate if it has high sensitivity and high specificity. However, in some situations either high sensitivity or high specificity might be of more importance. In this instance, we are most concerned about high sensitivity because we want to identify as many people as possible who will be at risk of developing future cognitive problems; on the other hand, we are not as concerned about falsely identifying small numbers of those people who will actually continue to function with normal levels of cognition. It is hoped that some intervention can be taken to slow down the rate of cognitive decline so it is important that we detect as many people as possible who are "at risk" of future cognitive problems. Moreover, it is unlikely that the intervention would cause any harm to individuals who were falsely identified as "at risk" so we are not to concerned about identifying a small number of these individuals. Nonetheless, it is important to note that sensitivity and specificity are closely linked, with an increase in one generally associated with a decrease in the other. Thus, although it might be acceptable to record a few false positive outcomes, the specificity must be reasonably high because otherwise the screening tool will identify nearly everyone as at risk and hence would be ineffective.

Results

To evaluate the efficacy of IT as a screening tool a number of different cut-off rules for initial and change scores were investigated. For instance, a cut-off value of 80 ms as an initial measure would predict that everyone with IT > 80 ms would be at risk for *low Gf* in the future.

For each cut-off value, sensitivity and specificity of the test and also the overall accuracy was examined. The results for initial IT scores are presented in Table 8.1.

In addition, Receiver's Operating Characteristics (ROC) curve analysis was run. This analysis examines every possible cut-off score, calculates sensitivity and specificity and then generates a plot of this information. The sensitivity values are plotted on the y-axis and 1 – specificity values are plotted on the x-axis. This allows for a test called Area Under the Curve (AUC) to be run, which tests whether the area under the ROC curve is significantly different to 0.5. Essentially, the AUC represents the proportion of individuals correctly classified and if this significantly greater than 0.5 (i.e. chance) then the screening test is deemed useful.

Table 8.1. Accuracy of Screening Tool (IT initial scores)

Cut-off Score	*Sensitivity*	*Specificity*	*Overall Accuracy*
60 ms	96%	12%	34%
70 ms	82%	26%	40%
80 ms	68%	53%	57%
90 ms	46%	74%	67%
100 ms	32%	85%	72%
110 ms	18%	92%	72%

The first point to be made is that overall accuracy increased as specificity increased because there were more people in the *normal-high Gf* group (n = 90) than in the *low Gf* group (n = 32). Thus, as specificity increased, more of the *normal-high Gf* group was classified correctly and therefore more of the total group were classified correctly and the overall accuracy increased. However, in the current context, high sensitivity was considered more important than high overall accuracy and it was the major focus. From Table 8.1, it is clear that the task had high sensitivity at low cut-off values. At 60 ms, the screening tool identified almost everyone that went on to develop cognitive problems in the future. However, the specificity was very low indicating that 88% (100% minus 12%) of people without cognitive problems were identified as at risk, which is clearly undesirable. If we consider a level that has slightly lower sensitivity (e.g. 70 ms), the screening tool fails to identify 18% of individuals who are at risk of cognitive problems but still falsely identify 74% of individuals as being at risk. Given the link between sensitivity and specificity, it is questionable whether there is a cut-off score that will give adequate sensitivity and specificity to be able to conclude that initial IT scores will provide an effective screening tool.

The AUC analysis provided a formal test of whether initial IT scores provided a useful screening test for low fluid ability at Time 3. The AUC was .613 (SE = .063) and this was not significantly different to .500 ($p > .05$). Thus, findings from the formal test are consistent with the above conclusion; initial IT scores did not provide an effective screening tool to detect low fluid ability.

In addition to initial scores, 6-month change scores were investigated as a screening tool for subsequent cognitive problems and these results are presented in Table 8.2. As described extensively in Chapter 5, change scores were calculated using residual change methods that statistically removed the effects of the initial scores. Therefore, it is not entirely clear what scale the resultant variable is on. However, it is clear that zero represents the mean change of the group and positive values indicate that an individual has slowed down more than the rest of the group over the 6-month period.

At a cut-off score of –20, the screening tool has high sensitivity and identifies 92% of individuals who are at risk of cognitive problems. However, the corresponding specificity is extremely low. In effect, the screening tool identifies almost everyone as being at risk of cognitive problems and is therefore ineffective. At the other extreme, a cut-off score of 20, identifies 31% of individual who are at risk of future cognitive problems, correctly identifies almost all individuals who are not at risk of cognitive problems and has an overall accuracy of 78%. In terms of the formal test from the ROC curve, the AUC was .596 (SE = .074) and this was not significantly different to .500 ($p > .05$). Thus, the 6-month IT change scores did not provide a useful screening tool for low fluid ability at the end of the study.

Table 8.2. Accuracy of Screening Tool (IT change scores)

Cut-off Score	*Sensitivity*	*Specificity*	*Overall Accuracy*
-20	92%	7%	29%
-10	77%	24%	38%
0	58%	57%	57%
10	35%	85%	72%
20	31%	95%	78%

Note. The cut-off scores represent change from Time 1 to 2 (i.e. ms) once differences in initial scores have been statistically removed (see text for clarification and Chapter 5 for discussion).

Discussion

In the previous section, examination of Tables 8.1 and 8.2 and the AUC analysis led to the conclusion that IT was not a highly successful screening tool. However, there are a number of issues that should be considered when interpreting this result and these will now be considered. First, as mentioned on page 174, the current sample is relatively high functioning and the difference between the "low" and "high" fluid ability groups is small. If IT were used to differentiate between a control group and a dementia group then we would expect the sensitivity and specificity to be much higher. Therefore, it is probably invalid to compare the sensitivity and specificity values from this study to other research that has aimed to distinguish between a control group and people with mild cognitive impairment or dementia.

Second, a number of studies in psychology have found high sensitivity and specificity levels when using questionnaires or neuropsychological tests to detect clinical disorders such as dementia or anxiety (e.g. De Jager et al., 2003; Devanand et al., 1997; Leyfer, Ruberg, & Woodruff-Borden, 2006). However, IT has been described as a biological task, in this thesis, with minimal demands on higher order cognitive processes. Therefore, it might be more valid to compare the performance of IT to other biological screening tests. Waaler (1980) examined the sensitivity and specificity of blood pressure and found that the optimal cut-off for males aged 50 – 59 years was at a level where the specificity approached 100% and the sensitivity was about 20%. Blood pressure estimates are used throughout the world to screen for hypertension and the sensitivity, from this study at least, was not particularly high. Therefore, there may still be some hope for IT as a screening tool.

Third, there is currently no prior warning for subsequent cognitive decline, therefore, with some refinement, IT may actually be quite useful. This IT screening tool can identify about one-third of people who are at risk of cognitive problems in the future, almost all individuals who are not at risk and accurately classify more than three-quarters of individuals overall. While higher sensitivity would be desirable, these results are encouraging and indicate that IT might have promise as a screening tool for future cognitive problems. To summarise, the formal tests indicated that IT was not an effective screen test. However, a number of point have been presented that indicate that IT might actual have some utility as a screening test and that it would be too early to discard the idea entirely. The following section will consider some possible improvements to the estimation procedure for IT that might improve the success of IT as a screening tool for cognitive problems or decline.

Why is IT not a more effective screening tool?[16] If we examine the distribution of IT scores for people in the *low Gf* and *normal-high Gf* groups there is a large degree of overlap. There are individuals in both groups who have quite long IT scores (> 140 ms) although there are slightly more in the *low Gf* group. The major difference between the distributions is at the other end of the distribution. In the *low Gf* group, just 15% of individuals had scores below 70 ms and the minimum value was 50 ms. In the *normal-high Gf* group, 27% had scores less than 70 ms and the minimum score was 32 ms. Therefore, it is valid to conclude that people with quick IT scores had a high likelihood of being in the *normal-high Gf* group but it is not equally valid to say that people with long IT scores had a high likelihood of being in the *low Gf* group. In effect, IT is a better predictor of membership into the *normal-high Gf* group than into the *low Gf* group and this means it has shortcomings in terms of its sensitivity as a screening tool.

Using the adaptive staircase procedure[17], it is very difficult to achieve a quick IT score because three items correct are necessary at any stimulus onset asynchrony (SOA) to descend to the next level of the staircase. Conversely, it is quite easy to get a slow IT score because just one mistake causes the program to ascend to the next higher level of the staircase. Furthermore, a slow IT score can occur for a number of reasons, including inattention, confusion when responding and general confusion with the task requirements, considerations that can be particularly relevant when attempting measurement in samples of older adults. For IT to be a useful screening task, it is necessary for long IT scores to clearly represent slow speed of processing. Although we can say with some degree of confidence that people with short ITs have quick speed of processing, unfortunately we cannot be nearly as confident with current methods that people with long IT scores have slow speed of processing.

Consider this example of two individuals with identical IT estimates of 107 ms. Person 1 registered eight reversals of the staircase at SOAs of 119, 136, 85, 119, 85, 119, 85 and 119 ms. Person 2 registered eight reversals of the staircase at 153, 170, 68, 136, 68, 136, 51 and 85 ms. The question that we need to consider is how well does the IT score of 107 ms represent these two individuals speed of processing? For Person 1, it is probably quite a good representation, given that he oscillated between SOAs of 85 and 119 ms several times; and 107 ms is about half

---

[16] This discussion focuses on initial scores rather than change scores. However, problems with initial scores will clearly be passed on to the change scores.

[17] Details of the adaptive staircase procedure for estimating IT are provided on page 68. For more detail see Wetherill and Levitt (1965)

way between these SOAs. However, for Person 2 the registered IT is not as reliable a representation. This individual achieved 68 ms on two occasions and 51 ms once. Therefore, he could clearly do the task at SOAs that were substantially shorter than 107 ms. Furthermore the reversals that occurred at the start were the largest values (170 and 153 ms) and could have reflected confusion with the task, simple responding mistakes or poor attention. Thus, for this individual, 107 ms may not be a good representation of his speed of processing and a quicker value, not currently specifiable, may be more representative. However, on current performance, the screening task would probably indicate that this person is at risk for *low Gf* in the future, even though it is clearly possible that he can do the IT task at quite low SOAs. Consistent with these post-hoc interpretations, Person 1 had a score on CCFT of 18 items correct (*low Gf* group), whereas Person 2 had a score of 25 items correct (*normal-high Gf* group).

For IT to be used as an effective screening tool, it is important to reconsider the estimation methods currently used. One method used in this thesis was to consider the variability of the eight reversals and remove extremely variable IT estimates. For example, if most reversals occurred around 85 and 119 ms but one occurred at 340 ms this would suggest that the overall mean estimate may not be a good representation of the true IT. The variability of the eight reversals is one aspect of the current method that might be used to examine the degree to which outliers exist in the eight values. Although this method (described extensively in Appendix E) has attempted to address the issue, arguably, it does not go far enough. Moreover, it depends on the exclusion of data points, which is clearly undesirable. However, there are a number of suggested improvements that might be made to the current method of estimating IT, which may make it a more effective screening tool. Theoretically, these methods should also make IT a better biomarker for individual differences in functional aging.

The first suggestion involves the start of the adaptive staircase procedure. The estimation phase starts at 320 ms and the SOA is reduced by one refresh rate (~ 17 ms) for each correct answer until the first mistake is made. When this occurs the staircase is activated and the participant must complete three items correctly at any SOA to recommence descending the staircase. It was observed on a number of occasions that, participants made an error right at the start of the estimation phase but then indicated, by attempting to correct the response, that it had been due to confusion or pressing the wrong button. In the current design, this requires three correct responses at each level before the program shortens exposure to the next SOA. If an individual makes an early error while still a long way from the final IT, then this error will extend

the testing period substantially; and therefore increase the likelihood of unreliability when responding because of boredom, inattention or visual problems. If the SOA is longer than 200 ms when the first error is made, then there is a high likelihood that it is due to something other than slower speed of processing. Thus, the first suggestion is to adjust the program to check whether the first mistake occurs at an SOA > 200 ms (or some other empirically determined upper limit). If so, the program could present a number of trials at that SOA and/or alert the experimenter to a potential problem, rather than activating the staircase. Once the participant had demonstrated competence at that level then the estimation procedure could resume. If the next mistake occurred when the SOA was less than 200 ms, then we could be more confident that this was due to speed of processing problems, rather than to confusion or inattention and the staircase should be activated. If further mistakes occurred at a SOA > 200 ms then it might indicate that the individual was still having trouble understanding the task or responding and the experimenter could take steps to rectify these problems before allowing the estimation phase to continue. This should result in a more reliable estimate for IT without substantially increasing the length of the estimation phase. Whether this was so or not would be readily tested by conventional test-retest trials.

The second suggestion is that the median of the eight reversals could prove to be a better measure of central tendency than the mean. Whenever extreme values are registered, inevitably at longer SOAs, the mean will be dragged towards them, resulting in a slower IT estimate. For example, consider the example presented earlier where Persons 1 and 2 registered mean scores of 107 ms. Calculating their median scores, Person 1's median IT score was 119 ms and Person 2's median score was 110 ms. This procedure differentiates between them more effectively but it is still problematic because both people have quite high scores. Use of the median rather than the mean therefore produces a small improvement, particularly when there are extreme outliers, but this adjustment on its own is probably still insufficient to differentiate between these two discernibly different distributions of performance.

The final suggestion involves outliers in the set of eight reversals. Consider Person 2 who has a number of reversals at short SOAs including 51 and 68 ms. It is clear that these shorter values tell us much more about the individuals IT than do the extremely high value, like 153 and 170 ms. Thus, the second suggestion therefore is to focus on the smaller values when estimating IT rather than the longer, outlier values. One possibility would be to exclude any reversals that occur at specified exposure duration from the minimum exposure achieved by the individual or,

for some, as set by the minimum refresh rate for monitor screens (e.g. three refresh times = 51 ms). This method would assume that any exposure values, this much longer and more beyond the minimum achieved exposure were outliers, caused by inattention or confusion, that do not provide reliably useful information about the individuals speed of processing.

For Person 2, this would lead to the exclusion of reversals at 136, 153 and 170 ms, leaving four values of 51, 68, 68 and 85 ms. The mean and median of these four values is 68 ms. On the other hand, the minimum reversal for Person 1 occurred at 85 ms, which would lead to the exclusion of any SOA greater than 136 ms. Since there were no values greater than 136 ms, this person's mean SOA would remain at 107 ms (median = 119 ms). If Person 1 had an IT score of 107 ms (or 119 ms) and Person 2 had an IT score of 68 ms, they would have been clearly differentiated and IT as a screening tool would have accurately identified Person 2 as not being at risk of cognitive problems in the future. Furthermore, this method would be superior to the variance method because it does not require in the exclusion of any data points.

These three suggestions are not intended to provide a concrete solution to the statistical problems with IT as a screening tool. However, they do highlight some of the issues that should be considered and could lead to a substantial improvement in the reliability of the IT measure. In order to provide prescriptive guidance, the theoretical and statistical issues surrounding the adaptive staircase procedure would need to be considered in depth. The suggestions presented above do not take into account the properties of the psychometric curve that maps accuracy against SOA and the level of accuracy at which the estimate is set (in the current method 79% accuracy). However, if these issues were considered and even a small number of improvements were achieved, it seems probable that predictive validity of the IT measure would be enhanced and a more effective screening tool for everyday and cognitive problems in the future provided.

Final Assessment of IT as a biomarker

At the end of Chapter 1, a detailed plan was set out for testing the validity of a biomarker. Essentially, this plan proposed that a biomarker should be assessed theoretically to see whether it met the requirements; a literature review should be completed to consider whether the biomarker met the more specific hypotheses of Birren and Fisher (1992); and finally the *ex-post facto* model (Ingram, 1991) should be utilised to evaluate the biomarker. In Chapter 1, we stated that the *ex-post facto* model could be evaluated for a number of possible lead indicators at all, or some, of the levels of importance defined by Birren and Fisher (1992). That is, the biomarker could be

assessed along with measures of physiological processing, cognition and everyday living. If the study was long-term, the predictive validity of the biomarker for mortality could also be examined. Finally, we stated that, if results were positive after the *ex-post facto* model had been examined, then the *ipso facto* model could be used to evaluate interventions such as an exercise program. The efficacy of IT as a biomarker will now be considered by examining each step in the plan detailed above in light of the findings from this investigation.

Theoretical Requirements

In Chapter 3, an extensive examination of IT was completed to see whether this measure met the theoretical requirements of a biomarker. It was argued that IT was indeed *biological in nature, reflected some element of normal aging*, had acceptably *high reliability* (not withstanding problems with the estimation procedure just described), was *stable across generations* and was *non-lethal to animals and minimally traumatic to humans.* At that time, there was insufficient evidence to conclude that IT *changed independently with the passage of time* and *exhibited reliable change over a relatively short period of time*. However, it is now possible to address both of these theoretical requirements based on the current dataset.

Table 8.3. Change over time in Inspection Time by Age Group

Age Groups	6-month change			18-month change		
	n	Mean	SD	n	Mean	SD
70 – 74 years	35	1.01	14.23	38	0.43	18.10
75 – 79 years	50	-2.69	19.01	42	-0.34	14.49
80+ years	28	3.54	17.56	23	-0.09	23.54
Total	113	0.00	17.34	103	0.00	17.97

Note. The scores represent the change over 6-months and 18-months in ms once differences in initial scores have been statistically removed (see Chapter 5 for discussion).

To test whether that IT *changes independently with the passage of time,* we will consider whether individuals in different age groups (70 – 74 years, 75 – 79 years, and 80+ years) show differential decline in IT over 6-months and 18-months[18]. Table 8.3 presents descriptive statistics for the change scores in each of the three age groups. Positive mean scores indicate that the group has declined more on speed of processing (i.e. increase in IT), than the sample as a whole,

[18] These analyses used the residual change scores for IT over 6-months and 18-months, respectively.

over the relevant time period. Over the first 6-months, there was a tendency for the oldest group (80+ years) to show the most decline, followed by the youngest group and finally the group aged 75 to 79 years. However, this difference was not statistically significant (F (2, 110) = 1.25, p > .05, partial η^2 = .02). Over 18-months, the change was near identical for each of the three age groups and therefore not significantly different to one another (F (2, 100) < 1.0, p > .05, partial η^2 = .00). It is interesting that differences between age group were larger for 6-month change scores than 18-month change scores. One possible explanation of this finding is that the individuals with the largest decline over 6-months, discontinued participating before the end of the study. Thus, the sample with 18-month change scores would have been more homogeneous than the sample with 6-month change scores. Nonetheless, there was insufficient evidence to conclude that IT *changes independently with the passage of time.*

The second issue was whether IT *exhibited reliable change over a relatively short period of time*. In Chapter 5 we examined the reliability of the 6-month change score for IT, reporting the coefficient as 0.53. Considering the lower upper-limit to reliability in change scores, we concluded that this value was acceptable. This conclusion was supported by the fact that 6-month change scores for IT predicted final scores on all three fluid ability tasks (i.e. demonstrated significant predictive validity despite the reduced reliability). Therefore, we can now conclude that IT does *exhibit reliable change over a relatively short (i.e. 6-month) period of time.*

Specific Requirements

In Chapter 3, an examination of the specific requirements of a biomarker detailed by Birren and Fisher (1992) was completed. Speed of processing was linked to *mortality*, *physiological aging*, *cognitive aging*, *life-style* and *disease*. More specifically, IT was linked to *cognitive aging* and *disease*. There was insufficient evidence to conclude that *non-human primates show decline in IT with age* or whether *older males show more decline in IT than older females*. Some of these issues have been clarified by the current investigation and will be discussed below. They include the link between IT and life-style factors, physiological aging and gender.

First, the relationship between IT and life-style factors has been clarified to some degree. Relationships between IT and four lifestyle factors were investigated: *cigarette smoking*, *alcohol consumption*, *exercise* and *nutrition*. We found no evidence to suggest that IT was related to cigarette smoking but there were some statistical issues that might have affected this result. The elderly sample consisted of just six current smokers, so that there was not adequate statistical

power to compare this group with non-smokers and ex-smokers. When non-smokers and ex-smokers were compared to one another, there was no significant difference between them on IT. Current and ex-smokers were asked about the number of cigarettes that they smoked per day, to try to estimate the amount of nicotine and tar, etc. to which they had been exposed but this question was not answered adequately. Therefore, the data on cigarette smoking were incomplete and the sample contained very few current smokers. Thus, although no relationship between IT and cigarette smoking was found, there were statistical and methodological shortcomings that might have influenced the result and further research on this topic is therefore required.

The relationship between IT and alcohol consumption was complex. We found a linear relationship between IT and alcohol consumption with abstainers having the slowest ITs and heavy drinkers having the quickest ITs[19]. This pattern was consistent with expectations except that we hypothesised that heavy drinkers would also have slower ITs. One problem was that the sample included very few self-reported heavy drinkers and thus this group was not well represented. One cannot be confident that shorter ITs are a general characteristic of older heavy drinkers. Rather, this might have been a spurious finding, limited to this particular small sample. The lack of heavy drinkers might not be so much of a sampling problem as a general tendency for elderly people to avoid heavy alcohol consumption. If so, any relationship between IT and alcohol consumption might need to be tested in a younger sample. Although no statistically significant link was found between IT and alcohol consumption, the pattern of results was generally consistent with expectations. Thus, sampling a larger (n = 200) group who were younger (e.g. 50 – 70 year olds) is recommended, in order to resolve the true relationship between IT and alcohol consumption. It may also be useful to examine past alcohol consumption, because a number of the elderly people in this sample were classified as abstainers despite the fact that they may have drank alcohol in the past. Given that elderly people often take multiple medications, many are unable to drink alcohol. However, a history of alcohol consumption would be expected to have had some impact on the central nervous system and therefore any valid biomarker.

There was some support for the notion that exercise was related to IT. People who reported doing the most exercise and those people who reported involvement in sports had the shortest ITs. However, there was also evidence that sedentary individuals had quick ITs,

---

[19] The alcohol consumption groups did not differ significantly on IT scores.

compared to the rest of the sample. This was unexpected and these contradictory findings with respect to the link between IT and exercise led to the conclusion that a reliable relationship had not been found. However, similar to the argument for alcohol consumption, it might be particularly important to investigate past exercise behaviour. Dik, Deeg, Visser and Jonker (2003) showed that early life exercise was predictive of cognition in later life. There are a number of reasons that an elderly person might currently be sedentary, many of which are unrelated to a conscious decision to avoid exercise. Thus, the relationship in older people between exercise and biological or cognitive outcomes, including IT, is more complicated than it initially appears.

An appropriate way to study this relationship would be to run an intervention study in an elderly sample. Dustman et al. (1984) ran an exercise intervention study on a group of 43 sedentary adults. He assigned participants to aerobic exercise, strength and flexibility or a control group and the exercise groups completed one-hour of exercise, three times per week. After four months, the aerobic group showed significant improvement in DS and RT, compared with the other two groups who registered no significant change. It would be very interesting to run a similar study with IT to test the prediction that the aerobic group should display shortened ITs. This would establish a causal link between exercise and IT and add weight to the argument that IT has potential as a valid biomarker.

No evidence was found for an association between nutrition and IT. Three minerals, seven micronutrients, three fatty acids and three antioxidants were examined but no relationship between IT and nutritional intake was found. This was surprising, given that the sample size (n = 115) was adequate and there were several nutritional indices. We also failed to establish a link between nutrition and perceptual speed, despite the fact that perceptual speed has been linked to folate and antioxidants in the literature.[20] One possible explanation is that the method of using food diaries followed here may not be adequate in a sample of this kind. Although the diet diary, if kept properly, gives a good indication of nutritional intake, it does not necessarily provide a reliable indication of nutritional levels in the body. Russell (2001) and others have shown that elderly adults tend to have more problems with malabsorption, due to a number of factors, including medication use and disease. So, an individual might have adequate folate intake but if

[20] One of the perceptual speed tasks, PC, was significantly related to iron intake. All other effects were non-significant.

this is not absorbed properly then s/he will have inadequate folate levels in the body. For this reason, it may be necessary to take blood samples and have these analysed for nutritional levels; but this would obviously be more intrusive; at it was not possible in the current investigation. To summarise, no link between nutrition and IT was found but theory would predict an association and further investigation using blood samples, rather than diet diaries, should therefore be considered.

Overall, no link was established between IT and life-style factors. Despite these null findings, this area of research has been extended and a number of recommendations for further research have been made. Furthermore, it seems plausible to conclude that, although not established here, a relationship might still exist between IT and life-style.

Second, this investigation represents the first attempt to examine the relationship between IT and physiological measures. Birren and Fisher (1992) stated that *the biomarker should correlate with physiological and anatomical indicators of aging* and gave a number of examples including lung vital capacity and hearing threshold. This study failed to establish a relationship between IT and any physiological indicators of aging. IT was not significantly correlated with grip strength, blood pressure, weight, height or visual acuity at Times 1 or 3. However, there are reasons to believe that this might be due to the homogeneity of the sample. On pages 95 and 96, a review of the literature suggested that grip strength, blood pressure, weight, height and visual acuity were valid biomarkers. However, in this study at Time 3, all significant correlations between the physiological measures disappeared once gender was partialled out. At least some of these measures have been proven to correlate in previous research (e.g. Anstey & Smith, 1999) but these results could not be replicated in this sample. Therefore, it is possible that the lack of relationship between IT and physiological indicators of aging is an artefact of this homogeneous sample.

Finally, the data collected in this study have allowed us to consider whether older males show more decline in IT than older females. We have estimates of change in IT over 6-months and 18-months and these can be compared for males and females. Throughout this investigation, residual change methods to generate change scores. However, in the current context it seems just as valid to consider difference scores when attempting to answer the question – do older men show more decline in IT than older women? There was very little difference between men and women in IT change over 6-months, whether indexed by difference scores or residual change scores. However, there were slightly larger differences in change over 18-month. Sixty-two

women (M = 77 years[21], SD = 4.3) and 41 men (M = 77 years, SD = 3.5) completed IT at Times 1 and 3. Men displayed an average slowing of 5 ms while women improved by an average of 1 ms. This difference was statistically significant at the 10% level when indexed by the difference score (t (101) = 1.72, $p < .10$, d = 0.35) but the residual change score was not. Furthermore, the statistical power for this comparison was low (.40) indicating a 40% chance of detecting a difference when one exists. It is possible that on average older males do show marginally larger decline in IT than older females but, because of low power, this could not be reliably confirmed.

Ex-Post Facto Model

In Chapters 4 through 7, the ex-post facto model was used to investigate whether IT was a valid biomarker (i.e. lead indicator) for individual differences in functional age. A sample of 150 elderly adults (M = 77.6 years, SD = 4.4) was recruited and observed during a period of 18-months, without experimental manipulations. The major aim was to test whether IT could predict "functional outcomes" in the future and this was investigated at a number of different levels, based on Birren and Fisher's model. At the *subordinate* level, whether IT was related to physiological processes was examined including grip strength, blood pressure, height, weight and visual acuity. However, we failed to establish a link between IT and physiological processes and possible explanations have been presented previously (see p. 96 and p. 188). Thus, this level is not considered here further.

Whether IT could predict performance in cognition, specifically fluid and crystallised ability, was examined at the *coordinate* level. At the *superordinate* level, the test was whether IT predicted performance in everyday living. Given the short time frame, IT could not be examined at the *general* level; whether IT could predict longevity or mortality could not be adequately tested for this sample, within this time frame. However, as argued in Chapter 3 (see p. 55), if IT predicted decline in crystallised ability then this might imply that it was predictive of mortality, given that decline in crystallised ability is a risk factor for death within the next five years (Cooney et al., 1988). In the following sections, IT will be evaluated at the coordinate and superordinate levels.

[21] Age at Time 1

Coordinate Level – Cognition

Evaluation of a biomarker at the coordinate level involves examining the relationship between the biomarker and cognitive functioning. Because some types of psychometrically defined cognitive abilities decline with age (e.g. fluid ability, short-term memory and visualisation), Birren and Fisher have proposed that an effective biomarker will predict future test performance and/or decline in test performance. It is important to note that cognitive functioning is not a valid outcome measure simply because it declines with age. Rather, it is important because cognitive problems impact on so many aspects of life, and, ultimately for some cases what begins as mild cognitive impairment can develop into dementia. Therefore, evaluation of a biomarker at the coordinate level is an extremely important issue, at both theoretical and practical levels.

This investigation focused on fluid ability and crystallised ability as outcomes variables because fluid ability shows a large decline with age, with some studies reporting decline from the late 20s (Schaie, 1994). Furthermore, fluid ability encapsulates problem solving, which inevitably would be expected to impact on day-to-day life for an elderly person. A principal aim was, therefore, to test whether IT could predict future performance on tests of fluid ability and decline in fluid ability over the 18-month period of the study. Crystallised ability was also of interest because, although it tends to be maintained into old age, when it does begin to decline this can be indicative of impending death (Cooney et al., 1988). A second aim was therefore to test if IT predicted decline in crystallised ability over the 18-month study period, consistent with prognosis of mortality.

Fluid Ability.

In Chapter 7, it was reported that both initial IT scores and 6-month change scores for IT correlated with subsequent performance on tests of fluid ability. The trend was that people with slower IT at the start of the study, and those who registered slowing IT over the first 6-months of the study, had poorer fluid ability 18-month after first assessment. These two IT estimates were statistically independent and, when entered into a regression model, both explained significant variance in fluid ability 18-months subsequently. Moreover, when age and gender were entered into the model first, initial and 6-month change IT scores remained significant predictors. These results have provided strong evidence that IT can be prognostic, i.e. able to predict future performance tests of fluid ability. However, evidence that IT was able to predict decline in fluid ability over time was much weaker. For RSPM, the predictive validity of initial IT scores

approached significance (p = .067); i.e., people with slow IT at the start tended to decline more in RSPM over the subsequent 18-months. However, this was not confirmed for the CCFT or CF. To summarise, IT is prognostic for future test performance but it cannot be concluded that it predicts decline in fluid ability.

Three possible explanations for why IT did not predict a decline in fluid ability are worth considering. First, there was minimal decline in mean fluid ability over 18-month, so that there was little variance for IT to predict. Support for this suggestion comes for the observation that RSPM was the only test of fluid ability that showed reliable mean decline over 18-months and initial scores for IT approached significance for this outcome, with moderate effect size. Second, the sub-sample of people with 18-month change scores for fluid ability tended on average to be higher functioning than the whole sample and restricted in range, which may have reduced the correlation between IT and the fluid ability change scores. There were substantial missing data for change scores over 18-months in CCFT and CF, because, at the first testing session, several participants failed to complete CCFT and/or CF, essentially because they were reluctant to persist in what was inevitably for some, a very prolonged testing session. Importantly, these people were not random; subsequent considerations of their testing profiles revealed a tendency for those not completing these tests to take longer and to do less well on other tests. As a consequence of losing data for these individuals, power was reduced and those participants registering change scores for CCFT and CF tended to be more homogeneous and restricted in range than the sample as a whole[22]. Thus, the sub-sample of people with 18-month change scores on CCFT and CF had a restricted range of fluid abilities and this may have reduced the correlation between IT and decline in fluid ability. Finally, the 18-month time frame may have been too short to determine reliably whether IT predicts decline in fluid ability. Nonetheless, evidence supported IT as prognostic of decline in RSPM and this was the only measure to show significant mean decline over 18-months. Further work is necessary to assess fluid ability after a longer period of time, sufficient to establish mean decline in the group, and then to test whether initial IT and short-term change in IT can predict this decline.

---

[22] In Chapter 6, whether people who failed to complete CCFT or CF at Time 1 had poorer RSPM or more decline in RSPM over 18-months was tested. The evidence was clearly in the expected direction but statistically non-significant.

Crystallised Ability.

The aim was to test whether IT would predict decline in crystallised ability over time although, as described in Chapter 6 and elsewhere, a null outcome was expected. As described in Chapter 6, it was clear that all three crystallised ability measures were very stable over 18-months. Although a small number of people showed decline, this was of very small magnitude. From the analyses in Chapter 7 it was clear that there was no evidence that IT could predict decline in crystallised ability. In retrospect, given that two of the fluid ability tests displayed stability over 18-months, it is not surprising that the crystallised ability tasks showed mean stability; and 18-months was probably too short to examine decline in crystallised ability. The original proposal was that decline in crystallised ability might be used as a proxy for mortality. Just two people passed away during the course of this investigation and this was clearly too short a time frame to study mortality.

Birren and Fisher (1992) stated that a biomarker is valid at the *coordinate* level if it is predictive of future test performance *or* decline in test performance. Results have shown that IT predicts future test performance on fluid ability but did not predict decline in fluid or crystallised ability perhaps because of an insufficient time frame for this investigation. Nonetheless, based on these findings, a justified conclusion is that IT has promise as a valid biomarker when evaluated at the coordinate level, because it can predict future test performance on fluid cognition.

Superordinate Level - Everyday Functioning

The *superordinate* level is the second highest level at which a biomarker can be evaluated. As argued in Chapter 1, the decision was to work through the levels, from least important to most important. Given that IT has been validated at the *coordinate* level, it is now necessary to consider the next highest or superordinate level. The superordinate level involves examining whether the biomarker can predict functioning in everyday life, for which the analysis of future functioning in everyday living and decline in everyday functioning were considered valid outcome measures.

Birren and Fisher (1992) recommended that everyday functioning should be operationalised by instrumental activities of daily living. Initially, a questionnaire that included items on basic and instrumental activities of daily living (ADL scale) was administered. In addition, because of perceived limitations with the ADL scale, a second scale on the cognitive aspects of everyday functioning (CDL) was included at the final testing phase. These

instruments therefore enabled testing whether IT predicted future performance on general and/or cognitive aspects of everyday functioning and whether IT predicted decline in ADL over 18-months.

As set out in Chapter 7, evidence confirmed that initial scores for IT predicted future performance on ADL. People with slow IT performance at the start were more likely to be dependent on others for ADL in the future. Moreover, 6-month and 18-month change scores for IT approached significance for predicting future performance on ADL, with moderate effect sizes. In both cases, people who slowed down over the course of the study were more likely to be dependent on others in ADL at the end. Similarly, for CDL, evidence suggested that people who slowed on IT over 18-months tended to have more cognitive problems in their everyday lives at the end of the study. Taken together, these results confirmed a prognostic trend; IT predicted everyday functioning 18-months in the future.

Unfortunately, the CDL scale was only measured at the end of the study, so that no measure of decline in cognitive aspects of everyday life was available. For ADL, there was a significant mean decline over 18-months for the whole group and this was largely due to a sub-group of 30 people who displayed decline in ADL. Initial scores for IT bordered on statistical significance as a predictor of decline in ADL ($p = .052$). That is, people with slow ITs at the start of the study were more likely to decline in their independence in everyday functioning over the subsequent 18-months.

At the superordinate level, a biomarker must be predictive of future functioning in everyday living or decline in everyday functioning. This research has shown that IT predicts future performance in general and cognitive aspects of everyday functioning and decline in general everyday functioning. Thus, it is concluded that IT is a valid biomarker at the superordinate level, as well as the coordinate level.

General level - Mortality

For IT to be a valid biomarker at the highest level, it must be predictive of longevity or mortality. However, it was beyond the scope of this investigation to examine IT at the *general* level. Nonetheless, this question could be considered in the future and most of the work for this has been done. Initial scores, 6-month and 18-month change scores for IT have been measured and it should be possible to collect information about longevity in the future. Comprehensive contact and next-of-kin information are available so that it should be possible to collect survival data at a later stage. Logistic regression analyses could obviously test whether initial or change

scores for IT predicted survival status in the future. Given that IT has been validated at the coordinate and superordinate level, validation at the general level remains the final requirement for IT to meet to be considered a successful biomarker.

Ipso Facto Model

The final step in the plan from Chapter 1 was to design an experiment using the *ipso facto* model, if results from the ex post facto model were encouraging, which is seems can be confidently claimed. This model involves measuring the biomarker, splitting the sample into an experimental and control group, applying an intervention and observing differences between the groups on the biomarker at the end. If the intervention is effective, then the biomarker should change to different degrees in the experimental and control group. Given the success of IT at the coordinate and superordinate levels, it would be appropriate to run an experiment and test whether the experimental group demonstrated reduced slowing over time on IT, compared with the control group. The question then becomes what type of intervention study should be conducted?

In the above section on "Specific Requirements", it was suggested that it would be interesting to run an exercise intervention study. There were a number of difficulties to evaluating the relationship between IT and exercise and an intervention study would allow this to be clarified. Furthermore, a similar study has already been conducted (see Dustman et al., 1984), showing that aerobic exercise had an effect on other speed of processing measures over a period of just 3-months. The current research has shown that decline over 6-months in IT is reliable and predictive of test scores on fluid ability. Therefore, it would be feasible to design a study where the exercise program was run over a period of 6-months and test whether the experimental group had a significant different IT change score when compared with the control group. A positive outcome would provide support for IT as a biomarker.

Limitations of the Current Investigation

The current investigation has found extensive support for the original suggestion by Nettelbeck and Wilson (2004) that IT might have some promise as a biomarker. However, there were a few limitations and problems with this investigation that should be noted for future biomarker and IT research. Major issues were the restricted range of abilities in the sample, the assessment of functional age, the time frame of the investigation, and shortcomings with the test battery.

Restricted Range. During the recruitment period, we sought elderly people over 70 years of age who lived independently and were free from dementia. One of the limitations of this study was that volunteers were highly educated, compared to average trends among persons aged over 70 years and, consistent with that, demonstrated above average pre-morbid levels of intelligence. Members of the sample were also very independent in their everyday lives, with approximately half of the sample reporting independence in all areas of their lives. Therefore, there were marked restrictions in range of their abilities, which reduced the likelihood of observing significant relationships among the variables of interest. As the study progressed, the sample experienced attrition, with 15% of the participants discontinuing before the end. Those people who dropped out were significantly older, had slower ITs and had poorer performance on all of the cognitive tasks and, consequently, the sample of 127 at the end of the study was more restricted in range than the initial sample.

There were a number of effects for IT that approached but did not achieve statistical significance (defined by $\alpha = .05$). These included the predictive validity of 6 and 18-month change scores for final performance on ADL. If the range of ADL scores had been wider, it is likely that more statistically significant relationships would have been observed and it is for this reason that interpretation of outcomes has also been guided by effect sizes.

The problem of homogeneous samples and restriction in range is certainly not unique to this current study. Rather it is a problem common to all aging research and is largely due to selection bias and selective survival or attrition. In cross-sectional and longitudinal research with elderly participants, it tends to be the healthier, more intelligent and enthusiastic people who volunteer. This problem is known as selection bias and often leads to a restriction in range. Furthermore, as pointed out by Rabbitt, Diggle, Smith, Holland and McInnes (2001) this problem is compounded in longitudinal research by selective attrition. Rabbit et al. (2001, p. 535) state that “during long studies, the oldest, frailest and least able participants are the most likely to withdraw so that the samples become more elite”. As a result, the samples in longitudinal research become even more restricted in range, which reduces the likelihood that significant correlations or relationships between constructs of interest will emerge. Despite the restricted range in this study, results have been promising and IT has been shown to be a lead indicator for functional outcomes at the coordinate and superordinate levels. This implies that IT does indeed have promise as a useful biomarker, with efficacy likely to be greater in a more heterogeneous sample.

Assessment of Functional Age. According to Birren and Fisher (1992) it is valid to assess a putative biomarker based on its relationship with cognition, everyday living and mortality. However, everyday living is a very broad concept and may not be adequately represented by the questions in the ADL scale. What aspects of everyday life are most important in describing an independent elderly adult? The ADL scale would suggest that managing finances, transportation, using a telephone and taking care of shopping are highest priorities. However, although these aspects of daily life are undoubtedly very important core activities that are critical to independent functioning, they also should be regarded as limited and there must be other important considerations, which have not been represented in scales of this kind. Perhaps the best people to ask about what should be included are elderly people themselves. It may therefore be useful to run focus groups with elderly people, to ask them what abilities or skills they feel are important measures of independent living. The major point here is the ADL scale may be too limited in its focus and research should be done to find or develop more comprehensive measures to represent everyday functioning outcomes in the elderly.

Short Time Frame. A major limitation of this study was the short 18-month time frame, which was determined by current constraints on candidature for the PhD degree within the Australian postgraduate context. Just two of the functional outcomes, RSPM and ADL, showed significant mean change over 18-months and IT was a significant predictor of decline in both measures. However, this time frame was too short to permit a significant decline in the other cognitive tasks, which may explain the failure to find a link between slow IT at the start and decline in these tasks. Furthermore, an 18-month period was not sufficiently long to test whether IT was related to mortality, so that IT could not be evaluated at the general level. Ideally, the biomarker would be measured annually, over a longer period of time (e.g. 5 years) and after a subsequent period the mortality status of the sample would be assessed. The longer time frame would provide more opportunity for larger change in the biomarker, and in functional outcomes, and therefore more reliable change estimates. More data points would allow for more complex statistical analyses (e.g. application of latent growth models). Nonetheless, the findings from this investigation have been extremely encouraging and they suggest that future research of this type is warranted.

Issues with the Administration of the Test Battery. There were a number of problems with the battery of biomarker tasks. First, the method of using a measuring tape to estimate height was inadequate. Changes in height over 18-months were so small that the measurements needed

to be extremely accurate. This would best have been achieved with a stadiometer, rather than a tape measure. Despite this, there was evidence that decline in height was related to performance on fluid and crystallised ability tasks. Future studies should continue to evaluate height as a biomarker, and height should be measured with a stadiometer. Second, the test-retest reliability of height, weight, and visual acuity should have been assessed, particularly since all three are so readily measured. This check would simply involve measuring the biomarkers twice, in a testing session, with a small sample of people; and this would have allowed the estimation of the reliability of initial and change scores. Third, the reliability of visual acuity was very low and this may have been, to some extent, due to differences in lighting in the testing room. The same individual would be expected to score better on the Snellen chart if the lighting was brighter. So, although visual acuity was always measured in a well-lit room, differences between lighting from one occasion to the next, mainly when measured in people's homes, might have introduced error variance that reduced the reliability of this measure.

There were some problems associated with the administration of the cognitive battery. First, some tests that were presented on the computer should have been presented in paper-and-pencil form. All of the fluid ability tasks were presented on a computer screen and the participants had to press or verbalise their answer. These tasks had previously been computerised to study correct and incorrect decision speed, constructs described by Horn (1988) as part of *Gf-Gc* theory and this method had the advantage that outcomes variables could be calculated automatically for each section and test. However, in the current study a number of the elderly people indicated that they were uncomfortable with the computer or they complained of experiencing visual problems when looking at the screen for sustained periods of time. Given that correct decisions speed was not important in this investigation, it would have been better to administer the fluid ability tasks in paper-and-pencil form. Second, despite initial piloting that attempted to confirm the adequacy of the protocol planned, the test battery proved too extensive for some elderly participants to complete comfortably. When designing the cognitive test battery, three measures of each construct were chosen (*Gf*, *Gc* and *Gs*) so that factor scores could be calculated. However, this led to a very long testing session and ultimately to missing data for some of the fluid ability tests. In hindsight, it would have been more appropriate to include fewer cognitive measures but to ensure that all participants completed all of the tests.

The final methodological issues relate to the IT task and for the most part these have already been discussed (see p. 179). To re-iterate, problems can arise when individuals make

errors near the start of IT estimation, because the estimation phase becomes extended, increasing the likelihood of additional errors because of fatigue or reduced vigilance. An additional sub-routine could be written into the program, to check whether the first mistake was at a lengthy SOA – e.g. longer than 200 ms. The other major issue was that the algorithm used to estimate IT takes into account all eight reversals whereby an increment or decrement to SOA is made according to whether the response was incorrect or correct, respectively. As discussed above, early errors that occur at long SOAs, well beyond the minimum SOA actually achieved, tend to lengthen the estimate for IT. If the participant completes three items correct at a SOA of 51 ms, then any reversals at larger SOAs (e.g. > 120 ms) do not really tell us much about the time required to make a single observation of sensory information. A suggestion has been made, to focus on reversals at the lowest SOAs, when estimating IT and thereby exclude those reversals that occur a pre-specified time from the minimum SOA. The major point here is that the main features of the method for estimating IT by Adelaide researchers have been unchanged for some 20 years, yet there are improvements that could be made to these procedures to make IT a better representation of the construct of inspection time.

Next Steps for IT

This dissertation has provided evidence that IT is a valid biomarker for functional age. Therefore, the next steps should be to build upon this initial study and provide more evidence for IT as a prognostic, lead indicator for individual differences in functional age. A number of issues need further investigation and suggestions have been made about how to proceed. First, the relationship between IT and life-style factors needs further work, based on the recommendations made in the *Specific Requirements* section. It would also be interesting to see if change scores for IT were related to life-style factors. A sensible question would be, do people with adequate nutrition tend to show less slowing in IT over time than people with less adequate nutrition? Second, the predictive validity of IT for longevity or mortality should be studied. A convenient way to begin would be to collect mortality data on the current sample prospectively. It may also be possible to find a previous study on IT in the elderly (e.g. Nettelbeck & Rabbitt, 1992) and attempt to get permission to collect mortality data from that sample. If the IT and mortality data were available, then the predictive validity of initial IT scores for mortality could be studied quickly, rather than waiting for 5 – 8 years to collect data from the current sample. The final suggestion is that the estimation procedure for IT be further investigated with the aim of

improving reliable discriminability of individual differences. Based on the ideas presented in the discussion of IT as a screening test (see p. 179 - 182), there are relatively small procedural changes that might be made to the estimation procedure, to make the task shorter, and to generate more reliable estimates of IT, that would be relatively straightforward to check. Research of this nature would be potentially beneficial for all future research on IT and should therefore be a high priority.

APPENDIX A. LIFE-STYLE QUESTIONNAIRE

1.	Have you ever smoked cigarettes? (if No, go to question 5)	□ Yes □ No
2.	Do you currently smoke cigarettes?	□ Yes □ No
3.	If Yes, how many cigarettes do you smoke per day?	…….
4.	How many years have you or did you smoke for?	…….
5.	Do you drink alcohol regularly (eg. once a week)?	□ Yes □ No
6.	If Yes, how many standard drinks per week? (1 standard drink = 1 glass of wine 150ml).	…….
7.	How many times per week do you exercise?	…….
8.	On average, how long do you exercise for each time?	…….
9.	What type of exercise do you do?	…….

APPENDIX B. FOOD DIARY

Dear _____

We would like you to write down EVERYTHING you eat and drink into this Food Diary for the days specified at the top of each page.

Please record the AMOUNT that you eat and drink in the first column. For example, 1 glass of milk, 2 teaspoons of sugar, 50gm chips.

We would also like to know the BRAND NAMES AND TYPES of food eaten if possible. For example, 2 slices of Tiptop Multigrain Bread, 1 teaspoon of Meadow Lea margarine, 1 teaspoon of Cottee's Apricot Jam.

It is easy to forget little things like sugar in tea or coffee, a glass of water, and margarine on bread. The best way to avoid this is to FILL IN YOUR DIARY AS SOON AS POSSIBLE AFTER YOU EAT. Avoid leaving it to the end of the day because you may forget things.

We want to get a measure of what you usually eat so please DO NOT CHANGE YOUR EATING HABITS just because they are being recorded.

Finally, we would also like you to record ALL MEDICATIONS that you take over the same time period. Please tell us how many you take (eg. 2 tablets), the name of the medication (eg. Panadol), and the reason (eg. headache).

An example of a filled in FOOD DIARY is presented in the following 4 pages.

Day Wednesday		**Date** 18th March	
Quantity or weight	**Food and Drinks Consumed**	**Type and Brand**	**Preparation or Cooking Method**
BREAKFAST			
4	Weetbix	Sanitarium	
20 gm	Dried fruit		
½ cup	Milk	Farmers Union	
1 teasp	Sugar		
1 cup	Multi-vitamin juice	Berri	
MORNING TEA OR SNACK			
1 piece	Cake - chocolate	Home made	
1 cup	Coffee	Nescafé	
1 teasp	Sugar		
	Milk	Farmer's Union	

Day Wednesday		**Date** 18th March	
Quantity or weight	**Food and Drinks Consumed**	**Type and Brand**	**Preparation or Cooking Method**
LUNCH			
	Sandwich		
2 slices	White bread	Buttercup Wonderwhite	
1 teasp	Margarine	Flora	
1 slice	Cheese	Coon slices	
1 leaf	Lettuce		
3 slices	Tomato		
1 slice	Ham		
600 ml	Coke		
AFTERNOON TEA OR SNACK			
50 gm	Salt and Vinegar chips	Thins	
1	Apple (small)		

Day Wednesday		**Date** 18th March	
Quantity or weight	**Food and Drinks Consumed**	**Type and Brand**	**Preparation or Cooking Method**
DINNER/ TEA			
	Lasagne		Baked in oven
3 sheets	Lasagne strips	San Remo	
1 cup	Tomato and beef mince sauce	Home made	Fried in oil
½ cup	Cheese sauce	Home made	
	Salad		
3 leaves	Lettuce		
4 slices	Tomato		
4 slices	Cucumber		
1 teasp	Salad dressing	Praise French	
1 glass	White wine	Wynns	
EVENING SNACKS (or other food)			
200 gm	Strawberry yoghurt	Yoplait	

In the following table please indicate all medications that you took today and any vitamins, minerals or herbal supplements.

Day Wednesday	**Date** 18th March
Quantity	**Medication**
2 tablets	Vitamin C
1 tablet	Panadol (Headache)
2 tablets	Accupril (High blood pressure)

Day Friday		**Date** 2nd April 2003	
Quantity or weight	**Food and Drinks Consumed**	**Type and Brand**	**Preparation or Cooking Method**
BREAKFAST			
MORNING TEA OR SNACK			

Day Friday		**Date** 2nd April 2003	
Quantity or weight	**Food and Drinks Consumed**	**Type and Brand**	**Preparation or Cooking Method**
LUNCH			
AFTERNOON TEA OR SNACK			

Day Friday		**Date** 2nd April 2003	
Quantity or weight	**Food and Drinks Consumed**	**Type and Brand**	**Preparation or Cooking Method**
DINNER/ TEA			
EVENING SNACKS (or other food)			

In the following table please indicate all medications that you took today and any vitamins, minerals or herbal supplements.

Day Friday	**Date** 2nd April 2003
Quantity	**Medication**

APPENDIX C. INFORMATION TEST

Item	Question	Multiple choices	Answer
1	What do we call a baby cow?	(a) Bull (b) Calf (c) Foal (d) Piglet	b.
2	How many things make a dozen?	(a) Twelve (b) Six (c) Eggs (d) Ten	a.
3	Who was Captain Cook?	(a) A prime minister (b) An explorer (c) An inventor (d) A cook	b.
4	Name two kinds of coins	(a) 5 Dollars and 10 Dollars (b) Money and Dollars (c) 20 cents and 50 cents (d) Indian and Corn	c.
5	On what continent is China?	(a) Asia (b) South Africa (c) South America (d) Europe	a.
6	Which month has one extra day every four years?	(a) February (b) January (c) May (d) December	a.
7	What is the capital of Greece?	(a) Rome (b) Athens (c) Crete (d) Cairo	b.
8	How is Oxygen returned to the air?	(a) By breathing (b) By plants (c) By the wind (d) By clouds	b.

Item	Question	Multiple choices	Answer
9	What is water made of?	(a) Minerals and chemicals (b) Rain (c) Helium and Oxygen (d) Hydrogen and Oxygen	d.
10	What are hieroglyphics?	(a) Ancient Greek letters (b) Roman numerals (c) Egyptian picture writing (d) Cave drawings	c.
11	What country has the largest population?	(a) India (b) Russia (c) North America (d) China	d.
12.	What is the main material used to make glass?	(a) Sand (b) Plastic (c) Hydrogen (d) Fibreglass	a.
13	In what direction does the sun set?	(a) North (b) East (c) South (d) West	d.
14	Who invented the electric light bulb?	(a) Albert Einstein (b) Thomas Edison (c) Benjamin Franklin (d) Thomas Jefferson	b.
15	Who wrote Hamlet?	(a) William Tell (b) Mark Twain (c) Ernest Hemingway (d) William Shakespeare	d.
16	Who was Prime Minister of England during the Second World War?	(a) Winston Churchill (b) Stanley Baldwin (c) Margaret Thatcher (d) Clement Attlee	a.
17	In what country did the Olympic Games originate?	(a) Egypt (b) Greece (c) Rome (d) Italy	b.

Item	Question	Multiple choices	Answer
18	What is a barometer?	(a) It measures air pressure (b) It measures wind speed (c) It measures rainfall (d) It measures earthquakes	a.
19	On what continent is the Sahara Desert?	(a) Africa (b) Europe (c) Arabia (d) Asia	a.
20	Who was Anne Frank?	(a) A singer (b) A pilot (c) A girl who wrote a diary (d) A teacher of deaf and blind	d.
21	Who was Charles Darwin?	(a) He was a poet (b) He developed the theory of evolution (c) He was a character in a Dickens novel (d) He discovered the structure of DNA	b.
22.	Who painted the Sistine Chapel?	(a) Botticelli (b) da Vinci (c) Raphael (d) Michelangelo	d.
23	How far is it from London to Sydney (approx.)?	(a) 500 km (b) 7,000 km (c) 17,000 km (d) 40, 000 km	c.
24	Who was Mahatma Gandhi?	(a) An Indian Prince (b) A cricket player (c) An Indian independence leader (d) A Buddhist monk	c.
25	Whose name is usually associated with the theory of relativity?	(a) Planck (b) Newton (c) Watson (d) Einstein	d.
26	What causes iron to rust?	(a) Acid (b) Salt (c) Oxygen (d) Minerals	c.

Item	Question	Multiple choices	Answer
26	What causes iron to rust?	(e) Acid (f) Salt (g) Oxygen (h) Minerals	c.
27	Name three kinds of blood vessels in the human body?	(a) Pulmonary, capillary and aorta (b) Artery, vein and capillary (c) Artery, aorta and vein (d) Capillary, jugular and vein	b.
28	Visual problems are most often caused by a deficiency in …	(a) Vitamin A (b) Vitamin B (c) Vitamin C (d) Vitamin D	a.
29	Who was Catherine the Great?	(a) A Roman Empress (b) A French Queen (c) A Russian Empress (d) An Egyptian Queen	c.
30	Which is the closest planet to our sun?	(a) Mars (b) Mercury (c) Earth (d) Venus	b.
31	What is the world population (approximately)?	(a) 4 billion (b) 6 billion (c) 8 billion (d) 10 billion	b.
32.	What is the capital city of Sri Lanka?	(a) Sinhal (b) Colombo (c) Tamil (d) Matale	b.
33.	What is the speed of light (approximately)?	(a) 300,000 km/sec (b) 258,000 km/sec (c) 362,000 km/sec (d) 524,000 km/sec	a.
34	What was Marie Curie famous for?	(a) She was a physicist (b) She was a missionary (c) She was a medical doctor (d) She was a biologist	a.

Item	Question	Multiple choices	Answer
35	What does turpentine come from?	(a) Ethyl alcohol (b) Varnish (c) Acid (d) Pine trees	d.
36	Who wrote Faust?	(a) Mann (b) Hesse (c) Nietzsche (d) Goethe	d.
37	What does the musical term 'piano' mean?	(a) To be played evenly (b) To be played fast (c) To be played softly (d) To be played smoothly	c.
38	How far above sea level is Mount Everest?	(a) 7,448 metres (b) 7,984 metres (c) 9,298 metres (d) 8,848 metres	d.
39	Who was the Greek muse of history?	(a) Urania (b) Thalia (c) Polymnia (d) Clio	d.
40	From what language does the word 'Ombudsman' originate?	(a) French (b) Swedish (c) German (d) Dutch	b.

APPENDIX D. ACTIVITIES OF DAILY LIVING SCALE

* Those items included in the shortened version used for the final testing session.

This aim of this questionnaire is to assess to what extent your day-to-day activities are independent of another's assistance, or similarly, to what extent your activities have been restricted or limited from what they use to be.

Please tick one box per section next to the sentence that you feel best/most frequently describes how you go about your daily living activities. Remember, the statements refer to what you are currently ABLE to do, not what you have previously done or would like to do.

*Food Preparation

Able to select, plan, prepare and serve meals independently, as required ☐

Able to prepare food if ingredients supplied/set out ☐

Unable to cook a meal, but capable of making snacks and reheating food ☐

Can prepare food if prompted step-by-step ☐

Need to have meals prepared and served ☐

Eating

Able to eat without assistance, using correct cutlery ☐

Able to eat without assistance provided food is made manageable

(that is, food is a particular form, consistency, or size) ☐

Find it necessary to eat food with fingers ☐

Need to be fed ☐

Drink Preparation

Able to select and prepare drinks as required ☐

Can prepare drinks if ingredients left available ☐

Can prepare drinks if prompted step by step ☐

Unable to make a drink even with prompting and supervision ☐

Drinking

Able to drink without any problems, and from an unmodified glass or mug ☐

Require aids to drink (e.g. specific type of cup, use a straw) ☐

Have difficulty drinking, even with aids ☐

Require drinks to be administered ☐

Dressing

Able to select suitable clothing and can dress/undress self unassisted ☐

Can dress/undress self, but sometimes put clothes on/take clothes off in the wrong order and/or back to front ☐

Unable to dress/undress self but move limbs to assist ☐

Require total dressing/undressing (unable to assist) ☐

Bathing

Able to bathe self (in tub, shower, sponge bath) regularly and without help ☐

Able to bathe self with help getting in and out of tub/shower ☐

Need bath/shower to be drawn/turned on, but wash independently ☐

Can wash face and hands only, cannot bathe rest of body ☐

Can wash self if prompted and supervised ☐

Unable to wash self and need full assistance ☐

Teeth

Able to clean own teeth/dentures regularly and independently ☐

Can clean teeth/dentures if given appropriate items ☐

Require some assistance, toothpaste on brush, brush to mouth etc. ☐

Need full assistance ☐

Toilet

Can manage independently whilst at toilet, no incontinence ☐

Need to be reminded, or need help in cleaning self,
or have rare (weekly at most) accidents ☐

Soil or wet self while asleep (occurs more than once a week) ☐

Soil or wet self while awake (occurs more than once a week) ☐

Have no control of bowels or bladder ☐

Grooming (neatness, hair, nails, hands, face, clothing)

Always well-groomed, without assistance ☐

Able to groom self with occasional minor assistance (e.g. shaving, lipstick) ☐

Need regular assistance or supervision in grooming ☐

Require total grooming care, but remain well-groomed thereafter ☐

*Transfers

Able to get in/out of most chairs unaided ☐

Can get into most chairs but need help to get out ☐

Need help getting in and out of most chairs ☐

Totally dependent on being put into and lifted from most chairs ☐

*Mobility

Able to walk independently ☐

Can walk with assistance (that is, furniture, or arm for support) ☐

Use aids to walk (for example, a frame, stick, walker etc) ☐

Unable to walk ☐

*Mode of Transportation

Can drive a car (without passenger assisting), and this is main form of transport ☐

Travel independently on public transportation (and cannot drive car) ☐

Can arrange own travel via taxis, but do not otherwise use public transport ☐

Travel on public transport when accompanied by another ☐

Unable to use transport even when accompanied ☐

Do not travel at all ☐

*Shopping

Able to take care of all shopping needs independently ☐

Shop independently for 1 or 2 items (small purchases), with or without a list ☐

Unable to shop alone, but participate when accompanied ☐

Completely unable to shop ☐

Communication

Able to hold appropriate conversation (listen and respond at correct times) ☐

Show understanding and attempt to respond verbally with gestures ☐

Can make self understood but have difficulty understanding others ☐

Do not respond to or communicate with others ☐

*Telephone

Able to operate telephone on own initiative (look up and dial numbers etc) ☐

Dial a few well-known numbers ☐

Use telephone if number given verbally/visually or pre-dialled ☐

Answer telephone but do not make calls ☐

Unable to use telephone at all ☐

*Housekeeping

Maintain house alone or with occasional assistance ☐

(e.g. ‘heavy work domestic help’)

Perform light daily tasks such as dishwashing, bed making ☐

Need help with all home maintenance tasks ☐

Not applicable ☐

Gardening

Able to do gardening without assistance ☐

Can garden but often require assistance ☐

Garden infrequently, even with lots of assistance ☐

Do not participate in any gardening ☐

Not applicable ☐

*Laundry

Able to do personal laundry completely ☐

Launder small items (e.g. rinse stockings or socks, etc) ☐

All laundry must be done by others ☐

Not applicable ☐

Responsibility for own medications

Responsible for taking medication in correct dosages at correct time ☐

Responsible for medication if it is prepared in advance in separate dosages ☐

Unable to dispense own medication at all ☐

Not applicable ☐

*Ability to handle finances

Able to manage financial matters independently, collect and keep tack of income (e.g. budget, write cheques, pay rent/bills, go to bank) ☐

Can manage day-to-day purchases, but need help with banking, and major purchases, etc. ☐

Unable to handle money or recognise money values ☐

Not applicable ☐

Games/Hobbies

Able to fully participate in previous pastimes/activities ☐

Can participate but need instruction/supervision to do so ☐

No longer able to join in ☐

Not applicable ☐

Orientation - Time

Are fully aware of what the day, date, and approximate time is ☐

Repeatedly ask the time, day and/or date, due to not knowing ☐

Mix up night and day ☐

Orientation - Space

Can fully interpret and understand/know my surroundings ☐

Can interpret/understand familiar surroundings only ☐

Get lost in home, need reminding where bathroom is etc ☐

Do not recognise home as own (attempt to leave) ☐

THANK YOU

APPENDIX E. VARIANCE METHOD FOR INSPECTON TIME

The adaptive staircase procedure (Wetherill & Levitt, 1965) uses the average of eight reversals of the staircase to calculate the IT estimate. In most cases, the eight reversals occur at exposure durations that are close together and the mean gives a good representation of IT. An example of a standard reversal pattern can be seen in the left panel of Figure E1. However, in some cases the reversals may be very variable, suggesting a problem on the part of the participant during the estimation phase of the task. In such cases there may be two or more reversals that occur at a very long stimulus onset asynchrony (SOA), probably due to inattention or confusion, so that the mean does not give a good representation of IT (see right panel of Figure E1).

In this study, a technique hereafter referred to as the *variance method* was used to remove the most variable estimates. These extremely variable estimates were considered to be poor estimates of the individuals' *inspection time* and to introduce error variance into the set of IT scores. The process will now be described using the Time 1 scores as an example[23].

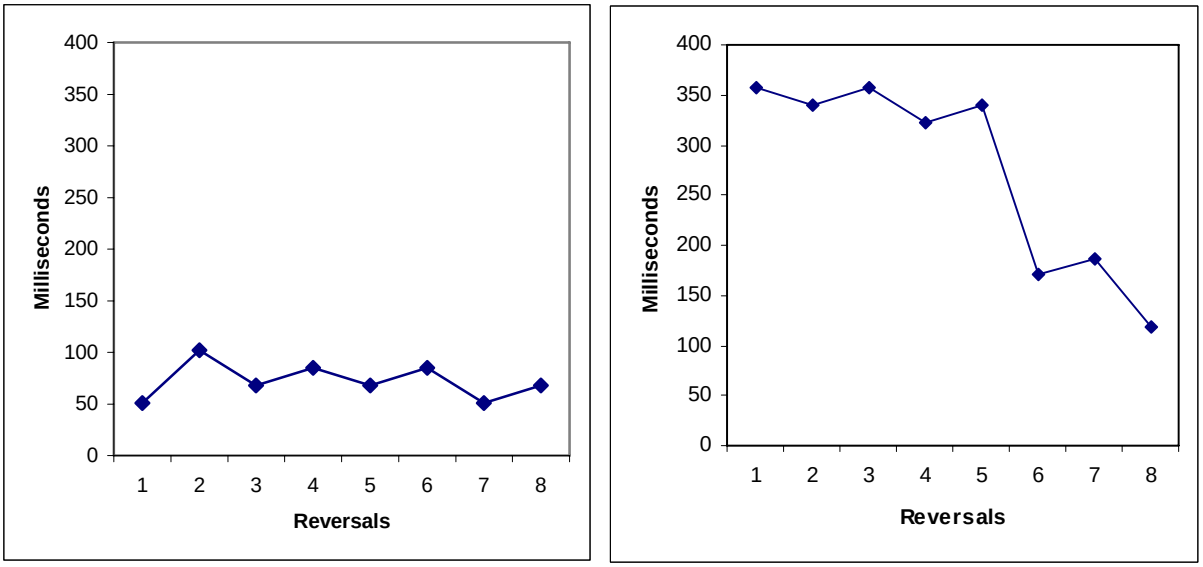

Figure E1. Standard reversal pattern (left) and variable reversal pattern (right)

One hundred and thirty nine people completed the IT task at Time 1 and for each person their eight reversals were examined. First, the standard deviation of the reversals was calculated,

[23] This method was used for the IT scores at all time points (i.e. Time 1,2 and 3).

for each person, to indicate how variable their reversals were. For example, Person X (shown in the left panel of Figure C1) registered his eight reversals at 51, 102, 68, 85, 68, 85, 51 and 68 ms (*M* = 72.25 ms, *SD* = 17.7). These reversals were all quite close together and the SD estimate is quite low as a results. In contrast, Person Y (shown in the right panel of Figure E1) registered his eight reversals at 357, 340, 357, 323, 340, 170, 187, and 119 ms (*M* = 274.13 ms, *SD* = 98.1). These reversals are much more variable and the corresponding SD estimate is high.

The second step was to calculate the average of the standard deviations across the whole group (average *SD* = 26.5 ms), to provide a representation of how variable the reversals were on average. Once, the average SD was available it was necessary to define a cut-off score to exclude those individuals with highly variable reversals. Consider the set of SD scores for the 137 participants as a new variable. The average of this new variable was 26.5ms and the SD was 14.1 ms. Any individual with a score greater than two *SD*s from the mean (i.e. 26.5 + 14.1 + 14.1 = 54.7) was considered to have a set of reversals that were excessively variable and hence were coded as missing data. For example, the IT score for Person Y, above, was excluded because the SD of his reversals was 98.1 and clearly more than 2 standard deviations from the mean. This method led to the exclusion of seven IT scores at Time 1, six IT scores at Time 2 and five IT scores at Time 3.

APPENDIX F. SHARED AND UNIQUE VARIANCE IN FLUID ABILITY

Inspection Time and visual acuity explained a significant amount of the variance in all three fluid ability tasks at Time 1. Therefore, the degree to which this variance was unique or shared was investigated using hierarchical regression analyses. The method used will be described using Raven's Standard Progressive Matrices (RSPM) as an example. However, the same method was used for all three fluid tasks.

Two hierarchical regressions were run with RSPM as the dependent variable. At Step 1, gender, education and age were entered as independent variables. In Model 1, visual acuity was entered at Step 2 and IT was entered last (see Table F1). Conversely, in Model 2, this order was reversed with IT entered at Step 2 and visual acuity entered last (see Table F2). These hierarchical regressions were used to estimate the shared and unique variance that IT and visual acuity explain in RSPM.

At Step 1, the independent variables explained 19.1% of the variance (i.e. R^2) in RSPM. At Step 3, the five independent variables explained 27% of the variance. Therefore, this implies that IT and visual acuity explain a total of 7.9% of the variance (27.0 – 19.1 = 7.9) after gender, education and age. From Model 1, we can see that IT explained 2.8% unique variance in RSPM (i.e. R^2 change from Step 3). Similarly, an examination of Model 2 shows that Visual Acuity explained 3.1% unique variance. Therefore, the shared variance that IT and visual acuity account for was 2% (7.9 – 2.8 – 3.1 = 2.0).

Table F1. Hierarchical Regression for Raven's Standard Progressive Matrices (Model 1)

Predictor	β	*t*	R^2	R^2 change
Step 1				
Gender	.313	3.62**		
Education	.198	2.28*		
Age	-.179	-2.06*	.191	
Step 2				
Visual acuity	-.237	-2.71**	.243	.052**
Step 3				
Inspection Time	-.179	-2.02*	.270	.028*

Note. * p < .05, ** p < .01

Table F2. Hierarchical Regression for Raven's Standard Progressive Matrices (Model 2)

Predictor	β	t	R^2	R^2 change
Step 1				
Gender	.313	3.62**		
Education	.198	2.28*		
Age	-.179	-2.06*	.191	
Step 2				
Inspection Time	-.228	-2.61*	.239	.048*
Step 3				
Visual acuity	-.191	-2.14*	.270	.031*

Note. * p < .05, ** p < .01

Table F3. Hierarchical Regression for Cattell Culture Fair Test (Model 1)

Predictor	β	t	R^2	R^2 change
Step 1				
Gender	.399	4.28**		
Education	.211	2.61*		
Age	-.191	-2.04*	.283	
Step 2				
Visual acuity	-.269	-2.85**	.347	.064**
Step 3				
Inspection Time	-.416	-4.89**	.495	.147**

Note. * p < .05, ** p < .01

Table F4. Hierarchical Regression for Cattell Culture Fair Test (Model 2)

Predictor	β	t	R^2	R^2 change
Step 1				
Gender	.399	4.28**		
Education	.211	2.61*		
Age	-.191	-2.04*	.283	
Step 2				
Inspection Time	-.457	-5.44**	.471	.188**
Step 3				
Visual acuity	-.166	-1.93	.495	.023

Note. * p < .05, ** p < .01

Table F5. Hierarchical Regression for Concept Formation (Model 1)

Predictor	β	t	R^2	R^2 change
Step 1				
Gender	.230	2.57*		
Education	.317	3.12**		
Age	-.275	-2.71**	.228	
Step 2				
Visual acuity	-.257	-2.49*	.288	.060*
Step 3				
Inspection Time	-.186	-1.76	.317	.029

Note. * p < .05, ** p < .01

Table F6: Hierarchical Regression for Concept Formation (Model 2)

Predictor	β	t	R^2	R^2 change
Step 1				
Gender	.230	2.57*		
Education	.317	3.12**		
Age	-.275	-2.71**	.228	
Step 2				
Inspection Time	-.241	-2.33*	.281	.053*
Step 3				
Visual acuity	-.208	-1.97	.317	.036

Note. * p < .05, ** p < .01

APPENDIX G. PREDICTIVE VALIDITY OF BIOMARKERS FOR COGNITIVE TASKS

Table G1. Predictors of Raven's Standard Progressive Matrices at Time 3

	Initial value				6-month change				18-month change			
Predictor variable	n	R^2 change	β-value	p-value	n	R^2 change	β -value	p-value	n	R^2 change	β -value	p-value
Age	125	.024	-.156	.078								
Inspection Time	111	.065**	-.260**	.006	102	.098**	-.312**	.001	103	.055*	-.236*	.016
Grip Strength	125	.018	.218	.124	121	.002	-.046	.619	123	.026	.164	.070
Systolic BP	113	.000	.000	.997	106	.003	-.054	.576	108	.008	-.091	.339
Diastolic BP	113	.007	.086	.353	106	.024	-.157	.103	108	.000	-.021	.830
Weight	125	.022	.159	.092	122	.001	-.024	.794	125	.002	.042	.635
Height	125	.004	.093	.485	122	.010	.099	.269	125	.000	.014	.878
Visual Acuity	125	.005	-.070	.435	122	.000	.003	.969	124	.004	-.059	.506
Digit Symbol	124	.244**	.494**	.000	121	.003	.051	.575	124	.016	.126	.154
Visual Matching	122	.175**	.418**	.000	119	.062**	249**	.006	122	.039*	.197*	.027
Pattern Comparison	123	.142**	.377**	.000	119	.052*	.227*	.012	122	.084**	.290**	.001

Note: Gender effects were removed before the examination of each biomarker. BP = Blood Pressure.
* $p < .05$, ** $p < .01$

Table G2. Predictors of the Cattell Culture Fair Test at Time 3

	Initial value				6-month change				18-month change			
Predictor variable	n	R^2 change	β -value	p-value	n	R^2 change	β -value	p-value	n	R^2 change	β -value	p-value
Age	122	.079**	-.282**	.002								
Inspection Time	109	.087**	-.307**	.001	101	.079**	-.282**	.004	102	.086**	-.296**	.003
Grip Strength	122	.048*	.351*	.014	118	.011	-.106	.262	121	.003	.055	.552
Systolic BP	110	.004	.043	.658	103	.000	-.020	.839	105	.001	-.026	.795
Diastolic BP	110	.021	.136	.155	103	.001	-.031	.754	105	.003	.051	.603
Weight	122	.035*	.200*	.039	119	.000	.010	.911	122	.002	.044	.627
Height	122	.009	.143	.291	119	.042*	.206*	.024	122	.018	.136	.136
Visual Acuity	122	.001	-.036	.693	119	.006	-.076	.414	122	.002	-.050	.584
Digit Symbol	121	.306**	.553**	.000	118	.006	.075	.417	121	.011	.105	.252
Visual Matching	119	.202**	.449**	.000	116	.074**	.272**	.003	119	.053*	.231*	.011
Pattern Comparison	120	.332**	.576**	.000	116	.061**	.248**	.007	119	.022	.147	.109

Note: Gender effects were removed before the examination of each biomarker. BP = Blood Pressure.
* $p < .05$, ** $p < .01$

Table G3. Predictors of the Concept Formation at Time 3

	Initial value				6-month change				18-month change			
Predictor variable	n	χ^2 step	Wald statistic	p-value	n	χ^2 step	Wald statistic	p-value	n	χ^2 step	Wald statistic	p-value
Age	122	16.15**	-3.72**	.000								
Inspection Time	110	4.95*	-2.11*	.035	101	5.66*	-2.38*	.023	102	4.63*	-2.08*	.038
Grip Strength	122	0.49	0.70	.484	119	1.23	1.09	.277	121	2.01	1.39	.166
Systolic BP	111	1.13	-1.06	.291	104	0.08	-0.28	.776	107	0.40	-0.64	.532
Diastolic BP	111	0.49	-0.69	.487	104	0.31	0.55	.580	107	0.20	-0.43	.656
Weight	122	3.28	1.75	.079	120	0.18	0.43	.668	122	0.03	-0.37	.865
Height	122	0.02	0.14	.902	120	0.13	-0.36	.715	122	0.97	-0.98	.327
Visual Acuity	122	0.95	-0.98	.328	120	2.04	1.37	.172	122	4.13*	1.89	.058
Digit Symbol	121	12.40**	3.28**	.001	119	1.49	1.21	.226	121	0.13	0.36	.718
Visual Matching	119	12.47**	3.30**	.001	117	0.73	0.85	.396	119	1.30	1.14	.256
Pattern Comparison	120	17.12**	3.71**	.000	117	0.06	-0.25	.802	118	0.90	-0.94	.344

Note: Concept Formation (CF) had a bi-modal distribution and was unsuitable for linear regression. Therefore, the CF measure was transformed to a factor with two groups to represent the low and high distributions. The *low* group (n = 42) had scores between 0 and 18 on the original measure and the *high* group (n = 80) had scores between 19 and 35. Logistic regression analyses were used to predict whether individuals were members of the *low* or *high* group for CF. Gender effects were removed before the examination of each biomarker. BP = Blood Pressure.

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

Table G4. Predictors of Information at Time 3

	Initial value				6-month change				18-month change			
Predictor variable	n	R^2 change	β-value	p-value	n	R^2 change	β -value	p-value	n	R^2 change	β -value	p-value
Age	123	.000	-.018	.836								
Inspection Time	110	.001	-.037	.688	100	.001	-.035	.703	102	.000	-.008	.934
Grip Strength	123	.044*	.337*	.014	120	.007	-.084	.351	122	.000	-.010	.908
Systolic BP	112	.001	.027	.767	105	.008	-.091	.323	107	.001	.027	.770
Diastolic BP	112	.006	.079	.385	105	.002	-.042	.654	107	.028	.169	.068
Weight	123	.001	.030	.746	121	.001	.023	.792	123	.003	.052	.547
Height	123	.009	.139	.281	121	.047*	.216*	.011	123	.000	.019	.829
Visual Acuity	123	.011	-.105	.226	121	.004	-.061	.482	123	.000	.000	.999
Digit Symbol	122	.034*	.185*	.032	120	.000	.001	.988	122	.000	-.020	.817
Visual Matching	120	.018	.134	.124	118	.007	.082	.347	120	.011	.105	.233
Pattern Comparison	121	.013	.113	.194	118	.018	.135	.123	119	.043*	.209*	.016

Note: Gender effects were removed before the examination of each biomarker. BP = Blood Pressure.
* $p < .05$, ** $p < .01$

Table G5. Predictors of Spot-the-Word at Time 3

	Initial value				6-month change				18-month change			
Predictor variable	n	R^2 change	β-value	p-value	n	R^2 change	β -value	p-value	n	R^2 change	β -value	p-value
Age	127	.001	-.027	.761								
Inspection Time	112	.006	-.080	.410	102	.000	.021	.833	103	.003	-.053	.597
Grip Strength	127	.000	.013	.930	123	.003	.059	.529	125	.023	.156	.089
Systolic BP	115	.003	.056	.557	108	.003	.059	.545	110	.014	.117	.226
Diastolic BP	115	.007	.086	.362	108	.002	.048	.622	110	.008	.087	.370
Weight	127	.003	-.055	.566	124	.001	-.028	.755	127	.001	-.030	.738
Height	127	.001	.046	.735	124	.004	.063	.491	127	.000	-.015	.866
Visual Acuity	127	.000	-.022	.808	124	.003	.055	.542	126	.007	-.083	.360
Digit Symbol	126	.036*	.190*	.033	123	.004	.066	.473	126	.002	.049	.589
Visual Matching	124	.114**	.338**	.000	121	.009	-.096	.296	124	.001	.032	.728
Pattern Comparison	125	.052*	.227*	.011	121	.005	.073	.430	123	.027	.163	.073

Note: Gender effects were removed before the examination of each biomarker. BP = Blood Pressure.
* $p < .05$, ** $p < .01$

Table G6. Predictors of Similarities at Time 3

	Initial value				6-month change				18-month change			
Predictor variable	n	R^2 change	β-value	p-value	n	R^2 change	β -value	p-value	n	R^2 change	β -value	p-value
Age	127	.001	-.031	.728								
Inspection Time	112	.004	.068	.478	102	.019	-.139	.157	103	.084**	-.292**	.003
Grip Strength	127	.008	.145	.311	123	.011	-.110	.236	125	.001	-.032	.731
Systolic BP	115	.001	-.027	.776	108	.004	.060	.532	110	.003	-.050	.600
Diastolic BP	115	.000	.007	.937	108	.003	.054	.576	110	.002	.044	.651
Weight	127	.007	.091	.339	124	.000	-.003	.971	127	.000	-.015	.865
Height	127	.002	.067	.621	124	.007	.083	.359	127	.000	.014	.874
Visual Acuity	127	.005	-.069	.445	124	.000	.019	.831	126	.005	.073	.419
Digit Symbol	126	.082**	.287**	.001	123	.003	-.052	.569	126	.000	.005	.959
Visual Matching	124	.035*	.188*	.035	121	.013	.113	.214	124	.023	.151	.094
Pattern Comparison	125	.069**	.263**	.003	121	.007	.084	.361	123	.001	.035	.702

Note: Gender effects were removed before the examination of each biomarker. BP = Blood Pressure.
* $p < .05$, ** $p < .01$

REFERENCES

Aartsen, M. J., Martin, M., & Zimprich, D. (2004). Gender Differences in Level and Change in Cognitive Functioning. *Gerontology, 50*, 35 - 38.

Australian Bureau of Statistics. (2001). *Census of Population and Housing (Adelaide)*: Retrieved March 2006, from http://www.abs.gov.au/.

Australian Institute of Heath and Welfare. (2004). *Australia's Health 2004*. Canberra: AIHW.

Aleman, A., Muller, M., de Haan, E. H. F., & van der Schouw, Y. T. (2005). Vascular risk factors and cognitive function in a sample of independently living men. *Neurobiology of Aging, 26*, 485 - 490.

Anstey, K. J. (1999). Sensorimotor Variables and Forced Expiratory Volume as Correlates of Speed, Accuracy, and Variability in Reaction Time Performance in Late Adulthood. *Aging, Neuropsychology, and Cognition, 6*(2), 84 - 95.

Anstey, K. J., Hofer, S. M., & Luszcz, M. A. (2003). A Latent Growth Curve Analysis of Late-Life Sensory and Cognitive Function over 8 Years: Evidence for Specific and Common Factors Underlying Change. *Psychology and Aging, 18*(4), 714 - 726.

Anstey, K. J., Lord, S. R., & Smith, G. A. (1996). Measuring Human Functional Age: A Review of Empirical Findings. *Experimental Aging Research, 22*, 245 - 266.

Anstey, K. J., Lord, S. R., & Williams, P. (1997). Strength in the Lower Limbs, Visual Contrast Sensitivity, and Simple Reaction Time Predict Cognition in Older Women. *Psychology and Aging, 12*(1), 137 - 144.

Anstey, K. J., Luszcz, M. A., Giles, L. C., & Andrews, G. R. (2001). Demographic, Health, Cognitive, and Sensory Variables as Predictors of Mortality in Very Old Adults. *Psychology and Aging, 16*(1), 3 - 11.

Anstey, K. J., Luszcz, M. A., & Sanchez, L. (2001). A Re-evaluation of the Common Factor Theory of Shared Variance Among Age, Sensory Function, and Cognitive Function in Older Adults. *Journal of Gerontology: PSYCHOLOGICAL SCIENCES, 56B*(1), P3 - P11.

Anstey, K. J., & Smith, G. A. (1999). Interrelationships Among Biological Markers of Aging, Health, Activity, Acculturation, and Cognitive Performance in Late Adulthood. *Psychology and Aging, 14*(4), 605 - 618.

Anstey, K. J., Smith, G. A., & Lord, S. (1997). Test-retest reliability of a battery of sensory, motor and physiological measures of aging. *Perceptual and Motor Skills, 84*, 831 - 834.

Anstey, K. J., Stankov, L., & Lord, S. (1993). Primary Aging, Secondary Aging, and Intelligence. *Psychology and Aging, 8*(4), 562 - 570.

Arking, R. (1991). *Biology of Aging: Observations and Principals*. New Jersey: Prentice Hall.

Babcock, R. L. (1994). Analysis of adult age differences in Raven's Advanced Matrices Test. *Psychology and Aging, 9*(2), 303 - 314.

Baddley, A., Emslie, H., & Nimmo-Smith, I. (1992). *Speed and Capacity of Language-Processing Test Manual*: UK Thames Valley Test, Bury St. Edmunds.

Baddley, A., Emslie, H., & Nimmo-Smith, I. (1993). The Spot-the-Word test: A robust estimate of verbal intelligence based on lexical decision. *British Journal of Clinical Psychology, 32*, 55 - 65.

Baker, G. T., & Sprott, R. L. (1988). Biomarkers of Aging. *Experimental Gerontology, 23*, 223 - 239.

Baltes, P. B., & Lindenberger, U. (1997). Emergence of a Powerful Connection Between Sensory and Cognitive Functions Across the Adult Life Span: A New Window to the Study of Cognitive Aging? *Psychology and Aging, 12*(1), 12 - 21.

Bashore, T. R. (1989). Age, Physical Fitness, and Mental Processing Speed. In M. P. Lawton (Ed.), *Annual Review of Gerontology and Geriatrics* (Vol. 9). New York: Springer Publishing Company.

Baxter, M. G., & Voytko, M. L. (1996). Spatial Orienting of Attention in Adult and Aged Rhesus Monkeys. *Behavioral neuroscience, 110*(5), 898 - 904.

Bell, B. (1972). Significance of Functional Age for Interdisciplinary and Longitudinal Research in Aging. *Aging and Human Development, 3*(2), 145 - 147.

Berg, L., Danziger, W. L., Storandt, M., Coben, L. A., Gado, M., Hughes, C. P., et al. (1984). Predictive features in mild senile dementia of the Alzheimer's type. *Neurology, 34*(5), 563 - 569.

Berr, C., Richard, M. J., Roussel, A. M., & Bonithon-Kopp, C. (1998). Systemic Oxidative Stress and Cognitive Performance in the Population-Based EVA Study. *Free Radical Biology & Medicine, 24*(7/8), 1202 - 1208.

Birren, J. E. (1965). Age Change in Speed of Behavior: Its Central Nature and Physiological Correlates. In A. T. Welford & J. E. Birren (Eds.), *Behavior, Aging, and the Nervous System: Biological Determinants of Speed of Behavior and Its Changes with Age* (pp. 191 - 216). Springfield, IL: Charles C. Thomas.

Birren, J. E. (1974). Translations in gerontology: From lab to life: Psychophysiology and speed of response. *American Psychologist, 29*, 808 - 814.

Birren, J. E., & Fisher, L. M. (1992). Aging and Slowing of Behavior: Consequences for Cognition and Survival. In T. B. Sonderegger (Ed.), *Nebraska Symposium on Motivation 1991* (pp. 1 - 37). Lincoln, NE: University of Nebraska Press.

Blumenthal, J. A., Madden, D. J., Pierce, T. W., Siegel, W. C., & Appelbaum, M. (1993). Hypertension Affects Neurobehavioral Functioning. *Psychosomatic Medicine, 55*, 44 - 50.

Borkan, G. A., & Norris, A. H. (1980a). Assessment of Biological Age Using A Profile of Physical Parameters. *Journal of Gerontology, 35*(2), 177 - 184.

Borkan, G. A., & Norris, A. H. (1980b). Biological Age in Adulthood: Comparisons of Active and Inactive U.S. Males. *Human Biology, 52*(4), 787 - 802.

Bosworth, H. B., & Schaie, K. W. (1999). Survival Effects in Cognitive Function, Cognitive Style, and Sociodemographic Variables in the Seattle Longitudinal Study. *Experimental Aging Research, 25*, 121 - 139.

Bosworth, H. B., & Siegler, I. C. (2002). Terminal Change in Cognitive Function: An Updated Review of Longitudinal Studies. *Experimental Aging Research, 28*, 299 - 315.

Botwinick, J., Storandt, M., & Berg, L. (1986). A Longitudinal, Behavioural Study of Senile Dementia of the Alzheimer Type. *Archives of Neurology, 43*(11), 1124 - 1127.

Brown, K. S., & Forbes, W. F. (1976). Concerning the Estimation of Biological Age. *Gerontology, 22*, 428 - 437.

Bryan, J. (2003). The Role of Nutritional Factors in Cognitive Ageing. In P.Sachdev (Ed.), *The Ageing Brain: The neurobiology and neuropsychiatry of ageing* (pp. 205 - 222). Lisse, The Netherlands: Swets & Zietlinger.

Bryan, J., Calvaresi, E., & Hughes, D. (2002). Short-Term Folate, Vitamin B-12 or Vitamin B-6 Supplementation Slightly Affects Memory Performance But Not Mood In Women of Various Ages. *Journal of Nutrition, 132*(6), 1345 -1356.

Bryan, J., & Luszcz, M. A. (1996). Speed of Information processing as a mediator between age and free recall performance. *Psychology and Aging, 11*, 3 - 9.

Bucks, R. S., Ashworth, D. L., Wilcock, G. K., & Siegfried, K. (1996). Assessment of Activities of Daily Living in Dementia: Development of the Bristol Activities of Daily Living Scale. *Age and Ageing, 25*(2), 113 - 120.

Bulpitt, C. J. (1995). Assessing Biological Age: Practicality? *Gerontology, 41*, 315 - 321.

Burbacher, T. M., & Grant, K. S. (2000). Methods for studying nonhuman primates in neurobehavioral toxicology and teratology. *Neurotoxicology and Teratology, 22*, 475 - 486.

Burns, N. R., Bryan, J., & Nettelbeck, T. (2006). Ginkgo biloba: no robust effect on cognitive abilities or mood in healthy young or older adults. *Human Psychopharmacology, 21*, 27 - 37.

Burns, N. R., & Nettelbeck, T. (2003). Inspection time in the structure of cognitive abilities: Where does IT fit? *Intelligence, 31*, 237 - 255.

Burns, N. R., Nettelbeck, T., & Cooper, C. J. (1999). Inspection Time Correlates with General Speed of Processing but not with Fluid Ability. *Intelligence, 27*(1), 37 - 44.

Burt, C. (1909). Experimental tests of general intelligence. *British Journal of Psychology, 3*, 94 - 177.

Burt, C. (1911). Experimental tests of higher mental processes and their relation to general intelligence. *Journal of Experimental Pedagogy, 1*, 93 - 112.

Calvaresi, E., & Bryan, J. (2001). B Vitamins, Cognition, and Aging: A Review. *Journal of Gerontology: PSYCHOLOGICAL SCIENCES, 56B*(6), P327 - P339.

Carroll, J. B. (1989). Factor analysis since Spearman: Where do we stand? What do we know? In R. Kanfer, P. L. Ackerman & R. Cudeck (Eds.), *Abilities, motivation, and methodology: The Minnesota symposium on learning and individual differences* (pp. 43 - 67). Hillsdale, NJ: Lawrence Erlbaum Associates.

Carroll, J. B. (1993). Cognitive Abilities: Methodology. In J. B. Carroll (Ed.), *Human Cognitive Abilities: A survey of factor analytic studies*. New York: Cambridge University Press.

Cattell, R. B., & Cattell, A., K. S. (1959). *The Cattell Culture Fair Test*. IL: IPAT.

Cerhan, J. R., Folsom, A. R., Mortimer, J. A., Shahar, E., Knopman, D. S., McGovern, P. G., et al. (1998). Correlates of Cognitive Function in Middle-Aged Adults. *Gerontology, 44*, 95 - 105.

Chodzko-Zajko, W. J., & Moore, K. A. (1994). Physical Fitness and Cognitive Functioning in Aging. *Exercise and Sport Sciences Reviews, 22*, 195 - 220.

Clark, J. W. (1960). The Aging Dimension: A Factorial Analysis of Individual Differences with Age on Psychological and Physiological Measurements. *Journal of Gerontology, 15*, 183 - 187.

Cobiac, L., & Syrette, J. (1995). *What is the Nutritional Status of Older Australians?* Paper presented at the Nutrition Society of Australia.

Cohen, J., & Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (2nd ed.). New Jersey: Lawrence Erlbaum Associates, Inc.

Cooney, T. M., Schaie, K. W., & Willis, S. L. (1988). The Relationship Between Prior Functioning on Cognitive and Personality Dimensions and Subject Attrition in Longitudinal Research. *Journal of Gerontology, 43*(1), P12 - P17.

Corsini, R. J. (1999). *The Dictionary of Psychology*. Philadelphia, PA: Taylor & Francis.

Costa, P. T., & McCrae, R. R. (1980). Functional age: A conceptual and empirical critique. In S. G. Haynes & M. Feinleib (Eds.), *Epidemiology of Aging. NIH Publications. No. 80 - 969* (pp. 23-46). Washington DC: U.S. Government Printing Office.

Costa, P. T., & McCrae, R. R. (1988). Measures and markers of biological aging: 'a great clamoring ... of fleeting significance'. *Archives of Gerontological Geriatrics, 7*, 211 - 214.

Crawford, J. R., Deary, I. J., Allan, K. M., & Gustafsson, J. E. (1998). Evaluating Competing Models of the Relationship Between Inspection Time and Psychometric Intelligence. *Intelligence, 26*(1), 27 - 42.

Damon, A. (1972). Predicting age from body measurements and observations. *International Journal of Aging and Human Development, 3*, 169 - 174.

Davis, J. W., Ross, P., Nevitt, M. C., & Wasnich, R. D. (1999). Risk factors for falls and for serious injuries on falling among older Japanese women in Hawaii. *Journal of the American Geriatrics Society, 47*(7), 792-798.

De Jager, C. A., Hogervorst, E., Combrinck, M., & Budge, M. M. (2003). Sensitivity and specificity of neuropsychological tests for mild cognitive impairment, vascular cognitive impairment and Alzheimer's Disease. *Psychological Medicine, 33*, 1039 - 1050.

de Lemos, M. M. (1995). *Standard Progressive Matrices: Australian manual*. Melbourne: ACER.

Dean, W., & Morgan, R. F. (1988). In defence of the concept of biological aging measurement - current status. *Archives of Gerontological Geriatrics, 7*, 191 - 210.

Deary, I. J. (1993). Inspection Time and WAIS-R IQ Subtypes: A Confirmatory Factor Analysis Study. *Intelligence, 17*(2), 223 - 236.

Deary, I. J. (2001). Wisdom from the ages. In I. J. Deary (Ed.), *Looking Down on Human Intelligence: From Psychometrics to the Brain* (pp. 223 - 261). Oxford: Oxford University Press.

Deary, I. J., Hunter, R., Langan, S. J., & Goodwin, G. M. (1991). Inspection Time, Psychometric Intelligence and Clinical Estimates of Cognitive Ability in Pre-senile Alzheimer's Disease and Korsakoff's Psychosis. *Brain, 114*(6), 2543 - 2554.

Devanand, D. P., Folz, M., Gorlyn, M., Moeller, J. R., & Stern, Y. (1997). Questionable Dementia: Clinical Course and Predictors of Outcome. *Journal of the American Geriatrics Society, 45*(3), 321 - 328.

Dik, M. G., Deeg, D. J. H., Visser, M., & Jonker, C. (2003). Early Life Physical Activity and Cognition at Old Age. *Journal of Clinical and Experimental Neuropsychology, 25*(5), 643 - 653.

Dirken, J. M. (1972). *Functional Age of Industrial Workers*. Groningen: Wolters-Noordhoff.

Dustman, R. E., Ruhling, R. O., Russell, E. M., Shearer, D. E., Bonekat, H. W., Shigeoka, J. W., et al. (1984). Aerobic exercise training and improved neuropsychological function of older individuals. *Neurobiological Aging, 5*, 35 - 42.

Elias, M. F., Robbins, M. A., Elias, P. K., & Streeten, D. H. P. (1998). A Longitudinal Study of Blood Pressure in Relation to Performance on the Wechsler Adult Intelligence Scale. *Health Psychology, 17*(6), 486 - 493.

Elias, P. K., Elias, M. F., D'Agostino, R. B., Silbershatz, H., & Wolf, P. A. (1999). Alcohol Consumption and Cognitive Performance in the Framingham Heart Study. *American Journal of Epidemiology, 150*(6), 580 - 589.

Evans, G., & Nettelbeck, T. (1993). Inspection time: a flash mask to reduce apparent movement effects. *Personality and Individual Differences, 15*(1), 91 - 94.

Finkel, D., Reynolds, C. A., McArdle, J. J., Gatz, M., & Pedersen, N. L. (2003). Latent Growth Curve Analyses of Accelerating Decline in Cognitive Abilities in Late Adulthood. *Developmental Psychology, 39*(3), 535 - 559.

Fleischmann, U. M. (1994). Cognition in humans and the borderline to dementia. *Life Sciences, 55*(25/26), 2051 - 2056.

Flynn, J. R. (1987). Massive IQ Gains in 14 Nations - What IQ tests really measure. *Psychological Bulletin, 101*(2), 171 - 191.

Flynn, J. R. (1999). Searching for Justice: The Discovery of IQ Gains Over Time. *American Psychologist, 54*(1), 5 - 20.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "MINI-MENTAL STATE" - A Practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12*(3), 189 - 198.

Fozard, J. L., Metter, E. J., & Brant, L. J. (1990). Next Steps in Describing Aging and Disease in Longitudinal Studies. *Journal of Gerontology: PSYCHOLOGICAL SCIENCES, 45*(4), P116 - 127.

Fry, J. (1985). What are the common diseases? In J. Fry (Ed.), *Common Diseases: Their Nature, Incidence, and Care* (pp. 16 - 26). Lancaster: MTP Press Limited.

Furukawa, T., Inoue, M., Fumihiko, K., Inada, H., Takasugi, S., Fukui, S., et al. (1975). Assessment of Biological Age by Multiple Regression Analysis. *Journal of Gerontology, 30*(4), 422 - 434.

Graham, D. P., Cully, J. A., Snow, A. L., Massman, P., & Doody, R. (2004). The Alzheimer's disease assessment scale - Cognitive subscale - Normative data for older adult controls. *Alzheimer Disease and Associated Disorders, 18*(4), 236 - 240.

Grudnik, J. L., & Kranzler, J. H. (2001). Meta-analysis of the relationship between intelligence and inspection time. *Intelligence, 29*, 523 - 535.

Guo, Z., Viitanen, M., & Winblad, B. (1997). Clinical correlates of low blood pressure in very old people: The importance of cognitive impairment. *Journal of the American Geriatrics Society, 45*(6), 701-705.

Gustafsson, J. E. (1984). A Unifying Model for the Structure of Intellectual Abilities. *Intelligence, 8*(3), 179 - 203.

Haan, M. N., Shemanski, L., Jagust, W. J., Manolio, T. A., & Kuller, L. (1999). The Role of *APOE* (e)4 in Modulating Effects of Other Risk Factors for Cognitive Decline in Elderly People. *Journal of American Medical Association, 282*(1), 40 - 46.

Hassing, L. B., Johansson, B., Berg, S., Nilsson, S. E., Pedersen, N. L., Hofer, S. M., et al. (2002). Terminal decline and markers of cerebro-and cardiovascular disease: Findings from a longitudinal study of the oldest old. *Journal of Gerontology, 57*(1), P268 - 276.

Heikkinen, E., Kiiskinen, A., Käyhty, B., Rimpelä, M., & Vuori, I. (1974). Assessment of Biological Age: Methodological Study in Two Finnish Populations. *Gerontologia, 20*, 33 - 43.

Hendrie, H. C., Gao, S., Hall, K. S., Hui, S. L., & Unverzagt, F. W. (1996). The relationship between alcohol consumption, cognitive performance, and daily functioning in an urban sample of older Black Americans. *Journal of the American Geriatrics Society, 44*, 1158 - 1164.

Heron, A. (1962). The Liverpool Age Project: Preliminary Communication on Population and Methodology. *Experientia, 18*, 472.

Heron, A., & Chown, S. (1967). *Age and Function*. London: Churchill.

Hill, R. D. (1989). Residual Effects of Cigarette Smoking of Cognitive Performance in Normal Aging. *Psychology and Aging, 4*(2), 251 - 254.

Hochschild, R. (1990). Can an Index of Aging be Constructed for Evaluating Treatments To Retard Aging Rates? A 2,462-Person Study. *Journal of Gerontology, 45*(6), B187 - B214.

Hollingsworth, J. V., Hashizume, A., & Jablon, S. (1965). Correlations between tests of aging in Hiroshima subjects. An attempt to define 'physiologic age'. *Yale Journal of Biological Medicine, 38*, 11 - 26.

Horn, J. L. (1988). Thinking about Human Abilities. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of Multivariate Experimental Psychology* (2 ed., pp. 645 - 685). New York: Plenum.

Horn, J. L. (1989). Remodelling Old Models of Intelligence. In B. B. Wolman (Ed.), *Handbook of Intelligence: Theories, Measurements, and Applications* (pp. 267 - 300). New York: Wiley.

Horn, J. L. (1990). Measurement of Intellectual Capabilities: A Review of Theory. In R. W. Woodcock & M. B. Johnson (Eds.), *Woodcock-Johnson Psycho-Educational Battery - Revised, Technical Manual.* (pp. 197 - 232). Chicago: Riverside.

Horn, J. L., & Cattell, R. B. (1967). Age differences in Fluid and Crystallized Intelligence. *Acta Psychologica, 26*, 107 - 129.

Horn, J. L., & Noll, J. (1994). A System for Understanding Cognitive Capabilities: A Theory and the Evidence on Which it is Based. In D. K. Detterman (Ed.), *Current Topics in Human Intelligence. Vol 4: Theories of Intelligence* (pp. 151 - 203). Norwood, NJ: Ablex.

Ingram, D. K. (1983). Towards the Behavioral Assessment of Biological Aging in the Laboratory Mouse: Concepts, Terminology, and Objectives. *Experimental Aging Research, 9*(4), 225 - 238.

Ingram, D. K. (1988). Key questions in developing biomarkers of aging. *Experimental Gerontology, 23*, 429 - 434.

Ingram, D. K. (1991). Is Aging Measurable? In F. C. Ludwig (Ed.), *Life span extension: Consequences and open questions* (pp. 18 - 42). New York: Springer Publishing Co.

Ingram, D. K., Nakamura, E., Smucny, D., Roth, G. S., & Lane, M. A. (2001). Strategy for identifying biomarkers of aging in long-lived species. *Experimental Gerontology, 36*(7), 1025 - 1034.

Judge, J. O., Schechtman, K., & Cress, E. (1996). The relationship between physical performance measures and independence in instrumental activities of daily living. *Journal of American Geriatric Society, 44*(11), 1332 - 1341.

Kail, R., & Salthouse, T. A. (1994). Processing speed as a mental capacity. *Acta Psychologica, 86*(2 - 3), 199 - 225.

Kalmijn, S., van Boxtel, M., P, J., Verschuren, M. W. M., Jolles, J., & Launer, L. J. (2002). Cigarette Smoking and Alcohol Consumption in Relation to Cognitive Performance in Middle Age. *American Journal of Epidemiology, 156*(10), 936 - 944.

Kaplan, E., Fein, D., Kramer, J., Delis, D., & Morris, R. (1999). *Wechsler Intelligence Scale for Children - III as a Process Instrument*. New York: The Psychological Corporation.

Kaufman, A. S., & Kaufman, N. L. (1993). *The Kaufman Adolescent and Adult Intelligence Test*. Circle Pines, MN: American Guidance Service.

Kersten, A., & Salthouse, T. A. (1993). Relations between time and accuracy in a continuous associative memory task, *Unpublished raw data*.

Kline, P. (1998). *The New Psychometrics: Science, psychology and measurement*. London: Routledge.

Kolb, B., & Whishaw, I. Q. (1996). *Fundamentals of Human Neuropsychology* (4th ed.). New York: W. H. Freeman.

Kranzler, J. H., & Jensen, A. R. (1989). Inspection Time and Intelligence: A Meta-Analysis. *Intelligence, 13*(4), 329 - 347.

Larrabee, G., Largen, J. W., & Levin, H. S. (1985). Sensitivity of age-decline resistant ("hold") WAIS subtests to Alzheimer's disease. *Journal of Clinical and Experimental Neuropsychology, 7*(5), 497 - 507.

Lawton, M. P., & Brody, E. M. (1988). Instrumental Activities of Daily Living (IADL) Scale - Self-Rated Version. *Psychopharmacology Bulletin, 24*, 789 - 791.

Leyfer, O. T., Ruberg, J. L., & Woodruff-Borden, J. (2006). Examination of the utility of the Beck Anxiety Inventory and its factors as a screener for anxiety disorders. *Anxiety Disorders, 20*, 444 - 458.

Lindeman, R. D., Romero, L. J., Koehler, K. M., Chi Liang, H., LaRue, A., Baumgartner, R. N., et al. (2000). Serum Vitamin B12, C and Folate Concentrations in the New Mexico Elder Health Survey: Correlations with Cognitive and Affective Functions. *Journal of the American College of Nutrition, 19*(1), 68 - 76.

Lindenberger, U., & Baltes, P. B. (1994). Sensory Functioning and Intelligence in Old Age: A Strong Connection. *Psychology and Aging, 9*(3), 339 - 355.

Lindenberger, U., & Baltes, P. B. (1997). Intellectual Functioning in Old and Very Old Age: Cross-Sectional Results From the Berlin Aging Study. *Psychology and Aging, 12*(3), 410 - 432.

Lindenberger, U., Mayr, U., & Kliegl, R. (1993). Speed and Intelligence in Old Age. *Psychology and Aging, 8*(2), 207 - 220.

Linn, R. L., & Slinde, J. A. (1977). The determination of the significance of change between pre and posttesting periods. *Review of Educational Research, 47*, 121 - 150.

Mackintosh, N. J., & Bennett, E. S. (2002). IT, IQ, and perceptual speed. *Personality and Individual Differences, 32*, 685 - 693.

Madden, D. J. (2001). Speed and Timing in Behavioral Processes. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the Psychology of Aging* (5 ed., pp. 288 - 312). San Diego: Academic Press.

Marsiske, M., Klumb, P., & Baltes, M. M. (1997). Everyday Activity Patterns and Sensory Functioning in Old Age. *Psychology and Aging, 12*(3), 444 - 457.

McClearn, G. E. (1997). Biomarkers of Age and Aging. *Experimental Gerontology, 32*(1/2), 87 - 94.

McFarland, R. A. (1956). Functional Efficiency, Skills and Employment. In J. E. Andersons (Ed.), *Psychological Aspects of Aging* (pp. 227 - 235). Washington: American Psychological Association.

McFarland, R. A. (1973). The need for functional age measurement in industrial psychology. *Industrial Gerontology, 19*, 1 - 19.

McFarland, R. A., & Philbrook, F. R. (1958). Job placement and adjustment for older workers: Utilization and protection of skills and physical abilities. *Geriatrics, 13*, 802 - 807.

McGrew, K. S. (1997). Analysis of the Major Intelligence Batteries According to a Proposed Comprehensive Gf-Gc Framework. In D. P. Flanagan, J. L. Genshaft & P. L. Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Tests and Issues.* (pp. 151 - 179). New York: Guilford Press.

Miller, R. E., Shapiro, A. P., King, H. E., Ginchereau, E. H., & Hosutt, J. A. (1984). Effect of Antihypertensive Treatment on the Behavioral Consequences of Elevated Blood Pressure. *Hypertension, 6*(2), 202 - 208.

Mohs, R. C., Rosen, W. G., & Davis, K. L. (1983). The Alzheimer's Disease Assessment Scale - An instrument for assessing treatment efficacy. *Psychopharmacology Bulletin, 19*(3), 448 - 450.

Morse, C. K. (1993). Does Variability Increase with Age? An Archival Study of Cognitive Measures. *Psychology and Aging, 8*(2), 156 - 164.

Mortensen, E. L., & Kleven, M. (1993). A WAIS Longitudinal Study of Cognitive Development During the Life Span From Ages 50 to 70. *Developmental Neuropsychology, 9*(2), 115 - 130.

Murray, I. M. (1951). Assessment of Physiological Age by a Combination of Several Criteria - Vision, Hearing, Blood Pressure and Muscle Force. *Journal of Gerontology, 6*, 120 - 126.

Nettelbeck, T. (1987). Inspection Time and Intelligence. In P. A. Vernon (Ed.), *Speed of Information Processing and Intelligence* (pp. 295 - 346). New Jersey: Ablex.

Nettelbeck, T., Edwards, C., & Vreugdenhil, A. (1986). Inspection Time and IQ: Evidence for a mental speed - ability association. *Personality and Individual Differences, 7*(5), 633 - 641.

Nettelbeck, T., & Lally, M. (1976). Inspection Time and Measured Intelligence. *British Journal of Psychology, 67*(1), 17 - 22.

Nettelbeck, T., & Rabbitt, P. M. A. (1992). Aging, Cognitive Performance, and Mental Speed. *Intelligence, 16*(2), 189 - 205.

Nettelbeck, T., Rabbitt, P. M. A., Wilson, C., & Batt, R. (1996). Uncoupling learning from initial recall: The relationship between speed and memory deficits in old age. *British Journal of Psychology, 87*(4), 593 - 607.

Nettelbeck, T., & Wilson, C. (1985). A Cross-Sequential Analysis of Developmental Differences in Speed of Visual Information Processing. *Journal of Experimental Child Psychology, 40*, 1 - 22.

Nettelbeck, T., & Wilson, C. (2004). The Flynn effect: Smarter not faster. *Intelligence, 32*(1), 85 - 93.

O'Connor, E. F. (1972). Extending classical test theory to the measurement of change. *Review of Educational Research, 42*(1), 73 - 97.

O'Connor, T. A., & Burns, N. (2003). Inspection time and general speed of processing. *Personality and Individual Differences, 35*(3), 713 - 724.

Osmon, D. C., & Jackson, R. (2002). Inspection time and IQ: Fluid or perceptual aspects of intelligence? *Intelligence, 30*, 119 - 127.

Pate, D. S., Margolin, D. I., Friedrich, F. J., & Bentley, E. E. (1994). Decision-making and attentional processes in ageing and in dementia of the Alzheimer's type. *Cognitive Neuropsychology, 11*(3), 321 - 339.

Rabbitt, P., Diggle, P., Smith, D., Holland, F., & McInnes, L. (2001). Identifying and separating the effects of practice and of cognitive ageing during a large longitudinal study of elderly community residents. *Neuropsychologica, 39*(5), 532 - 543.

Reff, M. E., & Schneider, E. L. (1982). *Biological Markers of Aging*. Washington, DC: Government Printing Office.

Regelson, W. (1983). Biomarkers in Aging. In W. Regelson & F. Marott Sinex (Eds.), *Intervention in the Aging Process - Part A: Quantitation, Epidemiology, and Clinical Research* (pp. 3 - 98). New York: Alan R. Liss.

Richards, M., Hardy, R., & Wadsworth, M. E. J. (2005). Alcohol Consumption and Midlife Cognitive Change in the British 1946 Birth Cohort Study. *Alcohol and Alcoholism, 40*(2), 112 - 117.

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

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A Growth Curve Approach to the Measurement of Change. *Psychological Bulletin, 92*(3), 726 - 748.

Rogosa, D., & Willett, J. B. (1983). Demonstrating the Reliability of the Difference Score in the Measurement of Change. *Journal of Educational Measurement, 20*(4), 335 - 343.

Rosen, W. G., Mohs, R. C., & Davis, K. L. (1984). A New Rating Scale for Alzheimer's Disease. *American Journal of Psychiatry, 114*(11), 1356 - 1364.

Rudinger, G., & Rietz, C. (2001). Structural Equation Modelling in Longitudinal Research of Aging. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the Psychology of Aging* (5 ed., pp. 29 - 52). San Diego: Academic Press.

Russell, R. M. (2001). Factors in Aging that Effect the Bioavailability of Nutrients. *The Journal of Nutrition, 131*, 1359 - 1361.

Saito, H., Yamazaki, H., Matsuoka, H., Matsumoto, K., Numachi, Y., Yoshida, S., et al. (2001). Visual event-related potential in mild dementia of the Alzheimer's type. *Psychiatry and Clinical Neurosciences, 55*(4), 365 - 371.

Salamon, M. J., & Conte, V. A. (1988). *Manual for the Life Satisfaction Scale (LSS): Formerly the Life Satisfaction in the Elderly Scale (LSES).* New York: Adult Development Center.

Salthouse, T. A. (1982). Psychomotor Indices of Physiological Age. In M. E. Reff & E. L. Schneider (Eds.), *Biological Markers of Aging*. Washington, DC: Government Printing Office.

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

Salthouse, T. A. (1991). Mediation of adult age differences in cognition by reductions in working memory and speed of processing. *Psychological Science, 2*(3), 179 - 183.

Salthouse, T. A. (1993). Speed mediation of adult age differences in cognition. *Developmental Psychology, 29*(4), 722 - 738.

Salthouse, T. A. (1996). The Processing-Speed Theory of Adult Age Differences in Cognition. *Psychological Review, 103*(3), 403 - 428.

Salthouse, T. A. (2000a). Aging and measures of processing speed. *Biological Psychology, 54*, 35 - 54.

Salthouse, T. A. (2000b). Pressing issues in cognitive aging. In D. C. Park & N. Schwarz (Eds.), *Cognitive aging: A primer* (pp. 43 - 54). Philadelphia, PA: Psychology Press.

Salthouse, T. A., & Babcock, R. L. (1991). Decomposing Adult Age Differences in Working Memory. *Developmental Psychology, 27*(5), 763 - 776.

Salthouse, T. A., Hambrick, D. Z., & McGuthry, K. E. (1998). Shared Age-Related Influence on Cognitive and Noncognitive Variables. *Psychology and Aging, 13*(3), 486 - 500.

Schaie, K. W. (1989). Perceptual Speed in Adulthood: Cross-Sectional and Longitudinal Studies. *Psychology and Aging, 4*(4), 443 - 453.

Schaie, K. W. (1994). The Course of Adult Intellectual Development. *American Psychologist, 49*(4), 304 - 313.

Schinka, J. A., Belander, H., Mortimer, J. A., & Borenstein Graves, A. (2003). Effects of the use on alcohol and cigarettes on cognition in elderly African American adults. *Journal of the International Neuropsychological Society, 9*(5), 690 - 697.

Singer, T., Verhaeghen, P., Ghisletta, P., Lindenberger, U., & Baltes, P. B. (2003). The Fate of Cognition in Very Old Age: Six-Year Longitudinal Findings in the Berlin Aging Study (BASE). *Psychology and Aging, 18*(2), 318 - 331.

Spearman, S. (1904). "General Intelligence", objectively determined and measured. *American Journal of Psychology, 15*, 210 - 293.

Spearman, S. (1927). Goodness and Speed of Response. In S. Spearman (Ed.), *The Abilities of Man: Their nature and measurement* (pp. 243 - 258). London: Macmillan.

Spirduso, W. W. (1975). Reaction and movement time as a function of age and physical activity level. *Journal of Gerontology, 30*, 435 - 440.

Spirduso, W. W. (1980). Physical fitness, aging, and psychomotor speed: a review. *Journal of Gerontology, 35*, 850 - 865.

Spirduso, W. W., & MacRae, P. G. (1990). Motor performance and age. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the psychology of aging* (3 ed., pp. 183 - 200). San Diego, CA: Academic Press.

Stip, E., Caron, J., Renaud, S., Pampoulova, T., & Lecomte, Y. (2003). Exploring Cognitive Complaints in Schizophrenia: The Subjective Scale to Investigate Cognition in Schizophrenia. *Comprehensive Psychiatry, 44*(4), 331 - 340.

Stough, C., Thompson, J. C., Bates, T. C., & Nathan, P. J. (2001). Examining neurochemical determinants of inspection time: Development of a biological model. *Intelligence, 29*, 511 - 522.

Swan, G. E., Carmelli, D., & Larue, A. (1998). Systolic Blood Pressure Tracking over 25 to 30 Years and Cognitive Performance in Older Adults. *Stroke, 29*(11), 2334 - 2340.

Taffe, M. A., Davis, S. A., Gutierrez, T., & Gold, L. H. (2002). Ketamine impairs multiple cognitive domains in rhesus monkeys. *Drug and Alcohol Dependence, 68*(2), 175 - 187.

Taffe, M. A., Weed, M. R., & Gold, L. H. (1999). Scopolamine alters rhesus monkey performance on a novel neuropsychological test battery. *Cognitive Brain Research, 8*(3), 203 - 212.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *Stanford-Binet Intelligence Scale: Fourth Edition.* Chicago: Riverside.

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago: University of Chicago Press.

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

Tully, C. L., & Snowdon, D. A. (1995). Weight change and physical function in older women: Findings from the Nun Study. *Journal of the American Geriatrics Society, 43*(12), 1394-1397.

Undheim, J. O., & Gustafsson, J. E. (1987). The Hierarchical Organization of Cognitive Abilities: Restoring General Intelligence Through the Use of Linear Structural Relations (LISREL). *Multivariate Behavioral Research, 22*(2), 149 - 171.

Van der Wal, E. A., & Sandman, C. A. (1992). Evidence for terminal decline in the event-related potential of the brain. *Electroencephalography and Clinical Neurophysiology, 83*(3), 211 - 216.

Verhaeghen, P., Borchelt, M., & Smith, J. (2003). Relation Between Cardiovascular and Metabolic Disease and Cognition in Very Old Age: Cross-Sectional and Longitudinal Findings From the Berlin Aging Study. *Health Psychology, 22*(6), 559 - 569.

Verhaeghen, P., & Salthouse, T. A. (1997). Meta-Analyses of Age-Cognition Relations in Adulthood: Estimates of Linear and Nonlinear Age Effects and Structural Models. *Psychological Bulletin, 122*(3), 231 - 249.

Vickers, D., Nettelbeck, T., & Willson, R. J. (1972). Perceptual indices of performance: the measurement of 'inspection time' and 'noise' in the visual system. *Perception, 1*, 263 - 295.

Vickers, D., & Smith, P. L. (1986). The Rational for the Inspection Time Index. *Personality and Individual Differences, 7*(5), 609 - 623.

Voytko, M. L., & Tinkler, G. P. (2004). Cognitive Function and Its Neural Mechanisms in Nonhuman Primate Models of Aging, Alzheimer's Disease, and Menopause. *Frontiers in Bioscience, 9*, 1899 - 1914.

Waaler, H. T. (1980). Specificity and sensitivity of blood pressure measurements. *Journal of Epidemiology and Community Health, 34*, 53 - 58.

Waldstein, S. R. (2000). Health Effects on Cognitive Aging. In P. C. Stern & L. L. Carstensen (Eds.), *The Aging Mind: Opportunities in Cognitive Research* (pp. 189 - 217). Washington, DC: National Academy Press.

Waldstein, S. R., Manuck, S. B., Ryan, C. M., & Muldoon, M. F. (1991). Neuropsychological Correlates of Hypertension: Review and Methodological Considerations. *Psychological Bulletin, 110*(3), 451 - 468.

Webster, I. W., & Logie, A. R. (1976). A Relationship Between Functional Age and Health Status in Female Subjects. *Journal of Gerontology, 31*(5), 546 - 550.

Wechsler, D. (1981). WAIS-R Manual: Wechsler Adult Intelligence Scale - Revised. New York: The Psychological Corporation.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale - III*. San Antonio, TX: Psychological Corporation.

Weed, M. R., Taffe, M. A., Polis, I., Roberts, A. C., Robbins, T. W., Koob, G. F., et al. (1999). Performance norms for a rhesus monkey neuropsychological testing battery: acquisition and long-term performance. *Cognitive Brain Research, 8*(3), 185 - 201.

Welford, A. T. (1977). Motor Performance. In J. E. Birren & K. W. Schaie (Eds.), *Handbook of the Psychology of Aging*. New York: Van Nostrand Reinhold Co.

Welsh, M. B. (2003). *Of mice and men: the structure and bases of murine cognitive abilities.* Unpublished doctoral dissertation, University of Adelaide, Australia.

Wetherill, G. B., & Levitt, H. (1965). Sequential estimation of points on a psychometric function. *The British Journal of Mathematical and Statistical Psychology, 18*(1), 1 - 10.

Weyer, G., Erzigkeit, H., Kanowski, S., Ihl, R., & Hadler, D. (1997). Alzheimer's Disease Assessment Scale: Reliability and Validity in a Multicenter Clinical Trial. *International Psychogeriatrics, 9*(2), 123 - 138.

Whalley, L. J., Fox, H. C., Deary, I. J., & Starr, J. M. (2005). Childhood IQ, smoking, and cognitive change from age 11 to 64 years. *Addictive Behaviors, 30*(1), 77 - 88.

Wicherts, J. M., Conor, V. D., Hessen, D. J., Oosterveld, P., Van Baal, G. C. M., Boomsma, D. I., et al. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence, 32*, 509 - 537.

Woodcock, R. W. (1990). Theoretical Foundations of the WJ-R Measures of Cognitive Ability. *Journal of Psychoeducational Assessment, 8*, 231 - 258.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Psycho-Educational Battery - Revised*. Allen, TX: DLM.