# COGNITIVE DYSFUNCTION AFTER CARDIAC SURGERY

Haydn Richard Till

B. Soc. Sci (Psyc) Hons

Submitted in partial fulfilment of the requirements for the

Doctor of Philosophy with the degree of Master of Clinical Neuropsychology

Macquarie Centre for Cognitive Science

Macquarie University

Sydney, Australia

February – 2005

Statement of Individual Work

I, Haydn Richard Till, acknowledge that this dissertation, completed under the conjoint supervision of Professor Max Coltheart, Clinical Associate Professor E. Arthur Shores, Dr David Cairns, and Dr Jennifer Batchelor, results from my own work and that the authorship of the document herein is mine. I acknowledge that this body of work has not been submitted for a higher degree to any other university or institution.

Signature: ________________ Date 10 / 07 / 06

# Table of Contents

# Acknowledgments

Postgraduate study and research is certainly an adventure. Probably more than one expects when first undertaking it. Fortunately, many people have supported me and placed their trust and faith in my capabilities as a student, tutor, practitioner, and researcher. To this end, I owe an enormous debt of gratitude to countless people who through their patience, guidance and conduct fostered my desire to learn and to achieve.

I would never have reached the end of this marathon without the support, guidance, and experience of my primary supervisor Professor Max Coltheart. Thank you Max, you provided me with substantial latitude and in doing so demonstrated your faith in my abilities as a researcher. To my associate supervisors Dr David Cairns, Dr Jennifer Batchelor, and Associate Professor E. Arthur Shores, Your confidence in my abilities have carried me through many difficult and tiresome times. I thank you for your encouragement, assistance and friendship. You are remarkable people and I feel privileged for having the opportunity to meet and work with each of you. An enormous debt of gratitude also goes to my sister Nina-Gai Till, whose editorial assistance was always timely and encouraging.

Finally, to my beautiful partner Romina, my loving family, and very supportive network of friends. Thank you for being gracious enough to provide help when requested and solitude when required. You have all contributed substantially to this achievement.

## List of Tables

Appendix C



Appendix D

# List of Figures

Study Four

# Abstract

**Overall Aim:** The overall aim of the current research was to evaluate the methodological efficiency and effectiveness of current neuropsychological research in the cardiac surgery context.

**Background:** Conducting a reliable assessment of cognitive impairment after cardiac surgery is crucial, particularly when treatment selection to remediate deficits depends on the conclusion drawn from the assessment. Studies examining possible neuropsychological decline following cardiac surgery have produced a variety of outcomes. Possible reasons for this include methodological differences across studies and inadequate or inconsistent definitions of what constitutes a significant change in functioning. The 1995 consensus statement on the assessment of post cardiac surgery neurobehavioural deficits attempted to rectify the methodological differences by proposing guidelines for conducting research and a core test battery for inclusion in all research within the field (Murkin, Newman, Stump, & Blumenthal, 1995). The current research consists of a series of studies designed to examine post consensus research in the field and to further elucidate methodological problems in current research strategies.

**Method and Results:** The current effort commenced with a review of outcomes published in the literature since the 1995 consensus statement's release. That review sought to examine current research methodology in the field and to highlight how various methodological practices influence conclusions drawn from the research. The review concluded that a large range of assessment and decision methodologies currently exist in the literature and that, as a result, consistency in outcomes and

conclusions has not been obtained. The review was followed by a series of four studies designed to further demonstrate the methodological inadequacies in the field. Study One consisted of a replication, within cardiac surgery related publications, of the power and effect size research conducted by Bezeau and Greaves (2001). *Aim*: The aim of that analysis was to establish the overall effect size and power of cardiac surgery clinical neuropsychological research. *Analysis*: The analytical method described by Zakzanis (2001) was used for between group comparisons. To avoid potential negative impact by higher correlations between pre and postoperative scores, the method used to derive power and effect size for within group studies was that described by Dunlap, Cortina, Vaslow, and Burke (1996). *Results*: The current analyses indicate that the statistical power and population effect sizes are underreported in cardiac surgery research publications. Analyses also revealed that very small effect sizes exist for both independent groups and repeated measures designs, and that the power of research findings was on average poor. *Conclusion*: It was concluded from those results that neuropsychological research into the effects of cardiac surgery requires substantial improvement to ensure that conclusions being drawn from investigations are reliably robust.

Study Two was an analysis of the structure of four test batteries used in previously published research studies for which raw data was made accessible by the studies' authors. *Aim*: The aim of this study was to examine the structure of those test batteries in order to establish if the originally hypothesised design was correct. *Analysis*: Facet theory was used to analyse the intercorrelations between test scores across each individual test battery. *Results*: The results of the facet analyses indicate that none of the batteries consistently supported their proposed methodological structures. *Conclusion*: It is concluded from those results that test batteries must be carefully considered and validated on the specific cohort before being used in research or recommended for use in clinical contexts.

Study Three involved an analysis of the effect sizes for the core test battery proposed in the 1995 consensus statement. Suitable published studies that were included in study two were used as the data source for this analysis. *Aim*: The aim of this study was two fold. Firstly, it sought to identify cohort specific effect sizes for each core test. Secondly, it sought to provide estimates of effect sizes that could be used by future researchers to optimise their research strategies, to ensure suitably powerful results are achieved. *Analysis*: The analytical methods used were the same as those used in study one. *Results*: The analyses indicate that in recent studies using the core neuropsychological test battery tests, the population effect sizes detected ranged up to -.7, which equates to distribution overlaps of at least 57%. Additionally, it was found that the effects being detected differed across assessment intervals. *Conclusion:* In accordance with the aim of the study, the results identified population effect sizes currently in operation in the cardiac research field for the core test battery tests. The estimates established provide valuable information for researchers planning research endeavours in the field using the core neuropsychological test battery.

Study Four was a preliminary investigation of the structural integrity of the consensus statement core test battery on a small sample of cardiac surgery patients. *Aim*: The aim of this study was to examine the structure of the core test battery to clarify the relationship between tests and to identify whether the structure of the battery proposed by the consensus group is replicable. *Analysis*: Facet Theory was used to examine the relationships (intercorrelations) between tests in the core test battery. *Results*: The results of the facet analysis indicate that the relationships between tests does not conform to the structure proposed by the consensus group, and that the structure of the battery varied across assessment intervals. Additionally, when both assessment intervals were considered contemporaneously, only two general facets of cognitive functioning could be consistently observed. *Conclusion*: The conclusion drawn from those results is that the core test battery requires careful consideration and validation to ensure that the resulting conclusions are appropriate to what is actually being examined.

**Overall Conclusions:** Reviewing the relevant literature reveals that deleterious effects due to cardiac surgery are continuously present in the cognitive domains of attention, psychomotor speed, memory, visuoperception, executive functioning, and general cognitive functioning. However, current investigations demonstrated that research methodologies in the field are insufficient for the purposes for which they are being used. While that finding does not altogether exclude the previous conclusions drawn in the literature, it does indicate that substantial methodological improvement is required to ensure the robustness of the conclusions being drawn. In an attempt to assist in the improvement of research methods in the field, statistical power and population effect sizes in research on cardiac surgery samples were identified. By identifying those values, foundation knowledge has been provided to assist future researchers to enhance their research design process.

# CHAPTER ONE

## Introduction

Normal neurological functioning is a dynamic but fragile process that is easily affected by malfunctioning bodily systems. Disruption to the cardiovascular system, such as that caused by heart disease and its therapeutic interventions, is acknowledged to have a significant influence on neurological functioning and has been since the very first published reports in the field in the late 1960's (Smith, 1995; Sontaniemi, 1995). The interrelatedness of neurological and cardiovascular functioning is evidenced by the fact that complications anywhere within the cardiovascular system can reduce neurological capacity, and that reduced neurological capacity can impede cardiac functions (Caplan, 1999; Sontaniemi).

Neurological complications arising from cardiac surgery range from catastrophic stroke causing death to transitory changes in cognitive capacities (Borowicz, Goldsborough, Selnes, & McKhann, 1996). A vast array and combination of complications exist along that spectrum and have been found in clinical research addressing the aetiology of cognitive impairment deriving from cardiac disease interventions (Borowicz et al.). Despite the extensive research on the topic, no set cognitive presentation has been identified for post cardiac surgery cognitive dysfunction.

A number of possible explanations exist for why a consistent phenotype has not been found for post cardiac surgery cognitive decline. It may be that heterogeneity in the personal factors of subjects restricts the ability of research to identify consistent patterns of decline. An alternative explanation might be that operative processes are so heterogeneous between surgeons and institutions that consistent outcome relationships cannot be identified.

In addition to the logistical difficulty arising from coordinating so many different elements, methods for assessing and describing outcomes have lacked consistency across the length and breadth of research within the field. For the last half a century, neuropsychological assessments have been used to quantify cognitive outcomes from cardiac surgery (Borowicz et al., 1996). Methods for assessing cognitive dysfunction arising from cardiac surgery have been a major discussion point, at least since the publication of the first consensus statement on the topic (Blumenthal, Mahanna, Madden, White, Croughwell, & Newman, 1995; Borowicz et al.; Mc Daid, Lewis, McMurray, & Phillips, 1994; Murkin, Newman, Stump, & Blumenthal, 1995; Murkin, Stump, Blumenthal, & McKhann, 1997; Newman, 1995; Slade, Sanchez, Townes, & Aldea, 2001). The consensus group discussion that resulted in the statement addressed a range of methodological issues pertinent to research within this field, and in doing so arrived at 14 points of consensus and a core test battery consisting of four neuropsychological instruments (Rey Auditory Verbal Learning Test, Trail Making Test Part A and B, Grooved Pegboard) (Murkin, Newman et al.). An additional statement of consensus, further addressing four points from the first statement, was published in 1997 (Murkin, Stump et al.). While widely recognised as important points of clarification for methodological practices of research in the field (Slade et al.), the statements also demonstrated that research within the field has lacked the cohesiveness required to build a solid theoretical foundation for understanding post cardiac surgery cognitive dysfunction. Despite the directions of the consensus statements (Murkin, Newman et al.; Murkin, Stump et al.), and the recognition by their key author that they were not the definitive methodological end-state (Murkin, 1995), very little effort has ensued outside of evaluating specific statistical methods for discerning significant performance change.

Point five in the 1995 consensus statement dealt specifically with assessment methodology, and in doing so considerably reinforced the importance of selecting the correct instruments for the purpose. In particular, point five addressed the "appropriate" selection of tests referencing the need to attend to issues such as sensitivity, reliability, and validity, as well as the individual and range of cognitive functions being assessed by the tests (Murkin, Newman et al., 1995; Newman, 1995). That point was supported and expanded by Blumenthal et al. (1995), who suggested that neuropsychological test batteries used in this field were chosen with regard for their reliability and validity, the availability of normative data and alternate forms, duration, ability to assess relevant cognitive domains, and their previous use. However, no literature exists to extend the validity of commonly used neuropsychological instruments to the cardiac surgery context, and at least one group of authors (Mc Daid et al., 1994) has questioned the generalisability of test validity to the cardiac setting. Given that the generalisability of test validity across settings is far from accepted (Messick, 1995), the selection of specific tests for use in the cardiac surgery context becomes a fundamentally important methodological step in research. Selecting tests and batteries solely on the basis of prior use, rather than context specific validity information, is not a suitable approach as it is not empirically defensible.

Therefore, despite a plethora of research, a number of reviews, and methodological consensus statements, the question remains, *how adequate are our efforts?* In an attempt to answer that question, neuropsychological research in cardiac surgery cohorts was evaluated. Chapter Two, titled "The Matter of the Heart – Overview of Structure, Illness, and Intervention ", while not strictly part of the review provides background knowledge about the anatomy of the cardiovascular system, the diseases that can affect that system, and the therapeutic interventions used to improve cardiac functioning. Chapter Three, titled "The Heart of the Matter – Examining Cognitive Decline in Cardiac Surgery", commences by reviewing the current understanding of the mechanisms underlying the different degrees of central nervous system complications following cardiac surgery. The discussion is then narrowed to specifically review the cognitive outcomes associated with cardiac surgery related central nervous system insults. The discussion then reviews the research methodology (i.e., instruments, samples, and decision criteria) used to draw conclusions in the field. Chapter Four, titled "Research Methodology and Results", commences with the overall aim and the general design of the study. It then presents the four studies investigating the overall aim. Finally, Chapter Five, titled "Discussion" provides a response to the question, *how adequate are our efforts?* It integrates and interprets the methodological criticisms identified in the review chapter with the results of the four studies, taking into consideration limitations in the current research approach. It concludes with recommendations for future research efforts before drawing an overall conclusion.

# CHAPTER TWO

## The Matter of the Heart – Overview of Structure and Function, Illness, and Intervention

### Introduction

Though not essential to understanding the effects of heart surgery on cognitive functions, which is the core focus of this dissertation, a general description of the circulatory system and its component parts is warranted. Bearing in mind that this system does not operate in isolation, the following description aims solely to provide a basis for understanding the following section on heart diseases and therapeutic interventions. Many anatomy texts and computer programs exist, providing an almost endless insight into the circulatory system and its component parts; therefore, interested readers are directed toward such texts and programs for more in-depth descriptions and discussions.

### Anatomy

#### Heart

The heart is located approximately eight to nine centimetres from the midline of the sternum, between the fifth and eighth vertebrae and behind the fifth and sixth ribs. It is cone shaped, with the base facing upwards, backwards, and to the right. This positions the apex downwards, forwards, and to the left. In an adult human, it measures approximately 12 to 13 centimetres in breadth and six to six and one half centimetres in thickness. It weighs between 280 and 340 grams in males and 224 to 280 grams in females. In general, the heart increases in length and weight with aging (Gray, 1995; Malhotra, Edelman, & Lily, 2003; Snell, 2004; Totora & Grabowski, 2003).

The heart is divided into right and left halves by a partition called the septum. The two halves are further divided into the upper and lower cavities, called "atria" and "ventricles" respectively. The atria each have a main cavity and an additional smaller cavity called the "atrial appendage"[1]. The four chambers are represented on the outside surface of the heart by the auriculo-ventricular sulcus running horizontally between the atria and ventricles, and the interventricular sulcus vertically between the ventricles (Gray, 1995; Malhotra et al., 2003; Snell, 2004; Totora & Grabowski, 2003).

In general, the Myocardium or heart muscle thickness varies according to the function of the chamber it constitutes. Within the atria, the walls are thin due to the muscles of these chambers only having to pump blood into their neighbouring ventricles. However, as the ventricles are required to propel blood throughout the whole body, they have thicker walls to facilitate more pumping pressure. A distinction can also be drawn between the ventricles. The right ventricle only circulates blood a short distance to the lungs (pulmonary circulation), whereas the left ventricle circulates blood throughout the remainder of the body. By comparison, the system supplied by the right ventricle is shorter and requires lower blood pressure to accomplish circulation. As such, the walls of the right ventricle are thinner than those of the left ventricle (Gray, 1995; Totora & Grabowski, 2003).

---

[1] Some discordance appears in the anatomical literature about the atria appendage nomenclature. Many modern texts use the word "auricle" when referring to the appendages, whilst others use "atrium". Morphologically speaking, the word "auricle" is incorrect (Netter, 1969). Therefore, in the interests of presenting an accurate anatomical representation of the heart, the current review will use "atrium".

The Cardiac Nervous System and Cardiac Cycle

Nervous system regulation of the heart originates in the cardiovascular centre within the medulla oblongata. That centre receives input from higher brain centres and from sensory receptors. The higher brain centres that provide input include the limbic system, hypothalamus, and the cerebral cortex. The sensory receptors that provide input include proprioceptors that monitor the position of the limbs and skeletal muscles, chemoreceptors that monitor chemical changes in the blood, and baroreceptors that monitor stretching of the major arteries and veins caused by alternations in blood pressure. The cardiovascular centre directs the pumping action of the heart by regulating the frequency of action potentials along the sympathetic and parasympathetic branches of the autonomic nervous system. The sympathetic and parasympathetic innervations carry both afferent and efferent fibres, allowing for a response-feedback loop between the driving centres within the brain and the heart musculature and surrounding vasculature (Gray, 1995; Malhotra et al., 2003; Netter, 1969; Opie, 2001; Totora & Grabowski, 2003).

The cardiac cycle is the cycle through which the atria and ventricles alternately contract and relax to circulate blood through both the systemic and pulmonary circulatory systems. Generally, the cardiac cycle takes approximately 0.8 seconds to complete when the heart is at 75 beats per minute. Specifically the cycle consists of atrial systole, ventricular systole, and a relaxation period. Atrial systole lasts approximately 0.1 seconds and consists of the atria walls contracting, which in turn places pressure on the volume of blood within their chambers, forcing it through the atrial-ventricular valve openings into the ventricles. During atrial systole, the ventricles are in relaxation or diastole period. At the end off the ventricular diastole period, the ventricles each contain approximately 130 millilitres of blood.

Ventricular systole lasts approximately 0.3 seconds. It involves an increase in pressure within the ventricular cavities, which forces blood against the atrial-ventricular valves causing them to close. For a brief period of about 0.05 seconds, the period of isovolumetric contraction, both the semilunar and atrial-ventricular valves are closed, trapping a full volume of blood within the ventricles, leaving no avenue for the muscle fibres to contract and hence no displacement of the blood volume contained within them. This results in a significant increase in intraventricular pressure. When the left ventricle pressure exceeds the aortic pressure of 80 millimetres of mercury (mmHg), the semilunar valves open allowing approximately 70 millimetres of blood to be displaced into the aorta for circulation through the systemic circulatory system. Similarly, when pressure within the right ventricle exceeds the pulmonary trunk pressure of 20 mmHg, the semilunar valve opens, and 70 millimetres of blood is circulated into the pulmonary vein for circulation through the pulmonary circulatory system. The semilunar valves are open for a period of approximately 0.25 seconds, which is called ventricular ejection. The relaxation period lasts approximately 0.4 seconds during which both the atria and ventricles both relax. As the work rate of the heart increases this period shortens, something that does not occur during the atrial and ventricular systole durations. As the ventricles relax, the pressure inside the chambers falls and allows blood in the aorta and pulmonary trunk to backflow. That backflow forces the semilunar valves to close. After they close, there is a brief interval where blood volume within the ventricles does not change. This is called isovolumetric relaxation. At the end of that period, the ventricles are approximately three quarters full. As the ventricles continue to relax, their pressure falls below that of the atria. That in turn allows the atria-ventricular valves to open and the blood to move from the atria to the ventricles (Gray, 1995; Snell, 2004; Totora & Grabowski, 2003).

Circulation

The blood-vascular system comprises the heart, blood vessels, and blood. The heart is the central organ of the circulatory system, and by its contractions, it drives blood throughout the body along circulatory pathways called blood vessels. The left side of the heart contains purified blood for circulation around the body, and the right half contains impure blood returning from circulating around the body (Gray, 1995).

As a rule, arterial blood vessels (arteries) carry blood away from the heart, while ventricular blood vessels (veins) carry blood back to the heart. The exception to that rule is blood circulating between the heart and lungs in a circuit called Pulmonary circulation. Blood passing from the heart to the lungs, called venous (impure) blood, travels through the pulmonary arteries. Once re-arterialised in the capillaries of the lungs, the blood travels back to the heart along the pulmonary veins (Gray, 1995; Snell, 2004; Totora & Grabowski, 2003).

Arterialised blood from the heart circulates throughout the body in what is called the Systemic circulatory system. Throughout the body, blood passes through the arteries into minute vessels called arterioles. Arterioles in turn open into capillaries, a close-meshed network of microscopic vessels within the body's organs where haemoglobin can pass from the vessels into the cells to deposit the nutrients they carry. After re-joining the capillary network, the venous (impure) blood returns to the heart through the venous system. Blood circulating through the spleen, pancreas, stomach, small intestine, and the greater part of the large intestine does not return directly to the heart for dispensing to the lungs. Instead, it is transported to the liver, via the portal vein, for cleansing. After cleansing, it returns to the heart through a network consisting of the hepatic veins and the inferior vena cava (Gray, 1995; Snell, 2004; Totora & Grabowski, 2003).

The first part of the systemic circulatory system is the Aorta. The aorta originates from the left ventricle, ascends briefly (Ascending Aorta), and then curves back and to the left of the left lung (Aortic Arch). From there it descends within the thoracic cavity (Descending Aorta) on the left side of the vertebral column into the abdominal cavity. In doing so it passes through the aortic opening within the diaphragm, which is situated opposite the lower border of the forth lumbar vertebra (Gray, 1995; Snell, 2004; Totora & Grabowski, 2003).

The ascending aorta delivers two branch arteries called the Right Coronary Artery, and the Left Coronary Artery, which provide blood supply to the heart muscle. Three arteries branch from the aortic arch, the Innominate Artery, the Left Common Carotid Artery, and the Left Subclavian Artery (Gray, 1995; Snell, 2004; Totora & Grabowski, 2003). The innominate (brachio-cephalic) artery extends approximately four to five centimetres before dividing into the Right Common Carotid Artery, and the Right Subclavian Artery. The right and left common carotid arteries extend into the base of the neck, and then climb to a point opposite the fourth cervical vertebra where they each divide to form the External Common Carotid Arteries and Internal Carotid Arteries. The external carotid arteries supply blood to the external parts of the head and face via four groups of arteries, the Anterior (Superior, Thyroid, Lingual, Facial), Posterior (Occipital, Posterior Auricular), Ascending (Pharyngeal, Ascending), and the Terminal (Superficial, Temporal, Internal Maxillary). As these aspects of the circulatory system do not perforate the cerebral matter, they will not be discussed further. The internal carotid arteries circulate blood to four segments, Cervical, Petrous, Cavernous, and Cerebral, of which the cerebral is the most relevant as it is the process through which blood circulates to the anterior portions of the brain (Gray; Snell; Totora & Grabowski). This aspect of circulation will be addressed in detail in the following section on cerebral circulation. As mentioned, the subclavian artery on the right side arises from the innominate artery, while on the left it arises directly from the aortic arch. Along their processes, both subclavian arteries give rise to four branches: the Vertebral Arteries, the Internal Mammary Arteries, the Thyroid Axis Arteries, and the Superior Intercostal Arteries (Gray; Snell; Totora & Grabowski). Of those, the vertebral arteries are pertinent to the current topic and will be discussed in the following section on cerebral circulation.

Cerebral Circulation

Blood supply to the brain derives from the internal carotid and subclavian arteries, which in turn are essentially direct extensions from the aortic arch. The internal carotid artery circulates blood to the brain via its cerebral branch. The cerebral branch perforates the dura matter on the inside of the anterior clinoid process. It then passes between the second and third cranial nerves to the anterior perforated space on the inner surface of the Sylvian fissure. At this point it gives rise to the Anterior Cerebral Arteries, the Middle Cerebral Arteries, the Posterior Communicating Arteries, and the Anterior Choroid Arteries (Gray, 1995; Snell, 2004; Totora & Grabowski, 2003).

The subclavian arteries give rise to the vertebral arteries, which enter the head via the foramen magnum, where they pass forward and upward to the front of the medulla oblongata. The vertebral arteries first give rise to the Posterior Meningeal Arteries that supply the dura matter in the cerebellar fossae and the falx cerebelli. Next, they give rise to the Anterior Spinal Artery, which feeds the pia matter and the substance of the cord. The Posterior Spinal Arteries, which supply the dorsal roots of the cord, also branch from these processes. Finally, they give rise to the Posterior Inferior Cerebellar Arteries, which provide the blood supply to the lower portions of the cerebellum, and which feed the bulbar artery that circulates blood to the medulla oblongata. The left and right vertebral arteries join near the midline, at the lower border of the Pons Varolii to become the basilar artery. The basilar artery extends under the arachnoid along the median sulcus from the posterior to the anterior of the Pons Varolii. It produces three branches, the Transverse Arteries, Anterior Inferior Cerebellar Arteries, and the Superior Cerebellar Arteries (Gray, 1995; Haines, 1995). At its ending it divides to form the two posterior cerebral arteries (Gray; Snell, 2004; Totora & Grabowski, 2003). Together, the common (internal) carotid arteries and the vertebral come basilar artery provide all blood flow pathways into the brain matter through a circuit called the Circle of Willis (Gray; Haines).

## Malfunction

Around the world[2], heart malfunction is a major cause of death and disability. During 1998, it caused approximately 29% of all deaths in Australia (Davies & Senes, 2001), and throughout the world its rates are increasing due to aging

---

[2] The epidemiological development of cardiovascular diseases and their treatments appears linked to economic, social, and demographic transitions like the development of western style cultures. As such, the discussion on heart failure may be considered generally applicable to all cultures. For the purposes of reading ease, however, only statistics for the Australian population are presented in this section.

populations and improved therapies to keep people with the condition alive (Dyer & Fifer, 2003). Heart malfunction can result from any number of conditions, the specific aetiology, manifestation, interventions, and outcomes for which are numerous and complex. Given the complex and diverse nature of heart malfunction, only the principal causal categories will be addressed in the following sections.

## Heart Failure

The pathophysiological state known as Heart Failure is characterised by the inability of the heart to circulate blood at a rate sufficient to meet the metabolic requirements of bodily tissues, or the ability to circulate sufficient blood but only from a significantly and continually increasing rate of operation. As such, heart failure covers a broad spectrum of outcomes from difficulties manifest during periods of high stress, to major effects where the heart cannot sustain life without external assistance (Dyer & Fifer, 2003; Winakur & Jessup, 2004).

Generally, heart failure is caused by defects in the contraction of the heart muscle. Heart failure is distinguishable from circulatory failure where some component of the circulation such as blood volume or concentration of oxygenated blood is responsible for the reduced cardiac performance. Severe myocardial failure produces heart failure, but the inverse is not always true as many conditions can produce heart failure without, at least in the beginning, there being damage to the myocardium. Similarly, heart failure always produces circulatory failure, but the inverse is not always true as many conditions can cause circulation to cease without having a basis in heart failure. In general, heart failure affects either the expulsion of blood from the heart cavities (systolic failure), or the filling of the heart (diastolic failure) (Anversa, Leri, & Kajstura, 2004; Gunasinghe & Spinale, 2004; Kannel & Vasan, 2004; Katz, 2004; Mann, 2004).

The causes of heart failure are varied and numerous, and there is a vast volume of work addressing those antecedents. Covering all the possible causes of heart failure is beyond the scope of this chapter; therefore, only a brief overview of the main clinical conditions that lead to cardiac surgical intervention is provided.

Coronary Heart Disease

Coronary Heart Disease (CHD) is a broad condition the core feature of which is myocardial muscle cell death caused by a reduction in blood flowing directly to those heart muscle cells.

Generally, circulation to the heart muscle is reduced due to narrowing of the coronary arteries resulting from obstructions such as atherosclerotic plaques and thrombi (Jamrozik et al., 2001). Arteries are constructed from three layers. The innermost layer, called tunica intima, consists of endothelial cells resting on a basement collagen membrane. The wall's middle layer, called tunica media, consists of well-developed and organised concentric layers of smooth muscle cells woven together with an elastin rich extracellular matrix. The outermost layer, the adventitia, consists of a lose array of collagen fibres, various nerve endings, fibroblast and mast cells. Each of the arterial wall layers contributes uniquely to circulatory functioning. The tunica intima contributes to the homeostasis of the vascular system through specifically regulated mechanisms. Primary amongst those mechanisms is the layer's ability to maintain prolonged contact with blood without stimulating the clotting process, something it achieves through secretion of various factors including heparin sulphate proteoglycan and thrombomodulin. The structure of the second layer facilitates the absorption of kinetic energy generated by the pumping action of the heart, thus reducing the rebound effects within the arteries. The third layer's structure contributes primarily to the overall structural integrity of the system (Gordon & Libby, 2003; Libby, 2001; Weissberg & Rudd, 2002).

Atherosclerotic plaques form at points where the innermost layer of the arteries, the tunica intima, is damaged. Though what causes the initial vascular damage to the tunica intima remains unclear, factors believed to contribute include low density lipoprotein levels, cytomegalovirus (herpes virus), prolonged high blood pressure, carbon monoxide in cigarette smoke, and high blood glucose levels (diabetes mellitus). Irrespective of the cause of damage to the artery wall, it is believed that cholesterol and triglycerides collect in the arterial wall's inner layer and through their continual contact with the smooth muscle cells they stimulate the abnormal reproduction of arterial cells to the point that blood flow is blocked. An additional danger is that the arterial wall surface abnormality accumulates platelets and subsequently, phagocytes to form a thrombus (Gordon & Libby, 2003; Libby, 2001; Sidawy, Arora, & Clowes, 2001; Totora & Grabowski, 2003; Weissberg & Rudd, 2002).

Sudden death directly traceable to CHD accounted for 21% (23,012) of all deaths of Australians between the ages of 40 and 90 years in 2000. This equates to a mortality rate of 285 deaths per 100, 000 people. Fortunately, the CHD mortality rate in Australia has decreased by 46% over the last decade (Mathur, 2002). The reduced overall per capita incidence of CHD related deaths may relate to greater awareness in the population of how to maintain a healthy heart, enhanced diagnostic procedures that allow for earlier diagnosis and intervention, and improved medical treatments across all stages of the disease process (Gaziano, 2001).

The development of CHD is multifaceted and complex. It involves a complex interaction of a vast array of fixed and modifiable risk factors. Common fixed risk factors include genetics, gender, and age, while common modifiable risk factors include behavioural, psychological, and sociological contributors (Ridker, Genest, & Libby, 2001). The 1995 Australian Bureau of Statistics National Health Survey found that approximately 81% of the population had one major modifiable risk factor, 43% had two or more major modifiable risk factors, and 13% had three or more major modifiable risk factors. When these numbers are broken down for men and women it was found that for men 18 years and older, approximately 85% had at least one risk factor, 49% had two risk factors, and 15% had three or more risk factors. The statistics for women were 76% having one risk factor, 38% having two risk factors, and 11% having three or more risk factors (Mathur, 2002).

Regarding genetic factors, at present there is no single direct genotype-phenotype relationship associated with the development of CHD. However, candidate genes specific to different stages of the disease process are believed to exist, and it is assumed that by identifying them, treatment processes could be altered to specifically suit the individual and the stage within the disease process to which they have progressed to (Herrmann & Paul, 2001). However, as Herrmann and Paul suggest, the difficulty in identifying clear-cut genotypic-phenotypic relationships derives from the challenge of phenotypic multiplicity, which is the proposition that several conditions with the same presentation exist but derive from distinctly different genetic bases.

Depending on the extent of the reduction in blood flow caused by atherosclerosis or the dislodgement of a thrombus, a range of conditions can occur from Angina Pectoris to Acute Myocardial Infarction resulting in death. Angina Pectoris, both stable and unstable forms, is typically the result of significant obstructive atherosclerosis in the coronary arteries (Cannon & Braunwald, 2001;

Fenster, Sox, & Alpert, 2004; Gupta, Sabatine, & Lilly, 2003; Kupersmith & Raval, 2004). Cannon and Braunwald suggest that five pathophysiological processes may contribute to the development of angina pectoris. They are (1) plaque rupture with resulting non-occlusive thrombus, (2) dynamic obstruction of epicardial or small muscular coronary arteries due to coronary spasm, (3) progressive mechanical obstruction, (4) inflammation/infection, and (5) secondary angina precipitated by increased oxygen demand or reduced oxygen supply to the myocardium. Specifically, Stable Angina Pectoris is characterised by poorly localised pain in the chest or arm discomfort that is reproducible with physical exertion or emotional stress, but which usually remits five to fifteen minutes later with rest or nitro-glycerine. Conversely, Unstable Angina Pectoris is characterised by the same symptoms, but with the addition of at least one of three other features. Those features include (1) occurs at rest and lasts more than 20 minutes, (2) is severe and described as a pronounced pain that has onset within the preceding hour, and (3) it occurs more frequently or for longer than previously experienced and has increased severity (Cannon & Braunwald). Acute Myocardial Infarction (AMI) generally results from coronary atherosclerosis and may involve the added complication of coronary thrombosis. The gradual development of atherosclerotic plaques within coronary arteries does not guarantee AMI, as along with their development a rich collateral circulatory network develops to compensate for the decreased blood flow. AMI generally develops when an abrupt change occurs such as plaque rupturing. After plaque rupture, there is exposure of platelet activation and aggregation promotion substances. This in turn results in the formation of thrombi. The resulting thrombi then interrupt blood flow leading to an imbalance between oxygen and nutrient supply and demand at the muscle. If the balance between supply and demand is sufficiently deficient and persistent, necrosis of the tissue being supplied by that process occurs (Antman & Braunwald, 2001; Berger, 2004). Based on the degree of tissue death, AMI's can be defined as either (1) transmural or (2) subendocardial. Transmural infarcts occur when the entire thickness of the myocardium necroses; whereas subendocardial infarcts involve only partial necrosis that does not extend all the way through the ventricular wall of the heart (Antman & Braunwald). The number of Australians 40-90 years old admitted to hospital with AMI in the 1999-2000 period was 28,002. In real terms, this marked a 12% decrease in incidence over the preceding estimate carried out five years earlier (Mathur, 2002).

Incidence rates provide an indication of degree of risk associated with being male or female. Of the total CHD cases for 1999-2000, 61.54% were males and 38.46% were females. These figures equate to incidence rates of 766/100,000 for males and 453/100,000 for females. This incidence rate of 1.7:1 (M:F) was found to hold across all age ranges, and indicates a substantially higher risk of developing CHD in males than females. Of the deaths in 1999-2000 directly attributable to CHD, 56.64% were males and 43.36% were females. Of the admissions to hospital for AMI within that period, 64.23% were males and 35.77% were females. These numbers equated to an incidence rate of 464/100,000 for males and 244/100,000 for females. Across all age groups, the incidence of hospital admissions for AMI was higher for males than females (Mathur, 2002)

In terms of age, CHD death rates increase in line with age such that a substantially higher proportion of CHD deaths occur in the 75-90 year old age range than in the 40-64 years old range. The incidence of AMI related hospitalisations also increases in line with age (Mathur, 2002).

Modifiable behavioural patterns contributing to CHD risk include dietary factors, insufficient physical activity, being overweight, and tobacco smoking. Of these, ceasing smoking reduces the relative risk of first AMI by 65%. Increasing physical activity and decreasing obesity have the positive effect of improving myocardium efficiency by decreasing its demand for oxygen, which in turn decreases the heart's base rate of operation (Ridker et al., 2001).

Modifiable psychological features found to contribute to CHD primarily relate to stress. Stress can produce vasoconstriction, which in the coronary arteries can further affect myocardial oxygen supply through arteries that may already have reduced blood flow due to atherosclerotic plaques. Another result of stress is increased catecholamine levels. Catecholamine is known to promote coagulation, which in itself may facilitate further development of thrombus at lesion sites within the coronary arteries, and increase the potentiality for a thrombolytic episode (Ridker et al., 2001).

Sociological modifiable risk factors, such as ethnic group and socio-economic status have also been shown through incidence statistics to have a strong relationship to CHD. For example, subpopulations such as indigenous Australians have been found to have a significantly higher death rate due to CHD (6-8:1) than non-indigenous Australians. Similarly, the socioeconomically disadvantaged (2:1) have significantly higher death rates from CHD. Fundamental to these statistics is that subpopulation membership and socio-economic status may be producing their impact in the areas of prevention and treatment rather than the actual disease process itself (Mathur, 2002).

Valvular Heart Disease

Valvular Heart Disease (VHD) is the term used to describe cardiac dysfunction caused by structural or functional abnormalities of the heart valves. Central causes of VHD include congenital defects such as absences, malformations, or additions, and acquired problems including calcification, and inflammatory diseases such as rheumatic fever (Carabello, 2004; Griffin, 2004; Liberthson, 2000; Smolens & Bolling, 2001).

Congenital valve defects generally relate to developmental syndromes such as Rubella and Williams syndromes (Griffin, 2004). Such conditions can affect the mitral and tricuspid valve development, although the incidence of congenital mitral valve defects is small. Tricuspid valve congenital abnormalities most commonly consist of Ebstein's anomaly, which occurs when part of the line of attachment of the valve leaflets is shifted downward toward the ventricle. That downward displacement results in part of the ventricle becoming atrialised. Congenitally abnormal valves sometimes also have abnormally developed leaflets, shortened or absent chordae tendineae, abnormal sizes and numbers of papillary muscles, or a combination of those defects. In addition to affecting the mitral and tricuspid valves, congenital defects may also result in the presence of an additional valve called the bicuspid aortic valve. However, the presence of the bicuspid aortic valve is usually associated with heart vasculature abnormalities on the left side, and aortic abnormalities (aortic root enlargement, aortic dissection, and histologic changes).

Acquired VHD is attributable to the effects of aging, diseases such as rheumatic fever, infection by micro-organisms, and systemic diseases such as connective tissue disorders. The effects of aging are often due to calcification, which is a process similar to the development of atherosclerotic plaques. Calcification begins with the depositing of microscopic calcium specs in the cusps of the valves. Over time, those deposits combine to form calcific nodes that reduce the aperture of the valve opening and effective flow of blood. When occurring in the aortic valve, the early stages of this process where obstruction is yet to occur is called Aortic Valve Sclerosis, while the obstructive stage is called Aortic Valve Stenosis. In Mitral valves, the process is referred to as Mitral Annular Calcification. Additionally, wear and tear may result in the accentuation of normal valve or surrounding structure architecture by way off thickening or thinning of some parts, and lengthening or increasing the prominence of others (Carabello, 2004; Griffin, 2004; Liberthson, 2000; Otto, 2004; Smolens & Bolling, 2001).

Rheumatic Fever is an inflammatory immunologic disease affecting the heart, joints, and skin. It occurs initially via pharynx infection due A Beta-hemolytic Streptococci, with the pathogenesis relating either to an autoimmune reaction initiated by the streptococcal infection or to enhanced immune response to streptococcal antigens that evoke antibodies that also react with human tissue antigens. In terms of the heart's valves, it primarily affects the left side by forming mild thickening of the leaflet or cusp, and the formation of small wart like growths on the aspects of the leaflets that come together to block the aperture. Although the acute lesions do not cause significant functional abnormality the subsequent scaring from the inflammatory process results in deformation of the valve (Dajani, 2001; Yachimski & Lilly, 2003).

## Intervention

Treatments for heart disease vary depending on the type and severity of the heart condition a person suffers. In many cases, treatments are applied in combination. More importantly to the issue of the effects of treatments on cognitive functions is that treatments applied can vary in invasiveness from medical management utilising pharmacological measures to direct interventions that involve open heart surgery and the use of heart bypass machines.

Pharmacological

Pharmacological interventions used to treat cardiac problems rank among the most frequently prescribed medications in the Australia. Collectively, the medications used to treat cardiovascular conditions accounted for approximately 17% of all medication prescriptions dispensed in Australia in 1994. In 1995-1996, approximately 125 million prescriptions for cardiac medications were filled in Australia, at a cost to the Pharmaceutical Benefits Scheme of over $2.5 billion (Waters, Armstrong, & Senes-Ferrari, 1998).

Drugs used to treat cardiac conditions have two general aims. The first is to progress the patient from a state of decompensated heart failure to a stable state with adequate blood flow to the myocardium and optimal filling pressure. Part of this aim is to transpose the patient to therapies suitable for ongoing chronic conditions. Drug classes used for this purpose are diuretics, vasodilators, and positive inotropic agents. The second aim is the ongoing maintenance of the patient to ensure survival and to improve functionality by reducing symptoms. Again diuretics, vasodilators, and positive inotropic agents are used, along with neurohormonal or cytokine inhibitors (Bristow, Port, & Kelly 2001; Ndumele, Friedberg, Antman, Strichartz, & Lilly, 2003).

Surgical

Surgical procedures used to rectify cardiac disease or malfunction include a broad range of techniques with differing degrees of invasiveness. At the less invasive end of the treatment spectrum are techniques that utilise catheters, which are passed through the skin into the blood vessels and fed through the venous system to the heart (Soltoski, Guzzo Lemke, Barolet, & Chisolm, 2004). At the more invasive end are procedures that require the chest cavity to be opened, the patient to be placed on mechanical bypass, and cardiac structures to be manipulated (Li, 2004a). In 1998, there were 22,253 cardiac surgeries carried out in Australia. This represented a rate of 1,188 per one million people in the Australian population (Davies & Senes, 2001). During 1999, the number of surgical procedures carried out to remediate cardiac problems was 20,791, or 1,088 per one million people in the Australian population (Davies & Senes, 2003).

Procedures utilising catheters to reduce the extent of an artery's blockage are called Percutaneous Coronary Intervention (PCI). One procedure within this type is Percutaneous Transluminal Coronary Angiography (PTCA). In this procedure, a catheter with a balloon positioned near its tip is inserted through the skin and into a major artery. It is then threaded through the circulation back towards area of the coronary artery that is obstructed. At that point, the balloon is inflated to disrupt the obstruction and open the passage for blood to flow. This type of surgery avoids the major trauma of bypass surgery because it does not require the patient's chest to be opened. Though this procedure is much less invasive, it is only useful for certain types of obstructions (Davies, 2003; Davies & Senes, 2002). At the inception of data recording for PCI procedures in Australia (1980) there were 11 procedures performed. In 1998, 19,444 PCI procedures were performed in Australia. Of those, the ratio of males to females was 2.6:1 or, in real terms 14,278 to 5,166 procedures (Davies; Davies & Senes, 2002).

At the more invasive end of the spectrum of surgical interventions are the procedures that involve placing the person on a bypass machine to continue the circulation of blood throughout the body, while the chest cavity is opened so that the heart or the vessels feeding it can be directly manipulated, bypassed, or replaced. There are many procedures within this category including Transplant surgery, Cardiac Artery Bypass Graft surgery, Valve surgery, Electrophysiological surgery, and Aortic surgery.

Heart transplant surgery in Australia is generally done in response to cardiomyopathy or ischaemic conditions. In 1998, 72 heart transplants and four heart lung transplants were carried out in Australia. Sixty-two heart transplant patients and all four of the heart and lung transplant patients survived (Davies & Senes, 2001). In 1999, 65 heart transplants and two heart lung transplants were performed. Of those surgeries, eight heart transplant and one heart lung transplant patients subsequently died (Davies & Senes, 2003).

Coronary Artery Bypass Graft surgery (CABG) involves using vessel grafts to construct bypass conduits for major arteries, to points beyond obstructions in the arteries feeding the heart muscle to restore adequate blood flow to nourish that muscle. The procedure usually requires the chest cavity to be opened and the circulation to be diverted away from the heart and lungs. In such instances, a bypass machine with an oxygenating pump is used to replenish the blood and to circulate it around the body. Graft material is generally taken from the person's saphenous vein in the leg or the internal mammary artery in the chest. In 1998, there were an estimated 17,448 CABG surgeries carried out. Of those 13,182 were conducted on males and 4,266 were conducted on females (Davies & Senes, 2001). In 1999, that number decreased slightly to 17,321 operations, with 13,170 operations being performed on males and 4,150 being performed on females (Davies & Senes, 2003).

Valve surgery (VS) is performed on all valves within the heart; however, it is most commonly performed on the aortic and mitral valves and less commonly on the tricuspid and pulmonary valves. As with CABG, VS procedures require the chest cavity to be opened and the circulation to be diverted away from the heart and lungs and into a bypass machine with an oxygenating pump before passing back into the body. VS can involve repairing the malfunctioning valve, but more commonly, it involves replacing the defective valve with a mechanical prosthesis, a pig bioprosthesis, or a human graft. In 1998, there were an estimated 4,578 surgeries carried out to repair or replace coronary valves (Davies & Senes, 2001). In 1999, this figure rose to 4,892 (Davies & Senes, 2003).

Electrophysiological surgery (ES) relies on the use of lasers to remove sections of heart muscle that are responsible for abnormal heart rhythms (arrhythmias) such as ventricular and supraventricular tachycardia. In 1998, there were 133 ES procedures carried out in Australia. None of those procedures resulted in patient death (Davies & Senes, 2001). In 1999, 191 such operations were conducted. Again, no deaths were associated with that cardiac surgical procedure (Davies & Senes, 2003).

Aortic surgery (AS) is generally carried out to repair or replace defective valves, or to replace dysfunctional aortic vessel wall sections. In 1998, 325 aortic operations were carried out in Australia (Davies & Senes, 2001). In 1999, there were 385 such procedures completed (Davies & Senes, 2003). The death rates associated with these surgeries were 40 in 1998 and 46 in 1999 (Davies & Senes, 2001; 2003).

Other surgical procedures requiring opening of the chest cavity include removing cardiac tumours, repairing the results of trauma, and surgery to the pericardium. In 1998, 116 other cardiac surgical procedures were carried out. Of those procedures, 18 resulted in patient death (Davies & Senes, 2001). In 1999, 136 such procedures were carried out, resulting in 24 associated deaths (Davies & Senes, 2003).

Cardiopulmonary Bypass

Human Cardiopulmonary bypass, first developed by Gibbon in 1954, utilises mechanical devices to bypass the heart and lungs so that circulation to vital organs (i.e., kidneys, brain) can be maintained while an operation to repair or replace damaged structures is carried out. Without cardiopulmonary bypass, surgeons would have only six minutes in which to perform their surgery before the brain suffered hypoxic damage. Bypass is used in most cardiac surgery procedures where stopping the heart or lungs is necessary (Seifert, 2002; Young & Dai, 2004a). The components of the mechanical bypass system include a pump to replace the pumping action of the heart, an oxygenator to act as an artificial lung for re-oxygenating the blood, heating and cooling elements to regulate the temperature of the blood, and tubing for patching into the circulatory circuitry (Seifert, 2002; Young & Dai, 2004b).

Cardiopulmonary bypass operates by draining venous blood from both the superior and inferior vena cava. As it is drained, the blood passes through an oxygenator where it is re-saturated with oxygen. The blood is then pumped back into systemic circulation where it progresses along normal circulatory pathways to nourish the system. Cardiopulmonary bypass can be a total or partial process with the essential difference being that in partial bypass some venous blood returns to the right atrium before being removed (Seifert, 2002; Young & Dai, 2004b).

Hypothermia is used as a safety measure during cardiac surgery. By reducing the body's temperature to below normal levels, the metabolic rate is reduced. This, in turn, reduces the body's need for oxygen. There are four levels of hypothermia (mild

= 32-35°C, moderate = 28-32° C, deep = 18-28° C, and profound = 15-18° C), and their use depends on the complexity of the surgical procedure required (Young & Dai, 2004b). The two methods for cooling the body are (1) surface cooling, and (2) core cooling. Inducing hypothermia, though considered a protective process, also caries the danger of ischemia for the body's organs. In order, the brain, heart, kidneys and liver are susceptible to damage (Li, 2004b). With profound hypothermia for up to an hour, there is a low risk of long term neurological injury (Li, 2004b). A number of methods, such as steroids, barbiturates, and retrograde cerebral venous perfusion are available to protect the brain during hypothermia (Li, 2004b).

When body temperature is reduced for surgery, gases become more soluble in the blood. This in turn causes the pH balance to rise. This process is corrected by one of two methods (1) alpha-stat, and (2) pH-stat. In alpha-stat, as the blood is cooled carbon dioxide is not added. This causes $pCO_2$ levels and progressive alkalosis to decrease along with the temperature. It is claimed that the alpha-stat method maintains the buffering capacity of the alpha-imidazole group of histidine, as well as cerebral autoregulatory capacity (Li, 2004b; Whitaker, Stygall, & Newman, 2002). In the pH-stat method, $CO_2$ is added in order to maintain a pH level of 7.40 throughout the temperature lowering process. This process causes cerebral autoregulation to decrease and the arteries to dilate (Li, 2004b; Whitaker et al.).

## Conclusion

The cardiovascular system, the diseases that affect it, and the interventions used to reverse or restrict those effects are a complex system of actions and reactions. When operating normally, the cardiovascular system services all the organs of the body at the level required to ensure their effective and efficient functioning. But what happens to organs such as the brain when that system fails and requires therapeutic intervention? In the following chapter, neuropsychological research pertaining to cardiac surgery such as CABG, VS and, CPB published since 1995 are reviewed.

# CHAPTER THREE

## The Heart of the Matter – Assessing for Cognitive Decline after Cardiac Surgery

### Introduction

The interrelatedness of neurological and cardiovascular functioning is evidenced by the fact that complications anywhere within the cardiovascular system can negatively impact neurological capacity, and that reduced neurological capacity can disrupt cardiac functions (Caplan, 1999; Sontaniemi, 1995). Disruptions to the cardiovascular system caused by risk factors, progressive or chronic disease, and acute insult are acknowledged as potentially significant influences on the central nervous system and subsequently cognitive functioning (Elias, Elias, Robbins, Wolf & D'Agostino, 2001; Everson, Kelkala, Kaplan & Salonen, 2001; Muldoon, Flory & Ryan, 2001; Phillips, 2001; Vingerhoets, 2001; Waldstein & Katzel, 2001). A full review of all potential influences on cognitive outcome associated with cardiovascular dysfunction is beyond the scope of this discussion. However, the interested reader is referred to Waldstein and Elias (2001) for a general overview. Disruption to the normal operation of the cardiovascular system, such as that caused by cardiac surgery, is recognised as a significant influence on central nervous system functionng and has been since the earliest published reports in the field (Benedict, 1994; Smith, 1995; Sontaniemi).

The estimated incidence of central nervous system complications associated with that treatment ranges up to 13%, depending on the surgical methods applied and the patient population examined (Barbut & Caplan, 1997; Mills, 1995; Newman & Stygall, 1999). Central nervous system complications suffered by patients undergoing cardiac artery bypass graft surgery (CABG) covers the full range of neurologic injury from major stroke causing death, through permanent cognitive decline, to temporary encephalopathy causing transient changes in cognitive capacities (Arrowsmith, Grocott, & Newman, 1999; Borowicz, Goldsborough, Selnes, & McKhann, 1996; Barbut & Caplan).

Death due to the effects of CABG surgery is an increasingly less common event due to improved intervention techniques (Barbut & Caplan, 1997; Newman & Stygall, 1999). Therefore, the utility of mortality rate as an outcome indicator has decreased (Arrowsmith, Grocott, & Newman, 1999; Mathes, Stone, & Dent, 2001; Pinna-Pintor, Bobbio, Giammaria, Suni, & Alfieri, 1998). In comparison, cognitive decline in any domain (i.e., memory, attention, psychomotor speed, and language) is recognised as a major determinant of patient wellbeing following cardiac surgery (Andrew, Baker, Kneebone, & Knight, 1998; Bruggemans, Van Dijk, & Huysmans, 1995; Mills, 1995; Newman & Stygall, 1999; Suksompong, Prakanrattana, Chumpathong, Sriyoschati, & Pornvilawan, 2002). Therefore, the incidence of such declines has been recognised as a potentially useful indicator of outcomes.

Some features that substantially detract from the utility of cognitive decline rates as outcomes, however, is the wide range of incidence values and the heterogeneous nature of declines observed in the literature. For example, the reported incidence of cognitive decline in the period immediately after cardiac surgery varies upt to 100%, while at 12 months post surgery it ranges up to 35%, and five years after surgery up to 42% (Arrowsmith, Grocott, Reves, & Newman, 2000; Mills, 1993; M. Newman, Grocott et al., 2001; M. Newman, Kirchner et al., 2001; Selnes, Royal, Grega, Borowicz, Quaskey, & Mckhann, 2001; Yates and Alstron, 2000). Compared to the incidence of stroke and permanent cognitive decline, encephalopathy following cardiac surgery has been reported in as many as 25% of patients four days post surgery (Barbut & Caplan, 1997).

The interrelatedness between the array of cognitive impairments demonstrated by cardiac surgery patients and the range of cardiac interventions indicates that an association exists between the two. However, that association does not equal an aetiological certainty. Aetiology is the specification of a causal relationship that is noticeably apparent at a cohort level (Sadegh-Zadeh, 1998). An example of an aetiological model in the current context might be, *cardiac surgery utilising cardiopulmonary bypass procedures results in a stroke due to the process of emboli flow during surgery through the circulatory system and into the left anterior communicating artery obstructing blood supply to neurons resulting in infarction.* While, many associations have been posited to explain the effect of cardiac surgery on the brain's functioning (Haddock et al., 2003; M. Newman, Croughwell et al., 1995; Newman, Stygall, & Kong, 2001; Roach et al., 1996), a definitive aetiological model of cardiac surgery related cognitive decline has not been fully and concisely demonstrated (Taylor, 1998a,b). Perhaps the lack of a concise aetiology can be attriguted the suspicion that a large number of patient and intervention specific factors contribute to post cardiac surgery cognitive dysfunction both individually and in combination, (Arrowsmith, Grocott, & Newman, 1999; Barbut & Caplan, 1997; Borowicz et al., 1996; Di Carlo et al., 2001; Gill & Murkin, 1996; Haddock et al.; M. Newman, Booth, Kaskowitz, Schwinn, Grocott, & Mathew, 2001; Newman & Stygall, 1999; Selnes, Goldsborough, Borowicz, & McKhann, 1999; Sontaniemi, 1995; Taylor). Another possible explanation for the lack of a concise aetiology, as suggested by Haddock et al. in a recent literature review, is the distinct lack of coherence in research outcomes due to the piecemeal and methodologically unregulated approach to assessing cognitive functioning and potential decline in cardiac surgery contexts.

# Methodology

In 1995, a consensus statement on the assessment of post cardiac surgery neurobehavioural functioning was published (Murkin, Newman et al., 1995). That consensus group discussion addressed a range of pertinent methodological issues before agreeing on 14 specific points and a core test battery consisting of four neuropsychological instruments. An additional statement of consensus, further addressing four points from the first statement, was published in 1997 (Murkin, Stump et al). While widely recognised as important stages of methodological clarification (Slade et al., 2001), the impact of the 1995 consensus statement was placed into context by the primary author, John Murkin, who acknowledged that it was not the definitive end point to improving research efforts within the field but rather a point of departure for further refinement (Murkin, 1995).

Three main methodological aspects have consistently been directly related to the lack of cohesiveness in research and subsequent divergent research findings. They are (1) samples, (2) assessment processes, and (3) decision criteria (Borowicz et al., 1996). In the following sections, each of those elements is addressed.

## Samples

A number of different elements are important and should be highlighted when discussing the population subgroups (samples) gathered together for research purposes. Those elements can be grouped together into two categories. They are (1) sample characteristics, and (2) sample size.

### Sample Characteristics

A myriad of sample characteristics can influence neuropsychological test performance and hence the outcome of research and the validity of assessment conclusions drawn. Age, gender, education, and IQ, for example, are well recognised as major influences on test performance (Bird, Papadopoulou, Ricciardelli, Rossor & Cipolotti, 2004; Lezak, 1995; Murphy & Davidshofer, 1998; Peters, Graf, Hayden, & Feldman, 2004; Rosselli & Ardila, 2003; Spreen & Strauss, 1998). Hence, test development and validation processes often incorporate an assessment of those potential influences on the test development or normative sample. Contexts such as culture and clinical group association, which have also been recognised as possible influences on test performance, by contrast may not have been similarly investigated on all tests (Kaufman, McLean, & Kaufman, 1988; Levav, Mirsky, French, & Bartko, 1998; Ogden, 2001; Ogden, Cooper, & Dudley, 2003; Ogden & McFarlane-Nathan, 1997; Roselli & Ardila, 1991, 2003)

Normal aging is generally considered to produce a gradual change in cognitive capabilities (Arrowsmith, Grocott, & Newman, 1999; Arrowsmith, Grocott, Reves et al., 2000; Lezak, 1995; Murphy & Davidshofer, 1998; Spreen & Strauss, 1998; Tulsky, Saklofske, Heaton, Bornstein, & Ledbetter, 2003; Tuman, McCarthy, Najafi, & Ivankovich, 1992). The rate of normal age related cognitive change, as evidenced by neuropsychological test performance (i.e., on Wechsler Memory Scale subtests) in the absence of cerebral insult, has been generally acknowledged as negative and relatively linear (Benedict, 1994; Heaton, Taylor, & Manly, 2003; Stump, James, & Murkin, 2000). When the addition changes due to surgical intervention are combined with pre-existing age related changes, reserve capacity may be exceeded, resulting in noticeable functional decline (Arrowsmith, Grocott, & Newman; Arrowsmith, Grocott, Reves et al., 2000).

Patient age at the time of surgical intervention has been strongly associated with a high risk of postoperative cognitive decline irrespective of the type of surgery patients undergo (Moller et al., 1998). In cardiac surgery populations, a significant relationship has been demonstrated between increasing age and the incidence of post surgery cognitive dysfunction (Andrew, Baker, Kneebone et al., 2000; Arrowsmith, Grocott, & M. Newman, 1999; Arrowsmith, Grocott, Reves et al., 2000; Benedict, 1994; Di Carlo, Perna, Pantoni, Basile, Bonacchi, & Pracucci et al., 2001; Hlatky et al., 1997; M. Newman, Grocott et al., 2001; S. Newman & Stygall, 1999; Rasmussen, 1999; Selnes, Royall et al., 2001; Vingerhoets, Van Nooten, & Jannes, 1997). In a review of the literature by van Dijk, Keizer, Diephuis, Durand, Vos, and Hijman (2000), the whole population incidence of age associated cognitive deficit after cardiac surgery was shown to vary from 4% to 47%. In terms of specific ageing stages, the reported incidence of post surgical cognitive deficit ranges between a fraction of a percent for patients 45 years or younger, to 10% for those over 70 years, and 9% for those over 75 years (Gardner et al., 1985; van Dijk et al., 2000; Yates & Alstron, 2000). In an investigation by Ahto, Isoaho, Poulijoki, Laippala, Sulkava, and Kivela (1999), the association between increasing age and increasing deficit rates was found to hold for both males and females.

The relationship between decline in cognitive functioning and normal aging has been reported to be non-linear (Stump, James et al., 2000). In the group of studies addressed in the preceding review section, the ages of participants ranged from the third to seventh decades of life, with the majority of studies being carried out on participants in the sixth decade. This represents a broad range of ages and life stages, all of which are likely to place different demands on cognitive functioning. The age bracket of patients undergoing and recovering from cardiac surgery is increasing due to an ageing population, something that will undoubtedly introduce additional effects relating to differing life stages. Despite the broad range of ages included in the research, most studies continue to treat age as a continuous variable, something that overlooks the impact of different life stages on cognitive demands.

Due to the consistently demonstrated relationship between age and cognitive functioning, age at time of surgery is often considered the most robust predictor of negative cognitive outcomes (Arrowsmith, Grocott, Reves et al., 2000; Benedict, 1994; Borowicz et al., 1996; Gill & Murkin, 1996; M. Newman, Croughwell et al., 1995; M. Newman, Kramer et al., 1995; S. Newman, 1995; Selnes, Goldsborough et al., 1999; Taylor, 1998a,b). To date, there has been no systematic large scale investigation of the influence of age on cognitive outcome following cardiac surgery. Therefore, despite the apparently strong association between increasing age and post cardiac surgery cognitive dysfunction, the parameters of the relationship remain largely unestablished (van Dijk, Keizer et al., 2000). Given that an association is known to exist between age, test performance, and cognitive outcome after cardiac surgery, and the fact that age is considered the most robust indicator of negative outcome, it is clear that further investigation of the relationship is warranted.

<u>Gender</u> is another demographic factor with significant links to cognitive functioning after cardiac surgery. In a brief review of unpublished cardiac surgery related data, Arrowsmith, Grocott, and Newman (2000) found that while women generally had a higher cognitive decline risk profile, being female was associated with improved survival rates five years after surgery. Hofste, Linssen, Boezman, Hengeveld, Leusink, and de Boer (1997) also found a significant relationship between gender and the risk of post operative delirium. However, they did not find an association with cognitive deficit. This result stands in contrast to Di Carlo et al. (2001) who found that being female resulted in a higher rate of cognitive decline at six months after surgery.

Some neuropsychological tests are known to encounter gender effects (Lezak, 1995; Spreen & Strauss, 1998), and it is presumably on the basis of that understanding that the 1995 consensus statement reiterated the need to utilise tests that are free from gender biases. Despite that understanding, the core test battery recommended by the consensus group included two tests with known gender bias effects (Rey Auditory Verbal Learning Test and the Trail Making Test; see Spreen and Strauss, 1998).

Gender has been found to be associated with cognitive dysfunction following cardiac surgery (Andrew, Baker, Kneebone et al., 2000; Hlatky et al., 1997; M. Newman, Grocott et al., 2001; Selnes, Royall et al., 2001; Vingerhoets,Van Nooten et al., 1997). Gender balances in the studies reviewed above appear to predominantly hover around an 80:20 male to female split. A distribution of that proportion is representative of the gender distribution (1.7:1, M:F) risk of cardiac disease in Australia (Mathur, 2002). Therefore, it could be considered representative from a census standpoint. Arrowsmith, Grocott, and Newman (1999) suggest that the preponderance of males requiring surgery is the main reason that gender differences have not been explored. Unfortunately, as is the case previously in the literature, very few of the studies presented in the above review addressed potential gender differences in cognitive outcomes following cardiac surgery. In one study that did investigate gender effects on specific neuropsychological tests in cardiac surgery samples, differences were found on Part B of the Trail Making Test (Vingerhoets, Van Nooten et al., 1997).

The effect of <u>education</u> on neuropsychological tests is also widely accepted (Lezak, 1995; Spreen & Strauss, 1998). However, in the outcomes research pertaining to cognitive decline following cardiac surgery, the exact nature of that effect remains generally unexplained (Haddock et al., 2003). In the studies reviewed above, only a small proportion actually presented education data, and only one study (Vingerhoets, Van Nooten et al., 1997) was found that directly examined the effects of education as an independent factor or covariate. Vingerhoets, Van Nooten et al. (1997) found that when compared to other demographic and health variables, education accounted for a greater percentage of score variance on verbal fluency (Controlled Oral Word Association Test) and motor speed and dexterity (Purdue Pegboard) in cardiac surgery samples.

In terms of the association between education and decline due to surgery, M. Newman, Croughwell et al. (1995) found that education had a protective effect on cognitive outcome after cardiac surgery, such that increasing levels of education are associated with less postoperative cognitive decline. Several other authors have also found similar associations between lower levels of education and increased levels of post cardiac surgery cognitive dysfunction (Di Carlo et al., 2001; M. Newman, Kirchner et al., 2001; Tardiff et al., 1997). Conversely, Smith et al. (2000) failed to find similar associations, and M. Newman, Grocott et al. (2001) reported that quality of life five years after surgery was associated with cognitive functioning but independent of education. Murkin, Martzke, Buchan, Bentley, and Wong (1995) reported that increasing levels of education correlated with increasing cognitive deficits only at seven days post operatively.

Studies mentioned in the preceding review section included samples with wide ranging age groups, differing levels of education, and a blend of males and females subjects. Unfortunately, in examining those variable samples, they also used uncorrected raw scores in their analyses, something that would based on the preceding discussion, appear to be a significant methodological inadequacy. Given the obvious uncertainties surrounding the effects of age, gender, and education on neuropsychological test performance in cardiac surgery populations, it is clear that systematic evaluation of their effects is required before assumptions about test reliability and validity can be extended to the cardiac surgery context.

Several studies examining cognitive decline in cardiac surgery samples have found significant preoperative neuropsychological test performance differences between subjects requiring cardiac surgery and normal individuals (Andrew, Baker, Kneebone et al., 1998; Fearn, Pole, Wesnes, Faragher, Hooper, & McCollum, 2001; Gugino, Chabot, Aglio, Maddi, Gosnell, & Aranki, 1997; Keith, Puente, Marks, Malcolmson, Tartt, & Coleman, 2002; Townes et al. 1989; Vingerhoets, Van Nooten et al., 1997). Additional evidence apparently supporting the suggestion that cardiac surgery candidates are distinctly different from normal individuals derives from neurophysiological (electroencephalogram) and cerebral imaging (Single Photon Emission Computed Tomography) studies (Gugino, Chabot, Aglio, Maddi et al.; Hall et al., 1999; Toner, Taylor, Newman, & Smith, 1998) research that have found preoperative abnormalities. The lack of conclusiveness surrounding the preoperative equality of neuropsychological test performances between cardiac surgery patients and normal individuals raises the question of whether they are samples from the same population or whether cardiac surgery candidates are a distinctively different population. Additionally, findings such as these in turn lead to the conclusion that performances on individual tests may change across cohorts in line with the cerebral compromise produced by the condition (Millis, Malina, Bowers, & Ricker, 1999; Price, Tulsky, Millis, & Weiss, 2002).

Sample Size

Determining the sample size required for an investigation is a complex process involving the consideration of many elements including the degree of precision in the instruments being used, the degree of confidence in the outcome that is desired, the degree of variability in the population being sampled, the complexity of the statistical analyses to be completed, and limitations on available resources required to achieve the optimum methodological level (Agresti & Finlay, 1997). If all of those elements were equal between studies, differences in the number of cases investigated would directly affect the reliability of research outcomes and conclusions. Additionally, in two studies of equal measurement methodology, the study with the larger sample size would obtain results with greater reliability (Cohen, 1988). The exception to that rule of thumb is investigations where the population being sampled is completely homogenous. In such a situation, a set of studies could have either one or 100 participants and they would produce the exact same result (Agresti & Finlay, 1997). However, in cardiac surgery research, it is unlikely that such a homogeneous sample would be achievable given the wide variety in both personal and interventional factors that exists. In the studies reviewed above, group sample sizes being used in the research ranged from small ($n = 7$) to large ($n = 308$). Given the diversity in the sample sizes utilised in the research and the high likelihood that both personal and interventional heterogeneity exists, it is understandable that different outcomes are obtained and that the reliability of those outcomes may also vary. It is also possible, by extension, that the conclusions being drawn from analysing those samples are unreliable as a result.

Assessment Methodology

Assessment methodology consists of two components, instruments or the individual tests used to assess cognitive functioning (instruments), and the scheduling of assessments (timings).

Instruments

At least the last half of a century of inquiry into cognitive functioning following cardiac surgery has seen neuropsychological assessments become the frontline of the investigative armory (Borowicz et al., 1996) and as a result very many neuropsychological tests have been used (Borowicz et al.; van Dijk, Keizer et al., 2000; Yates & Alston, 2000). The selection and use of appropriate neuropsychological tests, and the analysis and interpretation of their results has been widely addressed in the literature (Blumenthal et al., 1995; Borowicz et al.; Murkin, Newman et al., 1995; Murkin, Stump et al., 1997; S. Newman, 1995; Slade et al., 2001; Stump, 1995; Ryan & Hendrickson, 1998). Interestingly, despite significant discussion about and widespread use of neuropsychological tests in cardiac surgery research, the gold standard methodology for their use has not been defined (Blumenthal et al., 1995; S Newman, 1995; Slade et al., 2001; Stump, James et al., 2000).

The 1995 consensus statement (Murkin, Newman et al., 1995) represented a potentially significant methodological advancement in the investigation of the cognitive decline phenomenon in cardiac samples (Baker, Andrew, & Knight, 2001). Point five in the consensus statement specifically dealt with assessment methodology, and in doing so considerably reinforced the importance of selecting the correct instruments for the purpose. In particular, point five addressed the need to attend to issues such as sensitivity, reliability, and validity, as well as the individual and range of cognitive functions being assessed by the tests (Murkin, Newman et al., 1995; Newman, 1995). In an article directly associated with the consensus meeting, Blumenthal et al. (1995) also highlighted the necessity of selecting neuropsychological tests on the basis of their reliability and validity. However, they also emphasised additional considerations such as the need for available relevant normative data, alternate forms, and the importance of choosing tests that had previously been utilised in the research. In recognition of the diversity of methodologies being used in the research, and presumably in an attempt to fulfil the obligations identified in the consensus statement, a core test battery consisting of four neuropsychological instruments were recommended. The tests were parts A and B of the Trail Making Test, the Grooved Pegboard test, and the Rey Auditory Verbal Learning Test. In recommending the core battery, however, the consensus group did not elucidate any empirical or psychometric support for selecting those instruments. Nor were suggestions made about criteria for interpreting the results gained from using those instruments in the cardiac surgery context, although this has received some attention in recent times. Therefore, the consensus group failed to clarify the utility of the core tests by not providing evidence of the reliabilities and validities of the instruments in the cardiac surgery context. Examination of the literature both preceding and following the consensus statement indicates that this shortcoming has not been directly redressed. Despite the clear lack of empirical support for selecting any test, let alone the core test battery, for use within the cardiac surgery context, some authors continue to recommend that specific tests be utilised (Slade et al., 2001).

Tests currently being utilised in the research have generally been well examined for their psychometric properties in specific contexts. In recognition of this, several of the articles in the review section above have alluded to the tests used as being valid and reliable instruments, however, it is important to note that no evidence has been provided to support that assertion. While this may be the case for contexts in which the tests were developed or examined, at least one group of authors (Mc Daid et al., 1994) has questioned the generalisability of validity information to the cardiac surgery context.

Factor analysis is a commonly used procedure for examining the structure of neuropsychological tests and batteries designed to measure specific cognitive constructs and functional theories. Unfortunately, in some instances, re-examination of the factor structure of commonly used neuropsychological test batteries has led to alternate models for the structure of the batteries being proposed. For example, factor analyses of the Rey Complex Figure have resulted in a number of possible structural descriptions of cognitive constructs such as spatial abilities being identified (R. Guttman, Epstein, Amir, & L. Guttman, 1990). When the cognitive constructs tapped by a particular test are not clearly defined, it becomes difficult to relate performance on that test to theories of condition specific cognitive dysfunction (Chaytor & Schmitter-Edgecombe, 2003). Or in simpler terms, when the construct validity of a test is not known, substantial error is introduced into research using that test.

Some studies conducted in the cardiac research field used post hoc factor analysis to examine the structure of the test battery used, and in doing so, provided useful insight into how commonly used tests perform in the cardiac surgery samples. Greene and Sears (1994) examined the factor structure of a cognitive functioning test battery, which included the Wechsler Memory Scale – Form I and the Block Design and Vocabulary subtests from the Wechsler Adult Intelligence Scale – Revised. Their analysis accounted for 57% of the variance in scores on the tests with a three-factor solution. The factor solution was characterised as (1) a Cognitive Flexibility factor consisting of Block Design, Visual Reproduction, Digit Span, and Mental Control subtests forming; (2) a Retention of Verbal Information factor that included Information and Memory for Passages subtests; and (3) an Orientation factor that consisted of the Orientation subtest. Grigore, Mathew, Grocott, Reves, Blumenthal, & White et al (2001) factor analysed their research test battery and concluded that a four factor solution covering the domains of (1) immediate and delayed Verbal Memory and Language Comprehension (Short Stories from the Randt Memory Test); (2) immediate and delayed Visual Memory (Modified Visual Reproduction Test from the Wechsler Memory Scale); (3) Attention and Concentration (Digit Span from the Wechsler Adult Intelligence Scale – Revised), and (4) Visuospatial Orientation, Psychomotor, Processing Speed, and Attention (Digit Symbol subtest from the Wechsler Adult Intelligence Scale – Revised and Trail Making Test Part B). Their factor solution accounted for 83% of the variance in baseline scores. Importantly, neither Grigore, Mathew et al. (2001), nor Greene and Sears (1994), analysed postoperative data for their batteries to establish the reliability or consistency of their factor solutions across time (assessment intervals). Grigore, Mathew et al (2001) specifically addressed this shortcoming, stating that they did not analyse the battery at the post surgical interval to ensure that the structure of elements contributing to the combined score factors was the same at both intervals. While that logic may appear sound, it makes assumptions about battery structure that may not withstand rigorous scientific inquiry.

The proposition that context has important effects on a test's ability to accurately measure a phenomenon has resulted in suggestions that using clinical comparison data to interpret test scores is essential to developing specific understanding of the clinical population's cognitive phenotype (Peters et al., 2004). The collection of specific cohort normative data is an important component in developing an understanding of how a phenomenon effects a specific population. However, before such data can be collected, the constructs investigated by the tests must be fully elucidated to ensure appropriate construct validity and conceptual underpinning.

Messick (1995), in an excellent review of the meaning of validity in relation to psychometric instruments, concluded that the generalisability of test validity across settings can not be assumed. Messick (1995) described the generalisability of test constructs as "… a persistent and perennial empirical question …" and "… the reason that validity is an evolving property and validation a continuing process" (p. 741). Given the widespread recognition that construct validity is influenced by many factors, it appears that the generalisability of test validity to cardiac setting cannot be assumed. Therefore, it appears necessary to establish whether tests selected for use in the cardiac surgery context measure the same construct with the same level of accuracy as they did in the test development and initial validation context.

Over 100 different tests (including variations on tests) have been used to assess cognitive functioning in cardiac surgery research published since the release of the 1995 consensus statement. Despite the few previous efforts to establish test battery structure in a post hoc treatment, nowhere in the literature are there validation studies verifying the constructs of the individual tests, let alone the all important diagnostic efficiency statistics that facilitate our understanding of the meaning of both initial and altered test performances. The lack of definitive evidence of the structure of cognitive tests and batteries employed in research within the cardiac surgery context in turn raises the question of whether the conclusions being drawn from the methodology are in fact reliably obtained. The selection of tests for use in the cardiac surgery context is a fundamentally important methodological step in the research process. Therefore, adopting the appropriately rigorous methodological procedure of ensuring the construct validity of tests before using them in a particular context, selecting tests on the basis of prior use as suggested by Blumenthal et al. (1995) would appear to be methodologically unsound, irrespective of how many previous occasions the test had been used. Comprehensively understanding the meaning of test scores in the context of a particular theoretical rationale allows researchers to make valid interpretations of test results and in turn develop a greater understanding of the construct under investigation (Messick, 1995; Murphy & Davidshofer, 1998). At present, given the diversity and unsubstantiated nature of research methodology utilised to examine cognitive decline following cardiac surgery, this may not be feasible.

As has been sufficiently related in the cardiac surgery research literature, clinical test batteries often take several hours to administer as they attempt to assess all modalities for potential deficits (Stump, 1995). The difference between the clinical setting and cardiac research is the time available for examination. It has been stated in the cardiac surgery literature that neuropsychological assessments in cardiac research have a maximum time limit of about one hour (Stump). While the limited time to conduct assessments may be real, it does not discharge the researcher of the responsibility to use appropriate research design and methodology to ensure that their line of inquire is answered. Alternatively, it does not restrain them from refining their line of enquiry to fit the methodological parameters established by their situation. One test may be appropriate when the line of inquiry is to investigate a very specific and narrowly defined cognitive skill in isolation, although this would inevitably require knowledge that the test being used does in fact measure that skill, and as mentioned earlier there is no evidence to support the reliability and validity of neuropsychological instruments in the cardiac setting.

The core test battery suggested by the consensus group (Murkin, Newman et al., 1995), or any similar group of tests, is a battery that is easily administered within half an hour. Given that such a battery covers a range of cognitive skills, it is undoubtedly better suited to answering the general question of whether cognitive decline has occurred than any single test. Across the articles reviewed in the preceding section, 67% used at least one of the recommended core test battery tests. Interestingly, two of the studies that did not use any of the core test battery actually referred to the consensus statement's directions in their methodology. Unfortunately, in several of the studies discussed in the previous section, the batteries appear to have become too lean. That is, in some of the studies reviewed, a single test (usually the Mini Mental Status Examination) was used as the sole indicator of cognitive functioning. It is well recognised in the clinical neuropsychology field that cognitive functions, although often discussed as distinct domains, are really an array of interconnected and heavily inter-reliant skills (Spreen & Strauss, 1998; Stuss & Levine, 2002). Fortunately, in recognition of that knowledge, other studies reviewed have used extensive batteries designed to assess a broad range of cognitive domains.

Intervals

As the review in the preceding section shows, varying assessment intervals have been used in the research. Almost all of the studies discussed in the review sections incorporated a preoperative or baseline assessment to gage the pre-intervention functioning of participants. Postoperative and follow-up assessments were also almost always undertaken. Blumenthal et al. (1995), in their review of methodological issues in the assessment of post cardiac surgery cognition, comment that most studies have employed that same assessment interval format. They established that preoperative assessments were typically carried out the day before surgery, that postoperative assessments were typically carried out at seven to 10 days post-surgery, and that follow-up assessments were conducted at any time between six weeks and six months. Although minor differences are apparent between the intervals noted by Blumenthal et al. (1995), current intervals do not appear to be substantially different. Hence this aspect of the field's research methodology appears relatively stable and comparable.

Decision Methods

Methodological discussions about the use of neuropsychological tests in the cardiac surgery context have focussed on the ability of research to identify reliable individual change across time (Gill & Murkin, 1996; Kneebone, Andrew, Baker, & Knight, 1998; Murkin, Stump et al., 1997). The 1997 consensus statement (Murkin, Stump et al.) identified the importance of elucidating individual performance change over differences between group's means. The statement suggested that considering individual changes in postoperative cognitive functioning is the most sensitive means of discovering the clinical features of post cardiac surgery cognitive dysfunction (Murkin, Stump et al.). One piece of evidence in support of the impetus to focus on individual change is the knowledge that, when compared to normative data and even matched control samples, a reasonable proportion of candidates for cardiac surgery perform at abnormal levels in preoperative assessments (Selnes, Goldsborough et al., 1999). Hence, on a group analysis level, statistically significant differences may already exist between the treatment and control group (Andrew, Baker, Kneebone et al., 1998; Blackstone, 2000; Ebert, Walzer, Huth, & Herrmann, 2001; Fearn et al., 2001; Strooband, Van Nooten, Belleghem, & Vingerhoets, 2002). The individual difference view has been widely supported by agreement between researchers that any analyses using group mean scores invariably masks the true range of individual differences in the sample (Borowicz et al., 1996; Jacobson, Roberts, Berns, & McGlinchey, 1999; Jacobson & Truax, 1991; Murkin, Newman et al., 1995; Slade et al., 2001). Given these two streams of evidence, the only accurate method for determining whether change has actually occurred is to look for change within the individual rather than on a between group or within group means analysis.

Having accepted the general precept that change should be examined on an individual level, the next step is to determine the criteria that defines a meaningful change (Stump, 1995). However, as some authors have noted, this is a difficult task (Gill & Murkin, 1996; Newman, 1995). It is reasonable to expect that the proportion of patients showing decrements, improvements, or no change in performance across assessment intervals depends on the sensitivity and specificity of the decision criteria are used (Stump, 1995). Previous investigations into neuropsychological functioning post cardiac surgery have used a variety of methods to decide whether a change in functioning has occurred. Some of the methods used include arbitrary criteria such as a certain number of standard deviations change from preoperative mean level of functioning for a control group, or percentile reduction in performance on a certain number of tests in the battery (Borowicz et al., 1996; S. Newman, 1995; Slade et al., 2001). However, as demonstrated by Brown, Halligan, Wade, and Taggart (1999) who compared three different decision methods, different criteria produce different outcomes. In their analysis, Brown, Halligan et al. (1999) found that using more lenient criteria (i.e., ½ SD compared to 1 SD change) resulted in an increase in patients defined as declined from 23% to 51% at discharge and 16% to 33% at three month follow up. When using the criterion of 20% decline from baseline performance, the identification of the same patients was 48% and 31% at discharge and follow up respectively (Browne, Halligan et al.).

Despite that attention recently paid to decision criteria in the literature, there remains no consensus on the gold standard method for deciding whether or not a person has suffered a meaningful decline in cognitive functioning post surgery. This lack of consensus is adequately demonstrated in the current review where no less than seven different methods (many with several variations that were not added to the tally) were used across the studies. The fact that many different decision criteria are used in the one research field that is attempting to examine cognitive decline associated with cardiac surgery indicates a complete lack of standardisation across the field. To this end, the use of many varied methods significantly reduces the comparability of deficit incidence rates.

Many criticisms have been levelled at those methods (see Borowicz et al., 1996 for a full discussion of those points). The most pertinent of those criticisms to the current discussion is that using definitive criteria that dichotomises functions into impaired and non-impaired categories neglects the reality that premorbid and post-intervention functioning occurs along a continuum (Keith & Puente, 2002), and that changes along that continuum will produce specific effects for each individual.

Substantial research has also been devoted to the use of the statistical process known as Reliable Change (Hollon & Flick, 1988; Hsu, 1989; Jacobson, Follette, & Revenstorf, 1988; Keith & Puente, 2002; Tingey, Lambert, Burlingame, & Hansen, 1996). Hsu suggests that Reliable Change is a statistical method for concluding that the difference between results is in fact reliable. The primary reported benefit of this method is that it allows for the effects of measurement error and practice effects, outcome influences that are inherent in all psychometric assessments, to be factored into the measurement significance test (Hsu). However, to make use of those benefits, the test measurement error and practice effect for a relevant reference group must be known and the assumption must be made that the effect of practice will be the same for every individual (Sawrie, Chelune, Naugle, & Luders, (1996). Unfortunately, for many research areas where neuropsychological assessments form part of the research process, such specific normative data are not readily available (Heaton, Temkin et.al., 2001) and the full range of practice effects have not been elucidated.

As has been recognised since its earliest use, Reliable Change is not a meaningful measure of the clinical importance of outcomes. Rather it is a statistical significance measure that is akin to the *t* test (Hsu; Jacobson, Follette et al.; Jacobson, Roberts et al., 1999; Keith & Puente, 2002; Tingey et al., 1996). Determining the statistical reliability of detected differences in performance is an important step in any analysis (Maassen, 2000; Tingey et al.). However, statistical significance does not provide any indication of whether the change in performance is sufficiently important to raise suspicion or warrant further investigation (Jacobson & Revenstorf, 1988; Keith & Puente; Maassen). As such, it has also been suggested that test cut-off scores strongly correlated with group associations (such as impaired and non-impaired) are also required. However, as Hollon and Flick suggest, dichotomising subjects into dysfunctional and non-dysfunctional populations achieves nothing but a complicated and unsolvable dispute over the defining characteristics of each population, and as Cohen (1983) suggests dichotomising outcomes simply adds error around the true result.

It has been suggested that to quantitatively assess the degree of impact a treatment such as cardiac surgery has, an index reflecting the "pure magnitude of change" (Jacobson, Follette et al., 1984, p. 344) that is dependent solely on the population under investigation is required. Effect size is the measurement scale used to quantitatively describe the magnitude to which a phenomenon under study is present. When used in the context of investigations into changes in cognitive functioning due to cardiac surgical interventions, effect size equates to the degree of change in functioning. The advantage of the effect size statistic is that, irrespective of the unit of measurement used in gathering the data, the size of the outcome directly relates to the question of whether change has occurred. If no change has occurred, then the effect will be zero. If some change has occurred then the size of effect will directly represent the magnitude of the difference (Cohen, 1988).

It has been argued that even though an effect size estimate directly represents the observed difference between distributions of scores, it has no direct relevance to clinical significance, as even large effect sizes could be considered trivial by observers applying their own qualitative interpretation (Jacobson, Follette et al., 1984; Jacobson & Revenstorf, 1988). Cohen (1988) agreed with this proposition, and suggested that specifying non-validated and unstandardised (across the whole discipline) operational definitions to qualify the magnitude of the effect sizes found carries a significant risk of misapplication and misinterpretation. Follette and Callaghan (1996) also reiterated this point in their rebuttal to the proposition by Tingey et al. (1996). They suggested that the arbitrarily established effect size value of $d = .5$ while being sufficient to distinguish between impaired and non-impaired populations, provided insufficient descriptive information about the distinctive features of the dichotomies (Follette & Callaghan).

Derivation and explanation of changes in performance by way of effect size analysis appears useful to developing a greater understanding of a phenomenon under investigation. In recognition of the value of effect size estimates to research conclusions, some methodologists suggest that effect size analyses should be included in all assessments of research results where the view is to establish whether clinically significant change has occurred (Hageman & Arrindell, 1999). However, some statistical methods can be of limited use in particular research contexts (Prentice & Miller, 1992), and in general none have the ability to somehow resurrect a methodologically unconstrained piece of research and make it suitable for publication. Prentice and Miller specifically state that the utility of statistical methods unilaterally depends on the quality of the operational definition of the independent and dependent variables. That is, if the operational definition of the independent and dependent variables can easily be characterised in different ways, then the result of the statistical change analysis and any effect magnitude estimate becomes unreliable. In the review section presented earlier, it was identified that the piecemeal and methodologically unregulated approach to research in the field allows for variability in the operationalisation of independent and dependent variable definitions. Therefore, it can be suggested that until those methodological inadequacies are rectified, specifying which statistical analytical method to use and when to use it is premature.

Determining whether an individual or a patient group has suffered a clinically significant change in cognitive functioning is difficult, especially given that a single agreed upon operational definition of what represents a clinically relevant change in performance does not exist (Hollon & Flick, 1988). In recent years, a substantial literature has developed around the reliable change statistical method (Heaton, Temkin et al., 2001; Jacobson, Follete et al., 1984; Keith & Puente, 2002; Kneebone et al., 1998; Temkin et al., 1999). One reason for the development of that literature is that the reliable change method purportedly takes into consideration the effects of test reliability and practice effects when analysing for a significant difference between scores (Jacobson, Follette et al.; Kneebone, et al.). However, as suggest by several authors, the reliable change method only facilitates a decision that the change in performance observed is statistically real (or not as the case may be), it does not specify whether the detected difference in performance is clinically or functionally important (Jacobson & Revenstorf, 1988; Jacobson, Roberts, et al., 1999; Keith and Puente). Currently, no formalised and agreed guidelines exist to direct such decisions (Ryan & Hendrickson, 1998).

While many decision methods are being explored and utilised in the field, it is crucial to realise and acknowledge that simply establishing that a change has occurred is insufficient as an end point for research endeavours and clinical practice. The true meaning of a change is derived from the direction and size that it takes, as well as the degree of practical impact on the individual's ability to actually function in their individually specific daily activities at the level they did prior to undergoing surgery. To this end, none of the methods proposed by the various researchers in the field appear to be adequate without their relationship to life functioning being previously established. In the following section research outcomes associating personal and cardiac surgical factors with postoperative cognitive dysfunction, published since the release of the 1995 consensus statement, are reviewed. The review is not a critical appraisal of individual studies but rather a process of identifying general inadequacy themes currently affecting research in the field. As such, after presenting the research associated with each section, a conclusion paragraph summarises the collective methodologies of the studies discussed.

## Studies

### Emboli Studies

Circulation of emboli during cardiac surgery is widely agreed to be a major mechanism through which cardiac surgery produces cognitive dysfunction (Baker, Naser, Benaroia, & Mazer, 1995; Barbut, Hinton et al., 1994; Barbut, Yao et al., 1996; Eifert et al. 2003; Harrison, 1995; Scarborough, White, Derilus, Mathew, M. Newman, & Landolfo, 2003; Taggart, Browne, Halligan, & Wade, 1999). Generally, emboli are considered to consist of biological materials, gas, or inorganic particles. Biological material includes fat molecules and atherosclerotic debris. Fat molecules are thought to be quite small in size (10-70 $\mu m$) and to only last a short time after surgery. The effects left by the molecules, as found in autopsied tissue, have been called small capillary and arteriolar dilations (SCADS). SCADS have been considered to produce less severe decrement on cognitive functioning (Moody, Brown, Challa, Stump, Reboussin, & Legault, 1995). Atherosclerotic debris is generally considered to break from the ascending aorta during the clamping process, or from blood particles that have coagulated due to the normal physiological response occurring due to trauma to the system (Blauth, 1995; Borger & Feindel, 2002; Scarborough et al., 2003). Gas particles are thought to originate by way of several processes including ejection of residual air from heart chambers or the pulmonary veins when the aortic clamp is removed, cannulation of the aorta, introduction of bypass circuitry as trapped air, or dislodged from the blood due to turbulence ensuing from blood being pumped back into the system under pressure (perfusion) (Mitchell, & Gorman, 2002). Inorganic materials include particles of tubing from the bypass circuitry that are thought to break off due to wear and tear from the reperfusion pump (Blauth; Borger & Feindel).

Emboli are generally divided into two types according to their size. Those with a diameter larger than 200 $\mu m$ are called macroemboli, while smaller ones are called microemboli. As a general rule, the larger the embolic size and the greater the embolic load (the frequency of occurrence), the larger the neurological or neuropsychological deficits that result (Barbut, Hinton et al., 1994; Blauth, 1995). More specifically, however, it has been suggested that macroemboli may cause the overt neurological complications (stroke), that occur predominantly in the region of the middle cerebral artery (Blauth; Harrison, 1995). While microemboli are thought to produce progressively more diffuse patterns of cognitive damage depending on their size and number. That is, the smaller their size, the greater the likelihood that they will produce watershed area or isolated focal infarcts (Blauth; Borger, & Feindel, 2002; Harrison). Given that the amount, size, and frequency of emboli are recognised as important to post surgical cognitive outcome, it is understandable that considerable research attention has been paid to that association.

Some studies investigating cognitive functioning associated with the circulation of emboli have demonstrated greater declines in functioning postoperatively with greater embolic loads (Braekken, Reinvang, Russell, Brucher & Svennevig, 1998; Clark, Brillman, Davis, Lovell, Price, & Magovern, 1995; Hammon et al. 1997; Stump, Kon et al, 1996; Sylivris et al., 1998). However, the relationship between the number of emboli and the degree of cognitive decline has been found to be non-linear (Clark et al., 1995), and several studies have failed to find any statistically significant association at all (Eifert et al., 2003; Fearn et al, 2001; Jacobs et al., 1998; Mullges, Franke, Reents, Babin-Ebell, & Toyka, 2003; Neville, Butterworth, James, Hammon, & Stump, 2001).

The preceding discussion provides an example of the contradictory and inconclusive findings that are common to this field of inquiry. A possible explanation for the inconclusive findings may be the methodological differences between the studies. For example, amongst the ten studies discussed in this section, group sample sizes ranged from 12 to 203, group ages ranged from 43 years to 77 years old (M = 60 years, SD = 1.92 years). In the single study that reported education levels the range was seven years to 12 years. In addition to the sample differences, there were 33 different tests. In some cases different versions of a test were used, however, those variations were not counted in that tally. Assessment intervals also varied across the studies. Preoperative assessments, when accurately reported, were conducted one day before surgery. Postoperative assessments generally occurred between five and 12 days after surgery, with some studies conducting multiple assessments up to discharge from hospital. Follow-up assessments, where conducted, ranged from four weeks to six months, and again some studies conducted multiple follow-up assessments. In the above studies, approximately five different methods were used to establish if the change evident on testing was significant.

Extra versus Intra Cardiac Surgery Studies

Comparison of the cognitive effects of extra and intra cardiac surgery techniques has generally concluded that intra cardiac surgery techniques such as valve replacement surgery, carry more substantial effects than extra cardiac surgeries such as CABG (Andrew, Baker, Bennetts et al., 2001; Ebert et al., 2001). In their comparison of two groups, Ebert et al. found that while both groups suffered functional declines in fluency, arithmetic, and memory and learning skills in the immediate postoperative period, valve surgery resulted in greater reductions in memory functions. In terms of overall neuropsychological performance, the incidence rates of immediately postoperative cognitive decline were 57% for CABG and 71% for valve surgery. This difference was found to persist at one week follow-up assessments where the incidence rates were 19% and 36% respectively (Ebert et al,). In the study by Andrew, Baker, Bennetts et al. (2001) differences in deficit incidence rates were noted between valve and CABG groups depending on the test used. In that study, the incidence of deficits on at least one test at six months post surgery was 70% for valve surgery and 57% for CABG. When poor performance was required on at least two measures to indicate deficit, the incidence rates reduced to 40% and 27% respectively (Andrew, Baker, Bennetts et al., 2001). Despite the awareness that operation specific differences in cognitive outcomes exist, relatively little work has been undertaken to specifically explore the aetiology and outcomes of cognitive dysfunction in valve surgery (Zimpfer et al., 2003). In their study comparing biological and mechanical valve replacement outcomes, Zimpfer et al. found incidence rates of 52% and 45% at seven days after surgery and 50% and 12% at four months follow-up when using a neurophysiological measure (P300 evoked potentials) but no differences between the two surgery types on psychometric measures.

The preceding discussion again demonstrates the contradictory outcomes found in cardiac surgery research. Again, inconclusive findings may be due to the methodological differences between the studies. For example, in the three studies discussed, group sample sizes ranged from 29 to59, group ages ranged from 53 years to 73 years old (M = 64 years, SD = 6.5 years). Again, only one study reported education levels, which ranged from nine to ten years. In addition to the sample differences, there were ten different tests. Again, different versions of a test were used in different studies. As with the preceding section, those variations were not counted in that tally of tests. Assessment intervals also varied across the studies. Preoperative assessments, when accurately reported, were conducted one to two days before surgery. Postoperative assessments occurred between the second and seventh postoperative day, and follow-up assessments ranged from four to six months. In the above studies, at least two different methods were used to establish if the change evident on testing was significant.

Off-Pump Surgery Studies

Off-pump surgery, also known as beating heart surgery, is generally considered to carry fewer risk factors for poor cognitive outcome post surgery than surgery involving bypass. The reason for that conclusion is that the potential mitigating factors for positive cognitive outcome (perfusion rate, cannulation, aortic clamping, emboli, all of which are discussed in the on-pump section below) are removed from the process, allowing normal system functioning to continue comparatively unhindered (Dewey & Edgerton, 2003; Iglesias & Murkin, 2001; Murkin, Boyd, Ganapathy, Adams, & Peterson, 1999).

Despite purportedly carrying fewer risks, off-pump surgery has been associated with a preponderance of cognitive decline following cardiac surgery. However, the rate of occurrence of those declines has varied across studies, as has the reports of differences between on-pump and off-pump groups. For example, in examining post surgery outcomes, Murkin, Boyd et al. (1999) found an immediate postoperative incidence of cognitive decline of 66% in off-pump surgical patients compared to 90% for on-pump patients. When examining the same group at three months, they found the incidence to have decreased to five and 50% respectively

(Murkin, Boyd et al.). In a similar comparison, van Dijk, Jansen et al. (2002) found small non-significant differences in rates of cognitive dysfunction three months post surgery between on-pump and off-pump surgical groups (29.2% vs. 21.1% respectively), and even smaller non-significant differences at 12 month follow up assessment (33.6% vs. 30.8% respectively). Strooband et al. (2002), while demonstrating that 57% of on-pump and 63% of off-pump patients showed cognitive decline immediately postoperatively, failed to find significant differences between actual test performances for the two types of surgery. Conversely, at the six month follow-up assessment interval, they found the incidence rates to have reduced to 18% and zero percent respectively, and again, noted no significant differences in actual test performances between the groups (Strooband, et al.). Taggart et al. (1999) failed to detect any significant differences between on and off-pump groups at discharge and three month follow-up assessment. Similarly, Andrew, Baker, Kneebone et al. (1998) found no advantages between small groups of subjects undergoing either off-pump or on-pump techniques for single vessel grafts. However, they did detect small differences when comparing the off-pump single graft group and the on-pump single graft group to the on-pump multiple graft group (Andrew, Baker, Kneebone et al.).

The preceding discussion again demonstrates the contradictory outcomes found in cardiac surgery research. As demonstrated in previous examples, the inconclusive findings may be due to the methodological differences between the studies. For example, in the five studies discussed, group sample sizes ranged from seven to 50, group ages ranged from 57 years to 66 years old (M = 61 years, SD = 2.59 years), and education ranged from nine to 11 years (M = 10 years, SD = 0.7 years) in the studies that reported that data. In addition to the sample differences, there were 17 different tests. Again, different versions of tests were used in different studies. As with the preceding section, those variations were not counted in that tally of tests. Assessment intervals also varied across the studies. Preoperative assessments, when conducted or accurately reported, occurred one day before surgery. Postoperative assessments, when specified, took place between the fifth and seventh postoperative day, and follow-up assessments ranged from three to six months when conducted. In the above five studies, at least three different methods, some of which were variations of general decision method themes, were used to establish if the change evident on testing was significant.

On-Pump Surgery Studies

On-Pump, or Cardiopulmonary bypass (CPB), surgery involves the use of artificial circuitry outside of the body to temporarily replace the functions of the heart so that repairs can be made. For a cardiopulmonary bypass circuit to be advantageous it must function as if it is part of the intact cardiopulmonary system (DiNardo & Wegner, 2001). Unfortunately, and in spite of the significant advances made in such technologies, CPB cannot function exactly as the organic system does (Taylor, 1998a,b; Venn, Patel, & Chambers, 1995). This difference in functioning and the physiological actions it induces have been the basis of considerable investigation over the years since the release of the 1995 consensus statement. Several authors investigating the cognitive effects of CPB have failed to find any relationship between the use of CPB and declines in functioning (Kilo et al., 2001; Selnes, Grega, Borowicz, Royall, McKhann, & Baumgartner, 2003; Taggartet al., 1999; Wimmer-Greinecker et al., 1998). Conversely, other investigators have found that using CPB significantly increases the risk of immediate post surgery cognitive dysfunction (Bendszus, Reents, Franke, Mullges, Babin-Ebell, & Koltzenburg, 2002; Bruggemans et al., 1995; Chabot et al., 1997; Gugino, Chabot, Aglio, Maddi et al., 1997; Gugino, Chabot, Aglio, Aranki, Dekkers, & Maddi, 1999; Murkin, Boyd et al., 1999; Strooband et al., 2002; Suksompong et al., 2002; Toner et al., 1998) and more persistent medium term (up to 12 months) dysfunction (Chabot et al.; Gugino, Chabot, Aglio, Maddi et al.; Gugino, Chabot, Aglio, Aranki et al.; Murkin, Boyd et al.; Strooband et al.). In the studies that have found cognitive deficits associated with CPB, the cognitive domains that have shown effects included attention, psychomotor speed, memory, visuoperception, executive functioning, and general cognitive functioning (Bendszus et al.; Bruggemans et al.; Chabot et al.; Gugino, Chabot, Aglio, Maddi et al.; Gugino, Chabot, Aglio, Aranki et al.; Murkin, Boyd et al.; Strooband et al.; Suksompong et al..).

Various aspects of the CPB process have been investigated for their relationship to post surgery cognitive dysfunction. These include re-perfusion rates, filtering devices, blood gas management, body temperature, oxygen saturation, and anticoagulation protocols (Dewey & Edgerton, 2003; Gill & Murkin, 1996; Iglesias & Murkin, 2001; Kadoi, Saito, Goto, & Fujita, 2001; Murkin, Newman et al., 1995; Nollert et al., 1995; Roach et al., 1996; Robson, Alston, Deary, Andrews, Souter, & Yates, 2000; Taggart et al., 1999; Taylor, 1998; Yoshitani et al, 2001). In the following paragraphs two of the more pertinent of CPB process components, namely hypothermia and blood pH balance, that have been associated with cognitive outcomes are discussed.

Maintaining the correct blood pH is a crucial component of the CPB process. Maintaining the correct balance of blood gases (carbon dioxide, oxygen) generally occurs by one of two main methods. They are pH-stat and alpha-stat management. Studies investigating the differential cognitive effects of those two strategies have generally found varying results. Some studies have not found significant or important differences in the rates and degrees of impairment associated with each strategy (Engelhardt, Dierks, Pause, & Hartung, 1996; Murkin, Martzke et al., 1995), while one other group has found significant differences (Patel, Turtle, Chambers, James, S. Newman, & Venn, 1996). In the study by Patel et al., the rate of cognitive dysfunction immediately post surgery was 45.7% for the alpha-stat group and 68.6% in the pH-stat group using one criteria (deficit on two cognitive measures), and 20% and 48.6% respectively when a more stringent criteria was used. In a brief review article examining the influence of a variety of CPB related factors on cognitive functioning, Murkin (1995) suggested that pH management strategy procedural factors (i.e., cerebral autoregulation, cerebral blood follow) may be more pertinent points of investigation than the management strategy itself.

Temperature during the operating procedure has also been variously associated with alterations in cognitive functioning following cardiac surgery. Grigore, Mathew et al. (2001) and Grimm, Czerny, Baumer, Kilo, Madl, & Kramer et al. (2000) failed to find any differences in cognitive functioning associated with different temperature management strategies used during cardiac surgery. Grigore, Grocott et al. (2002) found no significant differences in post surgery cognitive performance when investigating cardiac surgery patients undergoing either slow or normal rewarming procedures. However, when assessment performance was considered in the context of a number of other predictors (i.e., diabetes, aortic clamp time, baseline cognitive functioning) the rate that patients were rewarmed was found to affect cognitive outcomes. However, in a related study from the same research centre, the association between postoperative temperature and cognitive outcome was found to be very small (Grocott et al., 2002). Conversely, Vingerhoets, Jannes et al. (1996) found a significant correlation between operative temperature and memory functioning, such that lower temperatures associated with less score change at follow-up assessment.

Contradictory outcomes are again demonstrated in the preceding discussion. Again, those findings may be due to the methodological differences in the research. In the 20 studies discussed, group sample sizes ranged from nine to 308, group ages ranged from 56 years to 73 years old (M = 62 years, SD = 4.73 years), and education ranged from ten to 13 years (M = 12.3 years, SD = 0.4 years) in the studies that reported such details. In addition to the sample differences, there were 32 different tests with many variations of single tests not being counted. Assessment intervals also varied across the studies with preoperative assessments being conducted anywhere between two weeks and one day before surgery. Postoperative assessments, when specified, took place between one day after surgery and the ninth postoperative day. Follow-up assessments ranged from one to 12 months, and up to 5 years in one study. In the above studies, at least seven different methods (not counting variations) were used to establish if the change evident on testing was significant.

<u>Biological Marker Studies</u>

There is a known association between organ damage (kidneys, pancreas, and liver) and the release of specific biochemical markers (BM) into the cerebral spinal fluid (CSF) (Johnsson, 1996). The suitability of BMs relates to the fact that organ systems are relatively homogeneous, therefore, increased levels of the markers usually retained within those systems in the CSF indicates with a good degree of reliability that damage has occurred within the associated system. The release of BMs becomes less predictive when damage occurs to the brain, due to differences between the brain and other bodily organs. Primarily, structures (i.e., cell types) and cognitive functions (i.e., language, memory) are somewhat diversely located within the brain, and direct associations between BMs and these aspects has not been established (Johnsson; Shaaban Ali, Harmer, & Vaughan, 2000). Vaage and Anderson (2001) suggest that for a biochemical marker to be considered truly useful, it should have high specificity to the brain, be rapidly deposited into the blood after injury and rapidly eliminated thereafter, be easy and inexpensive to test for, have high predictive abilities, and demonstrate a direct relationship between quantitative presence and the degree of cerebral insult suffered.

Biochemical markers of cerebral damage have been investigated in cardiac surgery populations. However, factors other than those already mentioned interact, rendering it difficulty to describe the relationship (Johnsson, 1996). For example, the use of bypass makes it difficult to obtain CSF samples through traditional methods such as lumbar puncture (Johnsson; Vaage & Anderson, 2001), and as there is limited knowledge about how well BMs transfer across the blood-brain barrier into the circulatory system, sampling them from blood is at present unreliable (Johnsson).

Johnsson (1996) carried out an extensive review of several BMs (Adenylate Kinase, Creatine Phosphokinase Isoenzyme BB, Lactate, Neuron-Specific Enolase, S100 Protein, Myelin Basic Protein, Lactate Dehydrogenase, Aspartate Aminotransferase, Glutathione, Vasointestinal Neuropeptide, and 7B2-Specific Neuropeptide) and concluded from their inconsistent use and varying outcomes that they were not yet a suitable replacement for neuropsychological assessment as a gold standard measure of cerebral damage due to cardiac intervention. However, he asserted that with further technological advancements some of those markers may become more relevant. Similarly, Shaaban Ali et al. (2000) comprehensively reviewed S100 protein research and concluded that it is potentially a suitable marker of cerebral injury, but that it requires further validation to elucidate the concentration required to be diagnostically indicative, and to isolate the topography of cerebral damage with which it is most closely associated. Since Johnsson's and Shaaban Ali et al.'s (2000) reviews two specific markers have received further investigation. They are S-100 Protein and Neuron-Specific Enolase.

S-100 protein is a calcium binding protein synthesised by astroglial cells in the central nervous system. It is usually retained within cells, but leaks out when the cells are damaged (Godet, Watremez, Beaudeux, Meersschaert, Koskas, & Coriat, 2001). S-100 protein levels have been variously associated and dissociated with patient demographics such as age (Farsak, Gunaydin, Yorgancioglu, Zorlutuna, 2003; Godet, Watremez, Beaudeux, Meersschaert, Koskas, & Coriat, 2001; Vaage & Anderson, 2001; Linsted, Meyer, Krop, Berkau, Tapp, & Zenz, 2002), type of surgery undertaken (Linsted et al.), and specific aspects of the surgical process such as number of emboli circulating and duration of hypothermic circulatory arrest (Kilminster, Treasure, McMillan, & Holt, 1999; Jonsson, Johnsson, Alling, Backstrom, Bergh, & Blomquist, 1999; Vaage & Anderson; Gibbs, Mahon, Newman, Prins, & Weightman, 2001). However, other investigations have also found strong associations between S-100 and non cerebral trauma such as broken bones, burns, and pathological conditions (Shaaban Ali et al., 2000; Vaage & Anderson). The association between level of S-100 protein and cognitive dysfunction as measured by neuropsychological scales appears to vary from non-existent to almost exact (Basile et al., 2001; Farsak, Gunaydin, Yorgancioglu, Zorlutuna; 2003; Herrmann, Ebert, Galazky, Wunderlich, Kunz, & Huth, 2000; Jonsson, et al., 1999; Kilminster et al., 1999; Rasmussen, Christiansen, Eliasen, Sander-Jensen, Moller, 2002; Robson, Alston, Deary, Andrews, & Souter, 2001).

Neuron-Specific Enolase (NSE) is an intracytoplasmic glycolytic enzyme enolase that is found in neurons and neuroendocrine tissue, and its presence in blood directly relates to neuron and neuroendocrine tissue death (Basile et al., 2001). The utility of NSE as a marker of cerebral damage in cardiac surgery has also received additional attention, though the concentration on this marker is much less. In some cases, the associations detected have been found over and above those for S-100 protein in the same subject sample, whereas in other cases they have been found to be less associated (Basile et al., 2001; Herrmann, Ebert, Tober, Hann, & Huth, 1999; Herrmann, Ebert, Galazky et al., 2000; Isgro et al., 1997; Rasmussen, Christiansen, Hansen, & Moller, 1999; Rasmussen, Christiansen, Eliasen et al., 2002). As with S-100 protein some associations with cognitive functioning have been found, but no consistent relationship has been identified.

As demonstrated in the above discussion, outcomes are again contradictory. These findings could be due, as demonstrated in previous examples, to the methodological differences in the research. In the nine studies discussed, group sample sizes ranged from 15 to 200, group ages ranged from 39 years to 68 years old, and education was not considered. Means and standard deviations were not calculatable due to the nature of the data presented in the source papers. In addition to the sample differences, there were 32 different tests with variations of single tests being used but not counted. Assessment intervals varied across the studies, with preoperative assessments being conducted anywhere between one and seven days (where detailed) before surgery. Postoperative assessments took place between day two and 14 days after surgery, with some studies using multiple postoperative assessments. Follow-up assessments ranged from two months to six months. At least five different methods were used in the preceding studies to establish if the change evident on testing was significant.

Genetic Studies

Even in the context of strong associations between the aetiological factors and cognitive dysfunction after surgery, aetiological factors do not account sufficiently for a large proportion of the variance in post surgery cognitive functioning (Haddock et al, 2003; M. Newman, Booth et al., 2001). This has led researchers to suggest that genetic predisposition may be an important contributor (Tardiff et al., 1997; M. Newman, Booth et al., 2001).

Research examining the genetic components associated with cognitive decline following cardiac surgery has been focused exclusively on the role of Apolipoprotein E (APOE) and in particular the ε4 allele. The rationale for the association between APOE and cognitive decline following cardiac surgery has been comprehensively described by M. Newman, Booth et al. (2001) and therefore, will not be extensively addressed here. In brief, however, APOE ε4 allele has been closely associated with the development of Alzheimer's type dementia. It has been closely associated with cognitive dysfunction in the acute stages after traumatic brain injury and impairment following stroke, such that those with the APOEε4 allele not only suffered higher mortality rates but also higher morbidity rates (M. Newman, Booth, Laskowitz, Schwinn, Grocott, & Mathew et al., 2001). Finally, APOE has also been associated with the metabolism of cholesterol and triglycerides in the blood (M. Newman, Booth et al., 2001). It is because of those associations that the APOEε4 allele has been investigated in relation to cardiac surgery based cognitive decline.

Excluding animal studies, until now, research on this area has been limited to just a few studies. In the earliest study, Tardiff et al. (1997) found that the presence of the APOE ε4 allele was associated with cognitive dysfunction. However, that association was strongly moderated by the effects of education. That is, over and above the effects of the allele's presence, decreasing levels of education are associated with increased risk of cognitive dysfunction. In a later closely related study by the same research group, APOE ε4 allele was found to be significantly related to psychomotor speed and attention concentration when covariates such as age and education level were controlled for prior to analyses (M. Newman, Booth et al., 2001). However, two subsequent studies failed to reliably detect any significant effects on cognitive functioning from the APOE ε4 allele (Steed et al., 2001; Robson, Alston, Andrews, Wenham, Souter, & Deary, 2002). Collectively, the results of the studies in this section indicate that direct causal association between post surgical cognitive dysfunction and the presence of genetic predictors such as the APOE ε4 allele is not currently established.

Again, the contradictory outcomes demonstrated in the preceding discussion may be due to methodological differences in the research. In the four studies discussed, group sample sizes ranged from 17 to 81, group ages ranged from 58 to 66 years old (M = 62 years, SD = 2.66 years). Education averaged 12 to 13 years in the two studies that reported those values. These details were drawn from studies that reported such information. In addition to the sample differences, there were 14 different tests with some variations of single tests not being counted. Assessment intervals also varied across the studies. Preoperative assessments, although conducted, were not described in any other studies in this section. None of the studies conducted postoperative assessments. Follow-up assessments were conducted between four weeks and one year. Five different decision methods were used in that small group of studies, with one study conducting a comparison of three different decision methods. In the above studies, one study compared of three different methods, and five different decision methods were used overall to establish if the change evident on testing was significant.

## Conclusion

Current research examining the impact of cardiac surgery on cognitive functioning appears to indicate the existence of a loose association between the two. However, in their review of the cardiac surgery literature, Haddock et al. (2003) concluded that there was a distinct lack of coherence in research outcomes due to a piecemeal and methodologically unregulated approach to research being used in the area. The preceding review, in line with Haddock et al.'s conclusions, clearly demonstrated the lack of coherence in methodologies and outcomes in the field. Therefore, it appears that clinical neuropsychological research within the cardiac surgery context requires substantial improvement before reliable and well-substantiated conclusions can be drawn about their efficiency and effectiveness. The preceding review indicates that there is a substantial range of methodologies in use and that all of these require re-investigation and clarification with a view to developing reliable assessment methodologies and a widely applicable operational definition of clinical meaningful change in cognitive functioning.

In the following chapter, a series of four studies is presented as a quantitative examination of the cardiac surgery research field. The first study examines the statistical power and effects sizes for research in the field conducted since 1995. The second study examines the facet structure of four test batteries previously used in research in the field. The third study addresses the effect sizes for each test in the core test battery. The final study examines the independence of the tests constituting the core test battery on a small sample of cardiac surgery patients.

# CHAPTER FOUR

## Research Methodology and Results

### Overall Aim

Through a series of studies addressing a number of smaller aims, the current research sought to statistically evaluate the methodology and, in particular, the neuropsychological tests and batteries used in current neuropsychological research in the cardiac surgery context. To this end, this series of studies forms the statistical counterpart to the review conducted in Chapter Three. The intended culmination of the current body of work was the clarification of instruments used to examining for cognitive decline and cardiac surgery associations.

### General Design

The protocol for the whole research project was approved by and conducted under the scrutiny of the Macquarie University Ethical Research Committee. The approval reference number was 30NOV2001-D067(JU). Four studies[3], utilising several different analytical processes, were included in the project. The first study was an analysis of statistical power and effect size values in relevant research published since the release of the 1995 consensus statement on the assessment of neurobehavioural functioning after cardiopulmonary bypass cardiac surgery. The second study was a facet analysis of test batteries used in studies previously published in the literature. Authors who had previously published their work in the area supplied the data for this study. The third study involved an effect size analysis

---

[3] Preliminary results from studies one and two were presented in 2003 at the Australian Psychological Society, College of Clinical Neuropsychologists pre-conference meeting in Perth, Western Australia.

of the consensus statement core test battery as used in research published in the field and examined in study two. The final study was a facet analysis of the core test battery tests, using data specifically gathered for the study.

## Analytical Methods

Two different statistical methods were used in the current research. The first was the calculation of Effect Size (ES) and Statistical Power (SP) values, and the second was Facet Theory (FT). Statistical package for the Social Sciences (ver. 11.1) and Microsoft Excel were used for effect size and statistical power analyses in the first study. Effect size analyses in the third study were conducted using Microsoft Excel and then confirmed by hand. SYSTAT (ver. 10) was used to conduct all facet theory analyses. The following section discusses the analytical approaches in detail.

### Statistical Power and Effect Size

Calculating statistical power for analyses is essential at both the research design and results interpretation stages (Cohen, 1990; Hair, Anderson, Tatham, & Black, 1995). In recognition of the importance of conducting statistical power analyses, the fourth edition of the American Psychological Association Publication Manual stipulates that such values should be reported in all research reports (American Psychological Association, 1994).

Conducting statistical power analyses and interpreting the results is a relatively easy task (Barlow, 1981; Cohen, 1962, 1990, 1992a, 1992b; Hammond, 1996; Rosnow & Rosenthal, 1989; Sedlmeier & Gigerenzer, 1989). Statistical Power is the probability that the null hypothesis will be rejected when it is false (Agresti & Finlay, 1997; Aron & Aron, 1994; Bezeau & Graves, 2001; Baguley, 2004; Cohen, 1962, 1965, 1988; Hair et al., 1995; Hammond; Howell, 1997; Sedlmeier & Gigerenzer; Tabachnick & Fidel, 1996). The calculation of Power requires the knowledge of three main elements, the size of the population effect for the phenomenon being investigated, the statistical significance level stipulated by the researcher, and the number of cases in the investigation. Irrespective of the inferential statistical model used (i.e., *t, F*), fixing any three of the four elements in the equation will automatically determine the fourth (Cohen, 1988, 1992a,b; Hair et al., 1995). For example, if researchers know from prior investigations that only very small effects are usually detected in their field of endeavour, and they know that they only have a budget for *x* subjects, they can easily calculate the likely power of their study. Similarly, if they know that they have *x* number of subjects and they seek to conduct a study with power equalling or exceeding .80, then they can calculate the likely effect that they will be required to detect to achieve that level.

Calculating the obtained power of research provides important insight into the overall utility of research findings. However, unless those obtained values are compared with a *benchmark* value of some widely accepted meaning, the value itself is limited. The rationale behind the benchmark adopted for the current research enterprise extended from the assumption that all research investigations into cognitive functioning deficits resulting from cardiac surgery seek to provide clinical understanding that could in turn be utilised in the preoperative risk stratification process to minimise potential harm. Cohen (1988) suggested that rejecting the null hypothesis when it is true produces worse consequences than accepting it when it is false (Cohen, 1965). The balance between those two consequences has come to be termed the "subjective relative seriousness" (Cohen 1965). Although several levels of relative seriousness have received attention in Cohen's (1988) power calculation tables, considerable use has been made of the 4:1 ratio. That is, it is roughly four times more erroneous to mistakenly reject the null hypothesis than it is to mistakenly accept it. Using that ratio, the power of the study would equal .80 (given $\alpha = .05$, $4^{*} \alpha .05 = \beta .20$. Power = $1-\beta$, therefore $1-.20 = .80$). Cohen (1965), however, also stated that the risk ratio of 4:1 and hence power of .80 was only suitable when *no other ratio* was indicated. To this, the present author would add the caveat that a ratio would be suitable *if it is theoretically or rationally defensible in the context in which it was being used.* To fulfil that caveat, it is suggested that the relative risk be based on an understanding of the real life risks associated with the dichotomous decision. In the cardiac treatment context, the real life risk is the withdrawal of a treatment option (surgery) due to the likelihood of suffering postoperative cognitive decline. If a theoretical understanding of consequences related to rejecting or accepting the null hypothesis is not available, then the rational approach of adopting a meaningful standard, such as that suggested by Cohen (1965) might be the next best option. For example, when no indicative information, such as incidence rates, is available to guide a researcher's decision, the expectation can be no more specific than a chance likelihood of decline or no decline. An intuitive researcher, might then consider the relative risk of falsely identifying a subject as likely to suffer cognitive decline following cardiac surgery, and subsequently withholding treatment, as roughly twice as erroneous as falsely identifying a subject as not likely to suffer a cognitive decline. This equates to a 2:1 ratio, which at the .05 alpha criterion produces a $\beta = .10$ and power of .90. Based on that rationale and Cohen's (1965) rationale, benchmark power values of .90 and .80 were selected for the purposes of comparing results obtained in the first study.

Effect Size was described by Cohen (1988) as "...the degree to which the phenomenon is present in the population, or the degree to which the null hypothesis is false." (p. 9-10). As such, it is an index of the degree of treatment effect. The null hypothesis represents the complete lack of the phenomenon in the population under investigation. In simple terms, when the null hypothesis is true, there is zero amount of the phenomenon present. Conversely, when the null hypothesis is false, there is some amount of the phenomenon present. As effect size values represent the degree to which the phenomenon under investigation is present, the larger the value of the effect size the more likely the alternative hypothesis becomes. In between-group designs, the value of the effect size represents the difference between the score distributions of each group when measuring from the chosen central point of one group to the corresponding point in the other group. In within-group designs, effect size values represent the difference between two time interval score distributions. In other designs, such as norm comparison studies, it represents the departure from a predetermined meaningful point. As effect size indicates the difference between distributions, it represents the degree of non-overlap between them. That is, when the null hypothesis is true, the effect size will equal zero, which directly equates to no difference between the distributions (Cohen, 1988). That is when the null hypothesis is zero, the effect size is zero indicating 100% overlap between the distributions (Zakzanis, 2001). The effect size scale ranges from zero to about 4.0. Cohen (1965) proposed the values of .2, .5, and .8 as conventions that indicate small, medium, and large differences between distributions, and those conventions have been widely adopted by researchers. However, effect sizes of those magnitudes respectively equate to 93, 66, and 53% overlap between distributions (taken from Table 1 in Zakzanis, 2001). Other effect size conventions have been proposed in the literature. For example, Bezeau and Graves (2001) describe an effect size of $d = 1.35$, representing approximately 33% overlap between distributions, as indicating a difference with clinical meaning in the greater field of Clinical Neuropsychological research. Conversely, Zakzanis (2001) has suggested that an effect size of the magnitude of $d = 3.0$, representing approximately 7% overlap between distributions, is necessary for differences to be considered clinically significant.

A significant problem in research is establishing the equality of findings between studies. That is, several studies might all investigate a single phenomenon, such as delayed memory in cardiac artery bypass graft surgery patients versus non-surgical controls, but use different scales (i.e., stories versus word lists) to collect their data. As the studies used different raw units of measurement to assess delayed memory functioning, and those different measures likely do not perfectly correlate together, the results obtained are not directly comparable. Specifically, the raw results are not comparable because they are on different scales. To make the studies comparable, Cohen suggests that a "pure estimate" (1988, p. 20) of the phenomenon's effect must be made. That is, an estimate must be made that is completely independent of the original measurement scale used. Calculating effect size achieves this requirement by standardizing on a single measurement scale the raw effects that were detected by the different tests. It does this by dividing the raw effect in original scale units of the dependent measure by the common standard deviation of the original population's distribution of scores on the measure used (Cohen, 1988).

There are a number of methods available for calculating the standardised effect size for studies, and the one chosen depends on the design of the study and the personal preference of the investigator using it (Hammond, 1996). The procedure used in study one to calculate the standardised effect size in <u>independent groups</u> (IG) designs was that illustrated by Zakzanis (2001). Zakzanis presented two measures of effect size, Cohen's *d* and Hedge's *g*. However, he suggested that Cohen's *d* was the appropriate estimate of effect size in neuropsychological research because it allows for the likely existence of heterogeneity of variance between the distributions to be compared. Repeated measures (RM) designs require a different approach when calculating the standardised effect size. In repeated measures designs the correlation of scores on the same measure at two points in time is generally higher than the correlation of scores between two groups on the same measure at one point in time. The increase in correlation reduces the standard error of the difference between the means. As the standard error of the difference across conditions is reduced, mean differences appear more substantial (Dunlap, Cortina, Vaslow, & Burke, 1996). Therefore, failure to distinguish between independent groups and repeated measures designs results in incorrectly inflated estimations of effect sizes and a false representation of the value of the phenomenon's effect (Dunlap et al.). Dunlap et al. illustrated the procedure used in the current analyses for repeated measure designs. The method adjusts for the impact of higher correlation values between test intervals and as such removes the error associated with those values. One difficulty with Dunlap et al.'s procedure, however, is that the user must know the correlation between pre and post measures. Unfortunately, across the literature, the correlation between repeated measure scores is generally not presented (Dunlap et al.). This was also the case for each published paper in the current sample of repeated measure design studies compiled for analysis. To manoeuvre around this difficulty, Dunlap et al. suggest that a reasonable estimate of the correlation can be used when it is considered representative of what could be expected to occur in the context under investigation. Anastasi (1988, as cited in Dunlap et al.) suggests that a correlation of $r_{12} = .75$ is a minimally appropriate estimate for test-retest reliability of psychometrically valid measures. Two articles (Bruggemans et al., 1997; Kneebone et al., 1998) directly dealing with the field of cognitive decline in cardiac surgery patients provided $r_{12}$ values ($r_{12} = .71$ and $r_{12} = .77$ respectively) for repeated measures designs. Averaging those correlations provided an estimated pre-post correlation of $r_{12} = .74$ for use in the current studies. As this was similar to the minimally appropriate value suggested by Anastasi (1988; as cited in Dunlap et al., 1996), an $r_{12} = .75$ was used in calculations for repeated measures designs in study one. In study three, where the Dunlap et al. (1996) formulae were also used, correlations were calculated on the raw data.

The interested reader is referred to the source articles by Zakzanis (2001) and Dunlap et al. (1996) for full explanations and examples of the two processes.

Facet Theory

To examine the structure of the neuropsychological test batteries used in research within the field, and the consensus statement core neuropsychological test battery (Murkin, Newman et al., 1995), a statistical process called Facet Theory was used. Facet theory is a flexible analytical approach that has been applied in a broad range of research disciplines (Brown, 1985). In psychological research, it has been used to develop theories about spatial abilities, child development, culture, and personality (Hans, Bernstein, & Marcus, 1985; Huismans, 1999; Huismans, & van de Vliert, 2001; Guttman et al., 1990; Guttman & Shoham, 1982; Karni & Levin, 1972; Maraun 1997; van de Vijver, 2001). It has also been applied in investigations into test structure and validity (Guttman & Zohar, 1999; Jann, 1999; Poreh & Shye, 1998). Facet theory was chosen as it enabled a preliminary examination of the relationships between tests without requiring the data collection efforts necessary to conduct a full factorial analysis. Given the poor response rate in the data collection phase of the project study, a facet theoretical evaluation seemed appropriate. Facet theory consists of two aspects (1) Facet Design and (2) Facet Analysis.

The facet design component of facet theory involves the construction of Mapping Sentence, which is a diagrammatic sentence that identifies and describes the full range of possible statements about the phenomenon to be investigated (Galliker, Weimer, & Wagner, 1995; Levy, 1993; Moreno & Carmona, 1999). Similar to constructing a theory, constructing a mapping sentence is unbounded with researchers being able to draw from any information source (Canter, 1995). If the mapping sentence constructed truly represents the phenomenon under investigation, then facet analysis of the data should result in a graphical representation of the mapping sentence. For example, a researcher may hypothesise, based on a sound theoretical underpinning drawn from the literature on the topic, that three *elements* A, B, and C combine to constitute phenomenon X, while a fourth *element* D does not form part of the phenomenon. Following that hypothesis, the researcher constructs a mapping sentence such as the one in the following example. In this example and in the mappng sentences generated in the following studies, bracketing and arrows indicate relationships between facets, while facets are presented in bold and the elements of each facet are presented in normal text.

Participant (X) will incur as a result of cardiac surgery measured as

$$\left| \begin{array}{c} \textit{Facet} \\ \mathbf{A} \\ \mathbf{B} \\ \mathbf{C} \\ \\ \mathrm{D} \end{array} \right| \begin{array}{c} \rightarrow \\ \\ \neq \end{array} \left| \text{Phenomenon (X)} \right|$$

*Figure 1*. Example mapping sentence.

With the aid of instruments specifically designed to examine for each element, the researcher collects data. After collecting sufficient data, the researcher uses facet analysis to examine the hypothesis.

Facet analysis, by way of a range of statistical models called Multidimensional Scaling (called Smallest Space Analysis in Facet Theory), represents variables (in the example A, B, C and D; and in the current research raw test scores for each case) as individual points in multi-dimensional space. The arrangement of those points in the multi-dimensional space is based on the intercorrelations between the variables. The arrangement of the points and the number of dimensions required represent the relationships between variables, such that points positioned closer together in space and with less dimensions (i.e., two dimensional space) represent closer and simpler associations than those positioned further apart or requiring more dimensions (Brown, 1985). Facet Analysis is similar to Factor Analysis in that it evaluates and ranks the associations between variables (i.e., test scores). However, it does not suffer the same limitations as Factor Analysis. For example, it is nonmetric and does not rely on the definitive rules that govern the extraction of key structural features (Karni & Levin, 1972). Factor analytical approaches utilise factor components, which are discrete divisions between groups of variables, to describe the structure of data. In contrast, facet analysis fosters continuity between variables by relying on meaningful gradients across the dimensional space. That continuity in turn ensures that many possible explanations about data relations are considered (Borg, 1993). In the example above, the researcher conceived that phenomenon X contained three elements (A, B, and C). The researcher also specifically excluded element D as a contributor to the phenomenon. If that hypothesis were correct then the multidimensional space containing the data would look like the multidimensional space in Figure 2.

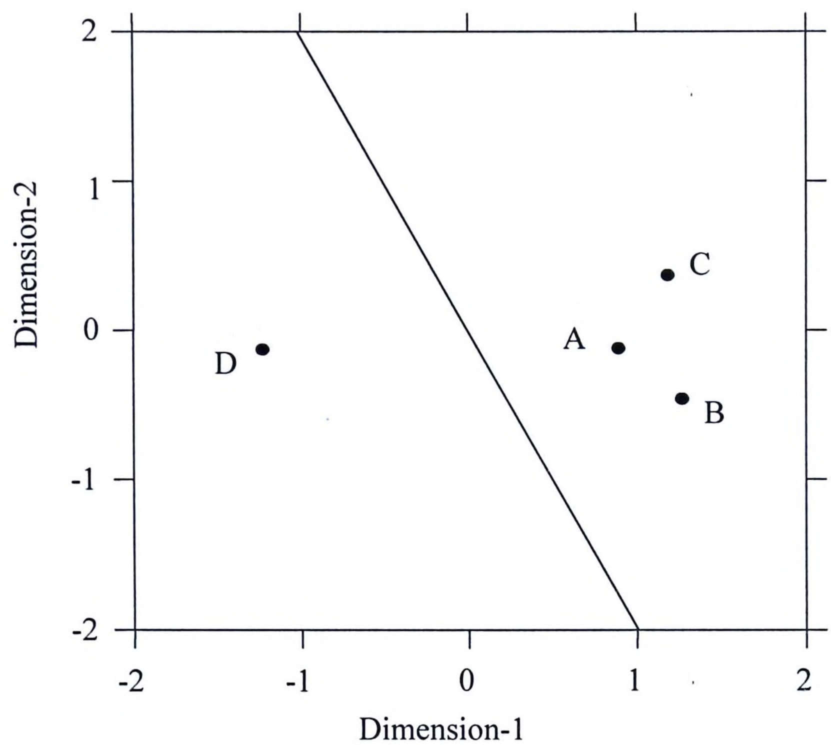


*Figure 2*. Example multi-dimensional space.

As can be seen in the space, elements A, B and C group closely together in the two-dimensional space and can be partitioned off from D, which is noticeably separated from the other intercorrelation points. Given this graphical representation, the facet representing phenomenon X, as proposed by the researcher, can be confirmed.

A key feature of facet analysis is that the spatial representation can be evaluated (Brown & Barnett, 2000; Guttman & Shoham, 1982). A measure, called the "Coefficient of Alienation" (or Attenuation), is calculaed to evaluate the degree with which the coefficient matrix and the graphical representation correspond, their *goodness of fit* (Brown & Barnett, Guttman & Shoham, 1982). As is the case generally in correlations, it ranges from 0 to 1.0. However, in this instance zero indicates that there is a perfect fit between the multidimensional graphical solution and the correlation matrix while a value of 1.0 indicates very poor fit. Therefore, the larger the value of the coefficient the less adequate the outcome obtained. The facet analysis literature does not diffinitively stipulate a set cutoff value for a significant coefficient of alienation. However, coefficient of alienation values <.15 have been widely used to indicate that the solution obtained was sufficient to describe the relationships within the data. As such, that rule-of-thumb will be applied in the current analyses.

# Study One – Statistical Power and Effect Sizes in Clinical Neuropsychological Research: The Cardiac Surgery Cohort


## Abstract

Many conclusions are drawn in research studies on the basis of analyses that may or may not have sufficient reliability for doing so. This article sought to evaluate the reliability of research into cognitive decline following cardiac surgery by examining the statistical power and effect sizes associated with the comparisons being conducted. The study consisted of a replication, within cardiac surgery related publications, of the power and effect size research conducted by Bezeau and Greaves (2001). Aim: The aim of that analysis was to establish the overall effect size and power of cardiac surgery clinical neuropsychological research. Analysis: The analytical method described by Zakzanis (2001) was used for between group comparisons. To avoid potential negative impact by higher correlations between pre and postoperative scores, the method used to derive power and effect size for within group studies was that described by Dunlap, Cortina, Vaslow, and Burke (1996). Results: The current analyses indicate that the problem of under reporting the statistical power and population effect sizes exists in cardiac surgery research publications. Analyses of the statistical power and effect size of studies revealed that, for both independent groups and repeated measures designs, the effect sizes were very small and the power of research findings was on average poor. Conclusion: It was concluded from those results that neuropsychological research into the effects of cardiac surgery requires substantial improvement to ensure that conclusions being drawn from investigations are reliably robust.

# Introduction

Statistical Power is the probability of correctly rejecting the null hypothesis when it is false (Agresti & Finlay, 1997; Aron & Aron, 1994; Bezeau and Graves, 2001; Baugley, 2004; Cohen, 1962, 1965; Hair et al., 1995; Hammond, 1996; Howell, 1997; Sedlmeier & Gigerenzer, 1989; Tabachnick & Fidell, 1996). Statistical power is part of an interdependent multi-dimensional relationship with the population effect size for the phenomenon being investigated, the statistical significance level used to conduct comparisons and the number of cases in the study. As any one of those four values can be calculated from the other three, a priori and post hoc estimation of the statistical power for a comparison are relatively easy tasks (Barlow, 1981; Cohen, 1962, 1990, 1992a,b; Hammond; Rosnow & Rosenthal, 1989; Sedlmeier & Gigerenzer). Calculating that value for each comparison in a study is an essential research design task, as it allows the research to formulate a sound research methodology. Calculating those values in the data analysis stage for the obtained data is also important, as it provides foundation information pertinent to the interpretation of research results (Cohen, 1990; Hair et al., 1995). In recognition of the importance of calculating statistical power, and the meaning given to statistical interpretations by such information, the fourth edition of the Publication Manual of the American Psychological Association stipulated that authors should routinely present information about the power of their research results (American Psychological Association, 1994).

In 1962, Jacob Cohen published his now historic power analysis study of psychological research in the Journal of Abnormal Psychology. Cohen calculated power for three hypothetical effect sizes – small = .2, medium = .5, large = .8, at $\alpha$ = .05 for observed samples. He found that for small effect sizes, the mean power of published studies was .18, while for medium and large effects the mean power values were .48 and .83 respectively. Cohen concluded that (1) the neglect of power analysis in the psychological literature was obvious, and (2) the power of studies to detect a medium effect ($d = .5$) was no better than a one in two chance (Cohen, 1962).

The basis for Cohen's proposition of $d = .5$ being a medium effect size was that it generally equated to half of a standard deviation difference between population means. Cohen's intent when specifying operational definitions for effect sizes was to clarify the statistical concept of population effect sizes, rather than to offer rigid reference points for blind usage. Cohen's definition of a medium effect size was an effect large enough to be observed by a careful investigator. His definition of a small effect was that it be noticeably smaller than a medium effect but not so small as to be trivial. His definition of a large effect was that it be as proportionately above the medium effect as the small effect size is below it (Cohen, 1992). An important caveat that Cohen (1965) placed on his operational definitions was that they would only be suitable substitutes when there was no alternate evidence about population differences available to the researcher. Then, perhaps to provide additional impetus for not unquestioningly accepting the example definitions he used to convey his point, Cohen suggested that a better course of action than accepting an arbitrary convention would be to adopt a heuristic process to identify appropriate values for the three operational definitions. To do this, he suggested that researchers base their process on a broad understanding of the phenomenon being investigated, a theoretical underpinning for the research, and past relevant research findings. Cohen also extended his warning against unilaterally accepting arbitrary conventions (even his own) for statistical power. He suggested that the convention of power = .80 be used only when “no other basis” was evident (1965, p. 98). Given Cohen’s warnings, it is clear that operational definitions for statistical power and effect sizes (small, medium, and large) are empirically dependent on the specific phenomenon being investigated and not some exemplar numbers that somehow apply to all fields of endeavour.

In a subsequent study replicating Cohen’s (1962) analysis, Sedlmeier and Gigerenzer (1989) found statistical power values of .14 for small population effects ($d$ = .20), .44 for medium effects ($d$ =.50), and .90 for large effects ($d$ =.80). These findings were remarkably similar to the results obtained by Cohen. In addition to replicating Cohen’s work, Sedlmeier and Gigerenzer compiled the results of similar research across a broad range of disciplines and found that statistical power values ranged from .10 to .55 for a small effect size ($d$ = .20), .37 to .89 for a medium effect size ($d$ = .50), and .73 to .98 for a large effect size ($d$ = .80). The wide range of statistical power values across the disciplines appears to support the proposition that the values of operational definitions for power and effect sizes are contextually specific and therefore should be established empirically.

Bezeau and Graves (2001) focussed their investigation of statistical power and effect sizes on clinical neuropsychological research. They studied 66 consecutive research articles, regardless of topic, in three major neuropsychological journals[4]. Their study sought to address five aims. The first was to identify the statistical power of clinical neuropsychological research to detect the arbitrarily

---

[4] The journals sampled by Bezeau and Greaves (2001) were the Journal of Clinical and Experimental Neuropsychology, the Journal of the International Neuropsychology Society, and Neuropsychology.

determined medium effect size of $d = .5$. Their analysis revealed that power for that level of effect ranged from .17 to .94 with a median of .45, a mean of .50 and standard deviation of .20. Their second aim was to calculate the power of the pooled studies to detect the arbitrarily defined large effect size of $d = .8$ and an effect size deemed sufficient to be of clinical relevance ($d = 1.35$). The latter value, they suggested, would be the lowest level of classification accuracy capable of providing clinically useful insight into treatment effects. Their analysis revealed that power for $d = .80$ ranged from .31 to .99 (median = .79, mean = .77, standard deviation = .19). The values of power for $d = 1.35$ ranged from .61 to .99 (median = .99, mean = .96, standard deviation = .08). The third aim was to identify the actual effect sizes present in the studies included in the analysis. Their analysis revealed that effect sizes ranged from .02 to 5.31 (median = .91, mean = 1.15, standard deviation = .84). The fourth aim was to calculate power for the *actual* average population effect size obtained. Their analysis revealed that for the effect sizes reported, power ranged from 0 to .99 (median = .93, mean = .85 standard deviation = .20). Their fifth and final aim was to describe the number of participants in the sample studies. They found that, on average, the sample studies consisted of 53.68 participants (S.D. = 38.49, range 10 to 187, median 39.5) and 29.5 neuropsychological tests (S.D. = 21.91, range 3 to 139, median 24). This equated to a mean of 2.84 subjects per test (S.D. = 3.36, range 0.26 to 20, median 1.45). Bezeau and Graves concluded that the effect sizes typically encountered in clinical neuropsychology research were larger than those reported in other disciplines. More importantly, however, they concluded that the sample sizes and statistical power employed were sufficient for the purposes of the research being undertaken. However, Bezeau and Graves emphasised that even in the context of achieving better overall performance compared to Cohen (1962) and Sedlmeier and

Gigerenzer (1989), only about three percent of studies conducted a priori power analyses, and only nine percent reported post hoc power analysis results. Despite their discouraging finding that power and effect sizes are still poorly reported in the literature, their analysis provided an important insight into the methodological integrity of clinical neuropsychological literature.

As a review, Cohen's (1962) work was seminal in identifying the inadequacies of the research methodological and reporting practices of that era. It may be expected that researchers would rectify their practices in light of such concerning revelations (Hammond, 1996). However, Sedlmeier and Gigerenzer's (1989) replication of Cohen's work revealed that more than 20 years later statistical power and population effect sizes continued to be overlooked as an essential component of research design and results interpretation. While Bezeau and Graves' (2001) work appears to indicate that there are better statistical power and population effect results for clinical neuropsychological research, their results also highlight that poor reporting of those values extends broadly across psychological research.

Rosnow and Rosenthal (1989) identified a series of faults in the "methodological spirit" of psychological research (p. 1276). The issues of liability they identified were (1) blind dedication to the reject/accept hypothesis testing approach, (2) perpetuation of lowly powered research designs, (3) raising statistical significance values (i.e., .05) to the level of descriptors, and (4) treating research findings as if one swallow *does* make a summer. In his contribution to this discussion, Hammond (1996) identified the incorrect interpretation of statistical significance tests, and the almost universal ignorance of statistical power as a means of qualifying results, as two primary inadequacies in current research practices. He attributed the misinterpretation of significance values to poor understanding of statistical methodology, and the lack of regard for the power of studies to the high cost of conducting appropriate investigations (Hammond). Rosnow and Rosenthal (1989) suggest that the knowledge base in psychological science can be improved by attending to three fundamental methodological issues. The first methodological issue they identify is de-emphasising the statistical significance test as the end of the research decision process. The second issue they highlight is the need to place greater emphasis on statistical power and population effect sizes as important points for interpretation in research. The third issue they raise is the value of replicating research to reinforce the reliability of findings.

The replication of earlier investigations conducted by Bezeau and Graves (2001) made an important contribution to our understanding of research methodology within the field of clinical neuropsychological research. However, Bezeau and Graves tempered that contribution when they acknowledged that averaging results across numerous research fields may have resulted in important cohort differences being neglected. The fact that they obtained different results to those of earlier similar research carried out in psychology and other disciplines appears to support that proposition. Therefore, exactly what constitutes appropriate operational definitions for statistical power and small, medium, and large population effect sizes remains unclear. The present study sought to extend investigations into statistical power and effect sizes by examining research within the specific clinical research cohort that examines cognitive functioning in cardiac surgery groups.

## Method

### Aims

The current research had several aims modeled on the work of Cohen (1962), Sedlmeier and Gigerenzer (1989), and Bezeau and Graves (2001). The first aim was to identify the extent to which a priori and post hoc power analyses have been reported in clinical neuropsychological research on cardiac surgery samples. The second aim was to identify the statistical power and estimated population effect sizes currently operating in clinical neuropsychological research on cardiac surgery samples. The third aim was to identify context specific small, medium, and large effect size benchmarks that can be used to plan future research within the field.

Procedure

The 1995 *Statement of Consensus on Assessment of Neurobehavioural Outcomes After Cardiac Surgery* (Murkin, Newman, Stump, & Blumenthal, 1995) could be considered a new start in the research methodology in this field, as it sought to provide a basic research structure for all studies in the field. Therefore, a search was conducted on electronic databases (Science Direct, Proquest, Ebesco Host, Psyclit, and Medline) of all journal articles published between 1996 and 2002. The search sought to identify all publications addressing cognitive functioning and cardiac surgery. The 1996 start date for the literature search was used as it allowed for the consensus statement to be published, and its recommendations to be widely disseminated amongst researchers and to be incorporated in research designs.

Database searches identified 90 papers published between the specified dates matching the broad search criteria. The papers were reviewed and then reduced to the final sample using the following exclusion criteria. Papers examining self-reports of cognitive decline were removed (n = 5). Then, abstracts and dissertation extracts were removed (n = 8). Next, studies where data collection was commenced prior to the publication of the 1995 consensus statement (n = 20) were removed. This was done in order to remove possible effects from outdated and no longer used surgical techniques (Baker, Andrew, & Knight, 2001). Several of the papers found were subsequent publications from a single research group and data collection process. Even though the research groups redefined their sub-samples in each of the articles, it was considered reasonable to suggest that the effects being detected would be relatively consistent due to all the participants emanating from the single data collection process. Therefore, to avoid redundancy in the data set, and potential bias by any single research group or cohort of cardiac surgery patients, only one publication from each research group was retained for analysis (van Dijk, Keizer et al., 2000). The paper that was retained from the group was the one that presented the most suitable data for the current analysis. Sixteen papers were removed following this exclusion criterion. Three other studies were removed because they formed the primary basis for analysis in another paper by the current author. Finally, 25 studies were excluded because no method could be found to change the results that were provided into data suitable for calculating statistical power and effect sizes for all time intervals.

As a result of the exclusion criteria $N = 13$ papers were deemed suitable for inclusion in the current analyses. Of those, $n = 7$ were suitable for analysis as independent groups designs. All were suitable for inclusion in the repeated measures design study. Table 1 in Appendix A lists the papers analysed in this study.
For both design types (independent groups and repeated measures), all possible comparisons were analysed. This was done irrespective of whether the authors treated such comparisons independently or pooled their results to answer their hypotheses. A significance level of $\alpha = .05$ was assumed for all studies, and all individual comparisons were treated as t-test analyses.

The procedure for calculating statistical power and population effect size (Cohen's *d*) used in the current analyses was that illustrated by Zakzanis (2001, discussed previously in the section headed Analytical Methods). To calculate the population effect sizes for comparisons, the means, standard deviations, and sample sizes for each group, neuropsychological test, and testing interval were entered into a Microsoft Excel database. Once the effect sizes were calculated, the statistical power was calculated for each comparison using the GPOWER program (Erdfelder; Faul; & Buchner, 1996).

As discussed in the general method section, Dunlap et al. (1996) suggest that power and effect size values will be incorrect if the analytical method used does not take into consideration the design differences. Specifically, in repeated measures designs scores on the same measure at two points in time are generally correlated. This is not generally the case in independent group's designs. The higher correlation in repeated measures approaches reduces the standard error of the difference between the means making the any mean differences appear more substantial. To protect against this Dunlap et al.'s formula was used to estimate Cohen's *d*. That formula requires the test-retest correlation for each comparison; however, as mentioned earlier such values are rarely available. Therefore, a test-retest correlation of $r = .75$ was adopted as acceptable representation of the overall correlation in the field based on the average of two related studies that showed such correlation values and the recommendation by Anastasi (1998, as cited in Dunlap et al.). Statistical power and population effect sizes were calculated using SPSS syntax (see Appendix B) specially written for this study[5].

---

[5] Special thanks to Dr. Alan Taylor, who adeptly transferred Dunlap et al.'s. process to SPSS syntax suitable for the current data.

When compared with a widely agreed upon *benchmark* value, the obtained power of research provides important insight into the overall utility of research findings. The rationale behind the benchmark adopted for the current research enterprise was addressed in some detail in the general method section. As a refresher, the benchmark was adopted on the basis that research into cognitive deficits due to cardiac surgery aims to provide clinical understanding that in turn could be utilised in the preoperative risk stratification process to minimise potential harm. Cohen (1988) proposed a benchmark value of .80 as it related to a subjective relative seriousness ratio of 4:1, for an alpha criterion of .05[6]. Cohen was cautious in specifying the arbitrary nature of the 4:1 ratio, suggesting that any other ad hoc ratio could be suitable. For the current study, the caveat, *if it is theoretically or rationally defensible in the context in which it is being used*, was added. Theoretically defensible meaning based on an understanding of the actual risks associated with the range of outcomes in the real life application of the dichotomisation, while rationally defensible means adopting a well founded standard. To do the later, one might consider the relative risk of falsely identifying a subject as likely to suffer cognitive decline following cardiac surgery, and subsequently withholding treatment, as roughly twice as erroneous as falsely identifying a subject as not likely to suffer a cognitive decline. This equates to a 2:1 ratio, which at the .05 alpha criterion produces a $\beta = .10$ and power of .90. Bezeau and Graves (2001), in their analysis of a broad spectrum of clinical neuropsychological research found that the median statistical power for population effect detected ($d = 1.35$) and sample sizes used ($N = 53$) was .93, a result similar to the value rationalised in the example above. For the

---

[6] Subjective relative risk of 4:1 at alpha .05 equals $\beta$ .20. Power = 1 - $\beta$, therefore, 1-.20 = .80.

purposes of comparing results obtained in the current study, benchmark power values of .90 and .80 were selected.

Following Cohen's (1962) suggestion that the utility of potential benchmark effect size values should also investigated, the current study analysed $d = .5$, .8, 1.35, 1.95, 2.7, 3.0, and 3.4. Those effect size values corresponded to an approximate overlap between group distributions or temporal measure distributions of 66%, 53%, 33%, 25%, 10%, 7%, and 5% respectively. The lowest two values ($d = .5$ and .8) were the values that have come to be called Cohen's medium and large effect sizes. The value $d = 1.35$ was described by Bezeau and Graves' (2001) as an effect size that could be considered clinically relevant, while the value $d = 3.0$ was described as the clinical cutoff criterion by Zakzanis (2001). The remaining effect sizes ($d = 1.95$, 2.7, 3.4) were chosen as reasonable analogies to common clinical decision base rates used in clinical neuropsychological practice.

## Results

The current study examined 13 clinical neuropsychological research studies conducted on cardiac surgery samples. Analyses were conducted separately for independent group and repeated measures comparisons. Table 1 presents the cumulative (across the studies examined) observed population effect sizes (Cohen's *d*) and statistical power values given those effect sizes. Table 2 presents the expected statistical power for various hypothesised population effect sizes. Table 3 presents the average number of tests and scores, the average number of subjects and the ratio of subjects to tests and scores. Table 4 presents the required sample sizes to obtain benchmark statistical power values, given observed and predetermined population effect sizes at alpha = .05.

Table 1

Observed Cumulative Population Effect and Statistical Power for Both Study Designs

| | IG* | | RM* | |
|---|---|---|---|---|
| | *d* | Power | *d* | Power |
| M | .31 | .24 | .14 | .41 |
| SD | .52 | .28 | .17 | .32 |
| Mdn | .17 | 1.00 | .07 | .32 |
| Min | .01 | .05 | .00 | .05 |
| Max | 4.50 | 1.00 | .90 | 1.00 |

*IG – independent groups design, RM – repeated measures designs.

Table 2

Results of Statistical Power Analyses on Predetermined Effect Sizes

| | Predetermined Effect Sizes | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $d = .5$ | | $d = .8$ | | $d = 1.35$ | | $d \geq 1.95$ | |
| | IG | RM | IG | RM | IG | RM | IG | RM |
| M | .57 | .72 | .89 | .95 | 1.00 | 1.00 | 1.00 | 1.00 |
| SD | .17 | .20 | .11 | .07 | .01 | .00 | .00 | .00 |
| Mdn | .52 | .69 | .90 | .98 | 1.00 | 1.00 | 1.00 | 1.00 |
| Min | .31 | .38 | .65 | .75 | .97 | .99 | 1.00 | 1.00 |
| Max | .88 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |

Table 3

Average Number of Tests and Scores Used in Studies, Number of Subjects, and Ratios

| | Number of | | *N* | | | | Ratios | |
|---|---|---|---|---|---|---|---|---|
| | Tests | Scores | M | SD | Mdn | Range | *N*:Tests | *N*:Scores |
| IG | 25 | 38 | 85 | 43.59 | 75 | 41 - 165 | 23.80:1 | 15.66:1 |
| RM | 45 | 95 | 96.39 | 77.15 | 76 | 16 - 308 | 27.84:1 | 13.19:1 |

Table 4

Sample Sizes Required to Achieve Benchmark Statistical Power for Obtained and Predetermined Effect Sizes at the Alpha Criterion of .05

| | | Observed Effect Sizes | | Predetermined Effect Sizes | | |
|---|---|---|---|---|---|---|
| | | IG | RM | | | |
| | | $M^{ID}_{d} = .31$ | $M^{RM}_{d} = .14$ | $d_{\text{Small}} =$ .80 | $d_{\text{Medium}} =$ 1.35 | $d_{\text{Large}} =$ 1.95 |
| Power | .80 | 268 | 1300 | 42 | 16 | 10 |
| | .90 | 366 | 1780 | 56 | 22 | 12 |

## Discussion

The non-reporting of statistical power values and population effect sizes has been extensively commented upon in the literature (Cohen, 1962, 1965, 1988, 1992; Sedlmeier & Gigerenzer, 1989; Bezeau & Graves, 2001). Despite all the commentary, many published research studies still do not provide that vital information. In a recent study by Bezeau and Graves, reporting of a priori power and effect calculations occurred in only three percent of research published in clinical neuropsychology. Similarly, post hoc analyses were only reported in nine percent of published studies. The first aim of the current study was to identify the extent to which a priori and post hoc power analyses were reported in the literature. In the current sample of 13 published studies reporting on cognitive functioning in cardiac surgery samples, none provided information on either a priori or post hoc power calculations. Additionally, none provided information on the estimated or observed effect sizes for the phenomenon under investigation or for the tests and scores used. The paucity of reporting statistical power and population effect sizes indicates that "Cohen's dilemma" continues to be an issue in clinical neuropsychological research. The current examination of the research pertaining to the cognitive outcomes of cardiac surgery samples indicates a distinct absence of a priori and post hoc statistical power and population effect size analyses results. This complete lack of reporting valuable information leaves the consumer with no means of transferring results from the research context to the clinical context (Barlow, 1981). That is, given current levels of reporting, researchers and clinicians reading the publications cannot be certain that the results observed are of practical significance when answering the central question of whether cognitive function is affected by cardiac interventions. In addition to making the interpretation of research findings more difficult, the lack of consideration for statistical power and population effects sizes leaves researchers with no contextually based indicators to facilitate the appropriate planning of future research.

The second aim of the current research was to calculate estimated population effect sizes and statistical power values occurring in clinical neuropsychological research on cardiac samples, and to heuristically examine the effects on statistical power for a range of predetermined effect sizes. The current results (see Table 1) indicate that clinical neuropsychological research on cardiac populations have a mean population effect size of $\underline{M}^{IG}{}_{d} = .31$ ($\underline{SD} = .52$) for independent groups comparisons. The standard deviation of the mean, the non-symmetrical distribution of effect sizes, and the range of values obtained (.01 – 4.50) indicate that the mean value may not be representative of the true population effect size. However, even considering the median value of the population effect size ($\underline{Med}^{IG}{}_{d} = .17$), the effects being reported in current sample of research comparisons appear to be very small. Given the mean population effect, the average statistical power of comparisons ($\underline{M}^{IG}{}_{Power}$) was .24 ($\underline{SD} = .28$). The average statistical power was well below the benchmark of .90 chosen for the current analyses. It was also well below the benchmark suggested forty years ago by Cohen (1965). In practical terms, the average statistical power of in the current research indicates that researchers will arrive at a correct rejection of the null hypothesis only 24 times out of every 100 attempts, a ratio that raises questions about the validity of research efforts within the field. Again, however, the range of values (.05 – 1.00) around the mean and the relatively large standard deviation indicate that interpreting the small mean power value may be inappropriate. Interpreting the median value, which was $\underline{Med}^{IG}{}_{Power} =$ 1.00, indicates that the statistical power of current independent groups design outcomes appears sufficiently large. The current results contrast with the findings of Bezeau and Graves (2001) in their broad investigation of clinical neuropsychological research. However, as they stated in their discussion averaging results across the whole of a research discipline likely results in a failure to observe important population specific effects. Therefore, perhaps the current disparity represents a real difference between research domains, which was regressed toward the mean in Bezeau and Graves' whole discipline analysis.

Analyses of repeated measures design comparisons found a strikingly small average population effect sizes ($\underline{M}^{RM}_{d}$ = .14, $\underline{SD}$ = .17). As with independent groups designs, repeated measures comparisons carried a broad range of population effect sizes (.00 - .90) and a large standard deviation. Examining the median population effect size ($\underline{Med}^{RM}_{d}$ = .07) indicated that very small effects are actually being demonstrated in the current sample of studies. Given the mean population effect size, the statistical power of findings was poor ($\underline{M}^{RM}_{Power}$ = .41, $\underline{SD}$ = .32). Again, a broad range of values (.05 – 1.00) and a large distribution standard deviation were obtained for power values. Examining the median value of power ($\underline{Med}^{RM}_{Power}$ = .32) in the studies did not improve the result of poor power.

Nowhere in the literature have the population effect size and statistical power for repeated measures comparisons in psychological research been investigated. Therefore, no studies exist to compare with the current findings. Comparing the repeated measures results with the independent groups analyses carried out by Cohen (1962) and Sedlmeier and Gigerenzer (1989), the population effect sizes and statistical power of the current research appear meaninglessly small. The difference between the results of prior research reviewing independent groups design studies and the current repeated measures results may reflect the different calculation processes used. However, this appears unlikely given the similarity between the results obtained in the current study for the two types of designs.

Several authors have suggested that effect sizes much larger than the values traditionally used are required for research results to be of any clinical relevance (Bezeau & Graves, 2001; Zakzanis, 2001). Given the size of the population effects currently operating in the research field, it appears that what we are measuring is "no more than negligible or trivial in size" (Cohen, 1988, p. 16-17). When the current power analysis results are included in the mix, it appears clear that larger effect sizes are also needed to ensure that results are of relevance to research as well. As such, the current findings indicate that clinical neuropsychological research examining phenomena in cardiac surgery samples is clearly not, as Bezeau and Graves said of the discipline as a whole, "better than we might have thought" (2001, p. 403).

The third aim of the present analysis was to identify context specific small, medium, and large effect sizes that could be used by researchers in designing future studies. The results (see Table 2) indicate that, assuming current research sample sizes and $\alpha = .05$, effect sizes of $d = .8$, 1.35, and 1.95 respectively represented reasonable estimations of small, medium, and large effect sizes for use in research on the cognitive functioning of cardiac surgery subjects. For both independent groups designs and repeated measures designs, when compared to the benchmark power values of .80 and .90, a population effect size of $d = 1.35$ produced sufficient power to be considered a reliable estimator. Upward extension of the effect size to $d = 1.95$ produced the maximum level of power attainable. As such, it was considered to indicate a large effect size in the current research field. Downward extension to next lowest effect size value investigated ($d = .8$) produced power values that ranged from below to above the benchmark values. Given that some of the time this effect size will produce reliable results, it was considered an adequate indicator of a small effect size in this research field.

In the research examined in the current investigation, the cognitive effects of cardiac surgery being detected are very small (see Table 1). Two possible explanations could equally well explain this outcome. It may be that the small effects being detected truly reflect the degree of functional change in cognitive abilities being experienced by cardiac patients after surgery. That is, the changes in cognitive functioning resulting from such surgery are so small that they may not have any functional correlate. Alternatively, methods currently used to assess for cognitive dysfunction after cardiac surgery may not be capable of detecting the changes that occur as a result of surgery and as such are underestimating the degree of impairment occurring. It would appear from the volume of research over the last several decades, that cognitive dysfunction after cardiac surgery remains an ever-present and pertinent clinical concern. Hence, it is reasonable to conclude that current methodologies may not be adequate for assessing those concerns. Given the divergence between results and the effort made to derive them, it would seem prudent that researchers review their research designs with a view to improving their examination of the phenomena and, subsequently, the reliability of their outcomes.

One method for improving on current methodology is to increase sample sizes in the studies undertaken. At present, the subject to variable ratio in studies appears suitable to fulfil the assumptions of the statistical significance tests being undertaken (see Table 3). However, given the parameters of current and predetermined population effect sizes, an alpha of .05, and seeking to obtain the statistical power benchmarks of .80 and .90, the average number of subjects (see Table 4) in independent groups studies would have to increase by a minimum of 350%. For repeated measures designs, an increase in by the order of 1700% would be required. Data collection processes of that magnitude would be, of course, logistically impractical to implement. However, if current sample sizes are the *practical logistical solution*, researchers must seek to alter other influential aspects of methodology to achieve results at the level required.

Another method for improving on current methodology is to improve the reliability of assessment instruments. As mentioned, the test-retest reliability of performance forms an important element of the calculation of effect size for repeated measures designs. The influence of test-retest reliability on Cohen's *d* was demonstrated by Dunlap et al. (1996) in their Monte Carlo simulation study. In that study, variance in Cohen's *d* behaved as a clear function of the correlation between measurement times. That is, the variance in *d* reduced as test-retest correlation increased. In the current analyses a value of $r_{12} = .75$ was chosen based on the recommendation of Anastasi (1998, as cited in Dunlap et al.) and two similar values reported in the cardiac surgery literature (Bruggemans et al., 1997; Kneebone et al., 1998). If all other aspects contributing to the determination of effect size were to remain the same, and test-retest correlation was increased to $r_{12} = .95$, the effect sizes being detected would more accurately reflect the true size of the phenomenon's effect on the population. Therefore, improving test-retest reliability of neuropsychological measures may be a pertinent first step in efforts to improve current research methodologies. Other possible negative influences inherent in current research methodologies include sources of obtained test score error and research design inadequacies.

Sources of error in deriving predicted scores such as measurement error and test practice effects, as well as systemic assessment biases such as age and subject's familiarity with the assessment process may affect the accuracy of measurements at each assessment point. For example, during sequential assessments, no change in test score at follow-up may actually represent a decline in performance on that measure masked by sources of error and bias. Given the potential influence of score error and systemic assessment bias, controlling their effects may improve the reliability of assessments (Bruggemans, van de Vijver, & Huysmans, 1999; Heaton et al., 2001; Jacobson & Roberts et al., 1999; Temkin et al., 1999). One potential solution is the development of statistical processes, (i.e., Reliable Change Index, standardized regression-based analyses) that take sources of bias and error into consideration when predicting performance on subsequent assessments (Bruggemans, van de Vijver, & Huysmans; Heaton et al.; Jacobson & Roberts et al.; Temkin et al.; Tulsky, Saklofske, Chelune, Heaton, Ivnik, & Bornstein et al., 2003). In the sample of papers examined in the current study, statistical processes to detect changes in performance beyond simple significance testing were not applied. Other researchers in the cardiac surgery field have used such statistical processes in their investigations (Andrew, Baker, Kneebone, & Knight, 1998, 2000, 2001; Kneebone, Andrew, Baker, & Knight, 1998). However, to date, empirically derived consensus about their utility for defining clinically significant change has not been established (Bruggemans, van de Vijver, & Huysmans; Heaton et al.; Jacobson, & Roberts et al.; Temkin et al.). Substantial validation, across the numerous tests used to assess the range of cognitive domains and capabilities in the various surgical interventions applied in cardiac surgery samples, is required before they can be uniformly applied in all areas of research.

Research design inadequacies such as insufficiently constraining group parameters according to different aspects of the various surgical processes may also influence the current findings. Previous investigations into the cognitive effects of specific aspects of surgical procedures within the cardiac surgery context indicate varied levels of dysfunction associated with the various peculiarities of surgery (Andrew, Baker, Bennetts et al., 2001; Andrew, Baker, Kneebone et al., 2001; Dewey & Edgerton, 2003; Ebert et al., 2001; Gill & Murkin, 1996; Iglesias & Murkin, 2001; Kadoi, Saito, Goto, & Fujita, 2001; Murkin, Newman et al., 1995; Nollert et al., 1995; Roach et al., 1996; Robson, Alston, Deary et al., 2000;

Strooband et al., 2002; Taggart et al., 1999; Taylor, 1998; Yoshitani et al., 2001). It is possible that the various influences of procedural differences in surgery within individual research studies influence the individual effect sizes and statistical power associated with the findings of those studies. By extension, those influences may have also affected the current findings. This being the case, further credence is added to the suggestion that general conclusions cannot be drawn about the cognitive outcomes of patients undergoing cardiac surgery unless methodological practices within the field are improved.

Some of the currently investigated papers included within group analyses. When combined for the current analyses, the within group results were similar to the between groups findings. This may have resulted from the authors of those studies not applying individual change analyses for each participant. Conducting those analyses in the current investigation was not possible, as the standard practice in published articles for group studies is not to provide data for each participant. As previously discussed, however, the use of individual change analyses has not been widely validated and as such there can be no certainty that such analyses would have lead to a different outcome to that achieved in the current inquiry.

A criticism that may be levelled at the current study is the exclusion of publications under the "redundancy" criterion suggested by van Dijk, Keizer et al. (2000). The redundancy criterion suggests that including samples that are already represented (i.e., a previous study in a series) may result in the series being overemphasized. Critics may argue that post hoc redefinition of a study population based on inherent subject criteria (i.e., from surgery type to medication type) results in a different study population and hence different population effect sizes. However, it is reasonable to propose that overall effect sizes produced by a phenomenon would include the effects produced within segments of the phenomenon. That is, the effect of undergoing treatment would include the segmented effects produced by the aspects of the intervention process itself. Whether this is the case, however, remains to be investigated once substantially more research within the field has been published.

An additional criticism, stemming in part from the first, is that only one person conducted the data search and selection process, including the exclusion of redundant papers. Unfortunately, the nature of Doctoral research is that studies are usually conducted without the benefit of grants and additional researchers. Perhaps, with the aid of additional reviewers, the papers included in the study would differ slightly to those currently included. However, by specifying stringent inclusion and exclusion criteria in the design phase of the current study, and collecting all relevant publications before commencing vetting, it was believed that negative data collection effects resulting from having only a single reviewer would be minimised. In future, however, researchers seeking to replicate the current investigation should use, where possible, multiple reviewers to ensure the validity of the data selection process.

## Conclusion

The aim of this paper was to examine research into the phenomenon of cognitive functioning in cardiac surgery samples, with a view to bolstering our understanding of the range of population effects operating. By providing measures of obtained effect sizes and power estimates, a basis is provided for future researchers to enhance their research design process, their findings, and the utility of those findings to consumers and future researchers. The current research goes some way to rectifying the paucity of information about statistical power and population effect sizes in research on cardiac surgery samples. However, it is fundamental to remember when interpreting the current findings that, as in all research endeavours, the only true measure of an outcome is its replication.

While the findings indicate trivial effect sizes and bring into question the likelihood that previous research results are accurate, they do not indicate that previous efforts have been wasted. Rather, they provide an impetus for new and more methodologically sound research endeavours in this field to find clinically meaningful effects that give credence to the subjective complaints of patients.

# Study Two – Structural Analysis of Neuropsychological Test Batteries Used in Cardiac Surgery Research.

## Abstract

Research studies are frequently undertaken using batteries of tests that historically have been considered to assess particular cognitive skills or functional domains. The assumption that a test will always assess a particular skill may not be well founded, given that few, if any of the test batteries used have been validated in the context in which they are being used. This article sought to evaluate the generalisability of validity assumption, by examining the structure of four test batteries used in previously published research. Facet analysis was used to examine the relationships between test scores at each assessment interval, to confirm whether the battery structure remained as planned by the researchers. The results of the present analysis indicate that none of the batteries analysed consistently supported the proposed methodological structure. The conclusion is drawn that test batteries must be carefully considered and validated on the specific cohort before being used in research or recommended for use in clinical contexts.

## Introduction

Cognitive decline, either temporary or permanent, is a recognised result of cardiac surgery (Ahto, Isoaho, Puolijoki, Laippala, Sulkava, & Kivela, 1999; Borowicz et al., 1996; S. Newman, Smith et al, 1987; Ross & Graham, 1993; van Dijk, Keizer et al., 2000). It is also well recognised that neuropsychological assessments can contribute to improved patient outcomes in cardiac surgery by identifying the incidence and specific clinical features of post surgery cognitive decline (Stump, Rogers et al., 1996). Given those insights, including neuropsychological assessments in cardiac surgery research is now a well-established practice (Murkin, 2001; Slade et al., 2001). Interestingly, however, the widespread use of neuropsychological assessments in that context has persisted despite a lack of consensus about the best methodology for the purpose (Blumenthal et al., 1995; S. Newman, 1995; Slade et al.; Stump, James et al., 2000). One methodological aspect that requires clarification is the construct validity of neuropsychological test batteries being used in the research.

Comprehensively understanding the meaning of test scores in the context of a particular theoretical rationale allows researchers to make valid interpretations of test results and in turn develop a greater understanding of the construct under investigation (Messick, 1995; Murphy & Davidshofer, 1998). In 1994, a group of international experts negotiated several key recommendations for conducting clinical neuropsychological research with cardiac surgery samples. The resulting consensus statement (Murkin, Newman et al., 1995) represented a significant potential methodological advancement in the investigation of the cognitive decline phenomenon in cardiac samples (Baker, Andrew et al., 2001). Point five in the consensus statement dealt with assessment methodology, and in doing so, considerably reinforced the importance of selecting the correct instruments for the purpose. In particular, point five dictated that when selecting tests for use in studies, researchers must attend to, amongst other things, the cognitive functions being assessed by each test, and the range of intellectual properties accessed by the battery (Murkin, Newman et al.; S. Newman, 1995).

A myriad of factors can influence test performance and hence the validity of assessment results. Age and IQ, for example, are well recognised as major influences on test performance (Bird et al., 2004; Murphy & Davidshofer, 1998; Peters et al., 2004; Rosselli & Ardila, 2003). Hence, test development and validation often incorporates an assessment of those potential influences. By contrast, contexts such as culture and clinical group association, which have also been recognised as possible influences on test performance (Kaufman, McLean, & Kaufman, 1995; Levav et al., 1998; Ogden, 2001; Ogden et al., 2003; Ogden & McFarlane-Nathan, 1997; Roselli & Ardila, 1991), have not been similarly investigated.

Several studies examining cognitive decline in cardiac surgery samples have found significant preoperative test performance differences between subjects requiring cardiac surgery and normal individuals (Andrew, Baker, Kneebone et al., 1998; Fearn et al., 2001; Gugino, Chabot, Aglio, Maddi et al., 1997; Keith, Puente, Marks, Malcolmson, Tartt, & Coleman, 2002; Townes et al., 1989; Vingerhoets, Van Nooten et al., 1997). Other studies have not demonstrated such differences (Kneebone et al., 1998; Shaw et al., 1987). Additional evidence apparently supporting the understanding that cardiac surgery candidates are distinctly different from normal individuals derives from research describing preoperative abnormalities detected in neurophysiological (electroencephalogram) and cerebral imaging (Single Photon Emission Computed Tomography) studies (Gugino, Chabot, Aglio, Aranki et al.; Gugino, Chabot, Aglio, Maddi et al.; Hall et al., 1999; Toner et al., 1998). The lack of conclusiveness surrounding the preoperative equality of neuropsychological tests between cardiac surgery patients and normal individuals raises the question of whether they are samples from the same population or whether cardiac surgery candidates are a distinctively different population. Despite the obvious uncertainty, the effect of context (i.e., cardiovascular disease) on the functions of specific tests remains unexplained. As a result, current assumptions about the constructs underlying neuropsychological tests may not be generalisable with empirical certainty.

The proposition that context has important effects on a test's ability to accurately measure a phenomenon, has resulted in suggestions that using clinical comparison data to interpret test scores is essential to developing specific understanding of the clinical population's cognitive phenotype (Peters et al., 2004). Recent investigations into the use of statistics such as the Reliable Change Index to discern changes in functioning have also emphasised the appropriateness of using clinical cohort specific normative data to facilitate understanding test performances (Follette & Callaghan, 1996; Jacobson, Roberts et al., 1999; Jacobson & Truax, 1991; Tingey, et al., 1996). The development of specific cohort normative data is an important component in developing an understanding of how a phenomenon affects a specific population. However, before such data can be developed, the constructs investigated by the tests must be fully elucidated to ensure appropriate construct validity and conceptual underpinning.

Factor analysis is a commonly used procedure for examining the structure of cognitive constructs, cognitive functioning theories, and neuropsychological tests and batteries. Unfortunately, in measuring cognitive constructs such as spatial abilities, factor analyses of particular instruments have resulted in a myriad of possible descriptions (Guttman et al., 1990). When the cognitive constructs tapped by a particular test are not clearly defined, it becomes difficult to relate performance on that test to theories of condition specific cognitive dysfunction (Chaytor & Schmitter-Edgecombe, 2003). Re-examination of the factor structure of commonly used neuropsychological test batteries using factor analyses has led to various models of the structure of the batteries used. This, in turn, has lead to the conclusion that relationships between performances on individual tests may change across cohorts in line with the cerebral compromise produced by the condition (Millis et al., 1999; Price et al., 2002). Some studies conducted have utilised post hoc factor analysis to examine the structure of the test battery used, and in doing so provide useful insight into how commonly used tests perform in the cardiac surgery samples. Greene and Sears (1994) examined the factor structure of a cognitive functioning test battery, which included the Wechsler Memory Scale – Form I and the Block Design and Vocabulary subtests from the Wechsler Adult Intelligence Scale – Revised. Their analysis accounted for 57.1% of the variance in scores on the tests with a three-factor solution. The factor solution was characterised as (1) Cognitive Flexibility that included Block Design, Visual Reproduction, Digit Span, and Mental Control subtests; (2) Retention of Verbal Information that included Information and Memory for Passages subtests; and (3) Orientation consisting of the orientation subtest. In analysing their test battery, Grigore, Mathew et al. (2001) arrived at a four factor solution covering the domains of (1) immediate and delayed Verbal Memory and Language Comprehension consisting of the Short Stories from the Randt Memory Test; (2) immediate and delayed Visual Memory consisting of the Modified Visual Reproduction Test from the Wechsler Memory Scale; (3) Attention and Concentration; consisting of Digit Span from the Wechsler Adult Intelligence Scale – Revised and (4) Visuospatial Orientation, Psychomotor, Processing Speed, and Attention consisting of the Digit Symbol subtest from the Wechsler Adult Intelligence Scale – Revised and the Trail Making Test Part B. Their factor solution accounted for 83% of the variance in baseline scores. Importantly, Grigore et al did not examine the structure of their battery at the follow-up assessment and thus do not provide any evidence of the consistency of the performance of their battery at different time intervals. Their reason for doing this was to ensure the structure of elements contributing to the combined score factors was the same at both assessment intervals. Results such as these provide support for the argument that the examination of tests in specific clinical cohorts is necessary to confirm their utility within that context.

Despite efforts to verify a posteriori the structure of batteries, there is still no clear understanding of the cognitive domains that should be tested in cardiac surgery contexts or the construct validity of specific tests purportedly able to assess those domains. Messick (1995) described the generalisability of test constructs as "... a persistent and perennial empirical question ..." and "... the reason that validity is an evolving property and validation a continuing process" (p. 741). Given the widespread recognition that construct validity is influenced by many factors, that test performances regularly differ between different contexts, it appears that the generalisability of test validity cannot be assured. Therefore, it appears necessary to establish whether a test measures the same construct with the same level of accuracy in the "current" context as it did in the test development and initial validation context. The lack of definitive support for the structure of cognitive tests and batteries employed in research in the cardiac surgery context clearly indicates the need for evaluation of research methodologies in the field.

## Method

### Aims

The general aim of the current study was to evaluate the relationships between neuropsychological tests when used in cardiac surgery cohorts, with a view to verifying the structure of test batteries selected by researchers. In Facet Analysis parlance, the aim of the study was to examine the intercorrelations between tests included in research batteries to support the appropriateness of the structure of the batteries as hypothesized by researchers in the field.

Procedure

A search of electronic databases (Science Direct, Proquest, Ebesco Host, Psyclit, and Medline) was conducted to identify all publications between 1996 and 2002 addressing the key areas of "cognitive functioning" and "cardiac surgery". The 1995 *Statement of Consensus on Assessment of Neurobehavioural Outcomes After Cardiac Surgery* (Murkin, Newman et al., 1995) provided a common structure for cardiac surgery research. Therefore, in order to allow the consensus recommendations to be incorporated in research efforts, 1996 was selected as the start date for the literature search.

The search identified 90 papers published between the specified dates matching the broad search criteria. The primary author or the listed correspondence person for each paper was contacted and asked to supply their raw data. Data requests were made via email or normal mail, depending on the listed contact details in the publications. Some authors had multiple publications from one research effort; therefore, 75 individual requests were made. Two further attempts were made to contact those who did not reply to the initial request. Several requests were returned unopened, and a larger number of others were not responded to at all. Three responses directly denied access to data for reasons such as legal constraints or continuing research. Seven favorable responses were received. However, only four datasets were actually supplied. It remains unclear why the other data sets were not forwarded. One of the four datasets received was later established to be from a study commenced before the consensus statement. Due to the extraordinarily poor positive response rate (3%) that dataset was retained for analysis. As there was insufficient overlap in research designs of the data received, construction of a meta-data set was not feasible. As such, the research plan was revised and each battery was analysed individually. To maintain data anonymity, datasets were randomly labelled as battery one through four.

Battery One was administered to $N = 50$ cardiac surgery patients preoperatively (at least 7 days), postoperatively (18 hours) and postoperatively (5 days). In the current study, the 18 hour postoperative assessment was not examined. It consisted of five neuropsychological instruments used to assess cognitive domains defined by the researchers as working/short term verbal memory, simple attention, psychomotor speed and coordination, and executive functioning. Those domains constitute the facets of the test battery. The tests associated with each facet (elements) were Logical Memory and Digit Span from the Wechsler Memory Scale - Revised edition, Trail Making Test Part A and Part B, and Controlled Oral Word Association Test. The mapping sentence for this battery was:

Participant (X) will incur as a result of cardiac surgery measured as

| *Facets (Elements)* | | *Range* | | *Time* |
|---|---|---|---|---|
| ***A* – Working/Short Term Verbal Memory**<br>(*a*1 – Logical Memory)<br>***B* – Simple Attention**<br>(*b*1 – Digit Span)<br>(*b*2 – TMT Part A)<br>(*b*3 – COWAT)<br>***C* – Psychomotor Speed/Coordination**<br>(*c*1 – TMT Part A)<br>***D* – Executive Functioning**<br>(*d*1 – TMT Part B)<br>(*d*2 – COWAT) | → | **No Decline** to **Some Decline** in | → | at **Preoperative** and **Post-operative** assessment intervals |

*Figure 1.* Mapping sentence for battery one.

Battery Two was administered to $N = 39$ cardiac surgery patients preoperatively (1-2 days) and at follow-up (3-4 weeks). It consisted of five computerised tasks and two standardised neuropsychological instruments. Specific cognitive domains examined were not identified in the original work. As such, each test in the battery was considered to assess a specific cognitive skill. The battery included computerised tasks of simple reaction time, choice reaction time, visual attention, visuomotor tracking, and visual spatial working memory. Digit Span backwards and Verbal Paired Associates from the Wechsler Memory Scale – Revised were also used. As each test was thought to tap into an individual cognitive skill, each was treated as an independent facet in the mapping sentence. The mapping sentence constructed for this battery was

Participant (X) will incur as a result of cardiac surgery measured as

| *Facets (Elements)* | | *Range* | | *Time* |
|---|---|---|---|---|
| ***A* – Simple Reaction Time**<br>***B* – Choice Reaction Time**<br>***C* – Visual Attention**<br>***D* – Visuo-motor Tracking**<br>***E* – Visuo-spatial Working Memory**<br>***F* – Auditory Verbal Working Memory**<br>***G* – Auditory Verbal Paired Associate Memory** | → | **No Decline** to **Some Decline** in | → | at **Preoperative** and **Post-operative** assessment intervals |

*Figure 2.* Mapping sentence for battery two.

Battery Three was administered to $N = 32$ cardiac surgery patients 9 to 15 months post surgery. It consisted of 10 neuropsychological instruments purportedly tapping the cognitive domains of executive functioning, speed of processing, attention, and learning and memory. Each of those cognitive domains was taken to constitute an individual facet in the mapping sentence. The elements of the battery constituting the facets were the Stroop Colour Word Test, Controlled Oral Word Association Test, Trail Making Test, Booklet Category Test, Grooved Pegboard Test, Digit Span and Visual Reproduction Tests from the Wechsler Memory Scale –

Revised, Paced Auditory Serial Addition Test, Symbol Digit Substitution Test, and Rey Auditory Verbal Learning Test. The mapping sentence constructed for this battery was:

Participant (X) will incur as a result of cardiac surgery measured as

| *Facets (Elements)* | | *Range* | | *Time* |
|---|---|---|---|---|
| ***A* – Executive Functioning**<br>(*a*1 – Stroop Colour Word Test)<br>(*a*2 – COWAT)<br>(*a*3 – TMT Part B)<br>(*a*4 – Booklet Category Test)<br>***B* – Speed of Information Processing**<br>(*b*1 – Grooved Pegboard Test)<br>(*b*2 – TMT Part A)<br>(*b*3 – Digit Span)<br>(*b*4 – Symbol Digit Substitution Test)<br>(*b*5 – Paced Auditory Serial Addition Test)<br>***C* – Learning and Memory**<br>(*c*1 – Rey Auditory Verbal Learning Test)<br>(*c*2 – Visual Reproduction) | → | **No Decline** to **Some Decline** in | → | at **Preoperative, Post-operative** and **Follow-up** assessment intervals |

*Figure 3.* Mapping sentence for battery three.

Battery Four was administered to $N = 130$ cardiac surgery patients preoperatively (1 day), postoperatively (7-8 days) and at follow-up (6 months). It consisted of 10 neuropsychological instruments assessing the cognitive domains of attention and concentration, verbal and nonverbal memory, language, visuospatial functions, executive functions, and motor and psychomotor speed. Again, the cognitive domains stipulated in the original research publication were adopted as the facets of the battery. The tests that constituted the elements of the facets were the Complex Figure Test, Rey Auditory Verbal Learning Test, Trail Making Test Parts A and B, Purdue Pegboard Test, Digits Test, Taps Test, Stroop Colour and Word Test, Bourdon-Wiersma Dot Cancellation Test, Line Bisection Test, Controlled Oral

Word Association Test, and Token Test. The mapping sentence constructed for this battery was:

Participant (X) will incur as a result of cardiac surgery measured as

<table>
<tr><th>Facets (Elements)</th><th></th><th>Range</th><th></th><th>Time</th></tr>
<tr><td>A – Attention and Concentration<br>(a1 – TMT Parts A and B)<br>(a2 – Digits Test)<br>(a3 – Taps Test)<br>(a4 – Stroop Colour Word Test)<br>(a5 – Line Bisection Test)<br>(a6 – Burdon-Wiersma Dot Cancellation Test)<br>B – Verbal and Nonverbal Memory<br>(b1 – Digits Test)<br>(b2 – Taps Test)<br>(b3 – Rey Auditory Verbal Learning Test)<br>(b4 – Rey Complex Figure Test)<br>C – Language<br>(c1 – Token Test)<br>(c2 – Controlled Oral Word Association Test)<br>D – Visuo-Spatial Functions<br>(d1 – Line Bisection Test)<br>(d2 – TMT Parts A and B)<br>E – Executive Functions<br>(e1 – Stroop Colour Word Test)<br>(e2 – TMT Part B)<br>F – Motor and Psychomotor Speed<br>(f1 – Purdue Pegboard Test)</td><td>→</td><td>No Decline<br>to<br>Some Decline<br>In</td><td>→</td><td>At<br>Preoperative, Post-operative<br>and<br>Follow-up<br>assessment intervals</td></tr>
</table>

*Figure 4.* Mapping sentence for battery four.

## Results

Tables showing the intercorrelations matrices of Pearsons coefficients for the interactions amongst the tests within each research battery (elements), which form the data for the current analyses, are presented as Tables 1 though 8 in Appendix C. As the intercorrelations form the data for each facet analysis, values presented in them are not interpreted. The results of each analysis, the derived multi-dimensional space, are shown in the following figures. Explanation and interpretation of the results for each analysis are presented after each figure. In the multi-dimensional spaces presented, straight lines in the represent the partitioning of data on the basis of the structural elements proposed in the mapping sentence for the battery. Bolded text represents the tests that combine to make the facet element proposed in the mapping sentence. Text that is not bolded represents the tests that could not be partitioned into the predefined elements.

The multi-dimensional spaces presented in Figure 5 and Figure 6 respectively represent the structure of the battery one at preoperative and postoperative assessment intervals. Battery one was originally proposed to contain four facets. The elements (tests) that constituted the battery were Trail Making Test Part A (TMTA), Trail Making Test Part B (TMTB), Logical Memory I (LM1), Logical Memory II (LM2), Controlled Oral Word Association Test (LF), and Digit Span (DS).

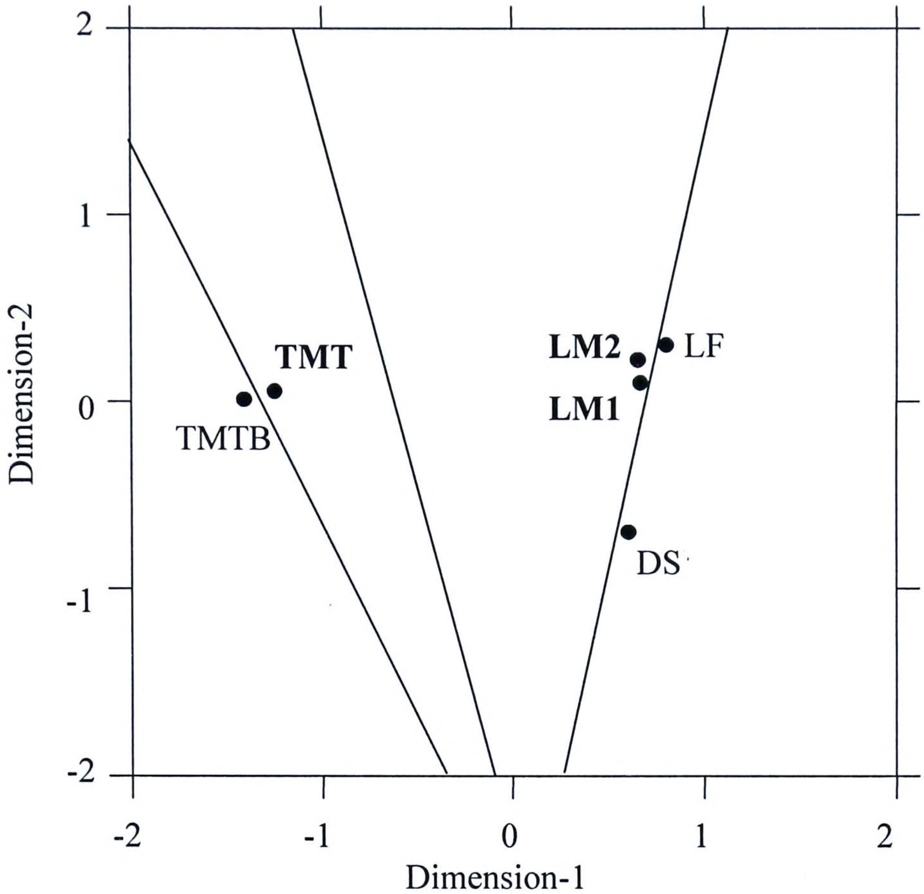


*Figure 5.* Multi-dimensional space for battery one at the preoperative assessment interval.

Facet analysis of preoperative data for battery one resulted in a two-dimensional spatial representation (see Figure 5) being achieved with adequate attenuation (<.15). That is, the two-dimensional representation was the simplest suitable representation of the data. It can be observed in that spatial representation that only two of the four proposed facets were elicited at the preoperative assessment interval. The facets that could be partitioned out were psycho-motor speed and working/short term memory. The elements constituting the other two facets (attention, executive functions) did not group closely enough in the two-dimensional spatial representation to be considered associates and subsequently facets. The results of the preoperative battery facet analysis indicated that the battery's performance did not match the research design.

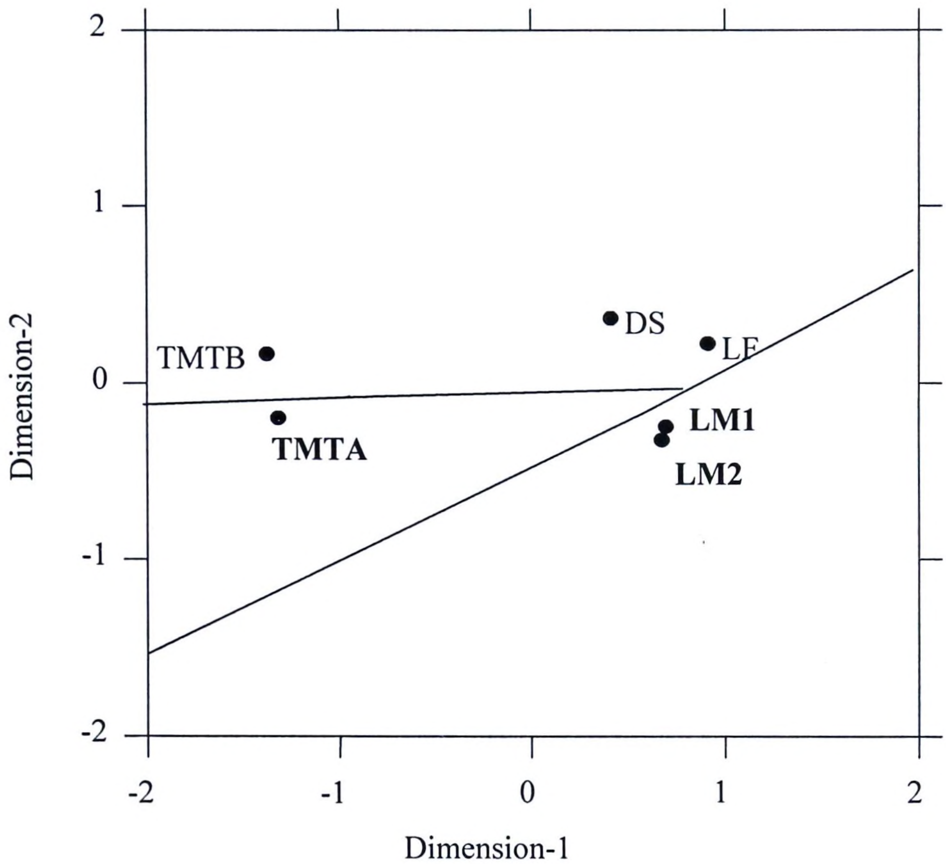


*Figure 6.* Multi-dimensional space for battery one at the postoperative assessment interval.

Facet analysis of postoperative data for battery one also resulted in a two-dimensional spatial representation (see Figure 6) being achieved with adequate attenuation (<.15). It can be observed in that spatial representation that, as with the preoperative data, only two of the four proposed facets were derived. The facets that could be partitioned out were psycho-motor speed and working/short term memory, the same ones as those partitioned out in the preoperative data. Again, the other elements did not group closely enough together to produce the remaining facets proposed. As with the preoperative data, the results for the postoperative data indicate that the battery does not function as originally proposed in the study methodology.

Importantly, both the preoperative and postoperative results produced the same facet structure. Producing the same structure indicates that the structure of the battery is relatively robust over serial assessments.

The multi-dimensional spaces presented in Figure 7 and Figure 8 respectively represent the structure of the battery two at preoperative and follow-up assessment intervals. Battery two was originally proposed to contain seven facets, each of which was based on one neuropsychological instrument. The facets were computerised simple reaction time – right and left hands (CSRTR and CSRTL), computerised choice reaction time (CCRT), computerised visual attention (CVFALU, CVFARU, CVFALL,CVFARL), computerised visuo-motor tracking (CVT), visuo-spatial working memory (VSWM), auditory verbal working memory – Digit Span backwards (DSB), auditory verbal paired associate memory (VPA).

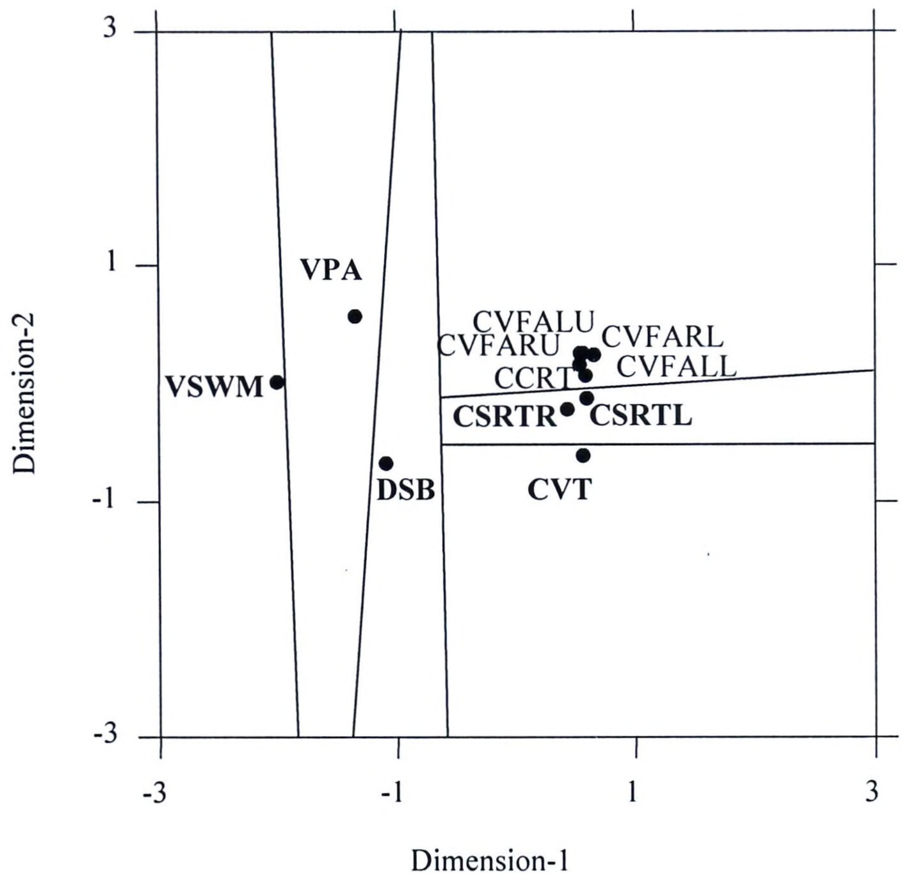


*Figure 7.* Multi-dimensional space for battery two at the preoperative assessment intervals.

Facet analysis of preoperative data for battery two also resulted in a two-dimensional spatial representation (see Figure 7) with adequate attenuation (<.15). In this instance, only five (bolded in figure) of the seven facets represented in the mapping sentence could be partitioned within the space. The facets that could be partitioned separately were computerised simple reaction time, computerised visuo-motor tracking, visuo-spatial working memory, auditory verbal working memory, auditory verbal paired associate memory. Computerised choice reaction time and computerised visual attention could not be partitioned separately, indicating that they closely resemble one another when used in the cardiac surgery context at the preoperative time interval.

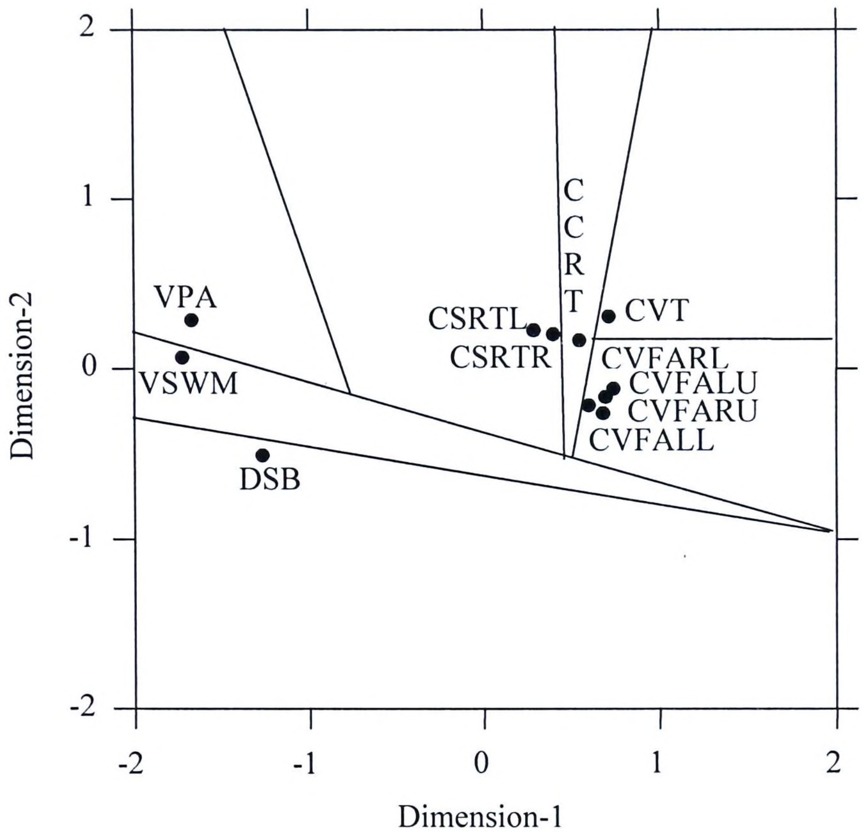


*Figure 8.* Multi-dimensional space for battery two at the follow-up assessment interval.

The two-dimensional spatial representation derived for battery two follow-up assessment data (see Figure 8) could, however, be partitioned into all the facets proposed in the study design and presented in the mapping sentence.

As shown in the preoperative and postoperative spatial representations (Figures 7 and 8 respectively), five of the facets proposed in the mapping sentence were consistently separately partitioned. Therefore, only those aspects of the original research methodology appear relatively temporally robust.

The multi-dimensional space presented in Figure 9 represents the structure of battery three at the follow-up assessment interval. Battery three purportedly contained three facets, executive functioning, speed of information processing and learning and memory. The tests constituting those facets were (SCWTI = Stroop Colour Word Test, HCT = Halstead Category Test, TMTB = Trail Making Test Part B, TMTA = Trail Making Test Part A, LF = Controlled Oral Word Association Test, GPL = Grooved Pegboard left hand time, GPR = Grooved Pegboard right hand time, AVLT1 = Rey Auditory Verbal Learning Test Trial 1, AVLT6 = Rey Auditory Verbal Learning Test Trial 6, DSB = Digit Span Backward, PASA2 = Paced Auditory Serial Addition Test two second interval, PASA4 = Paced Auditory Serial Addition Test four second interval, SDS = Symbol Digit Substitution Test, VRDR indicates Visual Reproduction delayed recall.

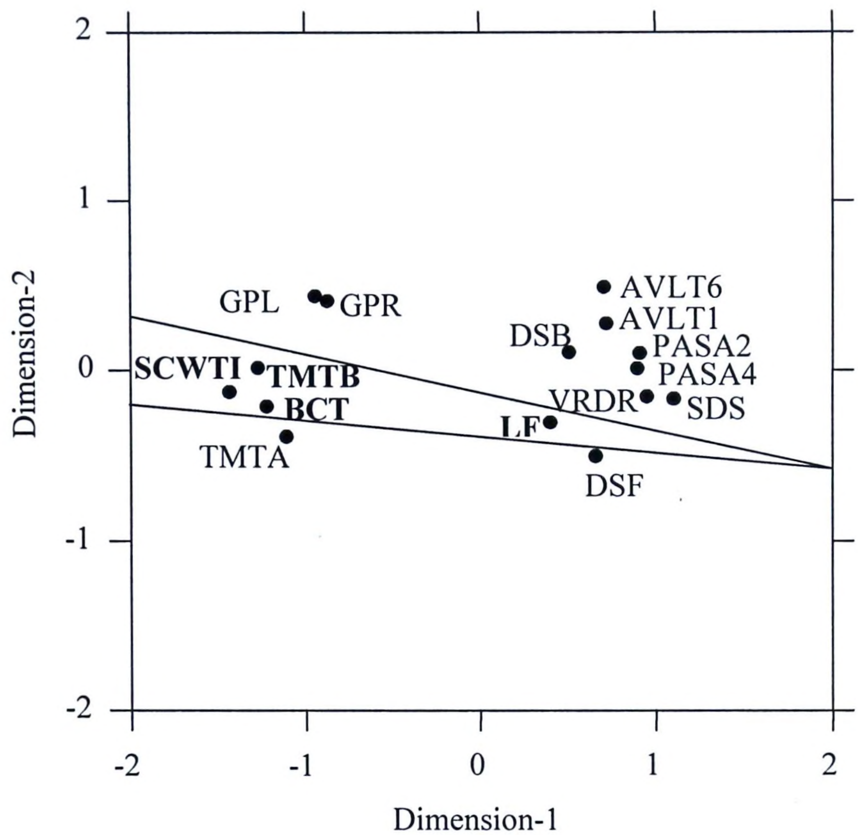


*Figure 9.* Multi-dimensional space for battery three at the follow-up interval.

Facet analysis of follow-up data for battery three resulted in a two-dimensional spatial representation with adequate attenuation (<.15). The dimensional space derived (see Figure 9), however, could only facilitate the partitioning of one facet, executive functioning. The remaining elements did not sufficiently associate together within the space to be considered cohesive facets.

The multi-dimensional spaces presented in Figure 10, Figure 11 and Figure 12 respectively represent the structure of the battery four at preoperative, postoperative and follow-up assessment intervals. Battery four was originally proposed to contain six facets. They were attention and concentration, verbal and nonverbal memory, language, visuo-spatial functions, executive functions, and motor and psychomotor speed. The elements (tests) comprising those facets were the Trail Making Test – Part A (TMTA) and Part B (TMTB), Digits Test (ODST), Taps Test (BTT), Stroop Colour Word Test – interference score (SCWTI), Line Bisection Test (LNBS), Bourdon Wiersma Dot Cancellation Test (BDC), Rey Auditory Verbal Learning Test – total recall (AVLTT) and delayed recall (AVLTD), Rey Complex Figure Test – immediate recall (CFTI) and delayed recall (CFTD), Token Test (TT), Controlled Oral Word Association Test – total score (LF), Purdue Pegboard Test (PPB30).

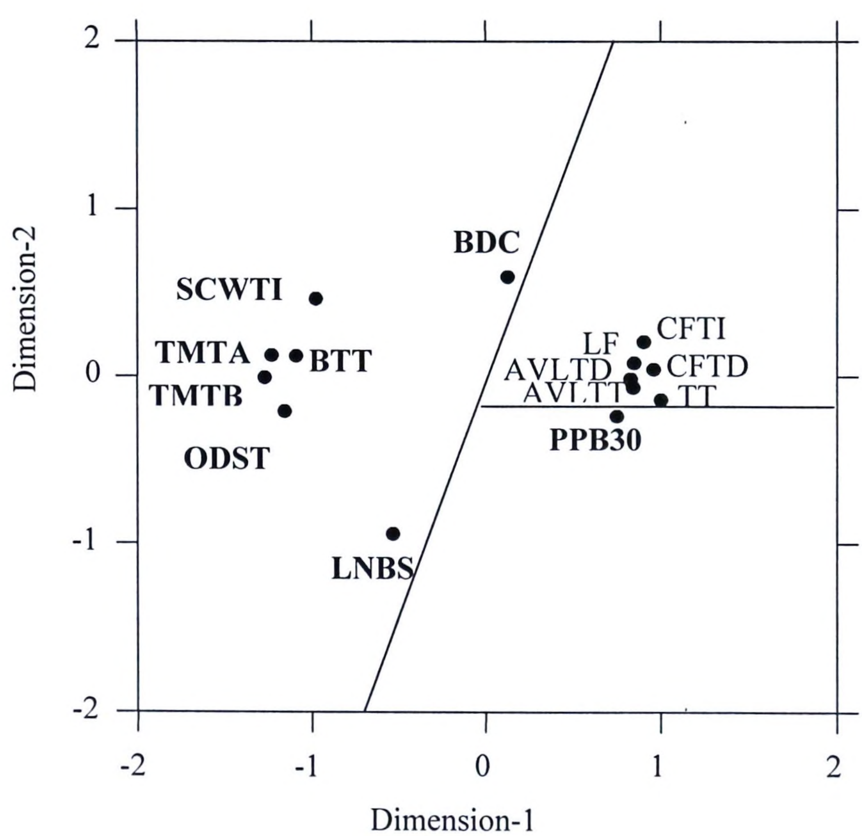


*Figure 10.* Multi-dimensional space for battery four at the preoperative assessment interval.

Facet analysis of preoperative data for battery four resulted in a two-dimensional spatial representation (see Figure 10) with adequate attenuation (<.15). The mapping sentence for Battery four initially proposed six facets, however, the spatial representation of the relationships between the elements supported only two (bolded in figure) of the originally proposed facets. The facets that could be partitioned separately were attention and concentration, and motor and psychomotor speed.

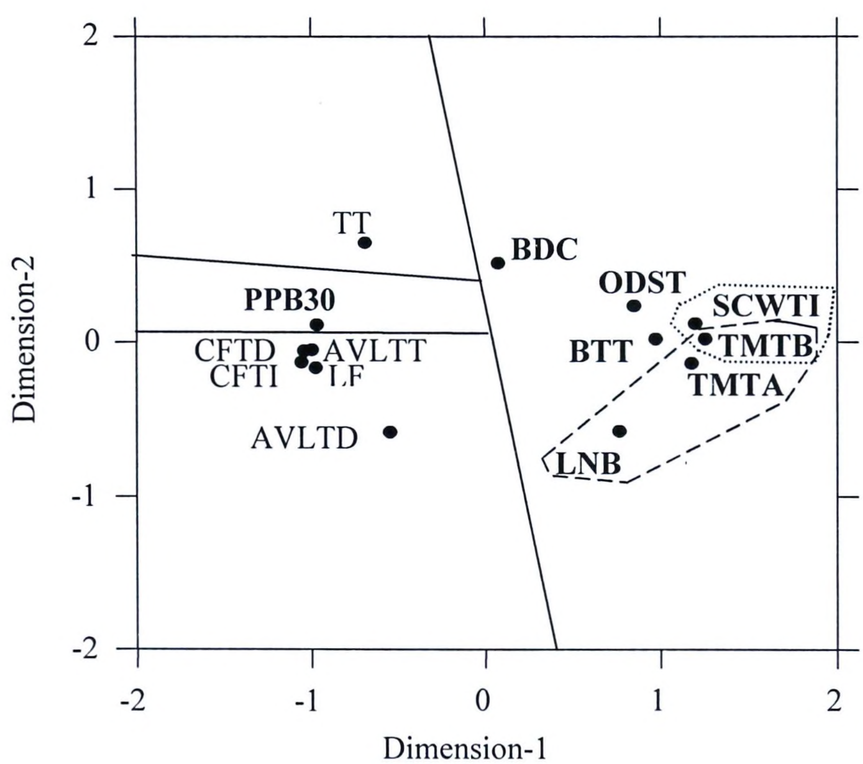


*Figure 11.* Multi-dimensional space for battery four at the postoperative assessment interval.

Facet analysis of postoperative data for battery four also resulted in a two-dimensional spatial representation (see Figure 11) with adequate attenuation (<.15). As with the preoperative data space, the facets that could be partitioned separately were attention and concentration, and motor and psychomotor speed. Elements constituting other proposed facets in the mapping sentence were identified as clusters within the partitioned space defined as attention and concentration. The clusters identified were executive functions and visuo-spatial functions.

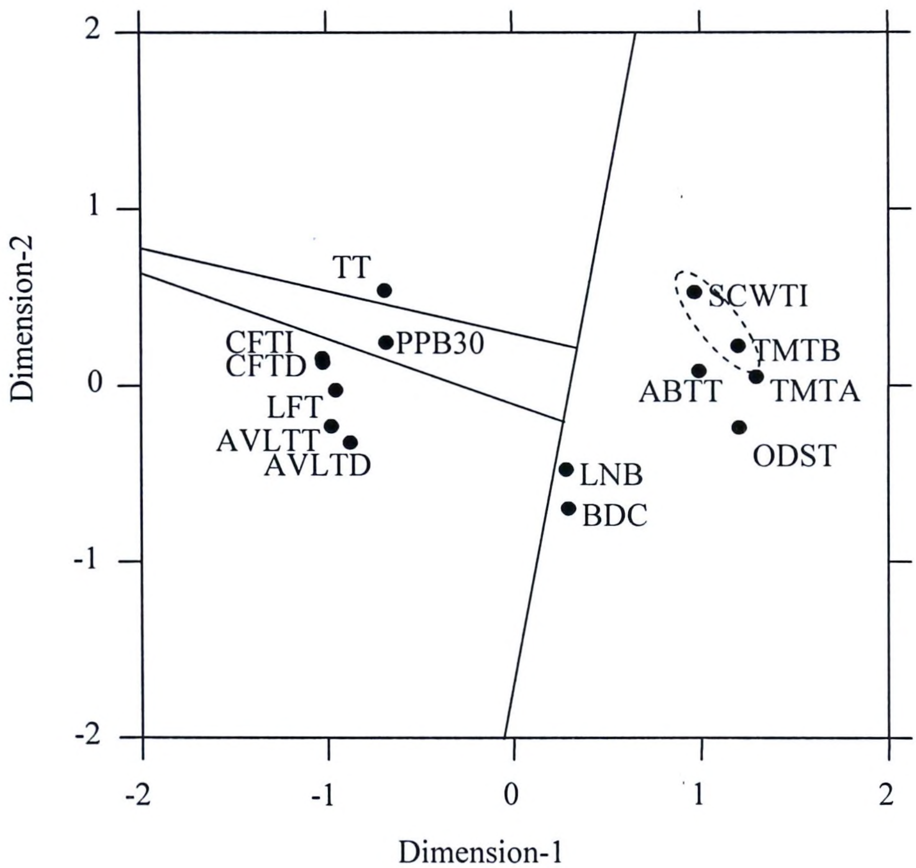


*Figure 12.* Multi-dimensional space for battery four at the follow-up assessment interval.

Facet analysis of battery four follow-up data resulted in a two-dimensional spatial representation (see Figure 12) with adequate attenuation (<.15). As with the results of the previous assessment intervals for battery four, only attention and concentration, and motor and psychomotor speed facets could be partitioned. As occurred in the postoperative spatial representation for battery four, elements constituting the proposed executive functions facet clustered together within the partitioned space for the attention and concentration facet.

As shown in the preoperative, postoperative and follow-up spatial representations (Figures 10, 11 and 12 respectively), only two of the facets proposed in the mapping sentence were consistently separately partitioned. Therefore, only those aspects of the original research methodology appear relatively temporally robust. One other facet was identified at two intervals; however, that facet was well encapsulated within another facet that was consistently partitioned on each occasion. The partitioning of one facet within another may indicate an interdependency of the skills behind the tasks that purportedly examine those facets. Interestingly, this appears to be the case in the mapping sentence describing the battery methodology where some tests are described as belonging simultaneously to several facets.

## Discussion

Researchers and authors who examine and report on cognitive decline following cardiac surgery have stressed the importance of selecting batteries based on an understanding of the cognitive functions assessed by the individual tests and the range of functions assessed as a whole (Blumenthal et al., 1995; Gill & Murkin, 1996; Murkin, Newman et al., 1995; Murkin, Stump et al., 1997; Slade et al., 2001; Stump, 1995). Despite those suggestions, however, previous efforts to examine the cognitive effects of cardiac surgery have used neuropsychological tests that have not been specifically validated for use in that context. In the few instances (Green & Sears, 1994; Grigore, Mathew et al., 2001) where the structure of research test batteries have been analysed, albeit in a post-hoc fashion and only for preoperative assessment data, the analyses have not provided confirmation of the structural reliability of the battery used. Rather, in each instance, the battery has been reducible to a smaller structure of combined scores drawn from combinations of tests.

Given that there is no published literature describing the construct validity of neuropsychological tests in cardiovascular context, the aim of the present study was to examine relationships between tests used in research batteries with a view to identifying whether the structure of those batteries was as they were designed. The facet analysis method used in the current examination was adopted as it provided an opportunity to investigate the structure of relationships between tests without being constrained by exacting assumptions and extraction rules normally associated with methods such as confirmatory factor analysis.

Regarding the structure of the test batteries in the current analyses, all five test batteries failed to conform to the predefined methodological structure at the preoperative assessment interval. For each battery, the original researchers reported that they chose tests because they were well-recognised measures of specific cognitive domains. However, as demonstrated in the mapping sentences, two of the four researchers assigned individual tests to multiple domains, and at the postoperative and follow-up intervals for Battery Four, those measures formed elements within multiple facets. Additionally, a number of purportedly different tests strongly associated together in some analyses, contradicting the methodology proposed by the researchers. The results of the current examination indicate two important points about neuropsychological research in the cardiac surgery context. Firstly, test batteries currently being employed to answer specific questions do not always comply with the structure detailed in the methodology of the studies. Therefore, researchers could not be certain that what they proposed to measure was what they actually measured and therefore they can not be certain that their conclusions based on those results are empirically sound. Secondly, the structure of some batteries, in instances where assessments occurred at multiple intervals, did not maintain the same structure on both occasions. Therefore, even if tests batteries conformed to design methodology initially, evidence exists to suggest that the constructs or domains being assessed within the battery may not be sufficiently temporally robust to facilitate sound interpretation of changes in performance that are purportedly being detected.

A possible explanation for the current findings is that while the tests used in each battery purportedly measure specific cognitive skills or domains in the context in which they were developed or normalised, when used in the cardiac surgery context they may not be measuring the same cognitive skill. The unplanned association between tests in the current analyses appears to support the importance of the role that context plays in interpreting the results of neuropsychological assessments. That is, when used in a different context to normal functioning individuals, such as cardiac surgery research, the tests may not actually be accessing the skills they were originally designed to assess.

Although any number of influences may affect performance on a test, perhaps the most explanatory influence in the context of cardiac surgery is that people requiring cardiac interventions may be a distinctly different population to normal healthy individuals. Support for this conclusion derives from the evidence that at the preoperative assessment, cardiac surgery patients perform in a manner that was significantly different to normal controls. Research examining people with cardiovascular disease has both supported and negated the proposition that people suffering cardiac conditions also have reduced cognitive functioning as a result of those conditions (Ahto et al., 1999; Ross & Graham, 1993). Additionally, studies comparing cardiac surgery cohorts with normal control subjects have found significant differences on cognitive tests during preoperative assessments. While preoperative assessments are often considered only as baselines, they may allude to a fundamental cognitive functioning difference between people requiring cardiac interventions and those who do not.

The current study contains a number of limitations that may influence the findings and subsequent interpretations. Firstly, the poor positive response rate for data requests (3%) may have resulted in an unrepresentative example of the neuropsychological test batteries currently being used in the research. To some extent, examining datasets individually may have controlled that difficulty. The current analysis was considered a valid process for evaluating the test batteries in the absence of the ability to fulfill the assumptions associated with conducting factor analyses. Secondly, there was little crossover in tests included in the test batteries investigated. While this may be representative of the test diversity in the broader research field, it makes generalizing from the current results difficult. The current results certainly imply that methodologically all may not be as planned. However, to verify this, additional analyses replicating the current findings will be required. Alternatively, future research could be undertaken to clearly elucidate the structure of test batteries being used.

Consistent construct validity of a test battery is important if researchers are to draw valid conclusions about the cognitive phenotypes associated with the application of specific cardiac interventions. Results of the current analysis demonstrate that in several instances, the structure of the test battery employed was not as originally proposed and that it varied across assessment intervals. The lack of consistency across assessment intervals in the current research indicates that researchers should be cautious when drawing conclusions about the cognitive domains affected by cardiac procedures. More importantly, the current results highlight the necessity of undertaking context specific validation of tests to ensure the methodological correctness of their use and an accurate conceptual understanding of the results obtained from using them.

## Conclusion

At present, neuropsychological assessments of cardiac surgery samples are undertaken with the belief that the instruments being used will function in the same way with the research sample as they do in generally unrelated test development and validation samples. Results of the current analyses, however, would seem to indicate that this might not be the case in cardiac surgery research context. Melling (2001) suggests that by observing constant conjunctions between events, general laws or theories can be generated about phenomenological happenings. However, as Levy (1993) correctly asserts, "… theory and method are inseparable in the process of theory construction…" (p. 260). Therefore, while there appears to be ample evidence in the literature to support the assertion that cognitive functioning is influenced by cardiac surgery, the methods by which that evidence has been gathered may not be sufficiently empirically valid to actually draw that conclusion. Collectively, the results indicate that test batteries currently employed in cardiac surgery research may not be performing as they were anticipated to in the research methodology and hence have insufficient construct validity for the purpose. If the impact of cardiac surgery on cognitive functioning is to be fully elucidated, future research efforts must, at a minimum, seek to verify the structural integrity of the test batteries being used before embarking on a process of data collection.

# Study Three – Meta Effect Size Analysis of the Consensus Statement Core Neuropsychological Test Battery

## Abstract

The 1995 consensus statement on neurobehavioural assessments in cardiopulmonary bypass surgery recommended a core battery of four neuropsychological instruments for inclusion in all research within the field. However, when recommending the battery, the consensus statement authors did not describe the ability of the tests to detect the types of changes in functioning that have typically been noted in the associated literature. Therefore, the current study involved an analysis of the effect sizes for the core test battery proposed in the consensus statement. Data was derived from suitable published studies that were included in study one of the current research. Aim: The aim of this study was two fold. Firstly, it sought to identify cohort specific effect sizes for each core test. Secondly, it sought to provide estimates of effect sizes that could be used by future researchers when planning their research to ensure it obtains suitable power. Analysis: The analytical methods used were the same as those used in study one. However, on this occasion, data was collated on an individual test basis instead of collating across all comparisons and all studies. Results: The analyses indicate that in recent studies using the core neuropsychological test battery tests, the population effect sizes being detected ranged from 0 to -.7, which equates to distribution overlaps of at least 57%. Additionally, it was found that the effects being detected differed across assessment intervals. Conclusion: In accordance with the aim of the study, the results identified population effect sizes currently in operation in the cardiac research field for the core test battery tests. The estimates established provide valuable information for

researchers planning research endeavours in the field using the core neuropsychological test battery.

## Introduction

Interest in post cardiac surgery cognitive deficits has persisted almost since the first cardiac surgery operations were undertaken (Borowicz et al., 1996). Amongst the proliferation of studies of this topic, a wide range of neuropsychological tests has been used (Borowicz, 1996; van Dijk, Keizer et al., 2000; Yates & Alston, 2000). The selection and use of appropriate neuropsychological tests, and the analysis and interpretation of their results has been widely addressed in the literature (Blumenthal et al., 1995; Borowicz et al., 1996; Murkin, Newman et al., 1995; Murkin, Stump et al., 1997; S. Newman, 1995; Slade et al., 2001; Stump, 1995; Ryan & Hendrickson, 1998). The *Statement of Consensus on Assessment of Neurobehavioural Outcomes After Cardiac Surgery* (Murkin, Newman et al., 1995), in its reference to the "appropriate" selection of tests, cited issues such as sensitivity, reliability, validity, and data analysis methods as important concerns for all researchers in the field. The statement proposed a set of three tests to be adopted as the core of all research tests batteries. The tests were parts A and B of the Trail Making Test (TMT-A, TMT-B), the Grooved Pegboard test (GPB), and the Rey Auditory Verbal Learning Test (RAVLT). In recommending the core battery, however, the consensus group did not elucidate the empirical or psychometric support for selecting those instruments, nor suggest any criteria for interpreting the results gained from using them in the cardiac surgery context.

Methodological discussions about the use of neuropsychological tests in the cardiac surgery context have focused on the ability of research to identify reliable individual change across time (Gill & Murkin, 1996; Kneebone et al., 1998; Murkin, Stump et al., 1997). The 1997 consensus statement, *Defining Dysfunction: Group Means Versus Incidence Analysis – A Statement of Consensus* (Murkin, Stump et al., 1997), identified the importance of elucidating individual performance change over differences between group's means. The statement suggested that considering individual changes in postoperative cognitive functioning is the most sensitive means of discovering the clinical features of post cardiac surgery cognitive dysfunction (Murkin, Stump et al., 1997). This view has been widely supported, with agreement that any analyses using group mean scores invariably masks the true range of individual differences in the sample (Borowicz et al., 1996; Jacobson, Roberts et al., 1999; Jacobson & Truax, 1991; Murkin, Newman et al., 1995; Slade et al., 2001).

One very real difficulty in examining individuals to discern changes in cognitive functioning is the definition of change used (Gill & Murkin, 1996; S. Newman, 1995). It is reasonable to expect that the proportion of patients showing decrements, improvements, or no change in performance across assessment intervals depends on the sensitivity of the decision criteria are used (Stump, 1995). Previous investigations into neuropsychological functioning post cardiac surgery have used a variety of methods to decide whether a change in functioning has occurred. Some of the methods used include arbitrary criteria such as *x* standard deviation change from preoperative mean level of functioning for a control group, or percentile reduction in performance on *x* tests in the battery (Borowicz et al., 1996; S. Newman, 1995; Slade et al., 2001). Many criticisms have been levelled at those methods (see Borowicz et al., 1996 for a full discussion of those points). The most pertinent of those criticisms to the current discussion is that using definitive criteria dichotomising functions into impaired and non-impaired categories neglects the reality that premorbid and post-intervention functioning occurs along a continuum (Keith & Puente, 2002), and that changes along that continuum will produce specific effects for each individual.

In recent years, substantial attention has been paid to the statistical process known as Reliable Change, a statistical method for concluding that the difference between results is in fact reliable (Hollon & Flick, 1988; Hsu, 1989; Jacobson, Follette et al., 1988; Keith & Puente, 2002; Tingey et al., 1996). Reliable Change, however, is not a meaningful measure of the clinical significance of research findings. Rather it is a statistical difference measure, akin to the *t* test (Hsu, 1989; Jacobson, Follette et al.; Keith & Puente; Tingey et al.), without clinical implication. While establishing statistical reliability is an important step in any data analysis (Maassen, 2000; Tingey et al.), it provides no indication of whether the change in performance is sufficiently important to raise suspicion or warrant further investigation (Jacobson & Revenstorf, 1988; Keith & Puente; Maassen). For that purpose, additional measures, such as test cut-off scores strongly correlated with group associations (such as impaired and non-impaired), have been proposed. However, as Hollon and Flick suggest, dichotomising subjects into dysfunctional and non-dysfunctional populations achieves nothing but a complicated and unsolvable dispute over the defining characteristics of each population.

To quantitatively assess the degree of impact a treatment such as cardiac surgery has on cognitive functioning, an index that reflects the "pure magnitude of change" (Jacobson, Follette et al., 1984, p. 344) is required. However, it is proposed that the index utilised should be stratified *in the context of the population or context under investigation*. Effect size is the measurement scale used to quantitatively describe the magnitude to which a phenomenon under study is present in the sample, and by extension, the whole population. When used in the context of investigations into changes in cognitive functioning due to cardiac surgical interventions, effect size equates to the degree of change in performance on neuropsychological testing and, by implication, the magnitude of impairment if present. The advantage of the effect size statistic is that it operates independently of the unit of measurement used in gathering the data. That is, where a large raw score on a test may not reflect good performance, the size of the effect directly equates to the presence of the phenomenon (i.e., change in cognition due to cardiac surgery). If no change has occurred, then the effect will be zero. If some change has occurred then the size of effect will directly represent the degree of change (Cohen, 1988).

However, it has been argued that even though effect size estimates directly represent the observed difference between distributions of scores, they have no direct relevance to clinical significance, as even large effect sizes could be considered trivial by observers who each apply their own qualitative interpretation of that value (Jacobson, Follette et al., 1984; Jacobson & Revenstorf, 1988). Cohen (1988) agreed with this proposition, suggesting that proposing operational definitions qualifying the magnitude of the differences found, that are not validated, and that are standardised across the whole discipline, carries a significant risk of misinterpretation. Follette and Callaghan (1996) also raise this issue in their rebuttal of the proposition by Tingey et al. (1996) that $d = .5$ was sufficient to distinguish between impaired and non-impaired populations. The use of cut-off points or qualitative descriptions to interpret effect sizes will, as happens when raw test performances are simply dichotomized, reduce the quality and quantity of information gained from the results. Despite that inherent danger, derivation and explanation of changes in performance provided by effect size analysis appears useful to developing a greater understanding of a phenomenon under investigation.

The later point has lead some methodologists to suggest that effect size analyses should be included in all assessments of research results where the view is to establish whether clinically significant change has occurred (Hageman & Arrindell, 1999).

Determining whether an individual or patient group has suffered a clinically significant change in cognitive functioning is difficult, especially given that there is no consensus in the literature about what represents a change in performance that is important enough to be of clinical relevance. A substantial literature has developed around the reliable change statistical method. However, that method only facilitates the decision that the change in performance is statistically *real*. It does not specify whether the detected difference in performance is clinically or functionally important. Currently, no formalised and agreed guidelines exist to direct such decisions (Ryan & Hendrickson, 1998). Therefore, the current investigation sought to describe the size of effects currently being detected by the core test battery, and to compare those values with the various conventions proposed in the literature, with a view to discerning whether effect size analysis is a suitable method for meaningfully describing clinical changes in performance.

## Method

### Aims

The current study sought to provide estimates of effect sizes for within group comparisons on individual tests in the consensus statement core neuropsychological test battery. Additionally, the study sought to compare those effects with benchmark conventions previously proposed in the literature, to establish whether effect size estimation could contribute usefully to the interpretation of changes in post cardiac surgery test performance on the tests specified in the 1995 consensus statement core neuropsychological test battery.

Procedure

The data search and collection process for the current study is the same as that described in the procedure section for study one. As a review, a search was conducted on electronic databases (Science Direct, Proquest, Ebesco Host, Psyclit, and Medline) of all journal articles published between 1996 and 2002. The 1996 start date for the literature search was used as it allowed for the consensus statement by Murkin, Newman et al. (1995) to be disseminated and incorporated in research designs. The search, for all publications with titles or keywords addressing cognitive functioning and cardiac surgery, identified 90 published papers. The papers were reviewed by the current author and, if necessary, excluded according to the following criteria.

1) Papers not using the core tests as part of their battery or with results published in a manner precluding the current calculations (n = 34).
2) Data collection commenced prior to release of the 1995 consensus statement (n = 29). Studies were exited on this strategy in an attempt to remove the confounding effects of changes in surgical techniques (Baker, Andrew et al., 2001)
3) Subsequent publications from a single research group and data collection process. This was done to avoid the difficulties of redundancy of effect sizes values and potential bias by larger research groups who have produced multiple works from one data collection process (van Dijk, Keizer et al., 2000). The publication retained was the one that presented the most suitable data for the current analysis (n = 14).

Application of the exclusion criteria resulted in $N = 13$ papers being deemed suitable for inclusion in the current analyses. The details of those studies are presented as Table 1 in Appendix A. Not all of the studies deemed suitable for inclusion utilized all of the core test battery in their procedures. As such, the current analyses consist of a different number of cases for each core test at each interval comparison. To ensure some strength in the analyses of this very restricted sample of papers, analyses were conducted for each test only if there were five or more comparisons from three or more papers. The analysis for the current study proceeded as follows, first population effect sizes were calculated for each core test comparison conducted in the 13 studies. Calculation of effect sizes followed the previously described approach (see Study One) of Dunlap et al. (1996). Then the results for the comparison of each test were collated and the average calculated.

The effect size scale ranges from zero to above 4.0, with larger numbers equating to larger differences between the central points of the score distributions being examined (Cohen, 1988; Barlow, 1981). Various effect size thresholds have been identified in the literature as useful categorical indicators, and previously described in the method section of Study One, were used as benchmark comparators for the effect sizes being detected in the current sample. The benchmarks were Cohen's $d = .2$, .5 and .8, (Cohen's, 1988), 1.35 (Bezeau & Graves, 2001), and 3.0 (Zakzanis, 2001). Those effect size magnitudes equated to 93, 66, 53, 25, and seven percent overlap between score distributions (percentages taken from Table 1 in Zakzanis, 2001).

## Results

The current study examined N = 13 clinical neuropsychological research studies with a view to estimating the actual effect size values for each of the core neuropsychological test battery instruments. Table 1 in Appendix A lists the individual papers included in the study.

Table 1

Effect Size Estimates for Core Neuropsychological Test Battery Tests

| | Preoperative to Postoperative Change | | | | |
|---|---|---|---|---|---|
| | GPB[1] | | RAVLT[2] | TMT[3] | |
| | Dominant | Non Dominant | Total T1-T5 | Part A | Part B |
| Dunlap et al's. *d* | .5 | .4 | .1 | 0 | 0 |
| approx % Overlap | 66 | 72 | 92 | 100 | 100 |
| | Preoperative to Follow-up Change | | | | |
| Dunlap et al's. *d* | .2 | .3 | .5 | .7 | .4 |
| approx % Overlap | 78 | 85 | 66 | 57 | 72 |

1. GPB = Grooved Pegboard Test
2. RAVLT = Rey Auditory Verbal Learning Test
3. TMT = Trail Making Test

The top part of Table 1 shows that effect sizes detected for the preoperative to postoperative comparison were in the lowest quartile of the possible range of such values, as elucidated by Cohen (1988) and re-affirmed by Zakzanis (2001). That is, the degree of overlap between the distributions of scores in the preoperative to postoperative comparison was substantial in all tests. In Cohen's (1988) original context, the degree that the two distributions differed was less than 34% and 28% for both parts of the Grooved Pegboard test, less than 8% for the total of trials one to five on the Rey Auditory Verbal Learning Test, and not at all for both parts of the Trail Making Test. When compared to the various proposed conventions in the literature, current effect sizes for the tests fell well short of the clinically meaningful and clinically significant levels respectively proposed by Bezeau and Graves (2001) and Zakzanis. Applying Cohen's (1988) conventions, only the Grooved Pegboard test (GPB) detected a meaningful difference in performance. However, at best, that difference was small to medium in size.

As reported in the bottom part of the table, the effect sizes detected for the preoperative to follow up comparison all resided in the lowest quartile of the possible range of values (Cohen, 1988; Zakzanis, 2001). The degree of overlap between the distributions of scores in this comparison was also large. Again, to use Cohen's (1988) original context, the degree that the two distributions differed was less than 22% and 15% for both parts of the Grooved Pegboard test, less than 34% for the total of trials one to five on the Rey Auditory Verbal Learning Test, and less than 43 % and 28% on both parts of the Trail Making Test. When compared to the various proposed conventions in the literature, current effect sizes for the tests again fell well short of the clinically meaningful and clinically significant levels respectively proposed by Bezeau and Graves (2001) and Zakzanis. Applying Cohen's (1988) conventions, the Grooved Pegboard test detected at least a small effect, while the effects detected for the other two tests approximated a medium size.

When examining the effect sizes detected by the tests across the two assessment intervals, preoperative to postoperative and preoperative to follow-up, it is apparent that the magnitude of change varies in accordance with the length of time of the assessment interval. On a purely descriptive level, with the exception of Trail Making Test Part A, the differences between interval effect sizes did not appear to be substantially meaningful.

# Discussion

The notion of deciding when a change in cognitive functioning has actually occurred, and whether that change is clinically meaningful, has received some attention in the literature in recent years (Ryan & Hendrickson, 1998). A particular focus of that discussion has been the utility of change indices to improve the reliability of statistical decisions (Hsu, 1989; Jacobson, Follette et al., 1988; Keith & Puente, 2002; Tingey et al., 1996). While improved reliability of decisions is an important step, simply saying that a difference in performance is reliable does not provide insight about whether the performance difference has clinical meaning (Jacobson & Revenstorf, 1988; Keith & Puente; Maassen, 2000). Arbitrarily defined test score cut points that dichotomize subjects as impaired or not impaired have been proposed as useful indicators of clinically meaningful change. However, even when used in conjunction with reliable change indices, such cut points do not allude to potential or real functional impacts resulting from altered cognitive capacity; that is, the association between the cutoffs and current functioning in the context of premorbid functioning has not been established.

With this in mind, the current study examined the utility of a continuous scale (the range of effect size values) marking the magnitude of difference in performance to assist in decisions about clinically meaningful change on consensus statement core battery test performance. The scale examined was the estimated within group effect sizes continuum proposed by Cohen (1988). In addition to that continuum, cut-off scores previously defined as indicative of clinically meaningful performance differences were evaluated for their utility to contribute to decisions about clinically meaningful change (Cohen, 1988; Bezeau & Graves, 2001; & Zakzanis, 2001).

Results of the present analysis indicated that estimated effects all fell in the lowest quartile of the range of possible sizes as suggested in Cohen (1988). When compared to that continuum, the current results indicate that the effects being detected were relatively small. When compared to cut-off points, such as the benchmark effect sizes suggested by Cohen (1988), Bezeau and Graves (2001), and Zakzanis (2001), each of the tests failed to dissociate interval assessment scores at clinically meaningful levels. The best performing test in this regard was the change in performance on Part A of the Trail Making Test from preoperative assessment to follow-up. That comparison demonstrated 57% overlap between the distributions of mean scores from a range of studies. A margin of overlap of that magnitude indicates that cardiac surgery patients could equally have clinically meaningful alterations in test performance or no meaningful change. Conversely, the greater polarization of results for the other core tests indicates that the differences are not likely to indicate a clinically meaningful change. In terms of the whole continuum, the effect size measures appear to provide limited assistance to the clinical decision process of whether cardiac surgery impacted neuropsychological functioning, unless they approximate either end of the continuum.

Prentice and Miller (1992) suggest that even small effect sizes can carry significant meaning, depending on the context applied to them. In the current analysis, the context applied was the benchmarks previously suggested in the literature. As discussed in the introduction, several benchmarks have been suggested in the literature. These include Cohen's (1998) effect size values of 2, .5, .8, Bezeau and Graves' (2001) effect size value of 1.35; and Zakzanis' (2001) effect size of 3.0. When the result for all core tests across both comparison intervals were compared to those previously defined values, it appeared that the alteration in performance demonstrated on all tasks was not of a magnitude that has significant clinical meaning. However, it is important to note that even though those values relate to a probabilistic magnitude their exact meaning has not been established because their positive and negative predictive powers and ecological validity have not been empirically established.

An alternative explanation of the current results may be that the core test battery contains neuropsychological instruments that are not sufficiently sensitive to detect meaningful cognitive deficits in cardiac surgery populations. The core test battery tests were chosen because of their extensive history in clinical neuropsychological evaluations and neuropsychological research (Murkin, Newman et al., 1995). However, historical use does not necessarily justify test selection. Specifically, there is no normative data or information on the diagnostic efficiency statistics of the core test battery in the cardiac surgery context. If current research efforts were interested in comparing cardiac surgery patients to normal controls, then context specific information may not be required. However, given that the majority of interest centers on within group[7] and within individual change, such information is pertinent if not essential to facilitate decisions about whether an individual's performance is clinically and meaningfully impaired.

The current investigation carries a significant limitation, namely the small sample of papers contributing to the analyses. This limitation is a direct relation to the strict inclusion / exclusion criteria applied in this study's methodology. A number of studies were excluded on the basis of potential redundancy. That is, they

---

[7] For discussion about methodological considerations such as within versus between group designs and the examination for decline using individual change analyses see Study One pp.101-3.

resulted from a range of analyses carried out by the same author group on a sample drawn from a single data collection process. While this is a valid exclusion criterion, its application significantly limited the available data for analyses. Therefore, the current results and conclusions must be considered in that context, and with caution, until they are verified by way of replication.

## Conclusion

The current results provide insight into the size of cardiac surgery treatment effects on cognitive functioning. In terms of the continuum of possible effects, those produced by cardiac surgery interventions fall within the lowest quartile, indicating that the surgery has little functional impact on cognition. When compared to previously suggested conventions for clinical decisiveness (cut-off points), the effects being detected appear to be of little clinical meaning. Importantly, given that questions remain regarding the validity of the core test battery for use in the cardiac surgery context, the validity of the clinical cut-off points, and the meaning of stages along the effect size continuum in terms of actual functional outcome, the utility of the effect size analysis approach to contribute to clinical decision processes remains unclear. At the very least, and as an absolute priority, research is required to address those points to ensure that investigative endeavours into the effects of cardiac surgery produce knowledge that is pertinent to the end user, the clinician.

# Study Four – A Preliminary Facet Analysis of the Consensus Statement Core Neuropsychological Test Battery

## Abstract

Recent discussion aimed at strengthening research methodologies in studies investigating cognitive decline following cardiac surgery has identified a number of important aspects for consideration. Among the points raised is the need for researchers to be aware of the cognitive functions being assessed by specific tests and combinations of tests. As part of those discussions, a core battery was suggested as a way to parallel research efforts and ensure comparability of outcomes. Despite the proposal of a core test battery, no evidence was presented in support of its validity for the context. The current study sought to evaluate the structure of core test battery in a small sample ($N = 12$) of patients undergoing cardiac artery bypass graft surgery. Facet Theory, consisting of defining a mapping sentence and conducting Facet Analysis (Smallest Space Analysis), was used to examine the relationships between test scores at both preoperative and postoperative assessment intervals, to confirm whether the battery structure assessed the cognitive domains emphasised by its proponents. The results of the present analysis indicate that the relationships between tests and scores vary across assessment intervals. Additionally, when both assessment intervals were considered contemporaneously, only two general facets of cognitive functioning could be observed. The conclusion drawn is that the test battery requires carefully consideration and validation to ensure that conclusions being drawn from results are appropriate to what is actually being examined.

## Introduction

In 1994, a group of international "experts" reviewed the field of research, addressing the question of whether a decline in cognitive functioning is a result of cardiac intervention. The result of that inquiry was the *Statement of Consensus on Assessment of Neurobehavioural Outcomes After Cardiac Surgery* (Murkin, Newman et al., 1995). The statement consisted of 14 points pertaining to the conductance of neuropsychological research in the field. Point five in that statement related to the selection of tests for use in studies. In the context of a range of points about research methodology, it stated that researchers should primarily consider the cognitive functions to be assessed, the ability of specific tests to assess those functions, and the range of functions assessed by the battery as a whole (Murkin, Newman et al., 1995). To this end, the suggestions of point five reinforce basic assessment and research methodology. Namely, they reiterate the concept of construct validity in tests and batteries employed to answer specific questions. In simpler terms, researchers must have a comprehensive understanding of test functions (what the test measures and how it measures it) in the context of a particular theoretical rationale, in order to make valid interpretations of test results and, in turn, develop a greater understanding of the phenomenon under investigation (Messick, 1995; Murphy & Davidshofer, 1998).

The consensus group also proposed a core test battery for use in all research conducted into cognitive decline following cardiac surgery. The tests chosen for the core test battery were the Rey Auditory Verbal Learning Test, Trail Making Test Parts A and B, and the Grooved Pegboard Test. Key authors of test review books, such as Lezak (1995), and Spreen and Strauss (1998), have identified the cognitive skills generally associated with specific neuropsychological tests. In doing so, they concluded that the Rey Auditory Verbal Learning Test measures verbal learning and memory skills, that the trail making test assesses attention, sequencing, mental flexibility, visual search, and motor functioning; and that the Grooved Pegboard Test taps into motor functioning.

The aim of identifying that core battery was to enhance parity between studies so that their data could be recast and re-analysed to allow direct comparison and integration of research results (Murkin, Newman et al., 1995). In essence, the proposal of a core test battery represented a significant effort to enhance the knowledge base within the field, by making research efforts comparable across the literature (Slade et al., 2001). The core test battery was proposed on the basis that it assessed four distinct cognitive domains (attention and concentration, memory, language, and psychomotor speed), the areas that the consensus group identified as important to assess "to determine if the cortical mantle is functionally intact (Murkin, Newman et al., p.1294). Despite suggesting that the core tests assessed the core domains of interest, individual tests were not specifically assigned to domains. As a result, the consensus group failed to address the recommendations of point five in their own consensus statement. As mentioned, the selected tests have historically been associated with the domains proposed (Spreen & Strauss, 1998). However, empirical support for the assumption that those associations would hold in the cardiac surgery context was not provided in the consensus statement, and this author has not been able to find any supporting evidence in subsequent related literature.

Recent studies examining cognitive dysfunction following cardiac surgery have used factor analytical methods as a process for reducing the array of results from the tests used to a few combination scores or a single general cognitive function score (Greene & Sears, 1994; Grigore, Mathew et al, 2001). In doing so, the authors of those studies aptly demonstrated that the structure of test batteries may not conform to standard expectations when used in the cardiac surgery context. Despite those few factor analytical efforts, there remains little if any direction about how neuropsychological tests commonly perform in the cardiac surgery context. Hence, by extension, there is no clear indication that neuropsychological instruments currently used in the field actually measure what they purport to measure.

Given the absence of empirical evidence supporting the construct validity of the core test battery, the current preliminary inquiry sought to assess the ability of the core neuropsychological test battery to measure the domains proposed in the consensus statement.

## Method

### Aims

The aim of the current study was to examine the performance of the core test battery, by way of facet analytical methods, with the primary view of identifying whether the three tests partition into the four cognitive domains, as proposed by the consensus statement group. When specifically stated in Facet Analysis terms, the aim of the study was to examine the intercorrelations between the core battery tests, with a view to identifying the structure of the battery and confirming whether it concurs with the structure proposed by the consensus group.

### Procedure

To examine the structure of the core test battery, preoperative and postoperative data were collected on a small sample of cardiac artery bypass graft patients. Twelve neurologically normal patients with no history of identified cerebral vascular disease or injury were enrolled in the study. All participants agreeing to participate in the study were scheduled for and underwent elective CABG under the same cardiac surgeon (PB[8]) within one week of attending the presurgical clinic. Participants each provided written informed consent, and the institution's Ethical Review Committee approved the study.

The sample consisted of 10 men and two women with a mean age of 57.8 years (SD = 16.1 years, range = 27 – 77 years). Education data, though collected, was not analysed, due to the small sample size. At the postoperative assessment interval, 4 participants (33%) chose not to continue in the study. The participants who exited the study were males aged 35, 61, 65, and 73 years. The most common reason for not continuing with the study was the participant's eagerness to be discharge from hospital and return home. One participant, the youngest of those exiting the study, refused to continue due to experiencing extreme concerns about his post surgery physical well-being.

Subjects in this convenience sample were recruited over a three-month period from the pre-cardiac surgery clinic at a tertiary referral hospital in Sydney Australia. The recruitment process involved a brief verbal introduction to the study, provision of an information sheet and a consent form, and a request that those interested in participating approach the researcher at the end of the clinic to schedule an assessment. All but one of the participants was scheduled for a baseline assessment on the day of the clinic. The patient who was not assessed on the day of the clinic was assessed the day before surgery. Baseline assessments were all conducted in the one-week period preceding each subject's operation. Postoperative assessments were conducted on the day before or the day of discharge from hospital. This equated to an average period of 11 days (S.D. = 2.7 days, range = 7 – 16 days) between

---

[8] PB – Dr Paul Bannon, Cardiac Surgeon, Royal Pince Alfred Hospital, Sydney Australia.

preoperative and postoperative assessments. All assessments were carried out in a quiet office on the hospital precinct and were conducted by a Clinical Neuropsychologist (the current author). The core neuropsychological test battery tests (Trail Making Test parts A and B, Grooved Pegboard, Rey Auditory Verbal Learning Test) were administered in accordance with standard instructions found in Spreen and Strauss (1998). Tests were administered as part of a larger battery being used for an ongoing study. The larger battery also included a line bisection task, the simple reaction time task from the California Computerised Assessment Package (Miller, 1990), the Digit Span, Logical Memory, and Verbal Paired Associates subtests from the third edition of the Wechsler Memory Scale (Wechsler, 1997), and the 21 item Depression Anxiety Stress Scale (Lovibond & Lovibond, 1995). The administration sequence of the core neuropsychological test battery was randomised within that larger battery.

Facet Analysis was the statistical method used for the current analyses. It is a flexible analytical approach that has been applied in a broad range of psychological research endeavours (Guttman et al., 1990; Guttman & Shoham, 1982; Guttman & Zohar, 1999; Hans et al., 1985; Karni & Levin, 1972; Maraun 1997; Poreh & Shye, 1998; van de Vijver, 2001). Facet analysis was chosen as it enabled a preliminary examination of the relationships between tests without requiring the data collection efforts necessary to conduct a full factor analysis. Like factor analysis, it evaluates and ranks the associations between variables (i.e., test scores). However, unlike factor analysis, it does not suffer the same limitations. It is a nonmetric procedure and as such does not have specific rules to govern the extraction of key features called facets (Karni & Levin). Where factor analysis describes discrete divisions between variables in order to group them into factors, facet analysis allows for continuity between groupings by describing meaningful gradients across the spaces between them. That is, by way of a range of statistical models called Multidimensional Scaling, facet analysis represents variables (i.e., tests) as individual points in multidimensional space. The arrangement of the points and the number of dimensions required represent the relationships between variables, such that points positioned closer together in space and dimension represent closer associations than those positioned further apart (Brown, 1985). The arrangement of points is based on the intercorrelations between the variables, and as such, allows the associations to be empirically evaluated (Guttman & Shoham, 1982). That empirical evaluation is the key feature of facet analyses that makes it useful for establishing whether the performance structure of a test battery accurately reflects the design structure. Therefore, in contrast to factor analysis, facet analysis ensures that many possible explanations about the relationships between data are evaluated (Borg, 1993).

As core battery tests purportedly assess four distinct cognitive domains, the mapping sentence for the current investigation listed each domain as an independent facet to be confirmed. Test scores utilised were number of seconds to complete Trail Making Test Part A (TMTA) and Part B (TMTB), number of seconds to complete Grooved Pegboard – dominant hand (GPD), non dominant hand (GPN) and both hands (GPB), and the trial one (RAVLT1), trial five (RAVLT5), total recall (RAVLTT), short term recall (RAVLT6) and long term recall (RAVLT7) on the Rey Auditory Verbal Learning Test. The facet sentence constructed for this battery was:

Participant (X) will incur as a result of cardiac surgery

| *Facet* | | *Range* | | *Time* |
|---|---|---|---|---|
| ***A* – Attention and Concentration**<br>***B* – Memory**<br>***C* – Language**<br>***D* – Psychomotor Speed** | → | **No Decline** to **Some Decline** in | → | at **Preoperative** and **Post-operative** assessment intervals |

*Figure 1.* Mapping sentence describing the proposed structural domains of the consensus statement core neuropsychological test battery.

If the proposition by the consensus group is correct, then the mapping sentence should be an accurate representation of the interrelatedness of the tests selected as the core neuropsychological test battery for inclusion in all research endeavours within the field. If the mapping sentence is correct, a spatial representation of the associations between test items similar to the example presented in Figure 1 should result. In the example figure, and the results of the actual analyses, straight lines in the dimensional space represent the partitioning of data on the basis of the structural elements proposed in the mapping sentence for the battery. Bolded text represents the tests that combined to make the facet element proposed in the mapping sentence.

The hypothetical multi-dimensional space presented in the example figure (Figure 1) represents the facet structure of the core neuropsychological test battery when the mapping sentence is confirmed.

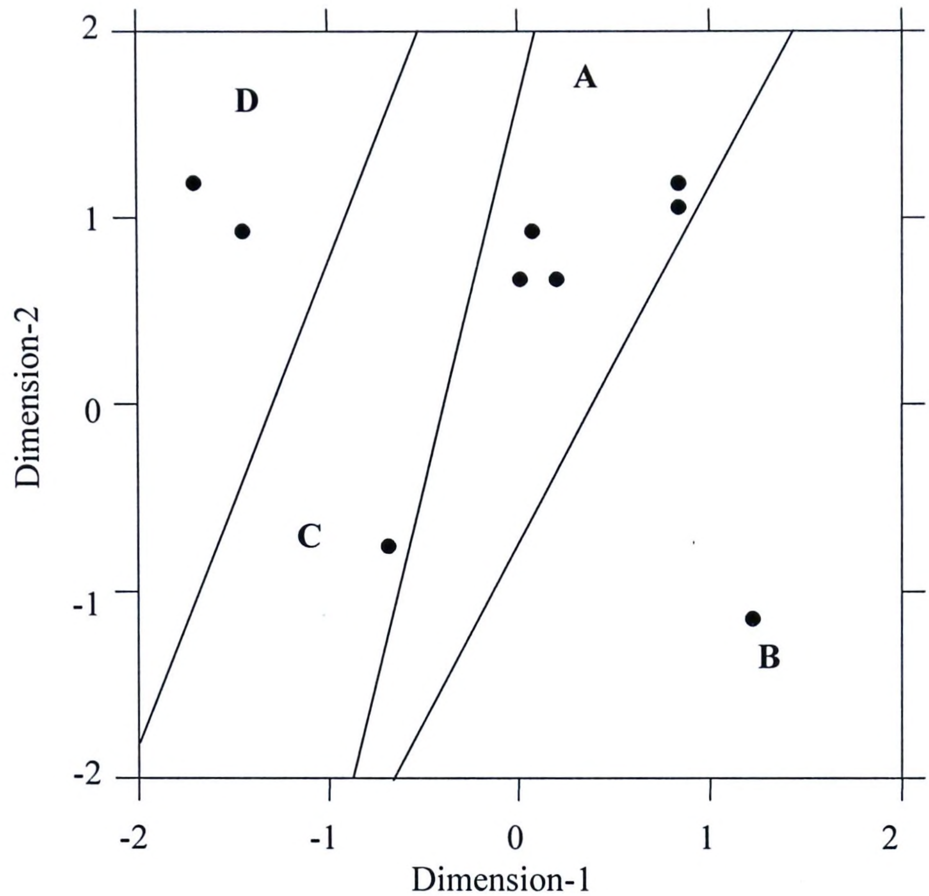


*Figure 2.* Example multi-dimensional space for the core neuropsychological test battery when the mapping sentence is confirmed.

In the example, facet analysis resulted in a two-dimensional spatial representation being achieved with adequate attenuation (<.15). Therefore, the two-dimensional representation was the simplest suitable representation of the data. As can be seen in the example figure, the spatial representation partitions into four distinct facets that are conceptually analogous to the divisions proposed in the mapping sentence. Therefore, it could be concluded from the example that spatial representation the mapping sentence accurately represents the constructs of the test battery and subsequently that the consensus group's proposed battery structure is an accurate expectation of how the battery will perform in the cardiac surgery context.

## Results

Table 1 and 2 in Appendix D respectively present the intercorrelations (Pearsons coefficients) matrices for the interactions amongst the nine raw test scores (elements) at the preoperative and postoperative assessment intervals. As the intercorrelations form the data for each facet analysis, values presented in them will not be interpreted. Explanation and interpretation of the results for each analysis are presented after each figure.

The multi-dimensional spaces presented in Figure 3and Figure 5 respectively represent the structure of the <u>core neuropsychological test battery</u> at preoperative and postoperative assessment intervals. The confirmed mapping sentences for those multi-dimensional spaces are presented in Figure 4 and Figure 6 respectively.

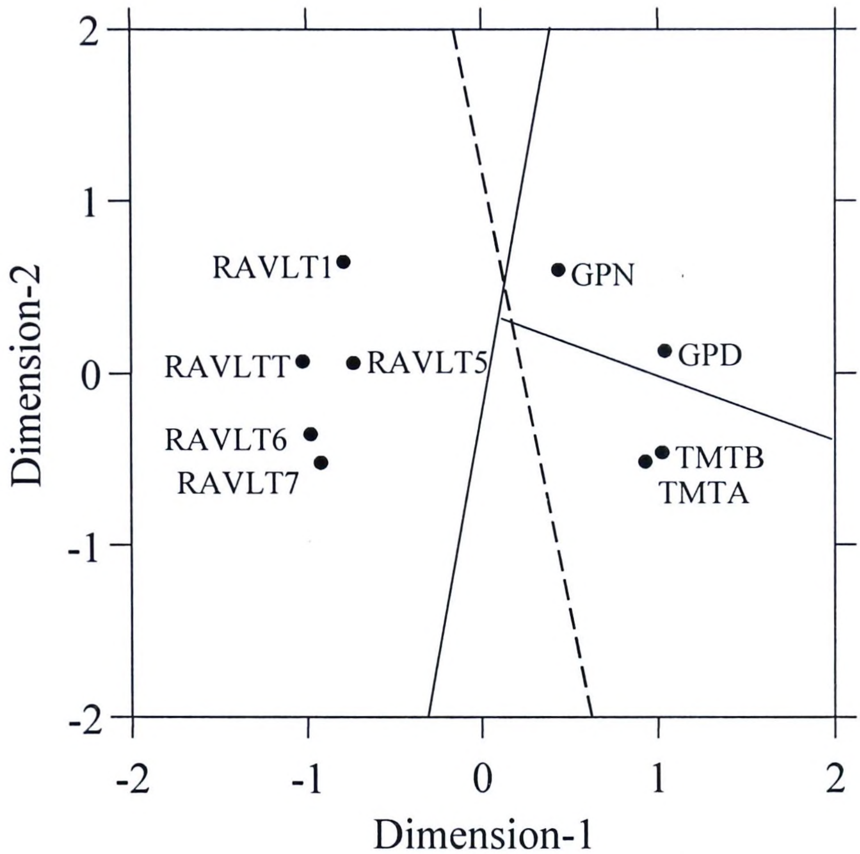


*Figure 3.* Multi-dimensional space for the core neuropsychological test battery at the preoperative interval.

Facet analysis of the preoperative data for the core neuropsychological test battery resulted in a two-dimensional spatial representation being achieved with adequate attenuation (<.15). Therefore, the two-dimensional representation was the simplest suitable representation of the data. As demonstrated in Figure 3, the spatial representation only partitioned into three discrete facets. Therefore, it could be concluded that the mapping sentence does not accurately represent the constructs within the test battery as proposed by the consensus group. Given the spatial representation shown in Figure 3, the confirmed mapping sentence is:

Participant (X) will incur as a result of cardiac surgery

| *Facet (Elements)* | | *Range* | | *Time* |
|---|---|---|---|---|
| ***A* – Attention and Concentration** | | | | |
| (*a*1 – TMTA) | | | | |
| (*a*2 – TMTB) | | | | |
| ***B* – Memory** | | **No** | | |
| (*b*1 – RAVLT1) | | **Decline** | | at the |
| (b2 – RAVLT5) | | to | | **Preoperative** |
| (b3 – RAVLTT) | → | **Some** | → | assessment |
| (b4 – RAVLT6) | | **Decline** | | interval |
| (b5 – RAVLT7) | | in | | |
| ***D* – Psychomotor Speed** | | | | |
| (*c*1 – GPN) | | | | |
| (*c*2 – GPD) | | | | |

*Figure 4.* Confirmed mapping sentence for the core neuropsychological test battery at the preoperative assessment interval.

As shown in the confirmed preoperative mapping sentence, the structure of the core neuropsychological test battery discretely accesses three out of the four facets (cognitive domains) proposed by the consensus group.

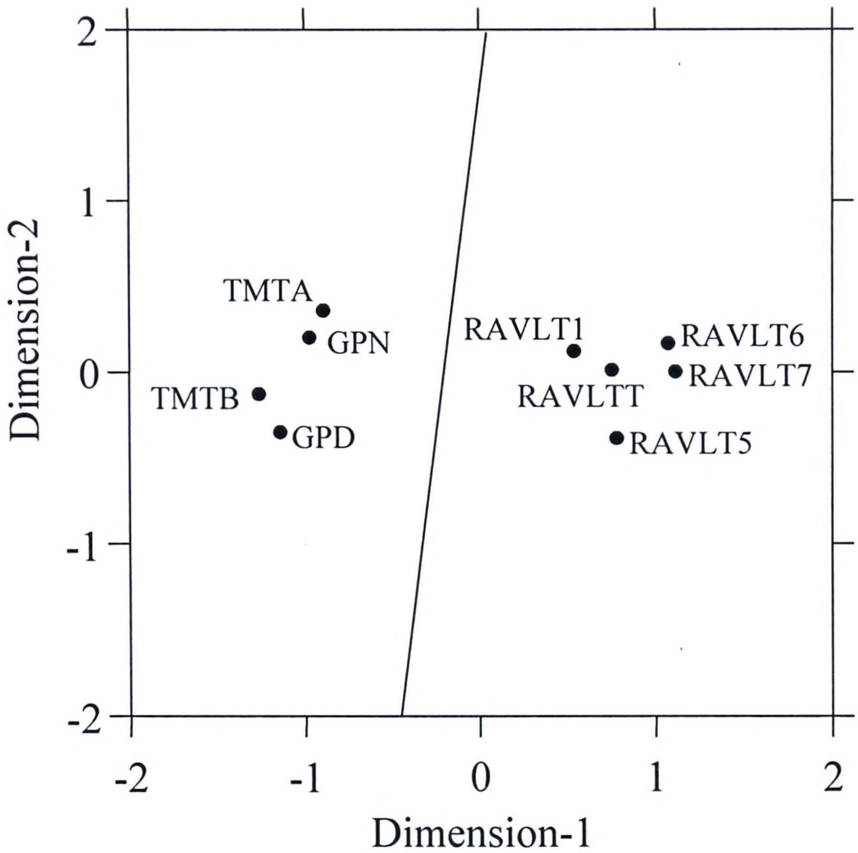


*Figure 5.* Multi-dimensional space for the core neuropsychological test battery at the postoperative interval.

In the example, facet analysis of the data resulted in a two-dimensional spatial representation being achieved with adequate attenuation (<.15). Therefore, the two-dimensional representation was the simplest suitable representation of the data. As demonstrated in the figure, the spatial representation partitions into two discrete facets. One facet, Memory, is the same as that proposed by the consensus group and found the analysis of preoperative data. The memory facet consisted of the various scores on the Rey Auditory Verbal Learning test, a test traditionally considered to assess learning and memory functioning. The other facet derived is a combination of the originally proposed attention and concentration facet and the psychomotor facet.

For the purposes of the current interpretation, the facet is labelled attention, concentration, and psychomotor speed. The tests that combined to form that facet were both parts of the Trail Making Test and the Grooved Pegboard Test. Both tests have a very strong psychomotor component, whilst the Trail Making Test also carries visual processing and attention/concentration components. Given the spatial representation shown in Figure 5, the confirmed mapping sentence is:

Participant (X) will incur as a result of cardiac surgery

| *Facet (Elements)* | | *Range* | | *Time* |
|---|---|---|---|---|
| ***B*** **– Memory**<br>(*b*1 – RAVLT1)<br>(b2 – RAVLT5)<br>(b3 – RAVLTT)<br>(b4 – RAVLT6)<br>(b5 – RAVLT7)<br>***E*** **– Psychomotor Speed, Attention and Concentration**<br>(*e*1 – GPN)<br>(*e*2 – GPD)<br>(*a*1 – TMTA)<br>(*a*2 – TMTB) | → | **No Decline** to **Some Decline** in | → | at the **Post-operative** assessment intervals |

*Figure 6.* Confirmed mapping sentence for the core neuropsychological test battery at the postoperative assessment interval.

As shown in the confirmed postoperative mapping sentence (see Figure 6), the structure of the core neuropsychological test battery only discretely assesses one out of the four facets (cognitive domains) proposed by the consensus group. Given this result, it can be concluded that the structure of the core neuropsychological test battery may not fit with that proposed by the consensus group.

Examining Figure 3 and Figure 5, it appears that the tests and scores separate sufficiently to indicate that they are not measuring exactly the same cognitive skills. That is, they do not appear to be contributing only to a general cognitive factor.

However, at both intervals the battery does not appear to divide consistently into the discrete structure proposed by the consensus group. An alternative explanation of the battery structure is evident when Figure 2 is partitioned (see dashed line) in the same manner as Figure 3. When examining the multi-dimensional spaces after they have been consistently partitioned, the facet structure of the core neuropsychological test battery could equally and perhaps more simply, be divided into two facets: (1) verbal, and (2) non-verbal functioning.

## Discussion

The core neuropsychological test battery was chosen by the consensus group (Murkin, Newman et al., 1995) to assess the cognitive domains of attention and concentration, memory, language, and psychomotor speed. Interestingly, though suggesting that the core battery assessed those four facets (cognitive domains), the consensus group did not specify which test fitted within each domain. The current facet analysis of raw data from a group of twelve cardiac surgery patients was conducted as a preliminary examination of the consensus statement core test battery, with a view of verifying the ability of that battery to assess the cognitive domains suggested by the consensus group as being crucial to proper functioning of the cortical mantle. Results of the current analysis demonstrate that when considering the test battery as a device to facilitate comparison of preoperative and postoperative functioning, as is often the case in research within the cardiac surgery field, it appears that the cognitive dimensions proposed by the consensus group were not consistently obtainable with the current tests. Rather, the analyses indicate that the core neuropsychological test battery is best described by only two elements, representing the distinct gross cognitive dimensions of verbal and non-verbal functioning.

It is important to recognise that when the consensus group stipulated the core test battery they did so with the proviso that it was a set of instruments for inclusion in research test batteries. While the current results indicate that the core test battery did not consistently discretely examine the cognitive abilities proposed by the consensus group, this does not mean that the battery does not access those skills. Rather, it indicates that the tests are measuring the common variance in cognitive functions rather than specific skill variance. Robson, Alston, Deary, Andrews, Souter, and Yates (2000), recognised that measuring common variance was a difficulty encountered in research examining cognitive functioning. As such, they suggested that a hierarchical investigation method examining "general" variance in cognitive functions first and then separable specific variances afterward is appropriate. As the battery appears to be measuring the common variance in cognitive functions, the current results indicate that the consensus statement core neuropsychological test battery is sufficient as a gross screen of general cognitive functioning. Then, as a second level of inquiry, the separate facets for verbal and non-verbal functions could be investigated.

Messick (1995) suggests that to understand a test's or battery's construct validity, that is the meaning of the scores obtained from it, there must be in place both evidence and sufficiently construed rationales. Though obviously based on the sound rationales previously used in the relevant field of research and a long-standing history in the field of clinical neuropsychology, the proposed core test battery lacked specific evidence supporting its application and interpretability in the cardiac surgery context. As such, the current preliminary investigation attempted to provide some insight into the construct validity of the core test battery when used in the cardiac surgery context. In doing so, the results obtained indicate that further empirical investigation is required to verify the utility of the tests to measure the cognitive functions identified by the consensus group as crucially indicative of adequate functioning of the cortical mantle.

Two main limitations require consideration when interpreting the current findings. The first is that the study involved only a small number of subjects, initially, and suffered a 33% drop out rate at the postoperative assessment interval. The second major limitation is that the partial effects of factors well recognised to influence neuropsychological test performance (such as age, IQ, gender, and culture) were not investigated or controlled in the analyses. Although the current analyses were a preliminary investigation only, and facet analysis is a less constrained analytical method than inferential statistics, the effects of the small sample size may have influenced the current findings. Therefore, the reader is cautioned that until further and more extensive investigation is carried out into the construct validity of the core test battery, it remains as useful as the other unverified neuropsychological test batteries currently used in the cardiac surgery research field.

## Conclusion

Collectively, the current results indicate that researchers should not faithfully accept that the core test battery discretely assesses the cognitive functions identified in the consensus statement as fundamentally important to investigate. Comprehensively understanding the meaning of test scores allows researchers to make valid interpretations of results, which in turn improves knowledge about the phenomenon investigated (Messick, 1995; Murphy & Davidshofer, 1998). As generalisability of test constructs is far from widely accepted (Messick), further extensive investigation into the construct validity of neuropsychological instruments utilised in the cardiac surgery field is required to ensure that interpretations of their results are valid and reliable. Intuitively, the lead consensus author, in his general introduction to the consensus statement, recognised this by suggesting that it was merely an initial step toward improving research methodologies (Murkin, 1995).

# CHAPTER FIVE

## Discussion

Disruption to the cardiovascular system such as that caused by heart disease and its therapeutic interventions has long been acknowledged to have a significant influence on neurological functioning (Smith, 1995; Sontaniemi, 1995). Neurological complications arising from cardiac surgery range from catastrophic stroke causing death to transitory changes in cognitive capacities (Borowicz et al., 1996). For the last half century, neuropsychological assessments have been used to quantify cognitive changes associated with cardiac surgery (Borowicz et al., 1996). Despite extensive research on the topic, no set cognitive phenotype has been identified for post cardiac surgery cognitive dysfunction. This limitation may in part be due to the fact that methods used in the research to assess and describe outcomes have lacked consistency across the length and breadth of research within the field. In 1995 and again in 1997, consensus statements were published with a view to improving on the methodologies used in the field (Murkin, Newman et al., 1995; Murkin, Stump et al., 1997). However, at the time the initial consensus statement was agreed upon, it was recognised by the key author that it was not the definitive methodological end-state but rather a first attempt at operationalising efforts (Murkin, 1995). Methods for assessing cognitive dysfunction arising from cardiac surgery have been a major discussion point, at least since the publication of the first consensus statement on the topic (Blumenthal et al., 1995; Borowicz et al.; Mc Daid et al., 1994; Murkin, Newman et al.; Murkin, Stump et al.; S. Newman, 1995; Slade et al., 2001). Not the least of the topics addressed is the issue of selecting the correct neuropsychological instruments for detecting changes that may occur as a result of undergoing cardiac surgery. Point five of the 1995 consensus statement suggested that when selecting tests for use in cardiac research, issues such as sensitivity, reliability, validity, and the range of cognitive functions being assessed by the tests, must be considered (Murkin, Newman et al.; S. Newman). Blumenthal et al. expanded on those issues, adding that the availability of normative data, alternate forms, and previous use of the test should be considered. Despite those directions, no literature exists to describe the validity of commonly used neuropsychological instruments in the cardiac surgery context. It appears that the intention was to rely on existing validity and reliability information for tests. However, that idea has been questioned by at least one group of authors (Mc Daid et al.). In a review of pertinent literature published since the release of the 1995 consensus statement, the current author concluded in line with other recent reviewers such as Haddock et al. (2003), that many conflicting research outcomes are being derived due to a piecemeal and methodologically unregulated range of investigative approaches being utilised in the field. Given that understanding, the current research endeavour, through a series of four studies, sought to evaluate the methodology behind neuropsychological research being conducted in the cardiac surgery context.

The first study incorporating three aims examined statistical power and effects sizes for neuropsychological research in the cardiac surgery field published since the release of the *Statement of Consensus on Assessment of Neurobehavioural Outcomes After Cardiac Surgery* (Murkin, Newman et al., 1995).

The first aim was to identify the extent to which a priori and post hoc power analyses were reported in the literature. Despite substantial commentary about statistical power values and population effect sizes not being reported in research literature (Bezeau & Graves, 2001; Cohen, 1962, 1965, 1988, 1992; Prentice &

Miller, 1992; Rosnow & Rosenthal, 1989; Sedlmeier & Gigerenzer, 1989), many published research studies still do not provide that vital information. In the current study, none of the 13 papers examined provided information on either a priori or post hoc power calculations, or estimated or observed effect sizes. The result indicates that the statistical shortcoming identified by Cohen (initially in 1962, and then continuously in subsequent discussions), Sedlmeier and Gigerenzer, and Bezeau and Graves continues to be an issue, even in a narrowly constrained area of research such as clinical neuropsychological investigations in cardiac surgery samples. Without the description of the practical significance of research outcomes, the consumer has no means of extrapolating findings and conclusions beyond the study in which they were published (Barlow, 1981). Hence, consumers reading the publications cannot be certain that the results reported carry any clinical relevance or that they are important enough to warrant further research. In addition to making the interpretation of research findings more difficult, not publishing statistical power and population effects sizes in every study leaves researchers with no contextually based indication of those values for use in planning future endeavours.

The second aim of the first study was to calculate estimated population effect sizes and statistical power values in clinical neuropsychological research on cardiac samples. For independent groups design comparisons effect sizes detected appear to be meaninglessly small ($\underline{M}^{IG}{}_{d} = .31$, $\underline{SD} = .52$; $\underline{Med}^{IG}{}_{d} = .17$; range = .01 – 4.50). This is confirmed when they are compared with the findings in a previous study examining the broader domain of all clinical neuropsychological research (Bezeau & Graves, 2001). Population effect sizes for repeated measures comparisons in neuropsychological research in the cardiac surgery context have not previously been investigated. Therefore, no studies exist to compare with the current findings. In the current analyses of repeated measures design comparisons, a strikingly small average population effect sizes ($\underline{M}^{RM}{}_{d}$ = .14, $\underline{SD}$ = .17; range = .00 - .90; $\underline{Med}^{RM}{}_{d}$ = .07) was again identified. Some authors have suggested that effect sizes much larger than the values traditionally used, and being found in the current examination, are required for research results to be of any clinical relevance (Bezeau & Graves, 2001; Zakzanis, 2001). Therefore, given the extraordinarily small effect sizes detected in the current batch of independent group and repeated measures comparisons, it appears that the phenomenon currently being studied is "no more than negligible or trivial in size" (Cohen, 1988, p. 16-17). The average statistical power observed for independent groups designs was well below the common and rigorous benchmarks of .80 and .90 respectively ($\underline{M}^{IG}{}_{Power\cdot}$ = 24, $\underline{SD}$ = .28; $\underline{Med}^{IG}{}_{Power\cdot}$ = 1.00; range = .05 – 1.00). Given the mean population effect, the average statistical power indicated that researchers would arrive at the correct conclusion only 24 times out of every 100 attempts, a ratio that raises meaningful questions about the validity of research efforts within the field. However, interpreting the median power value indicates that for independent groups design outcomes appear to be sufficiently robust. For repeated measures designs, mean statistical power was also poor ($\underline{M}^{RM}{}_{Power}$ = .41, $\underline{SD}$ = .32; range = .05 – 1.00; $\underline{Med}^{RM}{}_{Power}$ = .32), although this time it was stronger than its median reciprocal. Current power analysis results appear to indicate a need to improve the reliability of undertakings. The current effect size and statistical power findings indicate that clinical neuropsychological research examining phenomena in cardiac surgery samples is clearly not, as Bezeau and Graves said of the discipline as a whole, "better than we might have thought" (p. 403).

The third aim of the first study was to identify context specific small, medium, and large effect sizes that could be used by researchers in the design phase of future studies. The results indicate that, for both independent groups designs and repeated measures designs, a population effect size of $d = 1.35$ produced sufficient power to be considered a reliable estimator when compared to the benchmark power values of .80 and .90 at alpha of .05. By upward and downward extension, effect size values of $d = .8$ and $d = 1.95$ were respectively identified as small and larger effects size estimates for the cardiac surgery context. Those values were adopted as they represented the poles of the effect size range where sufficient power (compared to benchmark values) was maintained. To obtain the observed and predetermined population effect sizes at alpha = .05 and benchmark statistical power values, independent groups design studies would require sample sizes at least 350% larger, while repeated measures designs would have to expand their samples over 1700%. The logistical difficulties of data collection processes of this magnitude make such efforts improbable. If current sample sizes are the practical logistical solution, researchers must reduce the error associated with unconstrained operational definitions that fail to provide a definitive benchmark for making decisions, and with test batteries that have not been empirically established as being capable for the task.

Researchers and authors who examine and report on cognitive decline following cardiac surgery have stressed the importance of selecting batteries based on an understanding of the cognitive functions assessed by the individual tests and the range of functions assessed as a whole (Blumenthal et al., 1995; Gill & Murkin, 1996; Murkin, Newman et al., 1995; Murkin, Stump et al., 1997; Slade et al., 2001; Stump, 1995). Despite those suggestions, however, previous efforts to examine the cognitive effects of cardiac surgery have used neuropsychological tests that have not been specifically validated for use in that context. In the few instances (Green & Sears, 1994; Grigore, Mathew et al., 2001) where the structure of research test batteries has been analysed, the analysis has only been done for preoperative assessment data. As such, those analyses have not provided confirmation of the structural reliability of the battery. Given that there is no published literature describing the construct validity of neuropsychological tests in cardiovascular context, the aim of the second study was to examine relationships between tests used in research batteries with a view to identifying whether the structure of those batteries was as they were designed. The second study conducted examined four neuropsychological test batteries that had previously been used in research endeavours within the field to establish whether the neuropsychological constructs purportedly assessed by the batteries actually existed. Facet analysis was used as it allowed the structure of relationships between tests in the batteries to be examined without being constrained by the rules normally associated with factor analysis. For each battery, the original researchers reported that they chose tests because they were well-recognised measures of specific cognitive domains. However, analyses revealed that all test batteries failed to conform to their predefined methodological structure at the preoperative assessment interval and that their structure altered between the preoperative and postoperative periods. Therefore, researchers could not be certain that what they proposed to measure was what they actually measured. However, a number of purportedly different tests strongly associated together in some analyses, contradicting the methodology proposed by the researchers. The results of the current examination indicate two important points about neuropsychological research in the cardiac surgery context. A possible explanation as to why test batteries currently being employed do not always comply with the structure detailed in the methodology of the studies is that, while the tests used in each battery purportedly measure specific cognitive skills or domains in the context in which they were developed or normalised, the cardiac surgery context is distinctly different. Support for this conclusion derives from the evidence that cardiac surgery patients perform differently to normal controls during preoperative assessments. Additionally, research examining people with cardiovascular disease has both supported and negated the proposition that people suffering cardiac conditions also have reduced cognitive functioning as a result of those conditions (Ahto et al., 1999; Ross & Graham, 1993). While preoperative assessments are often considered only as baselines, they may allude to a fundamental cognitive functioning difference between people requiring cardiac interventions and individuals who do not.

Consistent construct validity of a test battery is important if researchers are to draw valid conclusions about the cognitive phenotypes associated with the application of specific cardiac interventions. Results of the current analysis demonstrate, that in several instances, the structure of the test battery employed was not as originally proposed and that it varied across assessment intervals. The results indicate that researchers should be cautious when drawing conclusions about the cognitive domains affected by cardiac procedures when using test batteries that have not been specifically validated for the purpose. At present, neuropsychological assessments of cardiac surgery samples are undertaken with the belief that the instruments being used will function in the same way with the research sample as they do in generally unrelated test development and validation samples. Results of the current analyses, however, would seem to indicate that this might not be the case in the context of cardiac surgery research. If the impact of cardiac surgery on cognitive functioning is to be fully elucidated, future research efforts must, at least, seek to verify the structural integrity of the test batteries being used before embarking on a process of data collection.

The *Statement of Consensus on Assessment of Neurobehavioural Outcomes After Cardiac Surgery* (Murkin, Newman et al., 1995), as part of its effort to improve on the piecemeal and methodologically unregulated nature of research in the field proposed a core battery of neuropsychological tests. However, in proposing the battery, the authors of the statement did not provide the empirical rationale for their choice of tests. Additionally, they did not provide any indications of how to interpret performances on the tests. There has been a great deal of discussion in the literature in recent years concerning the evaluation of changes in cognitive functioning, and whether the changes are clinically meaningful. A particular focus of that discussion has been the utility of change indices to improve the reliability of statistical decisions (Hsu, 1989; Jacobson, Follette et al., 1988; Keith & Puente, 2002; Ryan & Hendrickson, 1998; Tingey et al., 1996). While improved reliability of decisions is an important step, simply saying that a difference in performance is reliable does not provide insight about whether the performance difference has clinical meaning (Jacobson & Revenstorf, 1988; Keith & Puente; Maassen, 2000). Arbitrarily defined test score cut points that dichotomize subjects as impaired or not impaired have been proposed as useful indicators of clinically meaningful change. However, even when used in conjunction with reliable change indices, they do not allude to potential or real functional impacts resulting from altered cognitive capacity. That is, the association between the cutoffs and current functioning in the context of premorbid functioning has not been established. With this in mind, the third study sought to examine the core tests in terms of their ability to detect meaningful results in this field of research. Specifically, the current study examined the utility of a continuous scale marking the magnitude of difference in performance to assist in decisions about clinically meaningful change. The scale examined was the estimated effect sizes continuum proposed by Cohen (1988). In addition to a whole continuum approach, cut-off scores previously defined as indicative of clinically meaningful performance differences were evaluated for their utility to contribute to decisions about clinically meaningful change (Cohen, 1988; Bezeau and Graves, 2001; Zakzanis, 2001). In the current analysis, estimated effects for the core tests all fell in the lowest quartile of the continuum range. When compared to cut-off points such as the benchmark effect sizes suggested by Cohen (1988), Bezeau and Graves, and Zakzanis, each of the tests failed to detect interval assessment performance differences at clinically meaningful levels. The best performing test in this regard was Part A of the Trail Making Test, where the difference between preoperative and follow-up scores resulted in 57% overlap between the distributions. A margin of overlap of that magnitude indicates that cardiac surgery patients could equally have clinically meaningful alterations in test performance or no meaningful change. Conversely, the greater polarization of results for the other core tests indicates that the differences are not likely to indicate a clinical meaningful change.

In terms of the effect size continuum, values that approximate the upper end of the continuum appear to provide the best evidence of meaningful change in performance. However, Prentice and Miller (1992) suggest that even small effect sizes can carry significant meaning, depending on the applied context. In the current analysis, the benchmarks previously suggested in the literature were applied. As discussed in the introduction, several benchmarks have been suggested in the literature. These include Cohen's (1998) effect size values of 2, .5, .8; Bezeau and Graves' (2001) effect size value of 1.35; and Zakzanis' (2001) effect size of 3.0. When the result for all core tests across both comparison intervals was compared to those values, it appeared that the alteration in performance demonstrated on all tasks was not of a magnitude to imply significant clinical meaning. However, it is important to note that even though those values relate to a probabilistic magnitude, their exact meaning has not been established because their positive and negative predictive powers and ecological validity have not been empirically established.

An alternative explanation of the current results may be that the core test battery contains neuropsychological instruments that are simply not sufficiently sensitive to detect meaningful cognitive deficits in cardiac surgery populations. The core test battery tests were chosen because of their extensive history in clinical neuropsychological evaluations and neuropsychological research (Murkin, Newman et al., 1995). However, historical use does not necessarily justify test selection. Specifically, there is no normative data or information on the diagnostic efficiency statistics of the core test battery in the cardiac surgery context. If current research efforts were interested in comparing cardiac surgery patients to normal controls, then context specific information may not be required. However, given that the majority of interest centers on within group and within individual change, such information is pertinent, if not essential, to facilitate decisions about whether an individual's performance is clinically and meaningfully altered.

As an extension to the investigation carried out in the third study, the final study was a preliminary examination of the interrelatedness of the core neuropsychological test battery. A facet theoretical analysis of new raw data from a group of twelve cardiac surgery patients was conducted with a view to verifying the ability of the battery to assess the cognitive domains that were suggested by the consensus group to be crucial to proper functioning of the cortical mantle. Results of the current analysis demonstrate that when considering the test battery as a device to facilitate comparison of preoperative and postoperative functioning, as is often the case in research within the cardiac surgery field, it appears that the cognitive dimensions proposed by the consensus group were not consistently obtainable with the current tests. Rather, the analyses indicate that the core test battery consistently divided into only two elements representing distinct gross dimensions of cognitive functioning. It is important to recognise that when the consensus group stipulated the core test battery, they did so with the proviso that it was a set of instruments for inclusion in research test batteries. While the current results indicate that the important cognitive facilities proposed by the consensus group were not consistently examinable by the core test battery, this does not mean that they would not perform in that manner when larger scale batteries are used. By extension, however, the current results indicate that as a gross screen of general cognitive functioning, the core battery is sufficient. Messick (1995) suggests that to understand a test's or battery's construct validity (that is, the meaning of the scores obtained from it), both evidence and sufficiently construed rationales must be in place. Though obviously based on the rationales of prior use in the relevant field of research and a long standing history in the field of clinical neuropsychology, the proposed core test battery lacked specific evidence supporting its application and interpretability in the cardiac surgery context. As such, the current preliminary investigation attempted to provide some insight into the construct validity of the core test battery when used in the cardiac surgery context. The results of the current preliminary inquiry indicate that further investigation is required to verify the utility of the tests to measure the cognitive functions identified by the consensus group as crucially indicative of adequate functioning of the cortical mantle. Collectively, the current results indicate that researchers can not faithfully accept that the core test battery assesses the cognitive functions identified in the consensus statement.

A criticism that may be levelled at the first and third studies is the exclusion of publications under the "redundancy" criterion suggested by van Dijk, Keizer et al. (2000). The redundancy criterion suggests that including samples that are already represented (i.e., a previous study in a series) may result in the series being overemphasized. Critics may argue that post hoc redefinition of a study population based on inherent subject criteria (i.e., from surgery type to medication type) results in a different study population and hence different population effect sizes. However, it is the opinion of the author that overall effect sizes produced by a phenomenon would include the effects produced within segments of the phenomenon. That is, for example, the effect undergoing treatment would include the segmented effects produced by the aspects of the intervention process itself. Whether this is the case, however, remains to be investigated once substantially more research within the field has been published. An additional criticism, stemming in part from the first, is that only a single person conducted the data search and selection process, including the exclusion of redundant papers. Perhaps, with additional reviewers, papers that were excluded may not have been, and papers that were included may have been excluded. By specifying stringent inclusion and exclusion criteria in the design phase of the study, and collecting all relevant publications before commencing vetting, it was hoped that the effect of a single reviewer would be minimised. However, future research replicating the current investigation should, where possible, use multiple reviewers to ensure the validity of the data collection process.

Regarding the second study, the most obvious criticism is the small number of studies investigated and hence the unrepresentativness of the neuropsychological test batteries. This criticism is entirely valid, and also is entirely due to the poor response from authors publishing in the field. The second study had a positive response rate for data requests of three percent. To some extent, examining datasets individually may have minimized some of the bias associated with an unrepresentative sample. Additionally, the use of facet theory in the place of factor analysis may have also reduced the impact of an unrepresentative sample. A byproduct of the poor response rate was that there was little overlap in tests included in the test batteries investigated. While this may be representative of the test diversity in the broader research field, it makes generalizing results about specific instruments difficult. The current results certainly imply that methodologically, all may not be as planned. However, to verify this, replication analyses with multiple data sets using the same or very similar batteries is required. Alternatively, future research endeavors could include preliminary validation of the structure of their proposed test batteries prior to embarking on data collection for the study itself.

The fourth study has two main limitations that must be considered when interpreting the findings. The first limitation is that the study involved only a small number of subjects initially and suffered a 30% drop out rate at the postoperative assessment interval. The second major limitation is that the partial effects of factors well recognised to influence neuropsychological test performance (such as age, IQ, gender, and culture) were not investigated or controlled for in the analyses due to the limitations imposed by the small sample size. Although the study was a preliminary investigation only, and facet theory is a less constrained analytical method than inferential statistics, the effects of those two limitations may have influenced the current findings. Therefore, until further and more extensive investigation is carried out to investigate the construct validity of the core test battery, including a replication of the results of the fourth study, it remains as useful as the other unverified neuropsychological test batteries currently gaining use in cardiac surgery research.

The current results provide some degree of insight into the size of cardiac surgery treatment effects on cognitive functioning. While the findings indicate trivial effect sizes and substandard power, they do not indicate that previous research efforts in the field have been wasted. Rather, they provide an impetus for new and more methodologically sound research endeavours in this field to find clinically meaningful effects that give credence to the associations, reliable, important or otherwise, that have been found previously. That is, by providing that insight, future researchers have a basis upon which to enhance the quality of their research design process, and subsequently their findings and the utility of those findings to consumers and future researchers.

While there appears to be ample evidence in the literature to support the assertion that cognitive functioning is affected by cardiac surgery, the methods by which that evidence has been gathered may not be sufficiently accurate to actually draw that conclusion. At present, neuropsychological assessments of cardiac surgery samples are undertaken with the belief that the instruments being used will function in the same way within the research sample as they do in generally unrelated test development and validation samples. Results of the current studies would seem, however, to indicate that this might not be the case in the cardiac surgery research context. Collectively, the current results indicate that test batteries currently employed in cardiac surgery research may not be behaving in traditional manners, and as such, may not have sufficient construct validity to fulfil the purpose for which they are employed. Similarly, researchers can not faithfully accept that the core test battery assesses the cognitive functions identified in the consensus statement.

Comprehensively understanding the meaning of test scores allows researchers to make valid interpretations of results, which in turn improves knowledge about the phenomenon being investigated (Messick, 1995; Murphy & Davidshofer, 1998). As generalisability of test constructs is far from widely accepted (Messick), further extensive investigation into the construct validity of neuropsychological instruments, and in particular, the core test battery tests, utilised in the cardiac surgery field is required to ensure that interpretations of their results are valid and reliable.

Whilst the current results indicate very small population effect sizes and small statistical power values, they were derived by grouping results from a range of published studies. As such, they do not exclude the possibility that some individual participants may have suffered serious cognitive impairment. The current research goes some way to improving knowledge about statistical power, population effect sizes, and methodological integrity in research on cardiac surgery samples. However, if the impact of cardiac surgery on cognitive functioning is to be fully elucidated, future research efforts must, at least, seek to improve on these elements. As Levy (1993) correctly asserts, "… theory and method are inseparable in the process of theory construction…" (p. 260).

## Conclusion

The focus of the current research endeavour was the examination of research addressing the phenomenon of cognitive decline in cardiac surgery samples. This was undertaken with a view to answering the question, *how adequate are our efforts?* In the literature review of studies, the cognitive domains that have consistently shown deleterious effects due to cardiac surgery have included attention, psychomotor speed, memory, visuoperception, executive functioning, and general cognitive functioning (Bendszus et al., 2002; Bruggemans et al., 1995; Chabot et al., 1997; Gugino, Chabot, Aglio, Maddi et al., 1997; Gugino, Chabot, Aglio, Aranki et al., 1999; Murkin, Boyd et al., 1999; Strooband et al., 2002; Suksompong et al., 2002). In the analytical investigation, however, it was demonstrated that current research methodologies are insufficient for the purposes for which they are being used. While that finding does not exclude the understanding that cardiac interventions may produce cognitive decline, it does indicate that substantial methodological improvement is required to ensure the robustness of the conclusions being drawn in the literature. Given the current findings, it is not beyond reason to suggest that until the inequities identified by the current investigation are rectified, research endeavours within the field will continue to fall short of achieving the level of cohesive insight required to facilitate a complete description of the aetiology of cognitive outcomes due to cardiac surgery. Therefore, the question of how adequate our efforts are must be answered *not quite good enough yet*.

As with all research endeavours, the only true measure of an outcome is its replication or refutation by additional research inquiry. In an attempt to set the first foot on the road to improved research methods in the field, the current research identified statistical power and population effect sizes in research on cardiac surgery samples. By providing measures of those values, a basis has been set for future researchers to enhance their research design process, and subsequently their findings and the utility of those findings to consumers and future researchers.

References

Agresti, A., & Finlay, B. (1997). *Statistical Methods for the Social Sciences* (3rd Ed.). New Jersey: Prentice Hall.

Ahto, M., Isoaho, R., Puolijoki, H., Laippala, P., Sulkava, R., & Kivela, S. L. (1999). Cognitive impairment among elderly coronary heart disease patients. *Gerontology, 45,* 87-95.

American Psychological Association (1994). Publication Manual of the American Psychological Association (4th Ed.). Washington, DC: Author.

Andrew, M. J., Baker, R. A., Bennetts, J., Kneebone, A. C., & Knight, J. L. (2001). A comparison of neuropsychologic deficits after extracardiac and intracardiac surgery. *Journal of Cardiothoracic and Vascular Anesthesia, 15,* 9-14.

Andrew, M. J., Baker, R. A., Kneebone, A. C., & Knight, J. L. (1998). Neuropsychological dysfunction after minimally invasive direct coronary artery bypass grafting. *Annals of Thoracic Surgery, 66,* 1611-1617.

Antman, E. M., & Braunwald, E. (2001). Acute Myocardial Infarction. In E. Braunwald & D. P. Zipes & P. Libby (Eds.), *Heart Disease: A Textbook of Cardiovascular Medicine* (pp. 1114-1218). London: W. G. Saunders.

Anversa, P., Leri, A., & Kajstura, J. (2004). Myocardial Basis for Heart Failure: Role of Cell Death. In M. D. Mann (Ed.), *Heart Failure: A Companion to Braunwald's Heart Disease* (pp. 71-90). Philadelphia: Saunders.

Aron, A., & Aron, E. N. (1994). *Statistics for Psychology.* New Jersey: Prentice Hall.

Arrowsmith, J. E., Grocott, H. P., & Newman, M. F. (1999). Neurologic risk assessment, monitoring and outcome in cardiac surgery. *Journal of Cardiothoracic and Vascular Anesthesia, 13,* 736-743.

Arrowsmith, J. E., Grocott, H. P., Reves, J. G., & Newman, M. F. (2000). Central nervous system complications of cardiac surgery. *British Journal of Anaesthesia, 84,* 378-393.

Baguley, T. (2004). Understanding statistical power in the context of applied research. *Applied Ergonomics, 35,* 73-80.

Baker, A. J., Naser, B., Benaroia, M., & Mazer, C. D. (1995). Cerebral microemboli during coronary artery bypass using different cardioplegia techniques. *Annals of Thoracic Surgery, 59,* 1187-1191.

Baker, R. A., Andrew, M. J., & Knight, J. L. (2001). Evaluation of neurologic assessment and outcomes in cardiac surgical patients. *Seminars in thoracic and Cardiovascular Surgery, 13,* 149-157.

Barbut, D., & Caplan, L. R. (1997). Brain complications off cardiac surgery. *Current Problems in Cardiology, 22,* 449-480.

Barbut, D., Hinton, R. B., Szatrowski, T. P., Hartman, G. S., Bruefach, M., & Williams-Russo, P., et al. (1994). Cerebral emboli detected during bypass surgery are associated with clamp removal. *Stroke, 25,* 2398-2402.

Barbut, D., Yao, F. S., Hager, D. N., Kavanaugh, P., Trifiletti, R. R., & Gold, J. P. (1996). Comparison of transcranial doppler ultrasonography and transesophageal echocardiography to monitor emboli during coronary artery bypass surgery. *Stroke, 27,* 87-90.

Barlow, D. H. (1981). On the relation of clinical research to clinical practice: Current issues, new directions. *Journal of Consulting and Clinical Psychology, 49,* 147-155.

Basile, A. M., Fusi, C., Conti, A. A., Paniccia, R., Trefoloni, G., & Pracucci, G. et al. (2001). S-100 protein and neuron specific enolase as markers of subclinical cerebral damage after cardiac surgery: Preliminary observation of a 6-month follow-up study. *Eruopean Neurology, 45,* 151-159.

Bendszus, M., Reents, W., Franke, D., Mullges, W., Babin-Ebell, J., & Koltzenburg, M. (2002). Brain damage after coronary artery bypass grafting. *Archives of Neurology, 59,* 1090-1095.

Benedict, R. H. B. (1994). Cognitive function after open-heart surgery: Are postoperative neuropsychological deficits caused by cardiopulmonary bypass? *Neuropsychology Review, 4,* 223-255.

Berger, P. B. (2004). Acute Myocardial Infarction. In D. B. Federman & D. C. Dale (Eds.), *ACP Medicine* (pp. 17). New York, NY: WebMD.

Bezeau, S., & Graves, R. (2001). Statistical power and effect sizes of clinical neuropsychological research. *Journal of Clinical and Experimental Neuropsychology, 23,* 399-406.

Bird, C. M., Papadopoulou, K., Ricciardelli, P., Rossor, M. N., & Cipolotti, L. (2004). Monitoring cognitive changes: Psychometric properties of six cognitive tests. *British Journal of Clinical Psychology, 43,* 197-216.

Blackstone, E. H. (2000). Editorial: Neurologic injury from cardiac surgery – An important but enormously complex phenomenon. *The Journal of Thoracic and Cardiovascular Surgery, 120,* 629-631.

Blauth, C. I. (1995). Macroemboli and microemboli during cardiopulmonary bypass. *Annals of Thoracic Surgery, 59,* 1300-1303.

Blumenthal, J. A., Mahanna, E. P., Madden, D. J., White, W. D., Croughwell, N. D., & Newman, M. F. (1995). Methodological issues in the assessment of neuropsychologic function after cardiac surgery. *Annals of Thoracic Surgery, 59,* 1345-1350.

Borg, I. (1993). How to construct indices in facet theory. *Proceedings of the Fourth International Facet Theory Conference August 29 to September 1 1993. Prague, Czech Republic. Editor Dov Elizur.*

Borger, M. A., & Feindel, C. M. (2002). Cerebral emboli during cardiopulmonary bypass: effect of perfusionist interventions and aortic cannulas. *Journal of Extra-Corporeal Technology, 34,* 29-33.

Borowicz, L. M., Goldsborough, M. A., Selnes, O. A., & McKhann, G. M. (1996). Neuropsychologic change after cardiac surgery: A critical review. *Journal of Cardiothoracic and Vascular Anesthesia, 10,* 105-112.

Braekken, S. K., Reinvang, I., Russell, D., Brucher, R., & Svennevig, J. L. (1998). Association between intraoperative cerebral microembolic signals and postoperative neuropsychological deficit: Comparison between patients with cardiac valve replacement and patients with coronary artery bypass grafting. *Journal of Neurology, Neurosurgery, and Psychiatry, 65,* 573-576.

Bristow, M. R., Port, D. J., & Kelly, R. A. (2001). Treatment of Heart Failure: Pharmacological Methods. In E. Braunwald & D. P. Zipes & P. Libby (Eds.), *Heart Disease: A Textbook of Cardiovascular Medicine* (pp. 562-599). London: W. B. Saunders.

Brown, J. (1985). An introduction to the uses of facet theory. In D. Canter (Ed.). Facet Theory Approaches to Social Research. New York: Springer Verlag.

Brown, J., & Barnett, J. (2000). Facet theory: an approach to research. In G. M. Breakwell, S. Hammond, & C Fife-Schaw (2000). Research Methods in Psychology. Sage: London.

Brown, S. M., Halligan, P. W., Wade, D. T., & Taggart, D. P. (1999). Cognitive performance after cardiac operation: Implication of regression to the mean. *Journal of Thoracic and Cardiovascular Surgery, 117,* 481-485.

Bruggemans, E. F., van Dijk, J. G., & Huysmans, H. A. (1995). Residual cognitive dysfunction at 6 months following coronary artery bypass graft surgery. *European Journal of Cardiothoracic Surgery, 9,* 636-643.

Bruggemans, E. F., van de Vijver, F. J. R., &Huysmans, H. A. (1999). Defining neuropsychological deterioration after cardiac surgery. *Annals of Thoracic Surgery, 67,* 297.

Cannon, C. P., & Braunwald, E. (2001). Unstable Angina. In E. Braunwald & D. P. Zipes & P. Libby (Eds.), *Heart Disease: A Textbook of Cardiovascular Medicine* (pp. 1232-1263). London: W. B. Saunders.

Canter, D. (1985). How to be a facet researcher. In D. Canter (Ed.). Facet Theory Approaches to Social Research. New York: Springer Verlag.

Caplan, L. (1999). Cerebrovascular disease and neurologic manifestations of heart disease. In R. W. Alexander, R. C. Schlant, V. Fuster, R. A. O'Rourke, R. Roberts, & E. H. Sonnenblick (Eds.), *Hurst's The Heart, Ninth Edition, Companion Handbook.* Singapore: McGraw Hill.

Carabello, B. A. (2004). Valvular Heart Disease and Heart Failure. In M. D. Mann (Ed.), *Heart Failure: A Companion to Braunwald's Heart Disease* (pp. 415-426). Philadelphia: Saunders.

Chabot, R. J., Gugino, L. D., Aglio, L. S., Maddi, R., & Cote, W. (1997). QEEG and neuropsychological profiles of patients after undergoing cardiopulmonary bypass surgical procedures. *Clinical Electroencephalography, 28,* 98-105.

Chaytor, N., & Schmitter-Edgecombe, M. (2003). The ecological validity of neuropsychological tests: A review of the literature on everyday cognitive skills. *Neuropsychological Review, 13,* 181-197.

Clark, R. E., Brillman, J., Davis, D. A., Lovell, M. R., Price, T. R. P., & Magovern, G. J. (1995). Microemboli during coronary artery bypass grafting. *Journal of Thoracic and Cardiovascular Surgery, 109,* 249-258.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65,* 145-153.

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of Clinical Psychology* (pp.95-121). New York: McGraw Hill.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7,* 249-253.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (Second Edition). London: Lawrence Erlbaum Associates.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304-1312.

Cohen, J. (1992a). A power primer. *Psychological Bulletin, 112,* 155-159.

Cohen, J. (1992b). Statistical Power Analysis. *Current Directions in Psychological Science, 1*, 98-101.

Dajani, A. S. (2001). Rheumatic Fever. In E. Braunwald & D. P. Zipes & P. Libby (Eds.), *Heart Disease: A Textbook of Cardiovascular Medicine* (6th ed., pp. 2192-2198). London: W. B. Saunders.

Davies, J. (2003). *Coronary Revascularisation in Australia* (Bulletin No. 7). Canberra: Australian Institute of Health and Welfare.

Davies, J., & Senes, S. (2001). *Cardiac Surgery in Australia 1998* (Cardiovascular Disease Series No. 16). Canberra: Australian Institute of Health and Welfare & National Heart Foundation of Australia.

Davies, J., & Senes, S. (2002). *Coronary Angioplasty in Australia.* Canberra: Australian Institute of Health and Welfare and National Heart Foundation.

Davies, J., & Senes, S. (2003). *Cardiac Surgery in Australia 1999* (Cardiovascular Disease Series No. 21). Canberra: Australian Institute of Health and Welfare & National Heart Foundation of Australia.

Dewey, T. M., & Edgerton, J. R. (2003). Complications related to off-pump bypass grafting. *Seminars in Thoracic and Cardiovascular Surgery, 15,* 63-70.

Di Carlo, A., Perna, A. M., Pantoni, L., Basile, A. M., Bonacchi, M., & Pracucci, G., et al. (2001). Clinically relevant cognitive impairment after cardiac surgery: A 6-month follow-up study. *Journal of Neurological Sciences, 188,* 85-93.

Di Nardo, J. A., & Wegner, J. A. (2001). Pro: Low-flow cardiopulmonary bypass is the preferred technique for patients undergoing cardiac surgical procedures. *Journal of Cardiothoracic and Vascular Anesthesia, 15,* 649-651.

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1,* 170-177.

Dyer, G. S. M., & Fifer, M. A. (2003). Heart Failure. In L. S. Lilly (Ed.), *Pathophysiology of Heart Disease: A Collaborative Project of Medical Students and Faculty* (3rd ed., pp. 211-236). Sydney: Lippincott Williams & Wilkins.

Ebert, A. S., Walzer, T. A., Huth, C., & Herrmann, M. (2001). Early neurobehavioral disorders after cardiac surgery: A comparative analysis of coronary artery bypass graft surgery and valve replacement. *Journal of Cardiothoracic and Vascular Anesthesia, 15,* 15-19.

Eifert, S., Reichenspurner, H., Pfefferkorn, T., Baur, B., von Schlippenbach, C., & Mayer, T. E., et al. (2003). Neurological and neuropsychological examination and outcome after use of an intra-aortic filter device during cardiac surgery. *Perfusion, 18,* 55-60.

Elias, M. F., Elias, P. K., Robbins, M. A., Wolf, P. A., & D'Agostino, R. B. (2001). Cardiovascular Risk Factors and Cognitive Functioning: An Epidemiological Perspective. In S. R. Waldstein and M. F. Elias (Eds.), *Neuropsychology of Cardiovascular Disease* (pp. 83-104). London: Laurence Erlbaum.

Engelhardt, W., Dierks, T., Pause, M., & Hartung, E. (1996). Early cerebral functional outcome after coronary artery bypass surgery using different acid-base management during hypothermic cardiopulmonary bypass. *Acta Anaesthesiologica Scandinavica, 40,* 457-465.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers, 28,* 1-11).

Everson, S. A., Helkala, E., Kaplan, G. A., & Salonen, J. T. (2001). Artherosclerosis and Cognitive Functioning. In S. R. Waldstein and M. F. Elias (Eds.), *Neuropsychology of Cardiovascular Disease* (pp. 105-120). London: Laurence Erlbaum.

Farsak, B., Gunaydin, S., Yorgancioglu, C., Zorlutuna, Y. (2003). Elevated levels of S-100β correlate with neurocognitive outcome after cardiac surgery. *Journal of Cardiovascular Surgery, 44,* 31-35.

Fearn, S. J., Pole, R., Wesnes, K., Faragher, E. B., Hooper, T. L., & McCollum, C. N. (2001). Cerebral injury during cardiopulmonary bypass: Emboli impair memory. *Journal of Thoracic Cardiovascular Surgery, 121,* 1150-1160.

Fenster, P. E., Sox, H. C., & Alpert, J. S. (2004). Ischemic Heart Disease: Angina Pectoris. In D. B. Federman & D. C. Dale (Eds.), *ACP Medicine* (pp. 18). New York, NY: WebMD.

Follette, W. C., & Callaghan, G. M. (1996). The importance of the principle of clinical significance – defining significant to whom and for what purpose: A response to Tingey, Lambert, Burlingame, and Hansen. *Psychotherapy Research, 6,* 133-143.

Galliker, M., Weimer, D., & Wagner, F. (1995). The contribution of facet theory to the interpretation of findings: Discussing the validation of basic facets of the model of verbal discrimination. In J. J. Hox, P. G. Swanborn, and G. J. Mellenbergh (Eds.), *Facet Theory: Analysis and Design*. Zeist: Setos

Gardner, T. J., Horneffer, P. J., Manolio, T. A., Pearson, T. A., Gott, V. L., & Baumgartner, W. A., et al. (1985). Stroke following coronary artery bypass grafting: A ten-year study. *The Annals of Thoracic Surgery, 40,* 574-581.

Gaziano, J. M. (2001). Global Burden of Cardiovascular Disease. In E. Braunwald & D. P. Zipes & P. Libby (Eds.), *Heart Disease: A Textbook of Cardiovascular Medicine* (pp. 1-18). Sydney: W. G. Saunders.

Gibbs, N. M., Mahon, B. D., Newman, M. A. J., Prins, A., & Weightman, W. M. (2001). Low serum S100β protein levels following coronary artery surgery with or without cardiopulmonary bypass. *Heart, Lung and Circulation, 10,* 48-52.

Gill, R., & Murkin, J. M. (1996). Neuropsychologic dysfunction after cardiac surgery: What is the problem? *Journal of Cardiothoracic and Vascular Anesthesia, 10,* 91-98.

Godet, G., Watremez, C., Beaudeux, J. L., Meersschaert, K., Koskas, F., & Coriat, P. (2001). S-100β protein levels do not correlate with stroke in patients undergoing carotid endarterectomy under general anesthesia. *Journal of Cardiothoracic and Vascular Anesthesia, 15,* 25-28.

Gordon, M. B., & Libby, P. (2003). Atherosclerosis. In P. Libby (Ed.), *Pathophysiology of Heart Disease: A Collaborative Project of Medical Students and Faculty* (pp. 111-130). Sydney: Lippincott Williams & Wilkins.

Gray, H. (1995). *Anatomy: Descriptive and Surgical* (15th ed.). New York: Barnes & Noble.

Green, A. F., & Sears, S. F. (1994). Psychometric assessment of cardiac transplantation candidates. *Journal of Clinical Psychology in Medical Settings, 1,* 135-147.

Griffin, B. P. (2004). Valvular Heart Disease. In D. B. Federman & D. C. Dale (Eds.), *ACP Medicine* (pp. 19). New York, NY: WebMD.

Grigore, A. M., Grocott, H. P., Mathew, J. P., Phillips-Bute, B., Stanley, T. O., & Butler, A., et al. (2002). The rewarming rate and increased peak temperature alter neurocognitive outcome after cardiac surgery. *Anesthesia and Analgesia, 94,* 4-10.

Grigore, A. M., Mathew, J. P., Grocott, H. P., Reves, J. G., Blumenthal, J. A., & White, W. D., et al. (2001). Prospective randomised trial of normothermic versus hypothermic cardiopulmonary bypass on cognitive function after coronary artery bypass graft surgery. *Anesthesiology, 95,* 1110-1119.

Grimm, M., Czerny, M., Baumer, H., Kilo, J., Madl, C., & Kramer, L. et al. (2000). Normothermic cardiopulmonary bypass is beneficial for cognitive brain function after coronary artery bypass grafting – a prospective randomised trial. *European Journal of Cardio-thoracic Surgery, 18,* 270-275.

Grocott, H. P., Mackensen, B., Grigore, A. M., Mathew, J., Reves, J. G., & Phillips-Bute, B. et al. (2002). Postoperative hyperthermia is associated with cognitive dysfunction after coronary artery bypass graft surgery. *Stroke, 33,* 537-541.

Gugino, L. D., Chabot, R. J., Aglio, L. S., Aranki, S., Dekkers, R., & Maddi, R. (1999). QEEG during cardiopulmonary bypass: Relationship to postoperative neuropsychological function. *Clinical Electroencephalography, 30,* 53-63.

Gugino, L. D., Chabot, R. J., Aglio, L. S., Maddi, R., Gosnell, J., & Aranki, S. (1997). QEEG and neuropsychological profiles of patients prior to undergoing cardiopulmonary bypass surgical procedures. *Clinical Electroencephalography, 28,* 87-97.

Gunasinghe, S. K., & Spinale, F. G. (2004). Myocardial Basis for Heart Failure: Role of the Cardiac Interstitium. In M. D. Mann (Ed.), *Heart Failure: A*

*Companion to Braunwald's Heart Disease* (pp. 57-70). Philadelphia: Saunders.

Gupta, A., Sabatine, M. S., & Lilly, L. S. (2003). Ischemic Heart Disease. In L. S. Lilly (Ed.), *Pathophysiology of Heart Disease: A Collaborative Project of Medical Students and Faculty* (3rd ed., pp. 131-156). Sydney: Lippencott Williams & Wilkins.

Guttman, R., Epstein, E. E., Amir, M., & Guttman, L. (1990). A structural theory of spatial abilities. *Applied Psychological Measurement, 14,* 217-236.

Guttman, R., & Shoham, I. (1982). The structure of spatial ability items: A faceted analysis. *Perceptual and Motor Skills, 54,* 487-493.

Guttman, R., & Zohar, A. (1999). A structural analysis of reasoning ability tests: verbal, spatial, and mathematical. In Ruth Meyer-Schweizer (Ed.). Facet Theory Conference. Design and Analysis. Proceedings of the Seventh Facet Theory Conference 25-28 July 1999. Berne University. Die Deutsche Bibliothek.

Haddock, C. K., Carlos Poston, W. S., & Taylor, J. E. (2003). Neurocognitive Sequelae Following Coronary artery bypass graft: A research agenda for behavioural scientists. *Behaviour Modification, 27,* 68-82.

Hageman, W. J. J. M., & Arrindell, W. A. (1999). Establishing clinically significant change: increment of precision and the distinction between individual and group level of analysis. *Behaviour Research and Therapy, 37,* 1169-1193.

Haines, D. E. (1995). *Neuroanatomy: An Atlas of Structures, Sections, and Systems* (4th ed.). Baltimore, Maryland: Williams & Wilkins.

Hair, J. F. Jr., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate Data Analysis: with Readings* (4th Ed.). New Jersey: Prentice Hall.

Hall, R. A., Fordyce, D. J., Lee, M. E., Eisenberg, B., Lee, R. F., & Holmes, J. H., et al. (1999). Brain SPECT imaging and neuropsychological testing in coronary artery bypass patients. *Annals of Thoracic Surgery, 68,* 2082-2088.

Hammon, J. W., Stump, D. A., Kon, N. D., Cordell, A. R., Hudspeth, A. S., & Oaks, T. E., et al. (1997). Risk factors and solutions for the development of neurobehavioral changes after coronary artery bypass grafting. *Annals of Thoracic Surgery, 63,* 1613-1618.

Hammond, G. (1996). The objections to null hypothesis testing as a means of analysing psychological data. *Australian Journal of Psychology, 48,* 104-106.

Hans, S., Bernstein, V., & Marcus, J. (1985). Some uses of the facet approach in child development. In David Canter (Ed.). Facet Theory Approaches to Social Research. New York: Springer Verlag.

Harrison, M. J. G. (1995). Neurologic complications of coronary artery bypass grafting: Diffuse or focal ischemia? *Annals of Thoracic Surgery, 59,* 1356-1358.

Heaton, R. K., Temkin, N., Dikmen, S., Avitable, N., Taylor, M. J., & Marcotte, T. D., et al. (2001). Detecting change: A comparison of three neuropsychological methods, using normal and clinical samples. *Archives of Clinical Neuropsychology, 16,* 75-91.

Herrmann, M., Ebert, A. D., Galazky, I., Wunderlich, M. T., Kunz, W. S., & Huth, C. (2000). Neurobehavioral outcome prediction after cardiac surgery: Role of neurobiochemical markers of damage to neuronal and glial brain tissue. *Stroke, 31,* 645-650.

Herrmann, M., Ebert, A. D., Tober, D., Hann, J., & Huth, C. (1999). A contrastive analysis of release patterns of biochemical markers of brain damage after coronary artery bypass grafting and valve replacement and their associations with the neurobehavioral outcome after cardiac surgery. *European Journal of Cardio-Thoracic Surgery, 16,* 513-518.

Heyer, E. J., Lee, K. S., Manspeizer, H. E., Mongero, L., Spanier, T. B., Caliste, X. et al. (2002). Heparin-bonded cardiopulmonary bypass circuits reduce cognitive dysfunction. *Journal of Cardiothoracic and Vascular Anesthesia, 16,* 37-42.

Hlatky, M. A., Bacon, C., Boothroyd, D., Mahanna, E., Reves, J. G., Newman, M. F., et al. (1997). Cognitive function 5 years after randomisation to coronary angioplasty or coronary artery bypass graft surgery. *Circulation, 56* (Suppl. 2), II-11-14

Hofste, W. J., Linssen, C. A. M., Boezman, E. H. J. F., Hengeveld, J.S., Leusink, J. A., & de Boer, A. (1997). Delirium and cognitive disorders after cardiac operations: Relationship to pre- and intraoperative quantitative electroencephalogram. *International Journal of Clinical Monitoring and Computing, 14,* 29-36.

Hollon, S. D., & Flick, S. N. (1988). On the meaning and methods of clinical significance. *Behavioral Assessment, 10,* 197-206.

Howell, D. C. (1997). *Statistical Methods for Psychology* (4th Ed.). London: Duxbury.

Hsu, L. M. (1989). Reliable change in psychotherapy: Taking into account regression to the mean. *Behavioral Assessment, 11,* 459-467.

Huismans, S. (1999). Personal values and presentation of self to other people. In R. Meyer-Schweizer (Ed.), *Facet Theory: Design and Analysis* (pp.39-57). Bern: Institut für Soziologie der Universität Bern.

Huismans, S. F., & van de Vliert, (2001). Hofstede's cultural dimensions reanalyzed: Four independent dimensions of facets of one dimension called "Development". In D. Elizur (Ed.), *Facet Theory: Integrating Theory Construction with Data Analysis* (pp. 33-40). Prague: Matfyz.

Iglesias, I., & Murkin, J. M. (2001). Beating heart surgery or conventional CABG: Are neurologic outcomes different? *Seminars in Thoracic and Cardiovascular Surgery, 13,* 158-169.

Isgro, F., Schmidt, Ch. Pohl, P., & Saggau, W. (1997). A predictive parameter in patients with brain related complications after cardiac surgery. *European Journal of Cardio-thoracic Surgery, 11,* 640-644.

Jacobs, A., Neveling, M., Horst, M., Ghaemi, M., Kessler, J., & Eichstaedt, H., et al. (1998). Alterations of neuropsychological function and cerebral glucose metabolism after cardiac surgery are not related only to intraoperative microembolic events. *Stroke, 29,* 660-667.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15,* 336-352.

Jacobson, N. S., & Revenstorf, D. (1988). Statistics for assessing the clinical significance of psychotherapy techniques: Issues, problems, and new developments. *Behavioral Assessment, 10,* 133-145.

Jacobson, N. S., Roberts, L. J., Berns, S. B., & McGlinchey, J. B. (1999). Methods for defining and determining the clinical significance of treatment effects: Description, application, and alternatives. *Journal of Consulting and Clinical Psychology, 67,* 300-307.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology, 59,* 12-19.

Jamrozik, K., Dobson, A., Hobbs, M., McElduff, P., Ring, I., & D'Este, K., et al. (2001). *Monitoring the Incidence of Cardiovascular Disease in Australia.* Canberra: Australian Institute of Health and Welfare.

Jann, B. (1999). Examining validity of behavioural indicators: An application of smallest space analysis. In Ruth Meyer-Schweizer (Ed.). Facet Theory Conference. Design and Analysis. Proceedings of the Seventh Facet Theory Conference 255-28 July 1999. Berne University. Die Deutsche Bibliothek.

Johnsson, P. (1996). Markers of cerebral ischemia after cardiac surgery. *Journal of Cardiothoracic and Vascular Anesthesia, 10,* 120-126.

Jonsson, H., Johnsson, P., Alling, C., Backstrom, M., Bergh, C., & Blomquist, S. (1999). S100β after coronary artery surgery: Release pattern, source contamination, and relation to neuropsychological outcome. *Annals of Thoracic Surgery, 68,* 2202-2208

Kadoi, Y., Saitoi, S., Goto, F., & Fujita, N. (2001). Decrease in jugular venousoxygenation saturation during normothermic cardiopulmonary bypass predicts short-term postoperative neurologic dysfunction in elderly patients. *Journal of the American College of Cardiology, 38,* 1450-1455.

Kannel, W. B., & Vasan, R. S. (2004). Epidemiology of Heart Failure. In M. D. Mann (Ed.), *Heart Failure: A Companion to Braunwald's Heart Disease* (pp. 339-350). Philadelphia: Saunders.

Karni, E., & Levin, J. (1972). The use of smallest space analysis in studying scale structure: An application to the California Psychological Inventory. *Journal of Applied Psychology, 566,* 341-346.

Katz, A. M. (2004). Evolving Concepts of the Pathophysiology of Heart Failure. In M. D. Mann (Ed.), *Heart Failure: A Companion to Braunwald's Heart Disease* (pp. 1-9). Philadelphia: Saunders.

Kaufman, A. S., McLean, J. E., & Reynolds, J. C. (1995). The fluid and crystallized abilities of white, black, and Hispanic adolescents and adults, both with and without an education covariate. *Journal of Clinical Psychology, 51,* 637-647.

Keith, J. R., & Puente, A. E. (2002). Deficiencies in the detection of cognitive deficits. *Neuropsychology, 16,* 434-439.

Keith, J. R., Puente, A. E., Marks, H. F., Malcolmson, K. L., Tartt, S., & Coleman, A. E. (2002). Assessing postoperative cognitive change after cardiopulmonary bypass surgery. *Neuropsychology, 16,* 411-421.

Kilminster, S., Treasure, T., McMillan, T., & Holt, D. W. (1999). Neuropsychological change in S-100 protein release in 130 unselected patients undergoing cardiac surgery. *Stroke, 30,* 1869-1874.

Kilo, J., Czerny, M., Gorlitzer, M., Zimpfer, D., Baumer, H., & Wolner, E., et al. (2001). Cardiopulmonary bypass affects cognitive brain function after coronary artery bypass grafting. *Annals of Thoracic Surgery, 72,* 1926-1932.

Kneebone, A. C., Andrew, M. J., Baker, R. A., & Knight, J. L. (1998). Neuropsychologic changes after coronary artery bypass grafting: Use of reliable change indices. *Annals of Thoracic Surgery, 65,* 1320-1325.

Kupersmith, J., & Raval, A. (2004). Unstable Angina and Non-ST Segment Elevation Myocardial Infarction. In D. B. Federman & D. C. Dale (Eds.), *ACP Medicine* (pp. 19). New York, NY: WebMD.

Levav, M., Mirsky, A. F., French, L. M., & Bartko, J. J. (1998). Multinational neuropsychological testing: Performance of children and adults. *Journal of Clinical and Experimental Neuropsychology, 20,* 658-672.

Levy, S. (1993). Facet theory: An overview of cumulative theory construction. *Proceedings of the Fourth International Facet Theory Conference August 29 to September 1 1993. Prague, Czech Republic. Editor Dov Elizur.*

Lezak, M. D. (1995). *Neuropsychological Assessment* (3rd Edition). Oxford: Oxford University.

Li, C. M. (2004a). Coronary Surgery and Angina. In P. R. Soltoski, H. L. Karamanoukian, & T. A. Salerno (Eds.), *Cardiac Surgery Secrets* (2nd ed., pp. 95-96). Philadelphia: Hanley & Belfus.

Li, C. M. (2004b). Hypothermia. In P. R. Soltoski, H. L. Karamanoukian, & T. A. Salerno (Eds.), *Cardiac Surgery Secrets* (2nd ed., pp. 283-284). Philadelphia: Hanley & Belfus.

Libby, P. (2001). The Vascular Biology of Atherosclerosis. In E. Braunwald & D. P. Zipes & P. Libby (Eds.), *Heart Disease: A Textbook of Cardiovascular Medicine* (pp. 995-1009). London: W. B. Saunders.

Liberthson, R. R. (2000). Management of Valvular Heart Disease. In A. H. Goroll & A. G. J. Mulley (Eds.), *Primary Care Medicine* (pp. 16). Sydney: Lippincott Williams & Wilkins.

Linsted, U., Meyer, O., Kropp, P., Berkau, A., Tapp, E., & Zenz, M. (2002). Serum concentration of S-100 protein in assessment of cognitive dysfunction after general anesthesia in different types of surgery. *Acta Anaesthesiologica Scandinavica, 46,* 384-389.

Lovibond, S. H., & Lovibond, P. F. (1995). *Manual for the Depression Anxiety Stress Scales* (2nd ed.). Sydney, Australia: Psychology Foundation of Australia.

Maassen, G. H. (2000). Principles of defining reliable change indices. *Journal of Clinical and Experimental Neuropsychology, 22,* 622-632.

Malhotra, R., Edelman, E. R., & Lilly, L. S. (2003). Basic Cardiac Structure and Function. In L. S. Lilly (Ed.), *Pathophysiology of Heart Disease: A Collaborative Project of Medical Students and Faculty* (3rd ed., pp. 1-28). Sydney: Lippincott Williams & Wilkins.

Mann, D. L. (2004). Heart Failure as a Progressive Disease. In M. D. Mann (Ed.), *Heart Failure: A Companion to Braunwald's Heart Disease* (pp. 123-128). Philadelphia: Saunders.

Maraun, M. D. (1997). Appearance and reality: Is the big five the structure of trait descriptors? *Personality and Individual Differences, 22,* 629-647.

Mathes, D. D., Stone, D. J., & Dent, J. M. (2001). Preoperative cardiac risk stratification: Ritual or requirement? *Journal of Cardiothoracic and Vascular Anesthesia, 15,* 626-630.

Mathur, S. (2002). *Epidemic of Coronary Heart Disease and its Treatment in Australia* (Cardiovascular Disease Series No. 20). Canberra: Australian Institute of Health and Welfare.

Mc Daid, C. M., Lewis, S. A., McMurray, T., & Phillips, T. (1994). Heart surgery: What are the neuropsychological consequences? *The Irish Journal of Psychology, 15,* 110-125.

Melling, C. A. (2001). Critical realism in philosophical bases for facet theory. *Facet Theory: Integrating Theory Construction with Data Analysis. Facet Theory Conference July 15-18 2001 Prague, Czech Republic. Editor Dov Elizur.*

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Miller, E. N. (1990). *California Computerised Assessment Battery (CalCAP) Manual*. Los Angeles: Norland Software.

Millis, S. R., Malina, A. C., Bowers, D. A., & Ricker, J. H. (1999). Confirmatory factor analysis of the Wechsler Memory Scale-III. *Journal of Clinical and Experimental Neuropsychology, 21,* 87-93.

Mills, S. A. (1993). Cerebral injury and cardiac operations. *Annals of Thoracic Surgery, 56,* S86-S91.

Mills, S. A. (1995). Risk factors for cerebral injury and cardiac surgery. *Annals of Thoracic Surgery, 59,* 1296-1299.

Mitchell, S., & Gorman, D. (2002). The pathophysiology of cerebral arterial gas embolism. *Journal of Extra-Corporeal Technology, 34,* 18-23.

Moller, J. T., Gluitmans, P., Rasmussen, L. S., Houx, P., Rasmussen, H., & Canet, J., et al. (1998). Long-term postoperative cognitive dysfunction in the elderly: ISPOCD1 study. *Lancet, 351,* 857-861.

Moody, D. M., Brown, W. R., Challa, V. R., Stump, D. A., Reboussin, D. M., & Legault, C. (1995). Brain microemboli associated with cardiopulmonary bypass: A histologic and magnet resonance imaging study. *Annals of Thoracic Surgery, 59,* 1304-1307.

Moreno, R. & Carmona, J. (1999). Developing mapping sentences using methodological dimensions of concepts. In Ruth Meyer-Schweizer (Ed.). Facet Theory Conference. Design and Analysis. Proceedings of the Seventh Facet Theory Conference 25-28 July 1999. Berne University. Die Deutsche Bibliothek.

Muldoon, M. F., Flory, J. D., & Ryan, C. M. (2001). Serum Cholesterol, the Brain, and Cognitive Functioning. In S. R. Waldstein and M. F. Elias (Eds.), *Neuropsychology of Cardiovascular Disease* (pp. 37-60). London: Laurence Erlbaum.

Mullges, W., Franke, D., Reents, W., Babin-Ebell, J., & Toyka, K. V. (2003). Reduced rate of microembolism by optimizes aortic cannula position does not influence early postoperative cognitive performance in CABG patients. *Cerebrovascular Diseases, 15,* 192-198.

Murkin, J. M. (1995). Introduction. *Annals of Thoracic Surgery, 59,* 1288.

Murkin, J. M. (1995). The role of CPB management in neurobehavioural outcomes after cardiac surgery. *Annals of Thoracic Surgery, 59,* 1308-1311.

Murkin, J. M. (2001). Perioperative neuropsychological testing. *Journal of Cardiothoracic and Vascular Anesthesia, 15,* 1-3.

Murkin, J. M., Boyd, W. D., Ganapathy, S., Adams, S. J., & Peterson, R. C. (1999). Beating heart surgery: Why expect less central nervous system morbidity? *Annals of Thoracic Surgery, 68,* 1498-1501.

Murkin, J. M., Martzke, J. S., Buchan, A. M., Bentley, C., & Wong, C. J. (1995). A randomized study of the influence of perfusion technique and pH management strategy in 316 patients undergoing coronary artery bypass surgery. *Journal of Thoracic and Cardiovascular Surgery, 110,* 349-362.

Murkin, J. M., Newman, S. P., Stump, D. A., & Blumenthal, J. A. (1995). Statement of consensus on assessment of neurobehavioural outcomes after cardiac surgery. *Annals of Thoracic Surgery, 59,* 1289-1295.

Murkin, J. M., Stump, D. A., Blumenthal, J. A., & McKhann, G. (1997). Defining Dysfunction: Group means versus incidence analysis – A statement of consensus. *Annals of Thoracic Surgery, 64,* 904-905.

Murphy, K. R., & Davidshofer, C. O. (1998). Psychological Testing: Principles and Applications ($4^{th}$ Edition). New Jersey: Prentice Hall.

Ndumele, C. E., Friedberg, M., Antman, E. M., Strichartz, G. R., & Lilly, L. S. (2003). Cardiovascular Drugs. In L. S. Lilly (Ed.), *Pathophysiology of Heart Disease: A Collaborative Project of Medical Students and Faculty* (3rd ed., pp. 371-421). Sydney: Lippincott Williams & Wilkins.

Netter, F. H. (1969). *Heart* (Vol. 5). New York: CIBA Pharmaceutical Company.

Neville, M. J., Butterworth, J., James, R. L., Hammon, J. W., & Stump, D. A. (2001). Similar neurobehavioral outcome after valve or coronary artery operations despite differing carotid embolic counts. *Journal of Thoracic and Cardiovascular Surgery, 121,*125-136.

Newman, M. F., Booth, J. V., Laskowitz, D. T., Schwinn, D. A., Grocott, H. P., & Mathew, J. P. (2001). Genetic predictors of Perioperative neurological and cognitive injury and recovery. *Best Practice and Research in Clinical Anaesthesiology, 15,* 247-276.

Newman, M. F., Croughwell, N. D., Blumenthal, J. A., Lowry, E., White, W. D., & Spillane, W., et al. (1995). Predictors of cognitive decline after cardiac operation. *Annals of Thoracic Surgery, 59,* 1326-1330.

Newman, M. F., Grocott, H. P., Mathew, J. P., White, W. D., Landolfo, K., & Reves, J. G., et al. (2001). Report of the substudy assessing the impact of neurocognitive function on quality of life 5 years after cardiac surgery. *Stroke, 32,* 2874-2881.

Newman, M. F., Kirchner, J. L., Phillips-Bute, B., Gaver, V., Grocott, H., & Jones, R. H., et al. (2001). Longitudinal assessment of neurocognitive function after coronary artery bypass surgery. *The New England Journal of Medicine, 344,* 395-402.

Newman, M. F., Kramer, D., Croughwell, N. D., Sanderson, I., Blumenthal, J. A., & White, W. D., et al. (1995). Differential age effects of mean arterial pressure and rewarming on cognitive dysfunction after cardiac surgery. *Anesthesia and Analgesia, 81,* 236-242.

Newman, S. P. (1995). Analysis and interpretation of neuropsychologic tests in cardiac surgery. *Annals of Thoracic Surgery, 59,* 1351-1355.

Newman, S., Smith, P., Treasure, T., Joseph, P., Ell, P., & Harrison, M. (1987). Acute neuropsychological consequences of coronary artery bypass surgery. *Current Psychological Research & Reviews, 6,* 115-124.

Newman, S., & Stygall, J. (1999). Changes in cognition following cardiac surgery. *Heart, 82,* 541-542.

Newman, S., Stygall, J., & Kong, R. (2001). Neuropsychological Consequences of Coronary Artery Bypass Surgery. In S. R. Waldstein and M. F. Elias (Eds.), *Neuropsychology of Cardiovascular Disease* (pp. 189-218). London: Laurence Erlbaum.

Nollert, G., Mohnle, Tassani-Prell, P., Uttner, I. Borasio, G. D., Schmoeckel, M., et al. (1995). Postoperative neuropsychological dysfunction and cerebral oxygenation during cardiac surgery. *Thoracic and Cardiovascular Surgery, 43,* 260-264.

Nolte, J. (2002). *The Human Brain: An Introduction to Its Functional Anatomy* (5th ed.). Sydney: Mosby.

Nussmeier, N. & Fish, K. J. (1991). Neuropsychological dysfunction after cardiopulmonary bypass: A comparison of two institutions. *Journal of Thoracic and Cardiovascular Anesthesia, 5,* 584-588.

Ogden, J. A. (2001). First do no harm. Culturally-appropriate neuropsychological assessment for indigenous people: A position paper. *Brain Impairment, 2,* 1-10.

Ogden, J. A., Cooper, E., & Dudley, M. (2003). Adapting neuropsychological assessments for minority groups: A study comparing white and Maori New Zealanders. *Brain Impairment, 4,* 122-134.

Ogden, J. A., & McFarlan-Nathan, G. (1999). Cultural bias in the neuropsychological assessment of young Maori men. *New Zealand Journal of Psychology, 26,* 2-12.

Opie, L. H. (2001). Mechanisms of Cardiac Contraction and Relaxation. In E. Braunwald & D. P. Zipes & P. Libby (Eds.), *Heart Disease: A Textbook of Cardiovascular Medicine* (6th ed., pp. 443-478). London: W. B. Saunders.

Otto, C. M. (2004). *Valvular Heart Disease* (2nd ed.). Philadelphia: Saunders.

Patel, R. L., Turtle, M. R., Chambers, D. J., James, D. N., Newman, S., & Venn, G. E. (1996). Alpha-stat acid-base regulation during cardiopulmonary bypass improves neuropsychologic outcome in patients undergoing coronary artery bypass grafting. *Journal of Thoracic and Cardiovascular Surgery, 111,* 1267-1279.

Peters, K. R., Graf, P., Hayden, S., & Feldman, H. (2004). Neuropsychological characterization of cognitively-impaired-not-demented (CIND) individuals: Clinicl comparison data. *The Clinical Neuropsychologist, 18,* 208-228.

Phillips, N. A. (2001). Thinking on Your Feet: A Neuropsychological Review of Peripheral Vascular Disease. In S. R. Waldstein and M. F. Elias (Eds.), *Neuropsychology of Cardiovascular Disease* (pp. 121-142). London: Laurence Erlbaum.

Pinna-Pintor, P., Bobbio, M., Giammaria, M., Maisano, F., Cuni, D., & Alfieri, O. (1998). Comparison of three methods of presurgical risk stratification in cardiac surgery. *Asia Pacific Heart Journal, 7,* 92-97.

Poreh, A., & Shye, S. (1998). Examination of the global and local features of the Rey Osterrieth Complex Figure using faceted smallest space analysis. *The Clinical Neuropsychologist, 12,* 453-467.

Prentice, D. A., & Miller, D. T. (1992). When small effects are impressive. *Psychological Bulletin, 112,* 160-164.

Price, L. R., Tulsky, D., Millis, S., & Weiss, L. (2002). Redefining the factor structure of the Wechsler Memory Scale –III: Confirmatory factor analysis with cross validation. *Journal of Clinical and Experimental Neuropsychology, 24,* 574-585.

Pugsley, W., Klinger, L. Paschalis, C., Treasure, T., Harrison, M., & Newman, S. (1994). The impact of microemboli during cardiopulmonary bypass on neuropsychological functioning. *Stroke, 25,* 1393-1399.

Rasmussen, L. S. (1999). Perioperative cognitive decline: The extent of the problem. *Acta Anaesthesiologica Belgica, 50,* 199-204.

Rasmussen, L. S., Christiansen, M., Eliasen, K., Sander-Jensen, K., & Moller, J. T. (2002). Biochemical markers for brain damage after cardiac surgery – time profile and correlation with cognitive dysfunction. *Acta Anaesthesiologica Scandinavica, 46,* 547-551.

Rasmussen, L. S., Christiansen, M., Hansen, P. B., & Moller, J. T. (1999). Do blood levels of neuron-specific enolase and S-100 protein reflect cognitive dysfunction after coronary artery bypass? *Acta Anaesthesiologica Scandinavica, 43,* 495-500.

Ridker, P. M., Genest, J., & Libby, P. (2001). Risk Factors for Atherosclerotic Disease. In E. Braunwald & D. P. Zipes & P. Libby (Eds.), *Heart Disease: A Textbook of Cardiovascular Medicine* (pp. 1010-1039). London: W. B. Saunders.

Roach, G. W., Kanchuger, M., Mangano, C. M., Newman, M., Nussmeier, N., & Wolman, R., et al. (1996). Adverse cerebral outcomes after coronary bypass surgery. *The New England Journal of Medicine, 335,* 1857-1863.

Robson, M. J. A., Alston, R. P., Andrews, P. J. D., Wenham, P. R., Souter, M. J., & Deary, I. J. (2002). Apolipoprotein E and neurocognitive outcome from coronary artery surgery. *Journal of Neurology, Neurosurgery, and Psychiatry, 72,* 675-680.

Robson, M. J. A., Alston, R. P., Deary, I. J., Andrews, P. J. D., & Souter, M. J. (2001). Jugular bulb oxyhemoglobin desaturation, S100β, and neurological and cognitive outcomes after coronary artery surgery. *Anesthesia Analgesia, 93,* 839-845.

Robson, M. J. A., Alston, R. P., Deary, I. J., Andrews, P. J. D., Souter, M. J., & Yates, S. (2000). Cognitive after coronary artery surgery is not related to postoperative jugular bulb oxyhemoglobin desaturation. *Anesthesia Analgesia, 91,* 1317-1326.

Rosenow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44,* 1276-1284.

Ross, D. T., & Graham, D. I. (1993). Selective loss and selective sparing of neurons in the thalamic reticular nucleus following human cardiac arrest. *Journal of Cerebral Blood Flow and Metabolism, 13,* 558-567.

Rosselli, M., & Ardila, A. (1991). Effects of age, education, and gender on the Rey Osterrieth Complex Figure. *The Clinical Neuropsychologist, 5,* 370-376.

Rosselli, M., & Ardila, A. (2003). The impact of culture and education on non-verbal neuropsychological measurements: A critical review. *Brain and Cognition, 52,* 326-333.

Rubart, M., & Zipes, D. P. (2001). Genesis of Cardiac Arrhythmias: Electrophysiological Considerations. In E. Braunwald & D. P. Zipes & P. Libby (Eds.), *Heart Disease: A Textbook of Cardiovascular Medicine* (6th ed., pp. 659-699). London: W. B. Saunders.

Ryan, C. M., & Hendrickson, R. (1998). Evaluating the effects of treatment for medical disorders: Has the value of neuropsychological assessment been fully realized? *Applied Neuropsychology, 5,* 209-219.

Sadegh-Zadeh, K. (1998). Fundamentals of clinical methodology: 2. Etiology. *Artificial Intelligence in Medicine, 12,* 227-270.

Sawrie, S. M., Chelune, G. J., Naugle, R. I., & Luders, H. O. (1996). Empirical methods for assessing meaningful neuropsychological change following epilepsy surgery. *Journal of the International Neuropsychological Society, 2,* 556-564.

Scarborough, J. E., White, W., Derilus, F. E., Mathew, J. P., Newman, M. F., & Landolfo, K. P. (2003). Neurologic outcomes after coronary artery bypass grafting with and without cardiopulmonary bypass. *Seminars in Thoracic and Cardiovascular Surgery, 15,* 52-62.

Sedlmeier, P. & Gigerenzer, G. (1989). So studies of statistical power have an effect on the power of studies? *Psychological Bulletin, 105,* 309-316.

Seifert, P. C. (2002). *Cardiac Surgery: Perioperative Patient Care*. Sydney: Mosby.

Selnes, O. A., Goldsborough, M. A., Borowicz, L. M., & McKhann, G. M. (1999). Neurobehavioural sequelae of cardiopulmonary bypass. *Lancet, 353,* 1601-1606.

Selnes, O. A., Grega, M. A., Borowicz, L. M., Royall, R. M., McKhann, G. M., & Baumgartner, W. A. (2003). Cognitive changes with coronary artery disease: A prospective study of coronary artery bypass graft patients and nonsurgical controls. *Annals of Thoracic Surgery, 75,* 1377-1386.

Selnes, O. A., Royall, R. M., Grega, M. A., Borowicz, L. M., Quaskey, S., & McKhann, G. M. (2001). Cognitive changes 5 years after coronary artery bypass grafting: Is there evidence of late decline? *Archives of Neurology, 58,* 598-604.

Shaaban Ali, M., Harmer, M., & Vaughan, R. (2000). Serum S100 protein as a marker of cerebral damage during cardiac surgery. *British Journal of Anaesthesia, 85,* 287-298.

Shaw, P. J., Bates, D., Cartlidge, N. E. F., French, J. M., Heaviside, D., & Julian, D. G., et al. (1987). Neurologic and neuropsychological morbidity following major surgery: Comparison of coronary artery bypass and peripheral vascular surgery. *Stroke, 18,* 700-707.

Sidawy, A. N., Arora, S., & Clowes, A. W. (2001). Atherosclerosis and the Pathogenesis of Occlusive Disease. In L. J. Greenfield (Ed.), *Essentials of Surgery: Scientific Principles and Practice* (pp. 26). Sydney: Lippincott Williams & Wilkins.

Slade, P., Sanchez, P., Townes, B., & Aldea, G. S. (2001). The use of neurocognitive tests in evaluating the outcome of cardiac surgery: Some methodologic considerations. *Journal of Cardiothoracic and Vascular Anesthesia, 15,* 4-8.

Smith, M. H., Wagenknecht, L. F., Legault, C., Goff, D. C., Stump, D. A., & Troost, B. T., et al. (2000). Age and other risk factors for neuropsychologic decline in patients undergoing coronary artery bypass graft surgery. *Journal of Cardiothoracic and Vascular Anesthesia, 14,* 428-432.

Smith, P. L. C. (1995). Cerebral dysfunction after cardiac surgery: Closing address. *Annals of Thoracic Surgery, 59,* 1359-1362.

Smolens, I. A., & Bolling, S. F. (2001). Valvular Heart Disease. In L. J. Greenfield (Ed.), *Essentials of Surgery: Scientific Principles and Practice* (pp. 34). Sydney: Lippincott Williams & Wilkins.

Snell, R. S. (2004). *Clinical Anatomy* (7th ed.). Sydney: Lippincott Williams & Wilkins.

Soltoski, P. R., Guzzo Lemke, V., Barolet, A. W., & Chisolm, R. J. (2004). Interventional Cardiology in Ischemic Heart Disease. In P. R. Soltoski & H. L. Karamanoukian & T. A. Salerno (Eds.), *Cardiac Surgery Secrets* (2nd ed., pp. 75-81). Philadelphia: Hanley & Belfus.

Sontaniemi, K. A. (1995). Long-term neurologic outcome after cardiac operation. *Annals of Thoracic Surgery, 59,* 1336-1339.

Spreen, O., & Strauss, E. (1998). *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary* (2nd Edition). Oxford: Oxford University.

Steed, L., Kong, R., Stygall, J., Acharya, J., Bolla, M., & Harrison, M. J. G., et al. (2001). The role of apolipoprotein E in cognitive decline after cardiac operation. *Annals of Thoracic Surgery, 71,* 823-826.

Strooband, N., Van Nooten, G., Belleghem, Y. V., & Vingerhoets, G. (2002). Short-term and long-term neurocognitive outcome in on-pump versus off-pump CABG. *European Journal of Cardio-thoracic Surgery, 22,* 559-564.

Stump, D. A. (1995). Selection and clinical significance of neuropsychologic tests. *Annals of Thoracic Surgery, 59,* 1340-1344.

Stump, D. A., James, R. L., & Murkin, J. M. (2000). Is that outcome different or not? The effect of experimental design and statistics on neurobehavioural outcome studies. *Annals of Thoracic Surgery, 70,* 1782-1785.

Stump, D. A., Kon, N. A., Rogers, A. T., & Hammon, J. W. (1996). Emboli and neuropsychological outcome following cardiopulmonary bypass. *Echocardiography, 13,* 555-558.

Stump, D.A., Rogers, A. T., & Hammon, J. W. (1996). Neurobehavioural tests are monitoring tools used to improve cardiac surgery outcomes. *Annals of Thoracic Surgery, 61,* 1295-1296.

Stuss, D. T., & Levine, B. (2002). Adult clinical neuropsychology: Lessons from studies of the frontal lobes. *Annual Review of Psychology, 53,* 401-433.

Suksompong, S., Prakanrattana, U., Chumpathong, S., Sriyoschati, S., & Pornvilawan, S. (2002). Neuropsychological alterations after coronary artery bypass graft surgery. *Journal of the Medical Association of Thailand, 85* (Suppl. 3), S910-S916.

Sylivris, S., Levi, C., Matalanis, G., Rosalion, A., Buston, B. F., & Mitchell, A., et al. (1998). Pattern and significance of cerebral microemboli during coronary artery bypass grafting. *Annals of Thoracic Surgery, 66,* 1674-1678.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using Multivariate Statistics* (3rd Ed.). New York: Harper Collins.

Taggart, D. P., Browne, S. M., Halligan, P. W., & Wade, D. T. (1999). Is cardiopulmonary bypass still the cause of cognitive dysfunction after cardiac operations. *Journal of Thoracic and Cardiovascular Surgery, 118,* 414-421.

Tardiff, B. E., Newman, M. F., Saunders, A. M., Strittmatter, W. J., Blumenthal, J. A., & White, W. D., et al. (1997). Preliminary report of a genetic basis for cognitive decline after cardiac operations. *Annals of Thoracic Surgery, 64,* 715-720.

Taylor, K. M. (1998). Brain damage during cardiopulmonary bypass. *Annals of Thoracic Surgery, 65,* S20-S26.

Taylor, K. M. (1998). Central nervous system effects of cardiopulmonary bypass. *Annals of Thoracic Surgery, 66,* S20-S24.

Temkin, N. R., Heaton, R. K., Grant, I., & Dikmen, S. S. (1999). Detecting significant change in neuropsychological test performance: A comparison of four models. *Journal of the International Neuropsychological Society, 5,* 357-369.

Tingey, R. C., Lambert, M. J., Burlingame, G. M., & Hansen, N. B. (1996). Clinically significant change: Practical indicators for evaluating psychotherapy outcome. *Psychotherapy Research, 6,* 144-153.

Toner, I., Taylor, K. M., Newman, S., & Smith, P.L.C. (1998). Cerebral functional changes following cardiac surgery: Neuropsychological and EEG assessment. *European Journal of Cardio-Thoracic Surgery, 13,* 13-20.

Totora, G. J., & Grabowski, S. R. (2003). *Principles of Anatomy and Physiology* (10th ed.). Hoboken, New Jersey: John Wiley & Sons.

Townes, B. D., Bashein, G., Hornbein, T. F., Coppel, D. B., Goldstein, D. E., & Davis, K. B., et al. (1989). Neurobehavioural outcomes in cardiac operations: A prospective controlled study. *Journal of Thoracic Cardiovascular Surgery, 98,* 774-782.

Tulsky, D. S., Saklofske, D. H., CHelune, G. J., Heaton, R. K., Ivnik, R, J, & Bornstein, R. et al. (2003). *Clinical Interpretation of the WAIS-III and WMS-III.* Sydney: Academic Press.

Tuman, K. J., McCarthy, R. J., Najafi, H., & Ivankovich, A. D. (1992). Differential effects of advanced age on neurologic and cardiac risks of coronary artery operations. *Journal of Thoracic and Cardiovascular Surgery, 104,* 1510-1517.

Vaage, J., & Anderson, R. (2001). Biochemical markers of neurologic injury in cardiac surgery: The rise and fall of S100β. *Journal of Thoracic and Cardiovascular Surgery, 122,* 853-855.

van de Vijver, F. J. R. (2001). Facet theory in cross cultural research. *Facet Theory Conference July 15-18 2001 Prague, Czech Republic. Editor Dov Elizur.*

van Dijk, D., Jansen, E. W., Hijman, R., Nierich, A. P., Diephuis, J. C., & Moons, K. G. M. et al (2002). Cognitive outcome after off-pump and on-pump coronary artery bypass graft surgery. *Journal of the American Medical Association (JAMA), 287,* 1405-1412.

van Dijk, D., Keizer, A. M. A., Diephuis, J. C., Durand, C., Vos, L. J., & Hijman, R. (2000). Neurocognitive dysfunction after coronary after bypass surgery: A systematic review. *Journal of Thoracic and Cardiovascular Surgery, 120,* 632-639.

Venn, G. E., Patel, R. L., & Chambers, D. J. (1995). Cardiopulmonary bypass: Perioperative cerebral blood flow and postoperative cognitive deficit. *Annals of Thoracic Surgery,59,* 1331-1335.

Vingerhoets, G. (2001). Cognitive Consequences of Myocardial Infarction, Cardiac Arrhythmias, and Cardiac Arrest. In S. R. Waldstein and M. F. Elias (Eds.), *Neuropsychology of Cardiovascular Disease* (pp. 143-163). London: Laurence Erlbaum.

Vingerhoets, G., Jannes, C., De Soete, G., & Van Nooten, G. (1996). Prospective evaluation of verbal memory performance after cardiopulmonary bypass surgery. *Journal of Clinical and Experimental Neuropsychology, 18,* 187-196.

Vingerhoets, G., Van Nooten, G., & Jannes, C. (1997). Neuropsychological impairment in candidates for cardiac surgery. *Journal of the International Neuropsychological Society, 3,* 480-484.

Waldstein, S. R. & Katzel, L. I. (2001). Hypertension and Cognitive Function. In S. R. Waldstein and M. F. Elias (Eds.), *Neuropsychology of Cardiovascular Disease* (pp. 15-36). London: Laurence Erlbaum.

Wang, D., Wu, X., Li, J., Xiao, F., Lui, X., & Meng, M. (2002). The effect of lidocaine on early postoperative cognitive dysfunction after coronary artery bypass surgery. *Anesthesia and Analgesia, 95,* 1134-1141.

Waters, A. M., Armstrong, T., & Senes-Ferrari, S. (1998). *Medical Care of Cardiovascular Disease in Australia.* Canberra: Australian Institute of Health and Welfare.

Wechsler, D. (1997). *Manual for the Wechsler Memory Scale–Third Edition (WMS–III).* San Antomio, Tx: The Psychological Corporation.

Weissberg, P. L., & Rudd, J. H. F. (2002). Atherosclerotic Biology and Epidemiology of Disease. In E. J. Topol (Ed.), *Textbook of Cardiovascular Medicine* (pp. 24). Sydney: Lippincott Williams & Wilkins.

Whitaker, D. C., Stygall, J, & Newman, S. P. (2002). Neuroprotection during cardiac surgery: strategies to reduce cognitive decline. *Perfusion, 17,* 69-75.

Wimmer-Greinecker, G., Matheis, G., Brieden, M., Dietrich, M., Oremek, G., & Westphal, K., et al. (1998). Neuropsychological changes after cardiopulmonary bypass for coronary artery bypass grafting. *Thoracic and Cardiovascular Surgery, 46,* 207-212.

Winakur, S., & Jessup, M. (2004). Heat failure. In D. D. Federman, & D. C. Dale (Eds.), *ACP Medicine* [Electronic Version]. New York: WebMD Corporation. Rectieved May 2004 from STAT!Ref Online Electronic Medical Library.

Yachimski, P., & Lilly, L. S. (2003). Valvular Heart Disease. In L. S. Lilly (Ed.), *Pathophysiology of Heart Disease: A Collaborative Project of Medical Students and Faculty* (3rd ed., pp. 185-210). Sydney: Lippincott Williams & Wilkins.

Yates, S. L., & Alston, R. P. (2000). Neurological, psychological and cognitive sequelae of cardiac surgery. *Current Anaesthesia & Critical Care, 11,* 187-193.

Yoshitani, K., Kawaguchi, M., Sugiyama, N., Sugiyama, M., Inoue, S., Sakamoto, T., et al (2001). The associationof high jugular bulb venous oxygen saturation with cognitive decline after hypothermic cardiopulmonary bypass. *Anesthesia and Analgesia, 92,* 1370-1376.

Young, Z., & Dai, B. (2004a). Complications of Cardiopulmonary Bypass. In P. R. Soltoski & H. L. Karamanoukian & T. A. Salerno (Eds.), *Cardiac Surgery Secrets* (2nd ed., pp. 278-282). Philadelphia: Hanley & Belfus.

Young, Z., & Dai, B. (2004b). Principles of Cardiopulmonary Bypass. In P. R. Soltoski & H. L. Karamanoukian & T. A. Salerno (Eds.), *Cardiac Surgery Secrets* (2nd ed., pp. 271-277). Philadelphia: Hanley & Belfus.

Zakzanis, K. K. (2001). Statistics tell the truth, the whole truth, and nothing but the truth: Formulae, illustrative numerical examples, and heuristic interpretation of effect size analyses for neuropsychological researchers. *Archives of Clinical Neuropsychology, 16,* 653-667.

Zimpfer, D., Kilo, J., Czerny, M., Kasimir, M. T., Madl, C., & Bauer, E., et al. (2003). Neurcognitive deficit following aortic valve replacement with biological/mechanical prosthesis. *European Journal of Cardio-thoracic Surgery, 23,* 544-551.

# Appendix A

## Published Papers Included in Statistical Power and Effect Size Analyses

Table 1 lists the published papers used to generate the data for the analyses conducted in studies one and three. Full article references are provided in the reference list. Authors names marked with an asterisk signify studies where both within and between groups comparisons were analysed. Details are presented about the assessment intervals (preoperative, postoperative, and follow-up) and the range of statistical power and effect sizes for each study. Due to the method of analysis for within group design studies, the range of statistical power and effect sizes are not available for within group comparisons. None of the studies examined in the current research endeavour analysed individual subjects' performances.

Table 1

Methodological Details of Published Papers Included in Studies 1 and 3, and the Range of Statistical Power and Effect Sizes Obtained in Each

| Author | *N* | Intervals | | | Range of Power | Range of *d* | Positive Result Claimed by Author |
|---|---|---|---|---|---|---|---|
| | | Preop | Postop | Follow-Up | | | |
| Basile et al. (2001) | 16 | 1 – 3 days | 2 days | 6 months | - | - | Yes |
| *Braekken et al. (1998) | 41 | 1 – 2 days | - | 2 months | .05 – .84 | .07 – 1.0 | Yes |
| *Grigore, Grocott et al. (2002) | 165 | 1 day | 6 weeks | - | .05 – .59 | .01 – .35 | Yes |
| *Heyer et al. (2002) | 61 | Before | 5 days | 6 weeks | .05 – .83 | .00 – .84 | Yes |
| Jacobs et al. (1998) | 18 | 1 day | 8 – 12 days | 3 months | - | - | Yes |
| Kilminster et al. (1999) | 130 | Before | 8 weeks | - | - | - | Yes |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Kilo et al. (2001) | 308 | Before | 1 week | 4 months | - | - | No |
| *Robson et al. (2002) | 86 | 1 day | - | 3 months | .05 – .18 | .00 – .24 | Yes |
| *Strooband et al. (2002) | 49 | 1 day | 6 – 7 days | 6 months | .01 – .80 | .05 – .99 | Yes |
| Suksompong et al. (2002) | 110 | Before | 3 – 5 days | - | - | - | Yes |
| *Taggart et al. (1999) | 75 | Before | Pre-discharge | 3 months | .00 – .36 | .00 – .31 | Yes |
| *Wang et al. (2002) | 118 | 1 day | 9 days | - | .03 – .34 | .05 – .35 | Yes |
| Wimmer-Greinecker et al. (1998) | 76 | 1 day | 5 days | 2 months | - | - | Yes |

# Appendix B

SPSS Syntax for Calculating Within Group Effect Sizes by Dunlap et al.'s (1996) procedure

```
manova jobsat disceff/
 wsfactor=time(2)/
 matrix=out(*).

select if (~sysmis(m1)).
execute.

compute n1=min(n1,n2).
compute n2=n1.

** To generate hypothetical data.

compute poolsd=sqrt((sd1**2 + sd2**2)/2).
compute sd1=poolsd.
compute sd2=poolsd.
** If d is 1.5.
compute d=1.5.
compute m2=m1 + poolsd*d.a
execute.

compute studnum=studnum + 1.
leave studnum.

string NUM (a8).
compute NUM="N".
string MN (a8).
compute MN="MEAN".
string SD (a8).
compute SD="STDDEV".
string cor1 (a8).
compute cor1="CORR".
string cor2 (a8).
compute cor2="CORR".
string miss1 (a8).
compute miss1="".
string miss2 (a8).
compute miss2="".
string miss3 (a8).
compute miss3="".
string pre (a8).
```

```
compute pre="PRE".
string post (a8).
compute post="POST".
compute unit1=1.
compute c1=.5.
compute unit2=1.
compute c2=.5.
execute.

varstocases make rowtype_ from num mn sd cor1 cor2/
  make varname_ from miss1 miss2 miss3 pre post/
  make pre from n1 m1 sd1 unit1 c1/
  make post from n2 m2 sd2 c2 unit2/
  keep=studnum.

* save outfile='c:\projects\htill\msdn_mat.sav'/
  keep=rowtype_ varname_ pre post studnum.

split file by studnum.

manova pre post/
  wsfactor=time(2)/
  matrix=in(*)/
  power=f(.05) exact.

split file off.
```

# Appendix C

## Intercorrelation Matricies for Study Two

### Battery One

The elements (tests) of battery one were Logical Memory – immediate and delayed recall (LM1 and LM2 respectively) and Digit Span (DS) from the Wechsler Memory Scale - Revised Edition, Trail Making Test Part A and Part B (TMTA and TMTB respectively), and Controlled Oral Word Association Test (LF).

Table 1

Intercorrelation Matrix for Battery One at the Preoperative Assessment Interval

| | TMTA | DS | LM1 | LM2 | TMTB |
|---|---|---|---|---|---|
| DS | -0.29 | - | | | |
| LM1 | -0.25 | 0.4 | - | | |
| LM2 | -0.3 | 0.3 | 0.81 | - | |
| TMTB | 0.55 | -0.37 | -0.39 | -0.35 | - |
| LF | -0.34 | 0.26 | 0.48 | 0.52 | -0.47 |

Table 2

Intercorrelation Matrix for Battery One at the Postoperative Assessment Interval

| | TMTA | DS | LM1 | LM2 | TMTB |
|---|---|---|---|---|---|
| DS | -0.34 | - | | | |
| LM1 | -0.4 | 0.43 | - | | |
| LM2 | -0.4 | 0.41 | 0.89 | - | |
| TMTB | 0.65 | -0.37 | -0.5 | -0.5 | - |
| LF | -0.59 | 0.52 | 0.56 | 0.49 | -0.54 |

Battery Two

The facets of battery two were computerised simple reaction time – right and left hands (CSRTR and CSRTL respectively), computerised choice reaction time (CCRT), computerised visual attention (CVFALU, CVFARU, CVFALL,CVFARL), computerised visuo-motor tracking (CVT), visuo-spatial working memory (VSWM), auditory verbal working memory – Digit Span backwards (DSB), auditory verbal paired associate memory (VPA).

Table 3

Intercorrelation Mmatrix for Battery Two at the Preoperative Assessment Interval

| | CSRTR | CSRTL | CCRT | CVT | CVFALU | CVFALL | CVFARU | CVFARL | DSB | VPA |
|---|---|---|---|---|---|---|---|---|---|---|
| CSRTL | 0.72 | - | | | | | | | | |
| CCRT | 0.61 | 0.52 | - | | | | | | | |
| CVT | 0.43 | 0.43 | 0.37 | - | | | | | | |
| CVFALU | 0.62 | 0.42 | 0.66 | 0.26 | - | | | | | |
| CVFALL | 0.63 | 0.47 | 0.65 | 0.32 | 0.89 | - | | | | |
| CVFARU | 0.62 | 0.41 | 0.64 | 0.29 | 0.91 | 0.91 | - | | | |
| CVFARL | 0.64 | 0.43 | 0.61 | 0.31 | 0.88 | 0.92 | 0.91 | - | | |
| DSB | -0.07 | -0.01 | -0.08 | -0.07 | -0.13 | -0.09 | -0.11 | -0.17 | - | |
| VPA | -0.26 | -0.11 | -0.14 | -0.26 | -0.16 | -0.19 | -0.17 | -0.20 | 0.06 | - |
| VSWM | -0.38 | -0.33 | -0.46 | -0.48 | -0.34 | -0.32 | -0.35 | -0.37 | 0.21 | 0.27 |

Table 4

Intercorrelation Matrix for Battery Two at the Postoperative Assessment Interval

| | CSRTR | CSRTL | CCRT | CVT | CVFALU | CVFALL | CVFARU | CVFARL | DSB | VPA |
|---|---|---|---|---|---|---|---|---|---|---|
| CSRTL | 0.72 | - | | | | | | | | |
| CCRT | 0.61 | 0.52 | - | | | | | | | |
| CVT | 0.43 | 0.43 | 0.37 | - | | | | | | |
| CVFALU | 0.62 | 0.42 | 0.66 | 0.26 | - | | | | | |
| CVFALL | 0.63 | 0.47 | 0.65 | 0.32 | 0.89 | - | | | | |
| CVFARU | 0.62 | 0.41 | 0.64 | 0.29 | 0.91 | 0.91 | - | | | |
| CVFARL | 0.64 | 0.43 | 0.61 | 0.31 | 0.88 | 0.92 | 0.91 | - | | |
| DSB | -0.07 | -0.01 | -0.08 | -0.07 | -0.13 | -0.09 | -0.11 | -0.17 | - | |
| VPA | -0.26 | -0.11 | -0.14 | -0.26 | -0.16 | -0.19 | -0.17 | -0.20 | 0.06 | - |
| VSWM | -0.38 | -0.33 | -0.46 | -0.48 | -0.34 | -0.32 | -0.35 | -0.37 | 0.21 | 0.27 |

Battery Three

The elements of battery three were Stroop Colour Word Test – interference score (SCWTI), Halstead Category Test (HCT), Trail Making Test Part B (TMTB), Trail Making Test Part A (TMTA), Controlled Oral Word Association Test – total score (LF), Grooved Pegboard left hand time (GPL), Grooved Pegboard right hand time (GPR), Rey Auditory Verbal Learning Test Trial 1 (AVLT1), Rey Auditory Verbal Learning Test Trial 6 (AVLT6), Digit Span Backward (DSB), Paced Auditory Serial Addition Test – two second interval (PASA2), Paced Auditory Serial Addition Test – four second interval (PASA4), Symbol Digit Substitution Test – (SDS), Visual Reproduction – delayed recall (VRDR).

Table 5

Intercorrelation Matrix for Battery Three at the Follow-up Assessment Interval

| | GPR | GPL | TMTA | SDS | DSF | DSB | PASA2 | PASA4 | AVLT1 | AVLT6 | VR | TMTB | LF | HCT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| GPLT | 0.86 | - | | | | | | | | | | | | |
| TMTA | 0.26 | 0.22 | - | | | | | | | | | | | |
| SDSTC | -0.55 | -0.55 | -0.59 | - | | | | | | | | | | |
| DSF | -0.19 | -0.13 | -0.48 | 0.35 | - | | | | | | | | | |
| DSB | 0.09 | 0.03 | -0.14 | 0.32 | 0.41 | - | | | | | | | | |
| PASA2 | -0.25 | -0.28 | -0.4 | 0.63 | 0.26 | 0.45 | - | | | | | | | |
| PASA4 | -0.22 | -0.27 | -0.46 | 0.68 | 0.3 | 0.37 | 0.9 | - | | | | | | |
| AVLT1 | 0.01 | -0.02 | -0.34 | 0.31 | 0.18 | 0.43 | 0.68 | 0.67 | - | | | | | |
| AVLT6 | -0.18 | -0.04 | -0.32 | 0.27 | 0.26 | 0.41 | 0.41 | 0.33 | 0.6 | - | | | | |

| | GPR | GPL | TMTA | SDS | DSF | DSB | PASA2 | PASA4 | AVLT1 | AVLT6 | VR | TMTB | LF | HCT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| VR | -0.5 | -0.51 | -0.43 | 0.67 | 0.44 | 0.32 | 0.57 | 0.54 | 0.4 | 0.51 | - | | | |
| TMTB | 0.34 | 0.37 | 0.66 | -0.62 | -0.45 | -0.25 | -0.62 | -0.54 | -0.31 | -0.25 | -0.51 | - | | |
| LF | 0.09 | 0.07 | -0.1 | 0.32 | 0.38 | 0.41 | 0.34 | 0.37 | 0.38 | 0.07 | 0.37 | -0.12 | - | |
| HCT | 0.23 | 0.3 | 0.46 | -0.31 | -0.17 | -0.3 | -0.28 | -0.22 | -0.4 | -0.47 | -0.4 | 0.46 | -0.26 | - |
| SCWTI | 0.25 | 0.26 | 0.52 | -0.64 | -0.41 | -0.45 | -0.62 | -0.6 | -0.52 | -0.26 | -0.59 | 0.57 | -0.51 | 0.49 |

Battery Four

The elements (tests) comprising those facets were the Trail Making Test – Part A (TMTA) and Part B (TMTB), Digits Test (ODST), Taps Test (BTT), Stroop Colour Word Test – interference score (SCWTI), Line Bisection Test (LNBS), Bourdon Wiersma Dot Cancellation Test (BDC), Rey Auditory Verbal Learning Test – total recall (AVLTT) and delayed recall (AVLTD), Rey Complex Figure Test – immediate recall (CFTI) and delayed recall (CFTD), Token Test (TT), Controlled Oral Word Association Test – total score (LF), Purdue Pegboard Test (PPB30).

Table 6

Intercorrelation Matrix for Battery Four at the Preoperative Assessment Interval

| | TT | PPB30 | LNB | TMTA | ODST | ABTT | BDCCS | AVLTT | AVLTD | CFTI | CFTD | TMTB | LF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PPB30 | 0.17 | - | | | | | | | | | | | |
| LNB | -0.08 | -0.04 | - | | | | | | | | | | |
| TMTA | -0.3 | -0.44 | 0.07 | - | | | | | | | | | |
| ODST | -0.25 | -0.21 | 0.07 | 0.2 | - | | | | | | | | |
| ABTT | -0.26 | -0.13 | 0.01 | 0.26 | 0.27 | - | | | | | | | |
| BDCCS | 0.02 | 0.09 | -0.1 | 0.06 | 0 | -0.03 | - | | | | | | |
| AVLTT | 0.32 | 0.33 | 0.01 | -0.43 | -0.32 | -0.21 | 0.15 | - | | | | | |
| AVLTD | 0.25 | 0.3 | -0.02 | -0.33 | -0.24 | -0.25 | 0.15 | 0.78 | - | | | | |
| CFTI | 0.27 | 0.17 | -0.13 | -0.37 | -0.12 | -0.12 | 0.03 | 0.33 | 0.32 | - | | | |

| | TT | PPB30 | LNB | TMTA | ODST | ABTT | BDCCS | AVLTT | AVLTD | CFTI | CFTD | TMTB | LF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CFTD | 0.34 | 0.26 | -0.17 | -0.39 | -0.28 | -0.21 | 0.07 | 0.45 | 0.46 | 0.64 | - | | |
| TMTB | -0.46 | -0.32 | 0.09 | 0.69 | 0.38 | 0.43 | -0.09 | -0.47 | -0.33 | -0.39 | -0.41 | - | |
| LF | 0.32 | 0.37 | -0.13 | -0.46 | -0.33 | -0.15 | 0.17 | 0.57 | 0.47 | 0.25 | 0.38 | -0.52 | - |
| SCWTI | -0.22 | -0.09 | 0 | 0.23 | 0.12 | 0.2 | 0.08 | -0.2 | -0.2 | -0.19 | -0.26 | 0.26 | -0.19 |

Table 7

Intercorrelation Matrix for Battery Four at the Postoperative Assessment Interval

| | TT | PPB30 | LNB | TMTA | ODST | ABTT | BDC | AVLTT | AVLTD | CFTI | CFTD | TMTB | ALF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PPB30 | 0.18 | - | | | | | | | | | | | |
| LNB | -0.16 | -0.12 | - | | | | | | | | | | |
| TMTA | -0.35 | -0.6 | 0.26 | - | | | | | | | | | |
| ODST | -0.23 | -0.35 | 0.14 | 0.28 | - | | | | | | | | |
| ABTT | -0.22 | -0.24 | 0.15 | 0.45 | 0.34 | - | | | | | | | |
| BDC | 0.16 | 0.07 | 0 | -0.02 | 0.20 | 0.06 | - | | | | | | |
| AVLTT | 0.24 | 0.47 | -0.16 | -0.48 | -0.36 | -0.34 | -0.09 | - | | | | | |
| AVLTD | -0.11 | 0.13 | -0.14 | -0.08 | -0.06 | -0.14 | -0.06 | 0.24 | - | | | | |
| CFTI | 0.10 | 0.31 | -0.35 | -0.53 | -0.26 | -0.27 | -0.03 | 0.41 | 0.1 | - | | | |

| | TT | PPB30 | LNB | TMTA | ODST | ABTT | BDC | AVLTT | AVLTD | CFTI | CFTD | TMTB | ALF |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CFTD | 0.08 | 0.37 | -0.30 | -0.52 | -0.23 | -0.31 | -0.03 | 0.47 | 0.14 | 0.92 | - | | |
| TMTB | -0.28 | -0.51 | 0.13 | 0.71 | 0.24 | 0.35 | 0.03 | -0.55 | -0.38 | -0.44 | -0.43 | - | |
| LF | 0.31 | 0.4 | -0.26 | -0.46 | -0.39 | -0.34 | 0 | 0.57 | 0.44 | 0.47 | 0.44 | -0.60 | - |
| SCWTI | -0.26 | -0.44 | 0.12 | 0.44 | 0.33 | 0.33 | -0.05 | -0.4 | -0.27 | -0.42 | -0.4 | 0.45 | -0.48 |

Table 8

Intercorrelation Matrix for Battery Four at the Follow-up Assessment Interval

| | TT | PPB30 | LNB | TMTA | ODST | ABTT | BDC | AVLTT | CFTI | CFTD | TMTB | LF | SCWTI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| PPB30 | 0.2 | - | | | | | | | | | | | |
| LNB | -0.03 | 0.06 | - | | | | | | | | | | |
| TMTA | -0.36 | -0.36 | 0 | - | | | | | | | | | |
| ODST | -0.46 | -0.28 | 0 | 0.34 | - | | | | | | | | |
| ABTT | -0.06 | -0.19 | 0.16 | 0.37 | 0.44 | - | | | | | | | |
| BDC | -0.09 | 0.06 | 0.24 | 0.11 | 0.06 | -0.02 | - | | | | | | |
| AVLTT | 0.21 | 0.23 | 0.03 | -0.33 | -0.35 | -0.31 | -0.04 | - | | | | | |
| CFTI | 0.21 | 0.28 | -0.15 | -0.38 | -0.23 | -0.28 | -0.09 | 0.29 | - | | | | |

| | TT | PPB30 | LNB | TMTA | ODST | ABTT | BDC | AVLTT | CFTI | CFTD | TMTB | LF | SCWTI |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CFTD | 0.22 | 0.29 | -0.17 | -0.4 | -0.24 | -0.26 | -0.14 | 0.3 | 0.93 | - | | | |
| TMTB | -0.23 | -0.28 | -0.05 | 0.63 | 0.32 | 0.39 | 0.12 | -0.43 | -0.23 | -0.26 | - | | |
| LF | 0.37 | 0.39 | -0.05 | -0.42 | -0.35 | -0.18 | 0 | 0.51 | 0.44 | 0.4 | -0.41 | - | |
| SCWTI | -0.04 | -0.07 | 0.05 | 0.33 | 0.11 | 0.28 | -0.12 | -0.29 | -0.33 | -0.27 | 0.49 | -0.31 | - |
| AVLTD | 0.18 | 0.1 | -0.04 | -0.29 | -0.24 | -0.16 | 0.02 | 0.82 | 0.33 | 0.38 | -0.36 | 0.42 | -0.24 |

# Appendix D

## Intercorrelation Matricies for Study Four

### Consensus Statement Core Neuropsychological Test Battery

The elements (tests) comprising those facets were the Trail Making Test – Part A (TMTA) and Part B (TMTB), Rey Auditory Verbal Learning Test – trial 1 (AVLT1), Rey Auditory Verbal Learning Test – trial 5 (AVLT5), Rey Auditory Verbal Learning Test – total recall (AVLTT), Rey Auditory Verbal Learning Test – trial 6 (AVLT6), Rey Auditory Verbal Learning Test – delayed recall (AVLT7), Grooved Pegboard Test – non dominant hand (GPBN), and Grooved Pegboard Test – dominant hand (GPBD).

Table 1

Intercorrelation Matrix for the Core Neuropsychological Test Battery at the Preoperative Assessment Interval

| | AVLT1 | AVLT5 | AVLTT | AVLT6 | AVLT7 | TMTA | TMTB | GPN |
|---|---|---|---|---|---|---|---|---|
| AVLT5 | 0.44 | - | | | | | | |
| AVLTT | 0.73 | 0.8 | - | | | | | |
| AVLT6 | 0.28 | 0.72 | 0.78 | - | | | | |
| AVLT7 | 0.07 | 0.67 | 0.67 | 0.96 | - | | | |
| TMTA | -0.42 | -0.19 | -0.39 | -0.31 | -0.26 | - | | |
| TMTB | -0.56 | -0.19 | -0.54 | -0.41 | -0.29 | 0.83 | - | |
| GPN | 0.22 | 0 | -0.05 | -0.12 | -0.2 | 0.27 | 0.25 | - |
| GPD | -0.22 | -0.22 | -0.54 | -0.35 | -0.39 | 0.5 | 0.42 | 0.43 |

Table 2

Intercorrelation Matrix for the Core Neuropsychological Test Battery at the Postoperative Assessment Interval

| | AVLT1 | AVLT5 | AVLTT | AVLT6 | AVLT7 | TMTA | TMTB | GPN |
|---|---|---|---|---|---|---|---|---|
| AVLT5 | 0.69 | - | | | | | | |
| AVLTT | 0.93 | 0.86 | - | | | | | |
| AVLT6 | 0.7 | 0.75 | 0.88 | - | | | | |
| AVLT7 | 0.65 | 0.62 | 0.81 | 0.93 | - | | | |
| TMTA | -0.15 | -0.6 | -0.43 | -0.61 | -0.73 | - | | |
| TMTB | -0.52 | -0.78 | -0.78 | -0.96 | -0.84 | 0.69 | - | |
| GPN | -0.15 | -0.61 | -0.46 | -0.7 | -0.76 | 0.98 | 0.78 | - |
| GPD | -0.45 | -0.48 | -0.65 | -0.93 | -0.91 | 0.62 | 0.9 | 0.73 |