DEVELOPMENT OF AN ONLINE PSYCHOMETRIC TEST OF SPATIAL ABILITY

Kenneth John Sutton

Master of Educational Studies
Bachelor of Educational Studies
Graduate Diploma in Educational Studies
Diploma in Teaching (Industrial Arts)
Engineering Trade (Patternmaking)

Submitted in the fulfilment of the requirements for the degree of

Doctor of Philosophy (Psychology)

School of Psychology
Faculty of Science and Information Technology
The University of Newcastle, NSW Australia

February 2011

ABSTRACT

This thesis reports on the development of an online psychometric test of spatial ability for designers that measures choice accuracy and reaction times. The research identified five spatial skills that collectively contribute to a construct called spatial ability. The test consists of 20 test items divided equally into five subtests where each measures a separate spatial skill. The five spatial skills were appropriately named and descriptions of each are provided in the final chapter. Evidence from nine sequential studies based on detailed statistical investigation including item analysis and exploratory factor analysis was used to establish the test. The important psychometric properties of reliability, validity, correlation and effect sizes were constantly assessed throughout the studies, and both parametric and nonparametric procedures were used where appropriate. Participants who took part in the studies were mostly students undertaking courses at university level and were recruited from both design and nondesign disciplines. Sample sizes for the different studies varied, but reached 650 in a final study consisting of male and female participants spread across 15 design disciplines. Two versions of the test were developed and both provide instructions and feedback to the test taker, and participation is possible without the assistance of a supervisor. One version is meant to be used by novice designers or instructors for diagnostic purposes, while the second collects data and demographic information and is intended for research. This thesis established the importance of spatial ability in a design environment and the need for a specific test for designers. Other outcomes include gender differences, the opinion of subject matter experts, comparison between design and nondesign groups and the impact of practice effect on the assessment of any real learning that may occur in a classroom setting. Two methods of item analysis were applied to appropriate datasets, and the relationship between spatial ability and general academic ability was investigated. Choice accuracy and reaction time data were analysed, and the studies mostly report quantitative research, though some qualitative research is also reported. This research examined a large number of subtests and test items that were reduced to the final configuration after strict compliance to psychometric test development standards. Both laboratory and online studies were conducted to help achieve the final outcome.

STATEMENT OF ORIGINALITY

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. I give consent to this copy of my thesis, when deposited in the University Library, being made available for loan and photocopying subject to the provisions of the Copyright Act 1968.

Signed: _____

DEDICATION

The completion of this thesis is dedicated first and foremost to my loving wife Yvonne who has been my best friend, confidant and companion since I was 17 years of age. Yvonne is truly the most important human being in my life whose unquestionable support and devotion made impossible tasks achievable throughout our life together. Yvonne has been an essential partner throughout my journey of mature-age education from those early school days through to the present time. Always doing things to help, always understanding and always sharing the determination and sacrifice required. I owe Yvonne the greatest thank you of all time for her love, encouragement, confidence in my capacity to succeed, and for overlooking the many things I neglected along the way. Without Yvonne's love, this PhD endeavour would have no meaning beyond a mere academic achievement.

A second dedication is to my loving parents Allan and Isabel. Both are now in their 90s who never had the opportunity for higher education because of the economic and domestic difficulties they experienced during their adolescent years. Despite their denied opportunities, they nevertheless saw the value of education and did everything possible to encourage their four children to seek the security and satisfaction often associated with higher education. To them, I owe a deep sense of gratitude for being the parents they were, for providing the benefits of a loving, safe and secure home, and for giving me a start that they themselves never had.

ACKNOWLEDGEMENTS

My first acknowledgement is to my loving wife Yvonne for always being there when I needed her most, for making the many sacrifices that were imposed upon her, and for taking care of the many things I conveniently put aside. Without question, the completion of this PhD would not have been possible without Yvonne's understanding, cooperation and ongoing support.

My second acknowledgement is to my primary supervisor Professor Andrew Heathcote for inviting my PhD candidature. I appreciated being guided by Andrew's work ethic, his passion for research, the academic rigour he insisted upon and the learning I received from our discussions and scientific writing he encouraged. My appreciation extends to Andrew's lasting confidence in me as an academic within the School, and also as a PhD student. His continual encouragement and reference to a potentially successful academic career regardless of a late start had a significant influence on my motivation. I thank Andrew for his ageless attitude to postgraduate students, for his grounding in experimental psychology, and for the respect and enthusiasm I developed for the pursuit of new knowledge.

My third acknowledgement is to my second supervisor Dr Miles Bore for respecting me as a colleague, valuing my opinion, his patience in listening to my point of view, his compliments towards many aspects of the draft thesis, and his identification of my developing expertise. Miles was especially helpful because of his effective communication style, his suggestions towards the draft thesis, the priority he gave to me at different times, and his demonstrated professionalism on every occasion. I am also most grateful to Miles for providing lecturing and teaching opportunities that helped me develop a specialisation within psychology. Miles is truly dedicated to the discipline of psychology, and I thank him for the manner in which he approached our discussions, the enrichment he gave to my understanding of psychometrics, and most importantly, the confidence he expressed in my ability to achieve.

My fourth acknowledgement is to Kim Colyvas, the senior statistical consultant commissioned by the Faculty of Science and Information Technology to assist staff and students in their research activities. Kim stands out because of his ability to quickly identify what is concerning the student, his software expertise, his attention to detail and his plain English approach to explaining statistics. His approach and patience are highly praiseworthy, and his viewpoint that there are no rules, expectations or milestones when embarking on a new learning experience was reassuring. I thank Kim for the excitement I have developed for statistics, for embracing my project like his own, and for making it possible to be a student of statistics over a lengthy period of time.

My fifth acknowledgement is to Associate Professor Anthony Williams who appeared to recognise something in me that I did not recognise myself. I will always appreciate Tony's

ongoing collaboration, his willingness to co-author many publications, his assistance with recruiting suitable research participants, and his contribution to a successful Australian Learning and Teaching Council (ALTC) grant application. I thank Tony for his long friendship, his continual support, and the career aspirations he always maintained for me.

My sixth acknowledgement is to Associate Professor Don Munro for his willingness to be a reader of my draft thesis. Don achieved this task in a very short period at a very inconvenient time and provided invaluable advice that I adopted to strengthen the standing of the thesis. I thank Don for his thorough reading, his attention to detail, his sound contribution and his readiness to elaborate on any issue I wanted to raise.

I also thank the executive of the School of Psychology for allowing me as a new staff member after a workplace restructure to take on this candidature. I appreciated their confidence in my ability to overcome any disadvantages I might encounter. Thanks also to a succession of Heads of School for their genuine encouragement, and for the meaningful discussions that occurred from time to time. Also thanks to every member of staff who accepted me without qualification despite not having a traditional background in psychology. This gave me every incentive to succeed. Finally, my thanks extend to the Pro Vice-Chancellor of the Faculty of Science and Information Technology for his initiative in setting up a statistical support service to assist staff and students in their research endeavours.

PUBLICATIONS AND GRANTS

Publications

Year 2010

A Rationale for Developing Spatial Skills in a Design Environment
Sutton, K., & Williams, A.
Connected 2010 – 2nd International Conference on Design Education 28 June – 01 July 2010, University of New South Wales, Sydney, Australia

A Multidisciplinary Approach to Measuring Spatial Performance in Novice Designers
Sutton, K., & Williams, A.
International Congress of Applied Psychology 2010 (ICAP 2010), 11-16 July 2010, held at the Melbourne Convention and Exhibition Centre in Melbourne, Australia

Implications of Spatial Abilities on Design Thinking
Sutton, K., & Williams, A.
Design Research Society (DRS) international conference Design & Complexity held July 7-9, 2010 in Montreal (Quebec), Canada, at the School of Industrial Design, Université de Montréal

Year 2009

Improving Spatial Abilities with 3D Mini-Maps
Nesbitt, K., Sutton, K., Wilson, J., & Hookham, G.
The 6th Australasian Conference on Interactive Entertainment , Sydney, Australia 14-16 Dec 2009

Spatial Ability Performance of Female Engineering Students
Sutton, K., Williams, A., & McBride, W.
The Australasian Association for Engineering Education (AAEE) 20th Annual Conference Adelaide, Australia 6-9 December 2009

Exploring Spatial Ability and Mapping the Performance of Engineering Students
Sutton, K., Williams, A., & McBride, W.
The Australasian Association for Engineering Education (AAEE) 20th Annual Conference Adelaide, Australia 6-9 December 2009

Spatial Abilities and its Implication for Novice Designers
Williams, A., & Sutton, K.
The 5th Annual DesignEd Asia Conference, Wan Chai, Hong Kong 1 & 2 Dec 2009

Year 2008

Developing a discipline-based measure of visualisation
Ken Sutton & Anthony Williams
2008 National UniServe Conference 1st October - 3rd October 2008 The University of Sydney

Profiling Spatial Ability Performance of Novice Designers
Anthony Williams, Ken Sutton & Rebecca Allen
The 5th Biennial International Conference on Technology Education Research 27 - 29 November 2008 Gold Coast, Queensland, Australia

Spatial Ability: Issues Associated with Engineering and Gender
Anthony Williams, Ken Sutton & Rebecca Allen

The Australasian Association for Engineering Education (AAEE) 19th Annual Conference Yeppoon, Australia 7-10 December 2008

Year 2007

Measuring 3D understanding on the web and in the laboratory
Sutton, K., Heathcote, A. & Bore, M. (2007).
Behavior Research Methods, 39 (4), 926-939.

Research outcomes supporting learning is spatial ability
Sutton, K., & Williams, A. (2007).
Proceedings of the 2007 AaeE Conference, Melbourne, 9-12 December 2007

Spatial cognition and its implications for design.
Sutton, K., & Williams, A. (2007).
Proceedings of the International Association of Societies of Design Research, The Hong Kong Polytechnic University, 12-15 November 2007

Developing interactive computer-based learning experiences to improve spatial ability.
Williams, A., & Sutton, K. (2007).
Reflection paper, IADIS International Conference on Cognition and Exploratory Learning in Digital Age 2007, Algarve, Portugal, 7-9 December 2007 [E1]

Year 2006

Impact of Spatial Ability on Students Doing Graphics Based Courses
Sutton, K., & Williams, A.
17th Annual Conference of the Australasian Association for Engineering Education Auckland, New Zealand 10-13 December 2006

Performance Differences on Spatial Ability Tasks
Sutton, K.
Australian Conference on Personality and Individual Differences, Newcastle, 1-2 Dec, 2006

Grants

Australian Learning and Teaching Council (ALTC) Grant. Assessing and Improving Spatial Ability for Design-based Disciplines Utilising Online Systems. Agreement dated 19 October 2007 ($196,331).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

TERMS AND DEFINITIONS

3D Ability Test. The online psychometric test of spatial ability which is the central focus of this research is called the 3D Ability Test. The test is also known by its acronym (3DAT).

3D Understanding. Regarded as an alternative name for spatial ability but emphasizes the importance of 3D in the interpretation of spatial ability. 3D understanding is also described as visualizing in three dimensions.

Algorithmic Design. A position at one end of the design spectrum that could be described as structured and restricted. Design that is driven by procedures, formulae and precedents.

Analytic Solution Strategies. This approach to solving problems depends on a systematic step by step strategy. Analytic procedures handle spatial information by breaking it down into nonspatial list-like elements and essentially treat one element at a time. Fewer errors may occur this way, but responses may be slower than alternative strategies.

Classical Test Theory. The traditional and well established approach to item analysis is known as the *classical test theory* model and combines measures of item difficulty, discrimination and reliability to determine the acceptability of test items.

Course. An academic subject that is part of a tertiary education program such as a degree or diploma. Also applies to primary and secondary education as part of an award upon graduation.

Design. A general term that covers disciplines engaged in design processes at any level in a range from algorithmic design to heuristic design. Disciplines include engineering, architecture, mechatronics, construction management, industrial design, surveying and graphic design.

Effect Size. Effect size is a calculation that provides a standardized measure of a difference within a group or a difference between two groups. Effect size could also be described as practical significance and it allows easy comparisons to be made. Standard benchmarks have emerged and are expressed in terms of indexes representing small, medium and large effects. Where standard deviations (SD) are similar, effect size can be calculated using the formula: $(MEAN_1 - MEAN_2)/SQRT((((N_1-1)*SD_1^2)+((N_2-1)*SD_2^2))/(N_1+N_2-2))$. Where SDs differ greatly, the formula used is: $(MEAN_1 - MEAN_2)/SQRT((SD_1^2+SD_2^2)/2)$. The concept of effect size can also be applied to relationships between variables (correlations). Standard benchmark indexes distinguish between small, medium and large effects.

Graphical Communication. Graphical communication generally describes an academic course dedicated to developing communication skills linked to graphics. Includes notational systems, sketching, documentation, computer assisted design (CAD) and the production of technical drawings. Also generally includes some activities requiring spatial ability.

Heuristic Design. Positioned at the opposite end of the design spectrum to algorithmic design and is considered to be unstructured and creative. Design that is not generally driven by set procedures and constraints, but by general principles and objectives.

Holistic Solution Strategies. A holistic approach to solving problems is essentially seen as visualizing and reacting to a task as a whole. This approach is considered more efficient because responses to stimuli generally involve less time.

Item Difficulty. This is a measure of the degree of difficulty of a test item indicated by an index number calculated according to the proportion of test takers getting the item correct. The higher the index number, the easier the test item is.

Item Discrimination. This is a measure of how well an item in a test separates high scorers and low scorers on the total test as a whole. A test item is not performing adequately if high scorers on the test as a whole score poorly on the test item.

Item Reliability. Item reliability is a measure of the internal consistency of test items within a test, or the internal consistency of test items within a subtest. A high index number indicates high internal consistency. Cronbach's alpha is one measure of item reliability.

Item Response Theory. A method of item analysis gaining acceptance is known as the *item response theory* model or sometimes referred to as the *latent trait theory* model. In simple terms, it represents each test item under consideration as a plot on a graph where one axis represents ability while the other represents the probability of a correct answer. The shape of the plot provides a quick assessment of the standard of the test item. There are one, two and three parameter versions of the item response theory model.

Latent Trait. A present but not active or able to be seen distinguishing characteristic or quality of a person.

Program. A program is also known as an undergraduate degree or equivalent offered by a university. Programs consist of courses that typically lead to the award of a Bachelors degree or equivalent after successful completion of the courses.

Reaction Time (RT). A measure taken that accurately shows the time taken for a test taker to respond to a test item or stimulus. It is generally reported in milliseconds. Other terms that may be used are *response time* and *time taken*. Note that RT on occasions in this thesis can also refer to a subtest called *Transformation* and its abbreviation is also RT. The context in which RT is used will make it obvious which meaning is applicable.

Reliability. A measure of reliability which is generally in the form of an index number reports the level or repeatability of results for a test or test item. There are a number of forms of

reliability such as internal consistency and test retest reliability. A good test should show high repeatability to be a meaningful instrument. A test can be reliable independent of its validity.

Sociological Experiences. This refers to life experiences especially during developmental years that shape human attitudes, behaviour and personal skills. Many possible life experiences and activities engaged in during formative years are thought to have an impact on the development of spatial skills and understanding.

Spatial Cognition. Regarded as higher level spatial thinking that involves all aspects of related knowledge including: perception, thinking, imagining, reasoning, judging, remembering and communicating.

Spatial Ability. Defined as the performance on tasks that require the mental rotation of objects, the ability to understand how objects appear in different positions, and the skill to conceptualize how objects relate to each other in space. Of note is that one disadvantage of using the word *ability* in a descriptor is that it can have a connotation of something that is less predisposed to change and seen instead to be more innate. For this reason, there would be a preference among many in the discipline to using *spatial performance* rather than *spatial ability*.

Spatial Factors. Spatial factors are known as elements, classifications or components of spatial ability. A spatial factor can also be described as a spatial skill. Spatial ability consists of a number of spatial skills, and collectively they provide an assessment of spatial ability. A different test or a set of tests are required to measure separate spatial factors.

Spatial Tests. Instruments designed to measure spatial ability. For the most part, they consist of test items of single design that are used for a wide range of purposes across a variety of disciplines.

Subject Matter Experts. Professionals from industry employment or academia where spatial ability is a strong requirement for success in that profession. The term is also abbreviated and used throughout as SME.

Subtest. Essentially a spatial test that is one subset of a larger test of spatial ability. A subtest is intended to measure a single specific spatial skill or help identify particular learning difficulties.

Technical Drawing. A technical drawing is a collection of related views of an object (e.g., buildings, machine parts) and a notational system that combine to graphically represent and convey technical and structural information about that object. It is an essential skill for industrial professions and tradespeople. An alternative name is engineering graphics.

Test Item. Each test subtest of any ability is made up of stimuli designed to capture the performance of a test taker. *Test item* is an alternative name for stimulus or question.

Test of Single Design. Test items that make up the stimuli in a subtest are considered to be of a single design. This means that the test items or stimuli are similar in style and shape and they are simply variations of the same design and purpose. In contrast, a test of spatial ability in the context of this research consists of a range of subtests and therefore a mixture of test items.

Test Type. A term used to describe a type of test or a subtest.

University Admissions Index (UAI). A person wishing to enrol in a university degree is often given a score based on their prior academic achievements which is used or partly used to decide if a place will be allocated, especially where places are competitive. In some states in Australia, this score is called a University Admissions Index (UAI). There are generally equivalents to this in other states or in other countries. In many respects, UAI is a good measure of general academic ability.

Validity. The validity of a test is the degree to which it measures what it claims to measure. There are a number of validities and those relevant to this research are defined in the thesis as they appear. These validities are named: content validity, construct validity, convergent validity, discriminant validity, face validity and validity with known groups.

Visualization. The ability to call up or form mental images or pictures or to make perceptible to the mind or imagination. The process of forming mentally visual images of objects not present to the eye.

Visual Perception. The process of selecting, transforming, organizing and interpreting graphical information received through our visual sensory receptors. The ability to interpret what is seen.

TECHNICAL DRAWING CONCEPTS

A technical drawing is defined as a collection of views of an object and a notational system that combine to communicate technical information about that object. Technical drawing is usually based on a standard set of three axes (X, Y and Z) meeting at right angles at a point called the origin. The axes provide a cartesian coordinate system for locating points, lines and planes. Reference planes (also called viewing planes) are usually defined, one parallel to the XY axes (XY plane), another to the XZ axes (XZ plane) and the third to the YZ axes (YZ plane). The details of an object are normally represented by a set of 2D drawings where each represents a different view of the object. 2D views are referred to as orthographic views. To produce a set of orthographic views, an object is theoretically positioned with respect to the three axes and individual views are projected perpendicularly on to each of the reference planes (see Appendix A, Figure A01). This method of projection is termed orthographic projection. The view seen through the XY plane is called the Top View (TV), the view seen through the XZ plane is called the Front View (FV) and the view seen through the YZ plane is called the End View (EV). 2D views produced by orthographic projection are termed degenerate views because one axis is excluded. The distance of the object from the reference planes is not critical to the shape of the projected views as the shape is the same regardless of the distance in an orthographic projection (Sutton, Heathcote, & Bore, 2007). This is a convention that contrasts with perspective drawings where the distance behind the viewing plane is critical to what is seen.

There are three important concepts that are fundamental to orthographic projection. These are termed *true shape*, *true length* and *true angle* and all three are related. Essentially they imply that the true shape of a surface, the true length of an edge and the true angle of inclination of an edge or surface are not always seen in a 2D view. Instead, what is often seen are apparent shapes, apparent lengths and apparent angles of inclination. A true angle for example, is seen when a projection of a line or an edge is parallel to a viewing plane. Knowing the difference and the conditions under which true measures are seen is important to the understanding of a technical drawing.

Isometric drawings, which provide more obvious information about the 3D properties of an object than orthographic projections, are produced by a view that is not parallel to any axis, with the view being projected onto a plane perpendicular to the viewing direction. For an isometric drawing, the top view of a line representing the viewing direction is typically at 45°, 135°, 225° or 315° to the X axis. In the standard setting of most Computer-Assisted Design software the viewing direction will be a true angle of 35.3° to the XY plane. In isometric drawings, parallel edges running away from the viewing plane are always drawn as parallel lines. This contrasts with perspective drawings where parallel edges that run from the viewing plane are drawn as converging lines. However, for relatively small objects (as opposed to, say, landscapes) the

difference between isometric and perspective drawings is negligible (Sutton et al., 2007). Given this, and the dominant use of isometric rather than perspective as 3D representations in technical drawing, the focus in this thesis is on isometric.

A common approach to developing technical drawing competency is to experience both 2D and 3D representations concurrently in order to develop an understanding of the relationship between the two. It is not desirable to develop 2D or 3D skills in isolation and in many respects, working in 2D may be more important than working in 3D. James et al. (2001) report that participants in their experiments spent more time looking at the end and front views of objects rather than three-quarter or intermediate views. They suggest that these are the views where there is the greatest amount of difference in the visibility of object features. In contrast, the three-quarter views are perceptually similar. The process of working from 2D to 3D drawings, and working from 3D to 2D drawings, is the common way students build up their understanding of concepts. The ability to interpret a multi-view drawing is learnt by forming mental images from the 2D views and visualising what the object will look like in 3D. As the complexity of objects increase, extra views are generally necessary, including sectional views (planes cutting through objects), exploded views (magnified projections showing individual parts separated), and assembled views (working parts in position) (Sutton et al., 2007). The number and type of views to form a technical drawing will depend on how complex the object is and how much information needs to be communicated. Collectively, the 2D views convey precise details about an object intended for manufacture or machining. Technical drawings generally consist of a set of 2D drawings (views) because complex information is more easily represented this way. As a consequence, an object is less frequently drawn as a 3D representation and the visualisation of the object relies heavily on the ability to interpret a set of 2D views. This skill is regarded as being able to *read* a drawing. As touched on earlier, Salthouse (1991) leaves little doubt about the importance of being able to read a drawing. He states that the ability to understand a technical drawing is vital because technical drawings are necessary in the process of converting from a design concept to a physical structure of some form. He adds that the correct interpretation of a technical drawing is critical because they often serve as legal documents to indicate what will be constructed. Costly delays and litigation are possible when errors occur. While designers may not always be engaged in producing actual drawings, Salthouse considers graphical comprehension will nevertheless be an important factor throughout their careers.

PROLOGUE

The research reported in this thesis is about developing an online psychometric test of spatial ability for designers. The name given to this test was the 3D Ability Test (3DAT) and the final version consists of five subtests with four test items in each. These numbers fluctuated throughout the different studies and at one stage 25 subtests and 119 test items were being considered. Each subtest measures a different spatial factor and collectively they assess a construct called spatial ability. From the outset, the prime objective was to establish a 3D ability test for designers (e.g., architects, engineers) with a longer term objective of developing learning tasks to improve spatial ability for novice designers. This research demonstrated an example of applied psychology, and for many studies, a psychology/ design nexus was crucial to outcomes. In many respects, the strengths of the psychology discipline such as expertise in experimental design, psychometrics and statistics were exported to an external discipline to achieve research objectives beneficial to that discipline. In this case, the discipline was design which encapsulated mostly engineers, architects and construction managers in this research. The decision to develop the 3DAT as an online test was based on the easy access it would provide to institutions wanting to use it as part of their curriculum. There were other advantages such as the ease at which changes could be made, the customisation possible, the simplicity in managing, collecting and collating dependable data, and the potentially large samples that could be captured for educational and research purposes. To achieve the online version, a procedure was put in place that included several studies conducted in a lab environment using local software to evaluate the viability of going online with the 3DAT. It was a progressive undertaking that considered such things as timing issues, server implications, data collection, graphical displays and the testing of many variables. The lab studies provided benchmarks that the online version could be compared with during the different stages of development. The data collected from the different studies were mostly quantitative that measured choice accuracy (correctness of answers) and reaction times (RT). The accuracy data were always utilised, but the RT was not reported for every study. The 3DAT was never promoted as a speeded test since accuracy was always seen to be more important than speed. RT was therefore considered difficult to utilise on every occasion. However, RT data proved to be far more meaningful than first thought and was fully exploited in a major study towards the end. The collection of some qualitative data did take place in two particular studies during the later stages of development and they provided meaningful information from a different perspective on both occasions.

The first chapter establishes the importance of spatial ability and provides reasons why the development of the 3DAT was justified in view of the large number of so called spatial tests that are available. The existence of spatial ability elements which are generally referred to as spatial factors is a disputed issue in this field of research. Considerable effort therefore was

spent throughout this research to test for spatial factors. The identity of these and the subtests that would measure them were important issues because they primarily dictated what the profile of the 3DAT should be to adequately measure spatial ability.

Because *technical drawing* is an essential tool for designers for communicating technical and sometimes complex information to others, some time was dedicated earlier to explaining the fundamental concepts and the notational system used in technical drawing. The requirement to read a technical drawing and to understand its complexities goes to the heart of why spatial ability is such a central attribute for designers.

The steps in developing the 3DAT were sequential across nine studies and psychometric properties were investigated in different degrees at various stages. For example, initially in a preliminary context, then later in an expanded capacity, or repeated when sample sizes increased. The later stages made it possible to produce more convincing evidence in support of the 3DAT. Studies varied in design and in some cases they were exploratory, and in other cases they focused on specific things such as item analysis, factor analysis, expert opinion and attempts to identify spatial factors. Many participants in the studies were novice designers who came from a range of design fields of study. It was also necessary to include participants from disciplines other than design because they were needed to help demonstrate particular psychometric properties. Every study involved both male and female participants, but males were always in the majority. This condition alone was a reminder of the shortage of females attracted to the design discipline. A substantial part of the investigations examined the all important concepts of validity, reliability and correlation. The diversity of the studies meant that a range of statistical procedures could be applied, and they collectively provided the evidence that the aims of the research had been achieved.

A significant proportion of this research was dedicated to exploring gender differences. The literature consistently reports a difference in spatial ability and a difference favouring males is generally found. However, some studies refute this position and argue that factors that influence findings include the type of test used to measure performance, gender stereotyping, sociological experiences and the type of training (if any) that has occurred. Because gender is generally a factor in most studies related to spatial ability, and because it is an issue many researchers are interested in, it would have been remiss not to have given it due consideration in this research. Consequently, gender issues received coverage from a number of perspectives. Comparisons were made between a design group and a nondesign group, also within those two groups, between disciplines and across various subtests. Findings were mixed, but generally males out performed females, though not in all cases. Sample sizes for female cohorts were not always ideal, but this was a reflection of the state of the design disciplines in general, except say for architecture where numbers were reasonably balanced. This research supported the position that

gender difference is not reducing in terms of test scores, though one small study suggested otherwise. Also that the difference is not always robust, and that the group that test takers belonged to had an impact on that difference.

The principles of test construction were applied to the development of the 3DAT, and although these are described in a number of ways and in different levels of detail in the literature, they essentially come down to several fundamentals. For example, a need for the 3DAT was established, and this included firm ideas about what it should measure and what method of measurement would be used. Test item preparation followed which was a lengthy process because of the uncertainty about the design of test items, their psychometric properties, the number required and item variation. Next the 3DAT was subjected to a series of pilot testing and several features of the experimental design were revised accordingly. Item analysis was somewhat continuous throughout developmental stages and resulted in various degrees of item reworking, item replacement, layout changes and the retesting of psychometric properties. The more indepth side of item analysis where subtests and test items could be permanently discarded occurred in the latter stages of development after large sample sizes had been achieved. Consequently, this achievement allowed full confidence in the analysis undertaken. This research conformed to commonly held views of test construction principles which essentially became a set of guidelines that pervaded every stage of development. The position taken was that accurate tests do not simply happen, but instead, they are the result of a systematic approach to test development procedures established over time.

FORMAT AND STRUCTURE

This thesis consists of a foreword section, five chapters, a reference list and a set of appendices that are referred to in the chapters. The appendices serve a more significant purpose in this thesis than may be usual since they are seen as an alternative to adding large Tables and Figures to the main text. They also provide additional information that complement the key facts reported in the thesis which a reader may find of special note. Ideally, the appendices will prove to be convenient and will be reviewed alongside the main text when referred to. Chapter 1 introduces the spatial ability construct and examines concepts, issues and theories that all underpin spatial ability in some way. A literature review is part of this chapter, but it is not developed as a separate section. Instead, it is integrated with other reporting with the intent of supporting, challenging or clarifying points as they arise. Chapter 1 also provides a rationale for conducting the research and presents hypotheses to clearly state the direction of the research. Also included towards the end of this chapter is a brief overview of the stages devoted to developing the 3D Ability Test (3DAT) since this development was the fundamental objective of this research. Chapter 1 concludes with a brief description of the final 3DAT and a mention of its special features and qualities.

Chapters 2, 3 and 4 are dedicated to specific stages in the development of the 3DAT and they are titled: *initial*, *transitional* and *final* respectively. These chapters are sequential and each covers several studies where each study has a different focus. All three chapters have intentionally similar formats and start with the aim of the study followed by specific objectives, methodologies, analyses undertaken and results. Further, each chapter finishes with a summary of all study outcomes with major findings particularly addressed. Throughout these chapters, psychometric test development principles are paramount and the essential properties of validity, reliability and those revealed from item analysis are foremost in every consideration. As a consequence, a range of statistical procedures are also reported, and coverage is progressive in the sense that issues such as reliability are not examined in every study, but only where required. Sometimes these properties are revisited as part of that progression.

Chapter 5 is concerned with *discussion and conclusions* for the entire research and draws attention to outstanding matters, implications, notable issues and recommendations. In particular, the research hypotheses are evaluated against final outcomes. In simple terms, the evidence from this research is reviewed and particular aspects are reported.

One final point is that gender is a significant topic in the literature and it is a major consideration in this research. In many respects, gender is deserving of a dedicated section, however, the approach taken was to treat this topic contextually where it seemed appropriate to emphasize particular points. This means that gender issues surface in many of the studies, and reporting is sometimes part of the reporting of other findings.

CHAPTER 1

INTRODUCTION TO SPATIAL ABILITY

Introduction

One of the more difficult skills for technical drawing students to develop is the ability to understand three-dimensional (3D) concepts. Technical drawing is a graphical communication method used by designers (e.g., architects and engineers) and tradespeople to share technical information about objects such as buildings, machinery and engineering structures. Technical drawing is central to the development of products from the conceptual stage through to the manufacturing stage. Objects are normally represented by a set of related two-dimensional (2D) drawings where each drawing represents a different view of the object. These views can be generated for any number of viewing directions and they can be full views, part views, sectional views and can be drawn to any scale. The complexity of the object generally dictates how many views are required and the type they will be. 2D views are referred to as orthographic views. Objects are represented by sets of 2D views because it is easier to convey complex information about them. 3D understanding is the ability to extract information about 3D properties from 2D views (i.e., drawings) (Sutton, Heathcote, & Bore, 2005). In technical drawing, the ability to work from 2D to 3D, and from 3D to 2D is paramount. This skill requires spatial abilities to interpret what is seen and to mentally manipulate visual representations. Spatial ability can be defined as the performance on tasks that require the mental rotation of objects, the skill to understand how objects appear in different positions, and the skill to conceptualize how objects relate to each other in space (Sutton & Williams, 2007). Spatial ability is an essential skill for designers who typically undertake graphical communication courses.

The current situation is problematic. Spatial ability is acknowledged as an important skill required by designers and considered by many to be deserving of greater emphasis in engineering education, including adopting it as a core competency (Contero & Naya, 2006). However, there are many issues that are not being addressed. Miller and Bertoline (1991) see the importance of spatial ability as being accepted among designers and technicians but consider some colleagues inside and outside of the discipline still need convincing. There are concerns about mixed abilities on entry to undergraduate training because of differences in prior learning experiences despite high achievements in other academic areas (Blasko, Holliday-Darr, Mace, & Blasko-Drabik, 2004). This suggests that undergraduate training may require accelerated learning programs to ensure a consistent standard upon graduation (Akasah & Alias, 2006). The researchers also confirm a disparity in prior learning experiences and point to some students starting undergraduate training as experts in spatial ability while others find the subject overwhelming. Sexton (1992) adds to this concern and draws attention to an inequality in prior

visualization training and contends that the only exposure undergraduates may receive is via one or two courses at undergraduate level. Mixed entry abilities pose problems for design educators in a similar way to how mixed prior learning in mathematics might pose problems for post secondary education in physics. Not helping the situation is a trend away from indepth graphical communication education at many higher education institutions. Although these courses are the main source of spatial ability training for many undergraduates, they are becoming less emphasized in some situations and are being dropped from the curriculum in many others (Sorby & Baartmans, 1996). In other words, there is a contradictory situation where the importance of spatial ability is increasing, but at the same time a commitment to courses where learners are first introduced to spatial concepts is declining.

Another problem consistently reported is gender difference that favours males. Though differences vary, most tests of spatial ability used in a range of studies show males to perform better than females. There is evidence that females are up to three times more likely to be deficient in spatial ability than their male equivalents (Sorby & Baartmans, 1996). In another study (Coleman & Gotch, 1998), the researchers confirmed the results from earlier studies conducted by other researchers that indicated a disparity in spatial skills between males and females. Importantly, Coleman and Gotch also showed that the performance of females was reasonably fixed over a period of time and was consistently below that of male counterparts.

What often arises is the question of whether spatial ability is an innate ability or whether it is a developed ability. That is, can the difference in male and female performances be explained in terms of a given ability, or is it due to cultural and environmental differences, especially those experienced in early childhood?

Also debated is the type of test used to measure spatial ability since some appear to better suit the strategies preferred by males rather than the strategies preferred by females. Males are more likely to use a *holistic* approach to solve spatial tasks while females are more likely to use an *analytic* approach (Sorby, Drummer, Hungwe, & Charlesworth, 2005). These researchers describe the holistic approach as visualizing a task as a whole while the analytic approach depends on a systematic step by step strategy. The holistic approach is considered more adept since responses require less time, which is an advantage, especially for tests where time is an issue. Linn and Petersen (1985) regard strategy choice to be a factor in gender performance on mental rotation tasks and found that males outperformed females on these tasks. Mental rotation tasks appear to favour a holistic approach requiring a single strategy to determine a solution. Holistic approaches treat spatial information in a *spatial manner* and maintain a spatial connection between task essentials. In contrast, analytic approaches treat spatial information by breaking it down into nonspatial list-like elements (Gluck & Fitting, 2003).

Other gender issues include the type of training conducted and a concern that research into effective training is limited (Hartman & Bertoline, 2005). Where training exists such as that expected in a graphical communication course, the question is whether the training is relevant to females since instructional methods can suit one gender better than another. There is some evidence for example that females feel pressured in mixed classes and that they achieve better results if they are able to work from home (Blasko et al., 2004). It is also reported that females responded well to training based on repeated practice that includes feedback (Kass, Ahlers, & Dugger, 1998). Furthermore, no gender difference in spatial performance was found in this study. Interestingly, Kass et al. focused on mental rotation, a task often reported as demonstrating a strong gender difference in favour of males (Voyer, Voyer, & Bryden, 1995).

Another issue thought to impact on the spatial performance of females is known as *gender-stereotyping*. This is a phenomenon where societal expectations influence what social and educational activities a person will engage in based on their gender. In other words, males do things that are expected of males (e.g., engineering), and females do things that are expected of females (e.g., humanities). Thus, it promotes a self-fulfilling prophesy such that success is largely impaired by gender role expectations dictated by society. The implication is that this may impact on attitudes, career aspirations and academic progress if choices are made contrary to expectations. Since spatial ability is normally linked to career paths chosen by males, it means that gender-stereotyping may have a greater impact on the spatial performance of females than other factors such as biological and learning strategies. In support, Quaiser-Pohl and Lehmann (2002) point out that Nash (1979) and Horner (1972) contend that people do better on cognitive tasks if their self-perception agrees with the gender-stereotyping of the task. Further, Spencer, Steele, and Quinn (1999, as cited in Sorby et al., 2005) report that achievement by individuals may be affected if the requirements of a task fail to match gender-stereotyping for that individual. Perhaps more importantly, gender-stereotyping strengthens the perception that females are less capable in a wide range of mathematical and spatial competencies than their male counterparts. This perception persists despite reports that females achieve grades equal to if not better than males (Holliday-Darr, Blasko, & Dwyer, 2000).

There are several other spatial ability problems to highlight. The first is the attrition rates for design students which are commonly linked to poor spatial skills, and associated efforts to improve retention rates. Deno (1995) points to a deficiency in visualization skills as a reason for many students dropping out of graphics courses very early in their careers, and considers addressing shortcomings would improve retention rates for engineering students. In one detailed study (Blasko et al., 2004), a number of variables such as academic background, motivation, parental persuasion, verbal skills and spatial ability were tested to determine what might impact on student retention rates the most. The researchers found that scores on basic tests of spatial

ability (e.g., mental rotation) were the best predictors of retention. Osborn and Agogino (1992) add weight to concerns about attrition and contend that many students drop out of graphics courses simply because they lack the confidence to handle the content in modules such as descriptive geometry. Descriptive geometry is a traditional area taught in technical drawing which is often criticised for being too hard and irrelevant to many disciplines. This thinking has led to the removal of descriptive geometry from many graphical communication courses. However, the consequence may now be that many students are not developing the spatial skills they require. One argument for retaining descriptive geometry is the contribution it makes to spatial ability development. Fundamental to descriptive geometry is the understanding and application of many elements of spatial ability. This includes mental rotation, true length and the relationship between sets of 2D views and 3D properties.

On the positive side, other researchers have shown that remedial courses designed to improve spatial ability can increase retention rates. For example, Sorby (1999) showed that weaker students who participated in special courses to improve spatial ability were less likely to be disheartened by the difficulty of their graphical communication courses. Further, Sorby and Baartmans (1996) provided evidence from a five year longitudinal study that retention rates were better for weaker students who participated in a remedial course compared to weaker students who chose not to. This was especially so for females. Supporting this position, Boersma, Hamlin, and Sorby (2004) go one step further and reported that a remedial spatial visualization course increased student retention rates at both general university and engineering curriculum levels.

Another problem is the neglect often accorded to the teaching of spatial ability. Maier (1998) suggests that the deliberate training of spatial ability has not been seen as important in many curricula, and that any direct or intentional teaching of 3D shapes and 3D geometry has been ignored for decades. This position is reinforced by Adanez and Velasco (2002) who argue that activities in primary and secondary education do not adequately foster the development of spatial skills. In many respects, the development of spatial understanding is *incidental* since no direct effort or strategies are employed to bring about improvement. Ben-Chaim, Lappan, and Houang (1988) see spatial ability as not being part of the normal school curriculum and refer to this incidental learning as an *informally acquired skill*.

One final point is that of early diagnosis. Adanez & Velasco (2002) argue that the learning process would be more effective if those students with low spatial understanding could be identified early in their elementary education. They put the case for introducing a spatial diagnostic test to identify poor performance which would allow strategies to be developed in a dedicated training program to bring about improvement. Similarly, Potter & van der Merwe (2001) make a case for early detection of students in the higher education sector who are

deficient in spatial ability. They report substantial improvement in pass rates (64% to 88%) in first year engineering technical drawing courses where specific strategies were introduced to increase performance. Interestingly, this occurred during a period when the cultural and academic background of the student population increased in diversity.

Spatial ability is not without issues. Problems arise from an uncertainty about the importance of spatial understanding, how best to measure spatial ability, what training is appropriate to bring about improvement, and gender issues that are not being addressed. These concerns exist at a time when traditional courses that normally allow students to improve their spatial understanding are diminishing. Current learning is mostly incidental with only some evidence that serious attempts are being made to remediate spatial understanding. A confounding factor is a societal expectation that encourages academic and career choices based on perceived gender roles. For many, spatial ability is innate and to them it makes little sense to measure performance or to seek ways to improve performance. However, there is also a growing number of researchers producing evidence about the importance of spatial ability and suggesting ways about how it can be improved. For improvement to occur, spatial ability must first of all be measured accurately to help decide about appropriate learning tasks. The measurement of spatial ability is fundamental to this research.

Importance of Spatial Ability

The importance of spatial ability needs to be established. It is difficult, for example, to see how complex technical information about structures and mechanisms can be shared between designers and tradespeople without the ability to identify 3D properties from 2D drawings. As well, the complexity of computer assisted design (CAD) software used by most designers today places extra demands on users because of the advanced features the software has to offer. Users require spatial understanding to manipulate 3D models, utilise multiviews, visualize objects from different viewing directions, understand sectional views and to work with a mixture of drawing scales. Sorby (2000) considers that there is a correlation between spatial understanding and the ability to work in a 3D computer software environment. There is also the likelihood of onsite inspections that are conducted by project supervisors who identify and resolve production problems that emerge during the life of a project. Designers charged with these responsibilities require spatial skills to assess the physical progress of projects against technical drawing information. Further, a notable duty of many designers is to provide an estimate of the cost of a project. This occurs at an early stage before production begins and the demand on spatial skills increases. The estimator requires a good understanding of 3D concepts and needs to extract information from a set of technical drawings to produce the quotation required. In essence, designers mainly communicate graphically (Leopold, Gorska, & Sorby, 2001), but without good spatial skills, it is difficult to see how they can achieve this effectively.

Spatial Ability in the Workplace

The importance of spatial ability can be more precisely illustrated using several key examples. Foremost is spatial ability as an essential skill in most technical-based vocations. There are professionals such as architects, engineers and construction managers who require the ability to visualize 3D properties and communicate ideas during the conceptual and design stages of product development. Similarly, tradespeople such as builders, machinists and fabricators need to be able to visualise and understand graphical information before manufacture and machining is possible. People employed in these occupations work from sets of technical drawings which are often complex in nature. Consequently, these people require spatial skills to mentally manipulate objects, to visualize how objects look from different directions and to particularly understand the relationship between 2D and 3D representations. Salthouse (1991) sees the ability to interpret technical drawings as a critical skill since drawings are considered fundamental in the transition from design to the construction of most objects. Other perspectives can be identified. Jenson (1986, as cited in Bertoline & Miller, 1990) reports that both industry and academic representatives associated with engineering graphics ranked spatial understanding as the most important of 14 identified skills in a survey conducted. Another consideration is the cooperation that must occur between disciplines. Professionals and tradespeople do not generally work in isolation, so there needs to be a common understanding of a graphical language. Osborn and Agogino (1992) make the point that engineering disciplines such as mechanical and civil often share interrelated information about the design of 3D structures, and essential for this is spatial understanding.

Other research supports spatial ability as a necessary skill for designers. For example, Smith (1964, as cited in Sorby 2007) reports that there are at least 84 different occupations where spatial understanding is an important attribute. Holliday-Darr et al. (2000) report on another perspective. They see powerful computers and modern software as increasing the importance of spatial ability for designers because this combination is now capable of displaying comprehensive and detailed spatial imagery not considered possible even a short time ago. Other researchers such as Sorby and Baartmans (1996), McGee (1979) and Contero, Naya, Company, Saorin, and Conesa (2005) simply make the point that spatial ability is central to success in graphics courses and engineering in general.

Because spatial understanding can be considered a necessary attribute for designers, there is convincing argument why spatial learning should be a core component in any introductory engineering graphics course. In other words, making an earnest effort to develop spatial skills similar to what now occurs for sketching, line work, dimensioning and the use of CAD. Quite often spatial learning is not given the attention it deserves, and any learning that takes place might be considered coincidental. Contero and Naya (2006) see spatial understanding as an

essential competency for engineers and argue that it must be given prominence in any future curricula. They also point to this coincidental learning and describe spatial skills as a secondary objective that is acquired through the study of other objectives. Though seen to be important, spatial ability does not receive the commitment required, and research into improving these cognitive skills is therefore considered critical (Rafi, 2006). In a tribute to Richard E. Snow (a renowned researcher in what Snow himself called *cognitive differential* psychology), Kyllonen and Lajoie (2003) reflect on Snow's robust interest in the value of spatial understanding. They point to Snow being driven by an enthusiasm for empirical and quantitative research, and one particular focus he put forward concerned the poor progress of his students. Snow considered this poor progress to be related to an inability to comprehend his graphs and diagrams which he ascribed to an educational system that was bias against spatial learning in favour of verbal production and understanding. Snow considered spatial ability to be important but he also saw it as an under developed attribute. There is also some evidence that the development of spatial ability is not given sufficient priority in secondary education which suggests many students enter design programs at tertiary level without prior learning in spatial ability (Leopold et al., 2001). In essence, there exists a body of evidence that supports spatial ability as an essential skill for technical-based vocations which includes those disciplines with a focus on design.

Spatial Ability as a Predictor of Success

Success in design-based programs such as engineering can be linked to spatial ability. As previously mentioned, there are concerns about low achievement in introductory graphical communication courses which often lead to unacceptable student attrition rates (Blasko et al., 2004). These courses are generally at the front end of respective programs and at a critical time when students are attempting to adjust to tertiary study. A relationship appears to exist between poor spatial skills and the number of students who withdraw from courses where spatial understanding is an integral part of those courses. Consequently, spatial ability is regarded by some researchers as a predictor of success in these courses which in turn provides a guide to how many students will continue in design-based programs. For example, Adanez & Velasco (2002) report that spatial tasks (particularly those concerned with visualization), are capable of predicting the number of engineering students who will do well in graphical communication courses.

To extend the notion of spatial ability as a predictor of success a little further, D'Oliveira (2004) draws attention to using spatial ability as a measure of aptitude for technical-based occupations and refers to a review of spatial ability presented by Smith (1964). Smith conducted a comprehensive review of the predictive worth of spatial ability and looked at a range of studies dating back to the 1920s. Smith concluded that spatial ability had a positive impact on predicting performance in some technical training courses (e.g., engineering drawing), and on

the success in mechanical and engineering-based apprenticeships. D'Oliveira also considered that the number of occupations where spatial ability could be used as a measure of aptitude could now be updated to include other occupations such as those in the aeronautical domain.

As a slight but relevant digression, D'Oliveira (2004) makes the point that research reported by Smith (1964) was focused on the traditional area of spatial ability (static) and suggests that no studies exist on the predictive merit of dynamic spatial ability. D'Oliveira adds that research in this area would augment the literature and further highlight the importance of spatial ability. Dynamic spatial ability (i.e., where movement occurs) is likely to have more practical relevance than static spatial ability (i.e., where no movement occurs) to occupations such as aircraft flying and air traffic control because of the different spatial reasoning required. On the other hand, static spatial ability is considered more applicable to designers.

Returning to the predictability of spatial ability, and to be a little more specific, Pellegrino, Alderton, and Shute (1984) emphasize the validity of the spatial ability tests themselves. They point to the correlation that exists between spatial ability measures and course grades in academic and vocational technical training courses and provide examples such as mechanical drawing, workshop courses, mathematics and physics. Pellegrino et al. also comment on job performance in industry and assert that spatial ability measures have a proven track record in predicting success in design-based vocations such as engineering and drafting. In so doing, they essentially support D'Oliveira (2004). The authors also stress that the predictability of spatial ability tests is most often superior to tests of verbal ability and measures of general intelligence. Importantly, Pellegrino et al. consider spatial ability as a separate intellectual ability to verbal, quantitative and reasoning abilities and maintain that this separation is the result of psychometric theory and investigation. Other researchers also comment on the predictive validity of spatial ability measures. Gimmestad (1989, as cited in Sorby, 2000) and Medina, Gerson, and Sorby (1998) both report on particular spatial ability tests as being the most noteworthy predictors of student achievement in graphical communication courses. Sorby's own studies supported these findings. As well, Potter et al. (2009) provide evidence of a strong positive relationship between spatial ability and academic achievement from a study they conducted with a first year engineering graphics course. Though not associated with design-based disciplines, Workman, Caldwell and Kallal (1999) report that MacDonald and Franz (1989) found that spatial ability testing was the best predictor of success and performance in apparel design courses when compared to quantitative and logical reasoning tests. Workman, Caldwell, and Kallal (1999) raise the idea of some forms of spatial tests being used to measure aptitude to help select students for technology-based disciplines. What this implies is that students who score well on these tests could be encouraged into these disciplines because of

their spatial ability rather than rely on a self-selection process based on general academic achievement only.

Evidence presented here indicates an existing relationship between spatial ability and success in graphical communication and technology courses that are generally part of any design-based program. This relationship then is often regarded as a predictor of success in design related programs. Where poor performance is noted, it often serves as an indicator of the attrition rate that can be expected. On the other hand, poor performance also acts as an early warning that some treatment of spatial understanding may be necessary if improvement in retention rates is to occur. The positive correlation between spatial ability and success in design-related courses serves to further illustrate the importance of spatial ability. Moreover, the practical implications of attrition, retention and predictability amplify this importance only too well.

Relevance of Spatial Ability to Other Disciplines

There is also the question of spatial ability and its relevance to disciplines other than design such as those in the sciences and some areas of health. For these disciplines, spatial understanding is equally important because of the need to interpret graphical representations of objects and to be able to communicate ideas using images and diagrams. As well, conceptual thinking is often best conveyed using computer generated 3D models where spatial reasoning is also likely to be critical. Examples of science disciplines that immediately come to mind are mathematics, chemistry, physics and geology. All these disciplines entail some form of graphical communication and some form of 2D to 3D and 3D to 2D transformations to understand real world situations. A number of researchers consider spatial ability to be fundamental in the formation of many competencies in these disciplines which in turn help improve overall performance in those disciplines (Coleman & Gotch, 1998; De Lisi & Wolford, 2002; Hartman & Bertoline, 2005; Sorby et al., 2005).

In some health disciplines where medical imaging is used, professionals are required to interpret difficult 2D low resolution images and make critical decisions according to this interpretation. Orthopaedic surgeons for example, need to be able to visualize skeletal structures from X-ray images which are generally taken from two or three orientations only (Allahyar & Hunt, 2003). Dentistry is another example, and any misreading of radiographic images because of poor spatial ability can have serious implications. Garg, Norman, and Sperotable (2001) report spatial ability to be critical to students learning about anatomy, and they emphasize the importance of 2D views. They consider that there are optimum viewing directions that provide the spatial information required, and suggest for models that have multiple viewing directions, a few key 2D views will be favoured by most learners at the expense of other views. However, Garg et al. acknowledge the value of multiview models (e.g., skeletons) but suggest learners need to have

control over viewing directions to gain the maximum benefit. They also emphasize the importance of spatial understanding in medicine and suggest implications for recruitment, performance and counselling. Another perspective is presented by Wanzel, Hamstra, Anastakis, Matsumoto, and Cusimano (2002) who report interesting findings related to specific surgical procedures. That is, they suggest there is a relationship between spatial ability and the competency of residents to conduct spatially complex surgical procedures, and they emphasize this by pointing out that, as the complexity increases, so too does the reliance on spatial skills. Wanzel et al. also indicate that the residents with the highest spatial skills are able to transfer previously acquired skills to more complex procedures with less difficulty. In contrast, those residents with lower measures of spatial ability are likely to need additional training and counselling to meet this challenge. The researchers also comment on a positive correlation between spatial ability and the quality of work from the residents who performed those procedures. Furthermore, their research also supports the use of spatial diagnostic tests to help identify spatial problems with implications for training and skill development.

Whilst the main focus of this thesis is on design-based disciplines, what has been established here is that spatial ability is applicable to other disciplines as well. The disciplines that emerge in particular are those from the sciences and some areas of health, but there is every likelihood that spatial ability is relevant to other disciplines such as creative arts. Consistent with this thinking, spatial ability was shown earlier to be important to apparel design and product development. Evidence overall draws attention to the relevance of 2D views and reinforces the significance of being able to extract 3D properties from 2D drawings. This fundamental skill is critical to communicating ideas graphically. This also raises the prospect of having purpose-developed spatial tests to identify problems, appropriate training methods and suitable learning tasks. A purpose-developed test of spatial ability for designers is central to this research.

Improving Spatial Ability

Training as a separate focus is an important part of spatial ability because it addresses the question of whether spatial ability is an innate ability only, or whether it can be developed. A range of findings suggest spatial ability can be improved, which is an important statement in view of historical opinions to the contrary. There is evidence for example, that spatial ability is trainable and can be developed from childhood to adulthood (Potter & van der Merwe, 2001). This evidence is consistent with Piagetian Theory that claims perception and mental imagery are processes that can be trained at any age throughout the human lifespan (Potter & van der Merwe, 2001). Ben-Chaim et al. (1988) provide evidence of improvement for middle school students (grade 5 to grade 8) who benefited substantially from contextualised learning. They found that improvement was possible using appropriate tasks, and that improvement increased with age and occurred equally for both male and females despite there being a gender difference

in performance. Ben-Chaim et al. also found a retention effect in that performance gains on posttests persisted after four weeks and again after one year. Their results, however, were in conflict with Sedgwick (1961, as cited in Ben-Chaim et al., 1988) who considered spatial ability was likely to be an innate ability that could not be improved with dedicated training. On the other hand, results from Ben-Chaim et al. agreed with the findings of Brinkmann (1966, as cited in Ben-Chaim et al., 1988) who put forward an opposing view and maintained that spatial skills could be improved with appropriate instruction. Ben-Chaim et al. concluded that, given the chance, both sexes have the same potential to benefit from spatial ability training.

In another study that examined the effect of training on the ability of learners to see the relationship between 2D and 3D views of an object, Duesbury and O'Neil (1996) reported that instruction and practice that targeted visualization tasks produced improved performance on measures of spatial ability. Their learning tasks were based on the use of computer assisted design (CAD) software and its capability to allow manipulation of computer-generated objects in real time. This gave the learner the ability to control movement and rotation of an object and to observe two-way transformations between 2D and 3D views. The ability to see the relationship between 2D and 3D views of an object is a skill critical to designers. A conclusion of Duesbury and O'Neil was that CAD software could be used successfully as an instructional platform for spatial ability training. In contrast, Duesbury and O'Neil suggest that the improved posttest results found by Lajoie (1986), Waldron (1985) and Zavotka (1985) could not be attributed specifically to the use of CAD alone. Instead, they regarded the improved performance in these studies was a consequence of the instructional design rather than the delivery platform itself. One final feature of the studies conducted by Duesbury and O'Neil was their use of two treatment groups (rotation and nonrotation) and a control group. The treatment groups only differed in the level of manipulation they were allowed within in the CAD package. The rotation group were able to move, rotate and fold things and switch between 2D and 3D views. However, the nonrotation group were restricted to switching between 2D and 3D views and observing changes only. There was no intervention for the control group between pretesting and posttesting. Posttest results showed that the rotation group performed significantly better than the two other groups, while there was no significant difference between the nonrotation and control groups. These results infer that learning tasks that allow active exploration are superior to those tasks that only allow passive participation. Improvement from training is clearly evident in the studies conducted by Duesbury and O'Neil.

Sexton (1992) expresses a more cautious position and cites mixed findings from other researchers. For example, studies conducted by Faubion, Cleveland, and Harrell (1942) showed no significant improvement. However, studies conducted by Dorval and Pepin (1986) did show a significant improvement. Of note is that the results that indicated improvement are far more

recent. Despite mixed results from other researchers, Sexton maintains that there is sufficient evidence to support the idea that spatial ability can be improved with training. However, he does state two conditions for this to occur. First, the training needs to be appropriately designed, and second, it has to be implemented for a sufficient period of time. From the data reported, it was not possible to determine if Sexton's own work supported the idea of improving spatial ability using contextualised training. Only posttests results were reported which was sufficient to distinguish between the results of two teaching methods, but not sufficient to decide if improvement in performance occurred for either of the two study groups (control and treatment) between pretesting and posttesting.

Another training consideration that favours the nurture argument, and also emphasizes the value of training, is the remediation work carried out by several researchers. For example, Sorby and Baartmans (1996) developed a course for first year engineering students who were weak in spatial understanding based on a textbook and a computer lab manual written to take advantage of dedicated software. The researchers used a spatial screening test to identify low spatial performers (50 males and 46 females) who were then encouraged to enrol in this course. The average mark of the 96 students prior to entering the course was 51% correct , and after treatment, the average posttest mark was a statistically significant 86% ($t=12.53$, $p<.000$). The course appeared to have a positive impact on the spatial performance of students who were at first identified as not being strong in spatial ability. It is therefore difficult to see how spatial ability can be seen purely as an innate attribute when improvement of this order is possible.

If training is accepted as an important issue for spatial ability, then something further to consider is the gender difference reported earlier in this thesis. Females are disadvantaged in career choices and academic options because of a perception that they cannot do well where spatial understanding is required. In many cases, individual females have this view of themselves, and often those who influence them such as families and friends think likewise. In many respects – based on the experience of the writer – females may be victims of a self-fulfilling prophesy which stems from this perception. This may come from a belief that females are constrained by innate factors reported in early research. However, more recent research has challenged this view, suggesting that cultural and early life experiences are mainly responsible (Quaiser-Pohl & Lehmann, 2002). Furthermore, there is evidence that certain types of training can help narrow the gender difference and sometimes remove it altogether (Kass et al., 1998).

While there is some support for the nature argument based mostly on earlier research, more recent research overwhelmingly endorses the view that spatial ability can be improved with contextualised learning (Pellegrino et al., 1984; Sorby et al., 2005; Ullman & Sorby, 1995). Achieving success in training is important since it demonstrates spatial ability can be improved under certain conditions. Although some level of spatial ability will be innate, the potential to

improve deserves emphasis and should be clearly understood by academics and workplace supervisors. Miller (1992) makes the point that instructional strategies should be developed to help students increase their spatial skills. Without these, Miller adds that students may give up their goal to be engineers or fail to reach their potential. Holliday-Darr et al. (2000) press this point further and warn, if spatial ability is seen only as an innate ability, then any student with low spatial skills and is aware of this may avoid training designed to improve these skills.

There is sufficient evidence that particular training strategies and conditions can be used to improve spatial ability with lasting effect, which tilts the *nature versus nurture* debate in favour of nurture. The success of training has demonstrated that the gender difference in spatial performance may be narrowing and provides confidence that females can do as well as their male counterparts. In any review of spatial ability, it is likely that part of the analysis will consider the question of spatial ability as a factor in human intelligence. Thurstone (1950) specified seven factors of intellectual ability and considered three of these to be related to visualization in space. Gardner (1991, as cited in Maier, 1998), who advocated a very detailed hypothesis of human intelligence, considered spatial intelligence to be an indispensable asset in a competitive society. Several psychometric studies of human intelligence have identified spatial ability as a principle factor with general agreement that it is distinct from other factors such as verbal reasoning (Allahyar & Hunt, 2003). If spatial ability is to be regarded as a separate and primary component of human intellectual ability, then this alone demonstrates the importance of spatial ability.

Why Develop a Specific Test of Spatial Ability

Whilst there are a number of spatial tests available and used to measure spatial ability, they are not ideal for novice designers. Reasons for their unsuitability come from a range of issues such as being too restrictive in what they measure, not tapping into various factors of spatial ability, psychometric qualities not being evident and some having a focus on 2D concepts rather than 3D concepts. For the most part, the tests are considered to be generic but too often a test is assumed to be relevant to particular disciplines or applications although this is not necessarily the case. In many situations, a test is of a *single design* and is used for diagnostic purposes. However, because of its single focus, it is not able to identify the reason for the poor performance of many learners. Most tests are of the pencil and paper type and thus are labour intensive requiring detailed organisation, supervision and administration. Obviously such tests fail to take advantage of modern technology and the benefits of online delivery. As well, some tests are biased in favour of particular solution strategies (e.g., holistic versus analytic, referred to earlier) which suit some participants but not others, and often this bias is detrimental to females. There are some tests that are better described as measures of nonverbal ability rather than spatial ability, while others measure only 2D spatial skills without reference to 3D

properties. A true test of spatial ability must include measures of both 2D understanding and 3D understanding, and a measure of spatial ability would have no relevance to a novice designer without a strong emphasis on 3D understanding. Where more than one test type has been used with the same participants, correlation between the tests is often low, which suggests they are not measuring the same underlying factor. This provides argument why a measure of spatial ability for designers should contain a number of test types (subtests). That is, so that different spatial skills can be measured.

Existing Tests of Spatial Ability

At this point, it is important to mention that there are many tests available that measure some aspect of spatial ability in one form or another. In one respect, they can be broadly divided into commercially available tests, and tests that are mainly used in research. For the commercially available, many of the tests have been developed with different age groups in mind, and they typically look to address specific educational questions or to measure specific aptitudes. Most often, they are supplied with test manuals that outline the history of the test, the purpose of the test, the strict requirements for supervising the test, and the norms for the test which are often shown for different populations. For the research focused tests, they generally target particular research questions and particular disciplines or demographics. In almost all cases, for both commercial and research tests, the test items are of a single design (see *Terms and Definitions*) and generally target one specific spatial skill.

There are many spatial tests available, and they can be grouped into various categories and can be used for both educational and research purposes. To provide an overall perspective, a number of these categories are described below:

- Technical Test Batteries. These are a mixture of tests that include measures of visual estimation, spatial perception, the skill to visually compare objects, and the ability to identify two dimensional shapes (SHL Group Ltd http://www.shl.com/).
- Visual Perception and Visual Motor Skills. This category estimates a combination of skills such as hand-eye coordination, space relations, position in space, visual closure and visual motor speed (Pearson http://www.pearsonassessments.com/).
- Nonverbal Reasoning Abilities. This category generally tests the recognition of two dimensional shapes and the ability to manipulate two dimensional objects (Pearson http://www.pearsonassessments.com/).
- Spatial Reasoning Abilities. These tests measure performance on a mixture of cognitive tasks involving the mental manipulation of shapes and patterns (Australian Council for Educational Research Ltd https://shop.acer.edu.au/acer-shop/Home.page).

- Space Relations. Tests in this category measure the ability to mentally manipulate objects in space and require test takers to see the relationship between an unfolded and folded view of a 3D object. (ACER Ltd https://shop.acer.edu.au/acer-shop/Home.page)
- Visual Memory Tests. Tests in this category are criterion measures of visuospatial memory and can be used for screening within a neuropsychological battery. Also suited for documenting changes in neurocognitive abilities over a period of time (Psychological Assessment Resources http://www4.parinc.com/).
- Visual Discrimination Tests. This group of tests measure deficits in perceptual accuracy. A test taker is required to match a test item to a target visual design within an array of four similar designs (Psychological Assessment Resources http://www4.parinc.com/).
- Motor-Free Visual Perception Tests. Originally developed to evaluate visual perception performance in children of all ages with no motor required, the tests are now suited for test takers who may have learning, motor or cognitive disabilities (Western Psychological Services http://portal.wpspublish.com/).
- Measures of Visual-Motor Skills. These tests are intended for children and assess their ability to transcribe geometric shapes accurately by hand based on what they sighted. They provide a complete picture of the test taker's visual-motor coordination skills (ProEd Australia http://www.proedaust.com.au/index.htm).
- Information Processing Skills. This category of tests measures how well a test taker processes information that is presented visually and auditorally. They provide clinicians with a quick and reliable measure of information processing (Slosson Educational Publications, Inc http://www.slosson.com/index.html).

Many spatial tests from a number of these categories provided the starting point for the development of the 3DAT. These tests are reported and referenced throughout the following chapters, and examples are shown in various appendices. A case for a specific test of spatial ability is presented below.

Problems with Current Tests

A particular problem with existing spatial ability tests is that they are restricted in what they can reveal about the spatial understanding of test takers. Most tests are of a single design where each test item within the test is simply a variation of the other test items. While such tests may serve as generic measures of spatial ability, they are not really capable of identifying any more than one specific component of spatial ability. For design-based disciplines, this is not very helpful, especially if the test is intended to be used for diagnostic purposes. It is important to be able to establish where problems lie if the aim of testing is to lead to improvement. A case in point is a test designed to measure the ability to mentally rotate objects in space. This is an important skill

required by designers, but it is not the only skill. For example, designers also require the ability to reason in 3D from a set of 2D drawings. In other circumstances, they need to be able to conceptualise 3D properties looking from several viewing directions. Interestingly, tests concerned with mental rotation are those that are most often chosen by researchers. As important as mental rotation is, a test that only focuses on mental rotation will not identify problems across a spectrum of spatial components.

Branoff (1998) supports this view and emphasizes that separate tests (intended as measures of spatial ability), where each consists of a different single design task, do not all measure the same component of spatial ability. Paivio (1986) extends this view and attests that correlations between tests of spatial ability tend to be low, thus indicating different underlying factors in spatial ability. Branoff adds to this and states that previous research shows that all measures of spatial ability do not tap into the same component of spatial ability. D'Oliveira (2004) refers to this as the *dimensions of spatial ability* and contends that selecting a test of single design because it covers some aspect of spatial ability is not really sufficient. Instead, D'Oliveira argues that it is important to select a test that will measure the dimension in question. For designers, there will be a range of dimensions to cover, and therefore any test of spatial ability will need to be multi-dimensional. Single design tests will not measure all dimensions. Allahyar & Hunt (2003) express doubt in any case about the effectiveness of many existing tests to accurately assess spatial ability and refer to them as surrogate measures. Allahyar & Hunt advocate virtual reality as a new strategy for measuring spatial ability but they do recommend some caution. In defence of researchers who elect to use single design tests, Pellegrino et al. (1984) highlight the disagreement among major studies about the number of separate factors of spatial ability and the difficulty in trying to define them. With this uncertainty in mind, the use of a single design test is understandable.

To gain a true measure of spatial ability, a test should also allow for different *solution strategies* that may be used by individuals. For example, some tests are better suited to *holistic* solution strategies while others are better suited to *analytic* solution strategies. A complicating factor is that the holistic approach is generally favoured by males while the analytic approach is more likely to be favoured by females (Hsi, Linn, & Bell, 1997). For several other tests, a speeded response is important and these will benefit those individuals who prefer holistic approaches. Gluck & Fitting (2003) see individual preferences for particular solution strategies to be problems for assessing spatial performance. Gluck & Fitting also add that simple tests of spatial ability are likely to disadvantage those individuals with a preference for analytic approaches. One final point from Gluck & Fitting is that they see resources and administration associated with test delivery as encouraging the selection of holistic rather than analytic strategy solutions. That is, they consider holistic tests to be easier to manage. While there are other strategies that

could be considered (e.g., pattern-based), Duesbury & O'Neil (1996) maintain that there are two predominant strategies that separate strong and weak performers and they refer to these as *constructive* and *analytical*. The descriptions for these two strategies align with the holistic and analytical strategies reported by Hsi et al. (1997) and Gluick and Fitting (2003).

One last point touched on earlier in this section is that some tests used to measure spatial ability should be classified as nonverbal tests, or at best, tests of 2D spatial ability. Tests that fit this description are the Minnesota Paper Form Test and the Raven's Matrices Test. Examples are shown in Appendix A (Figures A8 and A12 respectively). If either of these tests of single design are used in isolation to measure spatial ability, they will not pick up on the 3D concepts critical to design. In summary, any test intended for designers has to be capable of identifying each factor of spatial ability and to accommodate a number of solution strategies. It also needs the capacity to measure both 2D and 3D properties.

Multiple Subtests Required

Spatial ability is considered to be a measure of several factors which are sometimes called components or elements. Each factor is intended to represent a different spatial skill and any measure of spatial ability should be an aggregate measure of these skills. Thus, an ideal spatial ability test for designers will consist of multiple subtests where each targets a separate factor and results are combined to produce an overall assessment of spatial ability. A subtest in this context is defined as a test of single design containing a number of test items where each is simply a variation of the other test items. That is, each test item within a subtest is of the same type and aims to measure the same factor of spatial ability. Voyer et al.(1995) are resolute on this subject and convincingly propose that spatial ability is not a unitary concept. Instead, they consider spatial ability to be a combination of spatial elements. Their contention is that a separate test is required to assess any one aspect of spatial understanding and agree that a range of tests are really needed to fully evaluate spatial performance. Their position is supported by Blasko et al. (2004) who also believe that a precise measure of spatial ability depends on the assessment of several spatial skills. Collectively, these views endorse the use of multiple subtests to acquire an accurate measure of spatial ability. Gluck & Fitting (2003) are also critical about the use of one test type to measure spatial ability and see the common choice of mental rotation to be inappropriate for many circumstances. Mental rotation tests are often chosen by researchers as a generic test who use them in any number of test environments. However, Gluck & Fitting disagree with this practice and believe that a test should be chosen on the basis of how well it matches the tasks and skills required by particular vocations or activities. They see architecture and engineering as good examples of professions that require other spatial skills besides the ability to mentally rotate objects in space.

In another study that focused on the development of a spatial diagnostic test in the field of apparel design and product development, Workman et al. (1999) identified specific spatial skills required in this industry. The researchers labelled these skills as *spatial products*, *spatial storage* and *spatial thought*. They concluded that the test they developed called the Apparel Spatial Visualization Test (ASVT) was a better measure of these spatial skills than an established test called the Differential Aptitude Test – Space Relations (DATSR). Participants in the Workman et al. study who received specialized training relevant to clothing construction and patternmaking scored significantly higher on the ASVT compared to participants without training. However, for the DATSR, the difference in performance between the trained and untrained groups was not significant. Workman et al. considered the challenge for clothing and textile educators was to determine and assess the spatial skills appropriate to the apparel industry. The researchers were cautious about their findings and referred to them as preliminary and thought further studies with a larger and more diverse sample were needed to validate the instrument. However, the results support the notion of dedicated tests to measure spatial skills specific to certain disciplines.

Gender difference is always an issue. A spatial ability test consisting of multiple subtests has the potential to reduce this bias because subtests can be chosen to balance the learning strategies favoured by either gender. Essentially this comes down to having a balance between tests suited to holistic strategies thought to favour males, and tests suited to analytic strategies thought to favour females. A test platform that is based on test types suited to holistic approaches only (e.g., mental rotation) which is a common scenario, is likely to maintain or increase the gap in the spatial performance thought to exist between males and females. As confirmation, Linn and Petersen (1985) found a significant gender difference on tests of mental rotation. However, they did not find a significant difference on tests that required complex, step by step manipulations of spatial information. These latter tests are likely to suit analytic solution strategies. Of relevance is that mental rotation tests are particularly noted for showing robust gender differences favouring males (e.g., Voyer et al., 1995).

Apart from the potential to identify specific factors of spatial ability, a test consisting of multiple subtests can still be useful in recognising weaknesses in spatial understanding even if the subtests cannot be linked to any particular factor. If there is agreement that spatial ability is a modifiable attribute, then the main value of any spatial test will be its diagnostic potential to influence future planning and training decisions (Pellegrino et al., 1984). Poor performance will be identified where valid measures of spatial factors have been established, but using a mixture of subtests is not wasted if a subtest cannot be linked to a factor. If nothing else, a variety of subtests will tease out learning difficulties and help design educators make informed curriculum decisions to bring about better learning outcomes. A spatial ability test consisting of one test

type only (e.g., mental rotation) will not achieve this on its own. Another consideration is the diagnostic value to the learner. If a well designed and validated instrument can be developed, a significant benefit is the self-help it offers. A test consisting of a number of subtests that is easy to access will provide a remediation service to learners since they are able to independently determine their own strengths and weaknesses and work to improve skills where necessary. With easy access, repeat self-testing is possible to gauge whether or not improvement is occurring. Appropriately, Pellegrino et al. (1984) consider a test capable of producing information about the strengths and weaknesses of individuals could be developed by combining task analysis and psychometric test construction standards. This combination is an overriding feature of this research.

Online and Computer-based

Most existing spatial ability tests are not computer-based and also not available online. This point is acknowledged by Holliday-Darr et al. (2000) who state they could not find existing software to measure spatial skills that was also web-based and available to the wider community. One advantage of this type of test is that it can be developed to operate as a self-directed test without the need for a test administrator. From a research perspective, it means the test can function 24 hours a day seven days per week and offer both financial benefits and time savings to researchers. Further, expediency is possible because data files can be conveniently created, stored, organised and backed up without input from researchers. From an educational perspective, it means learners are able to take advantage and measure their own improvements in spatial performance because the setup provides an unlimited retesting capability. However, a test taker would need to be mindful of the practice effect and ensure a reasonable interval between repeated uses of the instrument. An online test can also be designed in such a way that results are generated automatically to provide immediate feedback. This is consistent with good pedagogical practice.

There are other advantages as well. For example, renewal of an online test is possible. Revisions may be necessary because of what the data reveal, which might mean the replacement of some test items or perhaps some changes to the guiding text or the instructions to users. Computer-based instruments also allow customisation to suit particular test circumstances. In some cases, randomisation may be desirable, in other cases, the number of subtests or the number of test items may need to be reduced. In other circumstances, the data collected may need to vary, or something about the demographics may need to change. A computer-based test can be modified, duplicated or varied to suit any number of research questions or training scenarios. From a research standpoint, an online test creates potential for large sample sizes which should appeal to many researchers. Most researchers struggle to get ideal sample sizes and sometimes fail to capture the ideal demographic. An online test allows recruitment across national and

international institutions, and to the broader community in general. This diversity improves the quality, analysis and outcomes of research.

Somewhat related is the consistency that online tests can offer. Within reason, and under some level of control, the test environment could be the same for all participants regardless of their location or their actual time on task. Importantly, this means that explanations, instructions and practice trials will be identical for all participants. However, this needs some qualification. There is a need to consider a variety of end-user connections and to address timing differences because of network bandwidth. These issues might be accommodated by providing minimum computer specifications and software requirements, or by stipulating certain locations where participation can occur. Another consideration is how research participants and students in educational settings like to work. Most generations today are computer-literate and generally prefer the convenience of modern computer technology and being able to work from home. Online testing fits neatly with these preferences and in return may increase motivation levels and participation rates.

Data collected online using web-based applications is reliable. In a different but relevant research setting, McGraw, Tew, and Williams (2000) report data collected from web-delivered experiments using an online psychological experimental site (http://www.olemiss.edu/ psychexps) to be consistent with those collected under traditional research conditions. McGraw, et. al. argued that their comparisons showed textbook results were possible from web-based experiments, thus indicating web technology to be suitable for conducting certain types of psychology-based experiments. While the researchers conceded the potential for confounds brought on by a mixture of workstations used by participants, they felt that this was easily compensated for by the larger sample size possible from web delivery. By inference, it is reasonable to assume that the consistency reported by McGraw et. al. across two different experimental platforms would extend to the testing of spatial ability online. The findings of McGraw et. al. help consolidate online technology as a reliable method for data collection.

From a slightly different perspective, Strong and Smith (2001) consider the value of many studies linked to spatial visualization to be questionable because of their limited size and scope. They also comment on the variations in the testing methods used by many researchers and suggest this further questions the validity of these studies. Consequently, Strong and Smith advocate the development of an online computer-based test of acceptable standard to make the collection of large data sets possible. The researchers believe that there are literally hundreds of visualization tests in existence and suggest that the online test could be adapted from several of these. Whilst the existence of this number of tests is questionable, the idea has merit and this thesis is largely based on this idea. Strong and Smith conclude by drawing attention to emerging research opportunities in spatial cognition. They consider new technologies should be explored

further and regard merging these with cross-disciplinary approaches to be a good direction for future research.

The 3DAT as an online test may not be the solution for every researcher wanting to measure spatial ability because other forms such as paper tests and practical assessments make important contributions. The 3DAT, however, can operate as both a research instrument and a spatial diagnostic provided researchers and educators appreciate the limitations of each mode.

Need identified

In summary, the need for a specific test of spatial ability for designers is established. There is no known test that consists of multiple subtests purposely developed to evaluate spatial factors which will collectively provide an overall measure of spatial ability. Ideally, this test would be computer-based that operates from a web site and accessible to the broader research community with potential to be customised to meet specialised needs. The instrument should double as a test for researchers and as a diagnostic tool for learners and function as a self-administered instrument without the support of a supervisor or instructor. An online test offers ease of use, reduced administration, increased sample sizes and the versatility to accommodate different requirements of both researchers and learners. The number of subtests should be determined from research and statistical analysis and the test items themselves should be derived from detailed item analysis that investigates their properties. These actions are essentially an application of psychometric test construction standards which would naturally include the fundamental concepts of reliability and validity. These procedures and the rigor they entail could be described as *evaluating the test*. It would be imperative for any spatial ability test under development to address the gender difference reported for many spatial tests. Moreover, this test should accommodate different solution strategies so that it is equally fair to all test takers regardless of their preference. Any intent to develop a spatial ability test would also need to be based on several research initiatives, appropriate analytical procedures and built on previous research. Currently it is difficult to feel confident when comparing results from different studies because different forms of spatial tests have been used. A new test developed to target designers will produce a standardisation that is not really established in spatial cognition. In some cases, there is not even certainty about the consistency of item difficulty for tests used in different studies even though they carry the same name. Standardisation will offer some assurance and makes it possible to compare results with a greater degree of confidence and accuracy. Caplan, MacPherson, and Tobin (1985) confirm that comparison is difficult between studies because there is generally little agreement about the definition of spatial skills. Because this definition is lacking, Caplan et al. concede that it is near impossible to compare studies where different tests have been used since there is no certainty about their compatibility or what aspect of spatial ability is being tested in each.

Spatial ability is an essential attribute required by designers. A purpose-designed test will identify poor spatial performance which in turn will assist design educators to decide about curriculum changes that may be needed. Having the means to accurately measure individual factors of spatial ability is also important although its value is not always appreciated. A test that measures different aspects of spatial ability will help with the development of 3D learning tasks that should be available where any spatial diagnostic test is being used. It is one thing to identify weaknesses, but this should be coupled with learning opportunities to bring about improvement. This view is recognised by Miller and Bertoline (1991) who believe design professionals need to take responsibility for developing tests of spatial ability so that factors that lead to the successful training of engineers can be identified. Spatial understanding is vital to designers, but generic or single design tests will not capture particular weaknesses that may exist. A test that measures the unique spatial competencies required by designers is needed.

Factors of Spatial Ability

Earlier in this chapter, it was established that factors (classifications, elements or components) of spatial ability were thought to exist. Spatial factors are seen as separate spatial skills that collectively provide a measure of spatial ability. However, although there is general agreement about their existence, there is some discord among researchers about the actual number and how they might best be defined (Miller & Bertoline, 1991). For example, Maier (1998) reports five factors, Linn and Petersen (1985) report three factors, and Pellerino et al. (1984) report two factors. Interestingly, Lohman (1979, as cited in D'Oliveira, 2004) reported 10 factors of spatial ability. The uncertainty about the division of spatial ability into factors is the likely outcome of researchers not being able to agree on a comprehensive understanding of spatial ability. D'Oliveira (2004) reasons that there are four areas of conflict that have contributed to this. They are: (i) defining spatial ability, (ii) the number of spatial factors, (iii) the names given to factors, and (iv) the spatial tasks used to measure each of the factors. D'Oliveira also cites Pellegrino and Hunt (1989, 1991) who further argue that spatial skills should be classified into two domains, *static spatial ability*, and *dynamic spatial ability*. They consider most of the factors identified so far to be based on static stimuli and therefore belong to one unique domain, but believe another distinct domain should be defined and called dynamic spatial ability. This domain would recognize the ability to make judgements about moving objects and relative motion. Voyer et al. (1995) suggest that an explanation for not having a universally accepted definition of spatial ability is linked to the large number of spatial tests used in spatial studies. They also suggest, as a second explanation, that this problem may instead be due to a failure to replicate a consistent factor structure when a range of tests are used. In essence, the number and variety of spatial tests are seen to accentuate the problem of identifying spatial factors.

Names and Definitions of Spatial Factors

The central problems in naming and defining spatial factors relate to inconsistency, overlap and identification methods. There are examples where the same names have been given to spatial factors but the descriptions of those labels are different. In other cases, factors have been defined in similar terms but the names allocated to these have not been the same. Adding to the confusion, particular spatial tests thought to be measures of specific spatial factors and used in different studies have not produced similar results. The disagreement is complex and results in no universal agreement about the number of spatial factors or how best they might be described. Pellegrino et al. (1984) proposed two main spatial factors based on a re-analysis of major psychometric studies and called these *spatial relations* and *spatial visualization* ability. A description of each follows:

- Spatial Relations: "This factor seems to tap the ability to engage rapidly and accurately in mental transformation or rotation processes for judgments about the identity of a pair of stimuli" (p. 240).
- Spatial Visualization: "Is defined by tests that are relatively unspeeded and complex. Such tasks frequently require a manipulation in which there is movement among the internal parts of a complex configuration and/or the folding and unfolding of flat patterns" (p. 241).

Mental rotation tasks fit neatly with the first factor, while paper folding tasks are examples of tasks that align with the second factor.

Pellegrino et al. (1984) also provide a reminder that there is little consensus about the number of spatial factors and how they should be defined. The authors see things a little differently to D'Oliveira (2004) and regard the divergence in the literature to be attributable to differences in: factor analytic approaches, sample demographics, test battery compositions and test formats and administrative procedures.

Linn and Petersen (1985) conducted a meta-analysis to investigate gender differences in spatial ability and decided that there were three spatial factors to emerge from their analytical procedures. They classified and labelled these as *spatial perception*, *mental rotation* and *spatial visualization*. A description of each follows:

- Spatial Perception. "Subjects are required to determine spatial relationships with respect to the orientation of their own bodies, in spite of distracting information" (p. 1482).
- Mental Rotation. "The ability to rotate a two or three dimensional figure rapidly and accurately" (p. 1483).
- Spatial Visualization. "Tasks that involve complicated, multistep manipulations of spatially presented information. These tasks may involve the processes required for spatial

perception and mental rotation but are distinguished by the possibility of multiple solution strategies" (p. 1484).

A good example of a test for the first mentioned factor is one where a test taker is required to recognise that the water level remains horizontal after a container of water is tilted. For the second factor, a good test would be mental rotation tasks that require speeded decisions about whether or not two objects are the same or different. Folding tasks are good examples of test items suited to the third spatial factor.

Linn and Petersen (1985) see spatial ability as an important component of general intelligence but concede that its definition requires clarification. They reason this because they believe there is no agreement on the classification of spatial ability measures. They accept, however, that a number of schemes do exist. Linn and Petersen add as well that there is some consensus about spatial ability involving a range of skills. Of note is that Linn and Petersen are researchers who are regularly cited by other researchers. Their 1985 study in particular does appear to have become a benchmark for later studies conducted by other researchers who investigated the potential division of spatial ability into a number of spatial factors.

Maier (1998) acknowledges that spatial ability is a complex concept which is usually divided into three spatial factors by other researchers. However, Maier considers this division is not sufficient to gain a detailed understanding of spatial ability and believes a specification beyond this is required. Maier in his earlier papers (1994, 1996a, as cited in Maier, 1998) proposed five spatial factors and refers to these as the *five elements of spatial intelligence*. Each of these is briefly described below:

- Spatial Perception. "Spatial perception tests require the location of the horizontal or the vertical in spite of distracting information" (p. 70).
- Visualization. "Comprises the ability to visualise a configuration in which there is movement or displacement among (internal) parts of the configuration" (p. 70).
- Mental Rotation. "The ability to rapidly and accurately rotate a 2D or 3D-figure" (p.70).
- Space Relations. "The ability to comprehend the spatial configuration of objects or parts of an object and their relation to each other" (p. 70).
- Spatial Orientation. "Spatial orientation is the ability to orient oneself physically or mentally in space. Therefore, the person's own spatial position is necessarily an essential part of the task" (p. 71).

In brief, test examples for each of these factors in order are: (i) water level tasks, (ii) solids intersected by planes, (iii) classical mental rotation tasks, (iv) identification of objects drawn in different positions, and (v) recognising objects from different viewpoints.

Maier (1998) expresses concern that some researchers have inappropriately placed different tasks under the same labels and thinks this has led to misunderstandings and contradictions. Maier states that the derivation of his five elements of spatial intelligence are based on: "several theories of intelligence, meta-analyses and a number of studies of spatial ability" (p. 70). This is a theoretical list of procedures but there is no evidence of factor analyses being conducted on empirical data. This procedure does seem relevant in this instance.

Other researchers have also contributed to the uncertainty about the identification and naming of spatial factors, and what might be regarded as good descriptions for each. For example, McGee (1979) also put forward convincing evidence for there being two spatial factors. McGee called these *visualization* and *orientation*, and each is described below:

- Spatial Visualization. "Ability to mentally rotate, manipulate and twist two and three dimensional stimulus objects" (p. 896).
- Spatial Orientation. "Involves the comprehension of the arrangements of elements within a visual stimulus pattern and the aptitude to remain unconfused by the changing orientations in which a spatial configuration may be presented" (p. 897).

McGee (1979) conducted a comprehensive review of the spatial ability literature and his findings most likely provided the foundation for work carried out by other researchers who followed. His work is also regularly cited and it still provides a foundation for spatial research conducted today. McGee's treatment of spatial abilities was broad and detailed and dealt with diverse aspects such as psychometrics, genetic, hormonal and neurological influences. In his concluding comments, McGee points to a plethora of factor analytic studies since the 1930s that provide compelling evidence for the presence of at least the two spatial factors he proposed (visualization and orientation). Important to this thesis, McGee also indicates that measures of visualization and orientation have a stronger correlation with success in technical domains than verbal skills, which increases their relevance to applied psychology. However, spatial ability is still seen as a construct that is not easily defined.

In other studies, researchers comment on spatial factors. For example, Voyer et al. (1995) refer to the three spatial factors identified by Linn and Petersen (1985) even though they are critical of some aspects of the analyses they conducted. Voyer et al. considered Linn and Petersen's definition of spatial ability to be vague, and that their classification of various spatial tests (critical to determining spatial factors) was oversimplified. Voyer et al. saw this classification as ignoring critical differences between the tests which potentially neglected patterns of correlations that may have existed. Further comment comes from Tartre (1993) who agreed with McGee (1979) about the existence of two spatial factors (visualization and orientation). Using a rationale based on the mental processes likely to be involved in performing particular tasks,

Tartre defined *visualization* as mental movement of an object, and *orientation* as a mental movement of a viewpoint while an object remains fixed. Her descriptor for visualization disagrees somewhat with that of other researchers, but her descriptor for orientation is not too different to those put forward by others. Tartre gave recognition to *mental rotation* as a possible spatial skill and proposed that visualization should be divided into two categories which he called *mental rotation* and *mental transformation*. She identifies an important spatial skill (mental transformation) that is difficult to ignore. Tartre considered mental rotation was about mentally rotating an entire object in space, while mental transformation was more about mentally rotating part of an object. However, mental rotation and mental transformation are not generally reported as separate spatial factors despite frequent mentions in the literature about the use of tasks to measure mental rotation.

Many researchers acknowledge the existence of spatial factors, but only a small number have produced confirmation data, or data to support the presence of spatial factors not previously identified. Of note is that there is not generally a big difference in the names given to spatial factors, but the descriptors do vary considerably. A case in point is *spatial relations*. There is a noticeable variation for example in the descriptors provided by Pellegrino et al. (1984) and Maier (1998) and quoted earlier in this section. Other issues include the overlap that often occurs in the descriptors, and some spatial tests appear to measure more than just one spatial factor. For the most part, researchers design their studies around the findings of a small group of researchers without actually increasing knowledge about spatial factors. One particular objective of this thesis is to produce evidence (or otherwise) for the presence of the spatial factors most often reported. Another objective is to specifically test for other spatial factors that may also exist.

Current Predicament

Earlier in this section, it was stated that spatial factors were seen as separate spatial skills that collectively produce a measure of spatial ability. Dividing spatial ability into factors is necessary because spatial ability is a construct not accurately measured by any one spatial test. There is no agreement among researchers to a precise definition of spatial ability or about an ideal spatial test to measure spatial ability. This position is emphasized by several researchers (e.g., Blasko et al., 2004; Miller, 1992) who report low correlation values between particular spatial tests. Where correlation values are low, it indicates that the two measures are not pointing to the same underlying factor. Miller (1992) agrees and concludes that the spatial tests he used (*perception* and *mental rotation*) were measuring different aptitudes. Part of the problem with identifying spatial factors is that different procedures may be used. Further, a factor analytic approach does not always produce agreement across different studies according to Voyer et al. (1995). However, this may be caused by problems with the tests or the standard

of the test items rather than the analytic process itself. Carpenter and Just (1986) used an *information processing approach* to try and overcome difficulties with the wide range of tests used in psychometric studies. This approach is based on identifying the mental processes used to solve different spatial tasks rather than attempting to group the tasks into various categories. However, according to Barratt (1953, as cited in Voyer et al., 1995) there is some limitation to this approach because of the potential for test takers to use different strategies to solve similar tasks in the same test. Thus, the risk is that a classification based on different mental processes may result in the same test being placed into different categories (Voyer et al., 1995). McGee (1979) adds another perspective and noted from a detailed literature survey, a contrast in approaches before the 1960s compared to the 1960s and 1970s. The first period focused on factor analytic studies while the latter periods focused on correlational and experimental studies to decide about variance in spatial performance. Perhaps in this current period there is now a swing back to factor analysis because of the certainty this more empirical approach has to offer.

What is not mentioned very often in the literature is the importance that should be given to item analysis. The test items themselves within any spatial test are critical because some can be shown to be *good* items while others can equally be shown to be *bad* items. In psychometric terms, this appraisal is derived from item difficulty, item discrimination, item validity and item reliability, which are objective measures that can be statistically calculated. Part of the problem could be researchers using test items that have different psychometric properties even though they belong to the same test type (e.g., mental rotation). The variation, ineffectiveness and inconsistency of test items can influence the results determined from factor analysis. This in turn makes it difficult to identify unique spatial skills. McGee (1979) sums up this predicament quite accurately and advocates that further investigation into the impact of item difficulty on factor structure should occur.

What clearly needs to be done is to establish the link between the many issues that influence the identification of spatial factors. At the most fundamental level is the disagreement among researchers about a universally acceptable definition for spatial ability. This suggests that spatial ability can be divided into several distinct spatial skills, and the term used to group these is known as *factors of spatial ability*. Other names are spatial factors, spatial elements or spatial components, though s*patial factors* is the name most often used in this thesis. Part of the problem is finding agreement about names and meaningful descriptions for each possible spatial factor. To measure particular spatial factors, a suitable test is required, but this presents a further problem. There are a wide range of spatial tests reported in the literature and used in a variety of studies, however, classifying these tests appears to be a challenge. There is test overlap and test item characteristics to contend with, and the psychometric properties of various spatial tests are not generally stated. This then leads to deciding what might be the most suitable method for

identifying spatial factors because a number of approaches are reported. While factor analytic procedures are mostly favoured, there are issues of inconsistency to consider, probably caused by a range of spatial test concerns. A central focus of this thesis was to find the answers to the many questions surrounding spatial factors. Factor analysis was a significant part of this process because it represented the best chance of identifying specific spatial skills. This in turn helped decide how many subtests were needed to achieve the best possible measure of spatial ability for designers.

Theoretical Foundations of Spatial Ability

At this point, it is appropriate to outline the theoretical foundation for current research activity in cognitive abilities, and in particular, the subset of spatial ability. Cognitive abilities as a paradigm has been neglected to a large extent, and thus has not been well reported by researchers overall. However, John B. Carroll (1916 – 2003), a specialist in psychometrics and educational psychology first started work in this area back in 1939. In several publications, Carroll provided an excellent account of the theoretical and historical background to the current understanding of cognitive abilities, and the division of those abilities into domains such as auditory reception and visual perception. The treatment of cognitive abilities and the spatial ability subset in this thesis is largely based on the studies conducted by Carroll, and the coverage therefore acknowledges the outstanding work done by him. Perhaps Carroll's major findings are reported in Carroll (1993), a comprehensive book in which he details the reanalysis of factor-analytic studies across 460 datasets from the factor-analytic literature. 260 of these datasets revealed some aspect of spatial ability, but of these, 94 datasets provided the most useful information. The reanalysis conducted by Carroll used the method of exploratory factor analysis developed over a 60 year period, and this was a technique he favoured over the more popular confirmatory factor analysis method. Carroll considered that exploratory factor analysis was better suited to early identification of cognitive abilities and the structures within.

Intelligence measures compared to cognitive ability measures, and being able to differentiate between the two are questions that arise. There seems little doubt from a second publication by Carroll (1992) that cognitive abilities are part of a general factor of intelligence (general factor), and he provides examples from tests batteries such as verbal, spatial, quantitative, memory and reasoning abilities to reinforce this. In one respect, revisions and new versions of intelligence tests have continually appeared, but they have been introduced with little consideration to new development. However, in contrast, Carroll also reports that the debate over intelligence and cognitive abilities has motivated researchers in psychometrics and cognitive psychology to further investigate cognitive abilities including a possible separation from intelligence. To support an argument in favour of some level of difference, Carroll points out that the variance in cognitive ability performance is not totally explained by a general factor. He reports that many

factorial studies claim that the general factor contributes the largest proportion of variance, say up to 90% of a test battery. However, Carroll takes issue with this statistic and argues that it is misleading. He asserts that the amount of variance contributed by the general factor to a specific measurement is generally a lot less, and probably not greater than 50% on average. This leaves the likelihood that ability factors outside the general factor of intelligence are contributing to the total variance in particular variables when assessed by factor analysis. Establishing a measure of an ability outside the general factor requires the identification of the types of tasks, and what characteristics of those tasks should be involved in that measurement. Other properties such as reliability, validity and high homogeneity in accordance with psychometric standards are also required (Carroll, 1992).

Carroll (1992) indicated that the paradigm of cognitive abilities has not been central to science, and therefore has not been pursued with any enthusiasm from researchers. Carroll considered that supposedly inflexible controversies stem from measures of aptitude and achievement, the factor analytic procedure, and the evaluation of genetic and environmental influences. Carroll added that a significant proportion of public and professional opinion believed that individual differences in cognitive abilities are insubstantial, or that such differences can be explained totally in terms of environmental factors. Carroll (1993) defined *cognitive abilities* as those abilities that involve some form of mental functioning and reasoning, both in the understanding and the performance of a task, as opposed to say manual abilities that might require no more than physical strength and endurance. He also considered that cognitive abilities can be separated into categories which he called domains, and he identified one of these domains as *visual perception*. Importantly, Carroll recognised that visual perception could be described by many other names such as spatial ability, visualization and spatial cognition. He regarded visual perception and other spatial skills to be part of the ability to search the visual field, to decide about shapes and positions of objects in space, to form mental representations of those objects, and then be able to mentally manipulate those representations. *Visual perception* is the term Carroll decided upon to encompass these and all other possible spatial abilities even though he considered this term may not be ideally descriptive. The term *spatial ability* is preferred for this thesis, but both terms tend to be used to mean the same thing.

The history of spatial ability research was conveniently summarised by Eliot and Smith (1983) as cited by Carroll (1993). They described three phases, and the first of these (1904 – 1938) was characterised by a period when researchers assessed evidence in support, or not in support of spatial factors beyond a general factor of intelligence. The second phase (1938 – 1961) was marked by a period when researchers attempted to establish the degree to which spatial factors were different to each other. The third phase (1961 – 1982) was a period when researchers tried to place spatial ability within the complex arrangement of other cognitive abilities, and also a

period of investigation when researchers attempted to identify sources of variance which influenced performance on measures of spatial ability.

Essentially, these phases suggest that research activity that targeted spatial ability was not intense throughout most of these periods, and it mostly focused instead on spatial ability as part of a general factor of intelligence. At the time of writing, Carroll (1992) believed that most of the evidence in support of growth, development, change and the decline of cognitive abilities (from aging) had been derived from investigations using global measures, for example, verbal and nonverbal scales from the Wechsler measures. Carroll added that there was not sufficient knowledge about which abilities could be modified, or which abilities were less able to be modified. Carroll put forward several other points of view. First, he pointed to a large body of abilities (e.g., creative and spatial) about which very little was known in terms of genetic and environmental influences. Second, he insisted that cognitive ability tests must be shaped by what skills they tap into, and the types of practical tasks they identify. Third, Carroll conceded that there was more to be done towards test design, and further conceded that insufficient knowledge about the range and composition of cognitive abilities was a barrier to expanding a satisfactory theory of cognitive abilities. This included the preparation of comprehensive tests to measure different abilities.

In more recent years, spatial ability has received a good deal of attention from psychometric investigation. Since Spearman's announcement of a general factor of intelligence in the 1920s, a number of specialised abilities as part of the visual perception domain, and largely separated from a general factor of intelligence, came to be identified. The problem, however, was that the research produced contradictory and confusing evidence about what visual abilities existed and how they should be described and measured. Carroll (1993) reported that many of these abilities, however, fitted neatly with a description of spatial ability because they dealt with how individuals perceived objects in space, and how they mentally changed their own positions to adopt a mentally different viewing direction.

In an attempt to identify separate factors of spatial ability, Carroll (1993) used informed subjectivity to decide upon five factors of spatial ability. He then went through the datasets he compiled and matched any mention of spatial factors to the five factors he had selected. It was seen as a starting point, and the idea was to establish groups of skills so that definitions of factors could be more clearly defined. Of the datasets examined, 94 provided evidence of at least two factors that fitted the categories that Carroll settled on. Very few datasets revealed more than two of the five assumed factors listed by Carroll which he called *visualisation*, *spatial relations*, *closure speed*, *closure flexibility* and *perceptual speed*. Carroll concluded that there was evidence of separate factors in the spatial ability domain, but conceded that more research was required to establish a better understanding of these factors. A final comment from Carroll

pointed out that there was an increasing amount of knowledge about individual differences in spatial ability, but there were still gaps overall, and measurement procedures still needed some refinement.

Carroll's focus on cognitive abilities in general, and spatial ability in particular, revealed a number of issues that reflect favourably on the decision to develop the 3DAT. In brief, there was evidence of uncertainty about an adequate definition of spatial ability and about the number of spatial factors that comprise spatial ability. Also, test results did not load consistently on factors when subjected to exploratory factor analysis, and furthermore, tests with the same names were often not actually similar in the skills they measured, the stimulus they used, or in what test takers were required to do. As Carroll (1993) explained, it was often necessary during his analysis to exercise careful attention to the type of test tasks, and a lot less attention to the names given by researchers to those tasks. Also to emerge was a serious concern among researchers about the speeded or nonspeeded conditions applied to the running of spatial ability tests. Lohman (1979) cited by Carroll (1993) commented critically about the neglect of psychometricians towards the speed-power issue. Carroll added, that although the time limit on a spatial test may be known, just being aware of this did not allow any judgment about how speeded a test actually was. One final concern was the difficulty in factorial classification. This occurs because even the most straightforward of spatial tests are complex, and therefore involve a mixture of skills such as encoding of spatial structures, mental manipulation of those structures, comparative decisions and timely responses. This may mean that in test development, it would be difficult to emphasize individual differences in one skill set, while at the same time minimizing individual differences in other skill sets (Carroll, 1993).

The theoretical and historical framework presented here for cognitive abilities is intended to emphasize that issues critical to visual perception have emerged from studies concerned with the structure of intelligence. Also, and most importantly, that this framework is the foundation for the research presented in this thesis. The work on intelligence including visual perception dates back to research conducted by researchers like Spearman and Thurstone in the early part of the 20th century. A focus on cognitive abilities has not generally been a large area of attention, but perhaps the best summation of this paradigm was provided by John B Carroll in the several publications previously mentioned. Carroll was motivated by a concern for the dimensional analysis of cognitive abilities and wanted to present a personal perspective on the current state of the paradigm, what directions it should take, and what the research and development emphasis should be. To do this, he embarked on a long and tedious review of the factor-analytic literature which involved reanalysing many datasets collected over several decades using mainly exploratory factor analysis. Some advances have been made in this area since Carroll's publications, but evidence from the literature presented in earlier sections of this

chapter suggest that many of the problems identified by Carroll still exist today. To clearly demonstrate the state of the paradigm when Carroll released his publications, Carroll announced that, after his research, the best he could come up with was an incomplete and imperfect understanding of all domains of cognitive abilities because of large variations in variables which were not always well refined or differentiated. There is no doubt that the domain of visual perception has entered a period of increased activity, but there is still a great deal of research and investigation still to be done.

Rationale and Hypotheses

Rationale

A brief recap of the current situation reveals uncertainty about how spatial ability should be defined and what constitutes a valid and reliable measure of spatial ability for designers. Part of the problem is deciding whether spatial ability consists of different spatial skills, and if so, how many might there be and how can they be identified. In effect, the rationale for conducting this research can be summarised into several main points. These are:

- A dedicated test is needed for novice designers because of the direct relevance of spatial ability to them. Many existing tests are considered generic and don't particularly target designers or measure the exact spatial skills they require. A specific test that measures spatial ability in a design context is required.
- It is paramount to develop a test in accordance with psychometric test construction standards. Conforming to these standards is essential to demonstrate a quality product, particularly since evidence of item analysis is not generally reported in the literature. A test that has been developed and evaluated against psychometric test standards is required.
- It is quite likely that spatial ability can be divided into a number of spatial skills known as spatial factors. Knowing about these factors helps define spatial ability which allows researchers to gain a better understanding of this construct. There is a need to clearly identify, name and describe spatial factors and thus establish a complete measure of spatial ability. Accurate profiling of spatial ability is only possible if a comprehensive test is developed and made available to researchers.
- Gender difference may be induced by the design of tests currently used to measure spatial performance. The investigation of gender differences is necessary, and the development of a gender-neutral test is imperative.
- Most existing spatial ability tests are pencil and paper type tests. While some others may be computer-based, few if any are online and potentially available to many users. Most existing tests are singular in design such that test items are similar. Some are 2D only, suggesting that they are better classified as nonverbal tests. 2D is an element of spatial

ability when it is linked in some way to 3D understanding. What is desired is a spatial test that is of a mixture of appropriate test items that take advantage of modern technology.

- The establishment of a scientifically evaluated test of spatial ability for designers will assist in the detection of weaknesses in spatial performance and show where improvement is possible. Such a test requires a number of different subtests to highlight specific learning problems. This will guide the development of 3D learning tasks to improve spatial ability. A broad test of spatial ability that identifies learning difficulties and informs design educators is required.
- There is merit in developing a test of spatial ability that can easily be accessed by students in design programs at any location. This would provide opportunities for ongoing self-assessment to gauge improvement. To achieve this, a computer-based test with online access is required.
- The importance of spatial ability is overlooked and performance tends to be taken for granted. Attention needs to be drawn to the relevance of spatial ability. A test developed to psychometric standards will help create a focus on this critical attribute for designers.

This research advocates that spatial ability is best measured using a range of carefully selected spatial subtests. What is needed is a comprehensive test that has been validated against psychometric principles that measures skill subsets particularly relevant to designers. The test needs to be a diagnostic tool to assist educators and available online to allow maximum access.

Hypotheses

Expressed in terms of a research question, this research essentially focused on whether it was possible to develop a specific test of spatial ability for designers. As a result, the hypotheses for this research were:

- It is possible to develop a test of spatial ability for design-based disciplines.
- The establishment of a test of spatial ability that satisfies psychometric test development standards is achievable.
- A test of spatial ability that consists of a range of specialized subtests can be developed.
- The existence or otherwise of distinct components of spatial ability called spatial factors can be determined.
- A consistent gender difference in spatial performance that favours males will be found.
- A test that identifies areas of poor spatial performance can be developed.

Objectives that align with the hypotheses are given for each of the nine studies that follow.

3DAT Development Overview

Central to this research was the development of a test of spatial ability for designers known as the 3D Ability Test and referred to as the 3DAT. Its development occurred across nine different studies representing various stages of investigation. These stages were grouped into three phases called *initial development, transitional development and final development*. Various subtest possibilities were considered and aspects such as practice trials, instructions, number of test items and time on tasks were investigated. Part of the development included comparisons between lab studies and online studies to determine if the controlled conditions possible in a laboratory setting could be replicated in an online setting. A consideration was also whether there was a viable alternative to instructions and guidance normally delivered by research supervisors. Throughout the development process, psychometric test construction standards such as the various forms of reliability and validity were applied. This included item analysis to help decide about good or bad test items in terms of item discrimination and item difficulty. To achieve this, a well-established statistical procedure known as the *Classical Test Theory* was used with results cross-referenced against an emerging procedure known as the *Item Response Theory*. Factor analytic procedures were undertaken to help decide about spatial factors which in effect were expected to define spatial ability. Part of the development process included interviews with subject matter experts and surveys given to research participants to help decide about subtests and test items that could become part of the final version of the 3DAT. Foremost in any investigation was examining gender issues and deciding about various subtests as accurate measures of the skills appropriate to designers. From the outset, a large sample size was considered necessary at least in the final stages so that reliable item and factor analyses could be performed. A guide for item analysis is five participants for every test item under consideration (Tabachnick & Fidell, 1996). With 72 test items being investigated towards the end of the research, it meant that a sample in the order of 360 was needed. It was always assumed that this would be difficult task to achieve, but it became possible through the cooperation of design departments across several university campuses.

The 3DAT is also founded on established test construction principles. Cohen, Swerdlik, and Smith (1992) describes these as: *test conceptualization*, *test construction*, *test tryout*, *item analysis* and *test revision*. The 3DAT was conceptualized from industry experience of the writer where it was obvious in the manufacturing sector how important it was for tradespeople to be able to visualise what designers had in mind when they produced their technical drawings. The *initial* phase of development was essentially concerned with scoping initiatives and exploratory investigations where studies were conducted with nondesigners to assess general spatial performance. These studies occurred across both laboratory and online conditions though the online condition was somewhat primitive compared to the final outcome. This provided

important considerations for the later stages of development. The *transitional* phase focused mainly on subtests thought to be specific to designers and consisted of trials, an interview schedule and a comparison between design and nondesign groups as one form of validity testing. The *final* phase in the sequence and perhaps the most critical involved testing and analysis with a large sample to finally decide about subtests, test items, spatial factors and psychometric properties. All three phases were particularly concerned with the various forms of validity, reliability and correlation.

The SPSS statistical software package was used for almost all of the statistical procedures conducted in this research. In the earlier studies, SuperLab Pro (a stimulus presentation software package) was used to run the 3DAT, and it produced DAT files that could be converted to XLS files and imported into SPSS. The online version of the 3DAT saved CSV files to a server and these were merged into one file, saved as an XLS file and also imported into SPSS. Minitab software was deployed for one early study to carry out a statistical analysis, and the JMP statistical software package was used for the item response theory analysis. Excel was used to produce the classical test theory analysis.

The final version of the 3DAT is an online spatial ability test for designers that consists of five subtests and 20 test items that produces a performance summary and allows access to incorrect test items for users to review if they choose. The subtests are presented to the test taker in random order and the test items within each are randomised as well. A number of subtests are based on previous psychological research and modified according to findings, while others have been created to measure specific concepts thought to be relevant to designers but not captured in other test types. The test items were produced by a 3D modeller who worked according to the direction, scrutiny and evaluation provided by the writer. They are novel in design and intentionally do not resemble common objects to avoid possible cues from life experiences. Instructions to participants are facilitated by video movies linked to each subtest and practice trials are available to ensure participants understand the requirements of each subtest. Built into the 3DAT is an email option that lets participants email a summary of their performance to themselves if they wish. A research version of the 3DAT is available which captures both demographic information and performance data. There is also an education version that can be used by educators as a spatial diagnostic test, or it can be used by learners as a repeatable self-assessment tool if required. There are also seven subtests that were part of the final analysis sitting in readiness for future applications. These were rejected for mostly not satisfying a psychometric standard. However, some, if not all are considered relevant to designers, but they need reworking and retesting before they become part of the 3DAT. The 3DAT measures choice accuracy and response time, and sample test items are shown in various appendices referred to in the chapters that follow. For response time (RT), there are no issues with data collection, or

inaccuracies or inconsistencies in measurement caused by data transfer problems between the test taker's workstation and the server that hosts the 3DAT. Essentially start times and finish times for each test item shown on the test taker's workstation are transferred to the host server and the difference is calculated locally to derive a response time in milliseconds. Also, it does not matter how long it takes for test items to load onto the test taker's workstation because the *start time* is not recorded until all images are loaded. The *finish time* is recorded as the time the test taker responds to the test item. Thus, there are no network delays to account for, and slower bandwidths and CPUs are not factors.

CHAPTER 2

INITIAL DEVELOPMENT

Overview

This chapter reports on the first stages in the development of the 3DAT which were regarded as exploratory studies to test possibilities for the 3DAT. The main focus was mainly on a range of subtests that were originally considered possible for the 3DAT and the viability of the web as a suitable platform for testing spatial ability. This phase of development also included early probes into several psychometric properties considered relevant, and early probes into whether gender difference in spatial performance would be a factor. The primary issues of subtest suitability, web viability and item analysis are addressed in study 1, while the secondary issues of validity and gender matters are covered in study 2 and study 3 respectively. Thus, the broad aim of this initial development was to gauge general spatial performance in the online condition and to decide about the feasibility of subtests being considered for the 3DAT. Each of the studies are treated separately.

Study 1 Lab and Online Conditions Compared

This study compared the results of a lab experiment and a parallel online experiment using two different samples to decide whether the conditions that could be controlled in a lab environment could be replicated in a web or online environment. The research reported in this section relates to a published paper (Sutton et al., 2007) authored by the writer and two PhD supervisors. This study was seen as an essential starting point to decide about the practicality of the web as a suitable platform for collecting valid and reliable data. There was an expectation that this comparison would expose problems that would need to be addressed if the 3DAT were to become a permanent and respected online test. Hence, the objectives of study 1 were to:

- investigate the web as a viable alternative to a lab setting for testing spatial ability,
- examine the appropriateness of early-identified subtests for the 3DAT, and
- conduct procedures to evaluate several psychometric qualities.

The initial 3DAT consisted of 6 subtests although three of these were considered two-way subtests. By this it is meant that the logical sequence in some subtests can be reversed. An example is the *paper fold task* where the task taker could be required to identify the folded position, or on the other hand, to identify the unfolded position. The initial 3DAT addressed all of the skills emphasised in traditional training, such as understanding of different types of projections, the concept of true length, folding and unfolding and the properties of coordinate systems. 3DAT was delivered by a computer, enabling measurement of both accuracy and speed. Speed is particularly important to the full development of expertise, as the final stage of

skill acquisition is marked by a transition from mastery (a relatively error free performance but slow and deliberate), into effortless and fast performance, as exemplified by language fluency in an experienced native speaker (e.g., Fitts 1964; Speelman & Kirsner, 2005). Studies of the development of fluency in cognitive choice tasks show that participants are able to reduce response time (RT) markedly in the transition from mastery to fluency while maintaining a high and constant, or only slightly increasing, level of accuracy (e.g., Heathcote, Brown & Mewhort, 2000). Hence the measurement of both accuracy and RT enables 3DAT to remain sensitive to improvements throughout all stages of skill acquisition.

Computer delivery allows the 3DAT to be used in both laboratory and web-based settings. Laboratory studies can be problematic both because of the resources required to obtain a sample sufficient for statistical techniques to test psychometric properties (e.g., factor analysis), and to some extent, because it is difficult to sample a demographic representative of the general community. Web-based research provides a possible resolution to these problems. Steyvers and Malmberg (2003) and Birnbaum (2004) provide evidence that the reliability and validity of data from web studies compare favourably with data collected from parallel laboratory studies. This study provides a comparison of 3DAT performance in parallel laboratory and web based studies in order to compare their reliabilities and to validate the web delivery method.

Initial 3DAT Described

Blasko et al. (2004) emphasise the need to use multiple spatial cognitive tasks to assess 3D understanding. They report results from mental rotation and correct fold tasks similar to the initial 3DAT using a web-based presentation (http://viz.bd.psu.edu/viz/). The initial 3DAT consisted of 89 test items divided into 6 subtests. Five subtests were based on previous psychological research, including the *correct fold* and *mental rotation* tasks used by Blasko et al. Strictly speaking, one of these (*dot coordinate*), has not been used as a measure of spatial ability for designers. Instead, it was identified as suitable subtest from a battery of tests used in the selection of medical students. The sixth subtest is based on the idea of true length, an important concept in technical drawing. An edge of an object can be represented in any view of the object but its true length is not always seen; only edges parallel to a projection plane have their true length in a projection. The items are varied in form and most are novel in design. The items are constituted of straight lines and flat planes. 3D understanding for curved objects emerges later in the further development of the 3DAT. They were created using computer assisted design software (AutoCAD) and saved in bitmap and GIF formats for the lab and web studies respectively. Image resolutions were comparable and the different formats were required to suit the software used for the two studies. A description of each of the 6 subtests follows.

2D – 3D Recognition

Objects are presented as orthographic and isometric projections. Participants select which of two alternatives of one type matched a standard of the other type (Bertoline & Miller, 1990; Cooper, 1990). Subtests use either (A) an orthographic standard or (B) an isometric standard, with 8 and 9 items respectively (see Appendix A, Figure A1).

Correct Fold

Objects are presented as an isometric projection or as an unfolded view. Participants select which of two alternatives of one type matched a standard of the other type (cf. Blasko et al., 2004). Subtests use either (A) an isometric standard or (B) an unfolded standard, with 5 items for each (see Appendix A, Figure A2).

True Length Recognition

Objects are presented as isometric and orthographic projections. In one subtest, participants decide which view in a set of orthographic projections shows the true length of a labelled edge in an isometric projection (True Length Recognition A). In a second subtest, participants decide which of three isometric projections shows the true length of a labelled edge in a set of orthographic projections (True length Recognition B). There are 13 and 9 items respectively in the subtests (see Appendix A, Figure A3).

Mental Rotation

Participants decide if a rotated isometric projection of an object matches the isometric projection of a standard or its mirror image (Metzler & Shepard, 1988). The object on the left is always in the same position and is the referent. The object on the right can be the same or the mirror image of the referent and its orientation in the XY plane can be different. There are five matching and five mismatching items (see Appendix A, Figure A4).

Possible/ Impossible Structures

Participants decide if an isometric projection can represent a 3D object (Schacter & Cooper, 1990). The objects can be one of two types. The first (possible) is one where the projection can reasonably represent a true object. The second (impossible) displays some visual feature that could not reasonably represent an aspect of a true object. There are 6 and 13 items of each type respectively (see Appendix A, Figure A5).

Dot Coordinate

Participants are shown an isometric projection of a 3D Cartesian coordinate system and a text description of the position of a point in that system. From four orthogonal projections, participants choose the projection that corresponds to the description (Bore & Munro, 2002). There are 11 items (see Appendix A, Figure A6).

Methodology

LABORATORY STUDY

Participants worked through the 89 items organised as a set of computer-controlled activities. The study was created in SuperLab 2.01, an experimental software package used for psychological research. Participants had control over the initiation of each subtest, with each subtest preceded by instructions containing an example and advice about how to respond. Practice trials for all subtests were conducted before the actual study to allow familiarisation with the subtests and response procedures. The setup was explained by the writer and participants could ask questions. No feedback was given during practice or testing. The study was conducted with groups of approximately five participants who were taken through the practice trials to explain what was required, but no strategies to determine correct answers were discussed. Instructions emphasised the concept of true length and the relationship between an isometric drawing and orthographic projections.

Breaks were built into the study to safeguard against fatigue and they occurred at the start of each of the 9 subtests. Participants controlled the duration of the breaks by initiating the start of each subtest after reading through the instructions and studying the example provided. Excluding breaks, the study took about 60 minutes to complete. The subtests were presented in the same order as their descriptions shown in the last section, but the order of items within each subtest was randomised for each participant. Participants entered their responses using a six button response pad.

Participant eligibility criteria were: (i) 18 years of age or older, and (ii) no self-reported prior technical drawing experience. These criteria were made explicit in recruitment advertising, and no participants applied to do the experiment who did not meet them. The sample of 41 participants (32 females and 9 males) was drawn from a participant pool of psychology students in first year university classes who received course credit for participation.

WEB-BASED STUDY

The web study replicated the laboratory study as closely as possible, with differences explained below. It was developed to utilise ColdFusion MX using Mach – II methodology. As a measure to protect against poor web experimental design, the implementation was checked against the 16 standards suggested by Reips (2002a). Because web participants had to work independently, (laboratory participants' questions could be answered by the supervisor), additional explanations were considered necessary. Thus, detailed information was provided to explain the relationship between orthographic projection and isometric drawings, the experimental design and the concept of true length. Hence, participation in the web study was more demanding in terms of reading and understanding the test requirements than for participants in the laboratory study. As a consequence, and because of the additional demographic information collected, the

web study took longer (75 minutes on average) to complete. Participants recorded their responses by mouse-clicking a number using the same numbering system as for response buttons used in the laboratory study (e.g., 2, 3 or 4). The numbers were displayed on the screen but separated from the image choices. RT was measured on the client side and managed through the web browser. From the start of each image being displayed, a javascript counter recorded the time until a response was received (except for a short delay intentionally built in to accommodate image loading time). The time taken (RT) was logged with the response of the participant.

Krantz (2001) identifies stimuli as a potential confound for comparison of the laboratory and web results and emphasizes the need for calibration. Krantz provides reasons for calibration such as differences in monitor displays, image stability and inconsistency of colour across monitors. However, the laboratory and web formats of 3DAT differed only minimally because a sophisticated web interface was used that was equivalent to the laboratory format in most aspects. The interfaces were near identical with the exception of text position, and the laboratory study required the use of a response pad, while participants in the web study needed to mouse click on numbers. The delivery of the images included a time delay before each image displayed to allow for hardware differences (also allowed for in the laboratory study). The average image file size was only 8kb and the images were simple line figures without colour or rendering. One focus of the comparison was on completing the study in a quiet controlled environment versus completing the study over the web using a virtually identical interface.

To identify the profile of the web participants, a demographic section in the study asked about gender, country of residence, ethnicity and vocation. This section also asked participants if they were aged 18 or older and if they had previous technical drawing experience. Participants younger than 18 or those with technical drawing experience were excluded from analyses, although they were able to complete the experiment. When the test was completed, participants were provided with a score out of 89. The results from the excluded group were not recorded and the final sample size of 30 consisted of 23 females and 7 males. Of the 260 eligible participants who entered the demographics section of the study, 80 made a start on the testing phase and 48 completed it, with 18 more being excluded because of a technical problem reported later in this paper. No participants were excluded on any other basis. The final sample of 30 consisted of participants from several countries, mostly from the USA (67%), and from a range of vocations such as academic, service, professional and military, with the majority (60%) indicating that they were students. Recruitment of web participants was conducted through the Psychological Research on the Net site (http://psych.hanover.edu/Research/exponnet.html), and special psychology interest groups like those suggested by Birnbaum (2004) and Reips (2002a). Recruitment from interest groups was carried out by advertising on their web sites. Web

participants could nominate for a prize draw, with the prize being a Aus$40 gift voucher. The web version of the test can be viewed by linking to the web site at: http://webapps.newcastle.edu.au /2d3dsurvey/index.cfm. Participants were not able to proceed to the actual study without first completing the demographics section and the practice trials.

Results

Reliability was tested by comparing Cronbach alpha coefficients and consistency by comparing mean accuracy and mean response time for correct answers between our laboratory and web based samples. Reliability results are reported in Table 1. Generally, both web-based and laboratory subtest scores produced acceptable alpha reliability coefficients. Psychometric standards define acceptable coefficients as greater than 0.7 with values above 0.8 considered highly acceptable. Values closer to zero indicate poor consistency across items. The low alpha coefficients found for 2D-3D Recognition – A (1A) and Correct Fold – B (2B) in the laboratory sample were not found in the web sample. Noteworthy is that reliability is consistently greater for the web study across all subtests for accuracy and for all but one subtest for RT.

Table 1
Comparison of Cronbach Alpha Reliability Coefficients for Parallel Web and Laboratory Studies

	Web-based Study		Laboratory Study	
Subtests	Accuracy	RT	Accuracy	RT
2D-3D Recognition – A (1A)	.68	.76	.09	.69
2D-3D Recognition – B (1B)	.79	.74	.48	.82
Correct Fold – A (2A)	.42	.69	.38	.56
Correct Fold – B (2B)	.57	.78	-.02	.51
True Length Recognition – A (3A)	.89	.76	.80	.68
True Length Recognition – B (3B)	.80	.70	.54	.50
Mental Rotation (4)	.62	.83	.61	.60
Poss/Impossible Structures (5)	.83	.88	.74	.84
Dot Coordinate (6)	.92	.94	.82	.74

Note. Sample (web = 30, lab = 41). Alpha overall not reported for 3DAT. Refer Chapter 5 for explanation.

Correlations between the accuracy scores for each subtest for the web-based sample and the lab-based sample are shown in Table 2. The high internal reliability of the scores is reflected in the high correlations generally found between subtest accuracy scores. The exceptions, as would be expected given the Alpha reliability coefficients, were the 2D-3D Recognition – A (IA) and Correct Fold – B (2B) subtests for the lab-based sample, where correlations between these subtests and all other subtests of the tests were weak and mostly did not reach significance. This was not found in the web-based sample where strong correlations between the 2D-3D Recognition – A (1A) and Correct Fold – B (2B) subtests and all other subtests were observed. Of additional interest were the moderate to strong correlations for the Dot Coordinate (6) subtest. This particular subtest requires considerably more reading of instructions while also

being the most difficult subtest of the six subtests presented in the instrument (see Figure 1 for percentage correct by subtest). The correlations found suggest that the Dot Coordinate subtest is measuring the same skill as the remaining subtests. However, this is not necessarily a good outcome for a test being developed to identify different factors of spatial ability. Where there is a high correlation between subtests, it indicates that they are pointing to the same underlying factor, which is not ideal except to identify duplication. Such results are not surprising in the initial stages of test development, and thus serve a worthwhile purpose. The test developer is hence made aware that subtests require different characteristics if they are to measure different spatial skills. A little more is said about this in later studies.

Table 2
Correlations Between Subtest Accuracy Scores For Web Sample and Lab Sample.

	1A	1B	2A	2B	3A	3B	4	5	6
				Web-Based Study					
1B	.77**								
2A	.58**	.50**							
2B	.48**	.36*	.51**						
3A	.63**	.74**	.55**	.51**					
3B	.62**	.66**	.53**	.54**	.83**				
4	.53**	.54**	.33	.62**	.44*	.43*			
5	.59**	.56**	.60**	.70**	.64**	.71**	.66**		
6	.48**	.54**	.51**	.44**	.75**	.63**	.28	.53**	
				Lab-Based Study					
1B	.28*								
2A	-.13	.23							
2B	-.03	.19	.19						
3A	.15	.56**	.40**	.16					
3B	.13	.66**	.47**	.21	.65**				
4	-.08	.42**	.41**	.36*	.55**	.56**			
5	-.00	.54**	.38*	.25	.73**	.70**	.49**		
6	.21	.36*	.46**	.38*	.43**	.56**	.35*	.41**	

Note. *p is significant at the .05 level. ** p is significant at the .01 level.

On occasions, *item – total* correlations are reported in correlation studies, and for this research, the equivalent would be *subtest – 3DAT* correlations. However, correlation coefficients (r) for these relationships are not shown in any correlation table presented in this thesis. This decision is based on there being a phenomenon where a researcher can expect the correlation to be high. This occurs because the total score (the 3DAT in this case) is made up of its component scores (e.g., subtest1, subtest2, subtest3). That is, each subtest contributes to the 3DAT overall, and therefore each component will correlate strongly with the total score because the component itself is part of the total score (i.e., 3DAT). A condition, however, is that the subtests satisfy psychometric standards. There may be differences in the strength of correlations for the

different combinations (e.g., *subtest1 – 3DAT* versus *subtest2 – 3DAT*), but these simply provide the level of contribution each component makes to the total score overall.

In order to examine the relative difficulty of each subtest and to compare the difficulties between web and lab-based samples, the mean percentage of correct responses given for each subtest were calculated and plotted as shown in Figure 1 (standard error of the means are also shown). Lab-based participants achieved a higher mean percent correct across all nine subtests of the 3DAT compared to the web participants, although these differences reached significance in only four of the subtests: 2D-3D Recognition – B (1B), Correct Fold – A (2A), Mental Rotation (4) and Possible/Impossible Structures (5). For both web and lab samples the lowest mean percentage correct was for the Dot Coordinate subtest. For the web sample, a One-Way Analysis of Variance (ANOVA) of the percentage correct means of the nine subtests was significant, $F(8,261) = 8.5$, $p < .001$, with Tukey's pairwise comparisons (Family Error rate, $p = .05$) showing that the percentage correct mean for Dot Coordinate was significantly lower than all other subtests with the exception of the Correct Fold – A (2A) subtest. In the lab sample, significant differences were also found, $F(8,360) = 16.9$, $p < .001$, with the Dot Coordinate percent correct mean being significantly lower than all other subtests percent correct means.

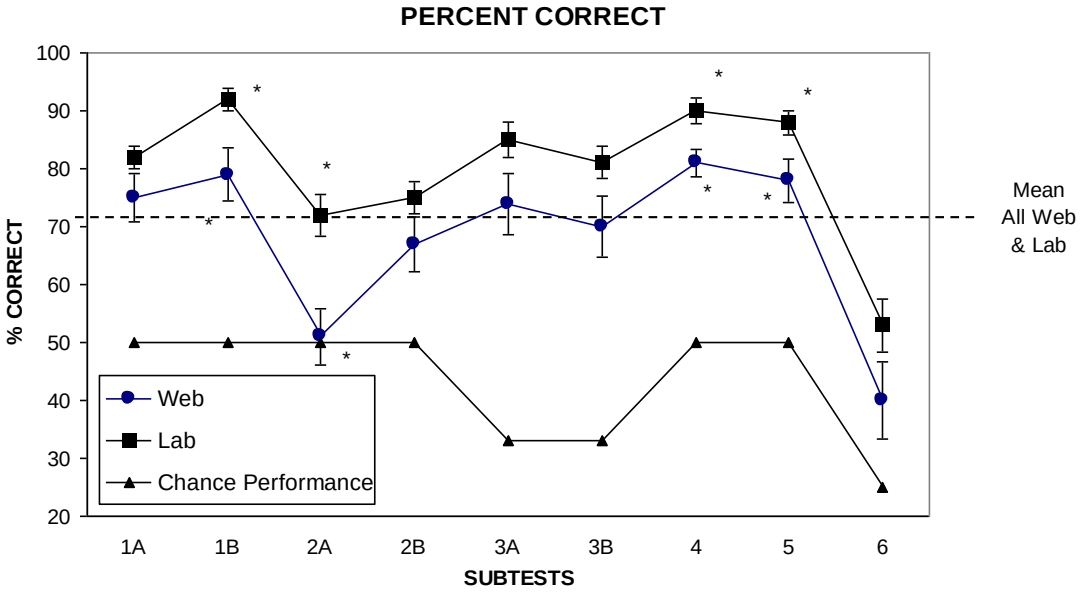


Figure 1. Plot of mean correct responses and standard error per subtest for Web and Lab samples. The line labelled *chance* indicates guessing performance given the number of test alternatives. The interpretation of laboratory results for Tests 1A and 2B should be qualified by their low reliability (see table 1). Significant difference ($p < .05$).

The time taken to give a correct response was recorded for both web and lab-based samples and a mean time (in seconds) and the standard error of means calculated for each subtest. As shown in Figure 2, the pattern of mean correct RTs was highly similar for both web and lab samples though web participants took longer than laboratory participants for 6 of the 9 subtests. These differences were significant for two subtests: Mental Rotation (4) and Possible/Impossible

Structures (5). The input method for the web study (mouse movements and clicks on a web page) may account for the slower response times. A One-way ANOVA of the web sample mean RT for correct responses revealed significant effect of subtest, $F(8, 248) = 18.8$, $p < .001$, with Dot Coordinate times being significantly longer than True Length Recognition – A (3A), Mental Rotation (4) and Possible/Impossible Structures (5) as indicated by Tukey's pairwise comparisons. A One-Way ANOVA on the laboratory-based sample mean RT for correct responses also revealed a significant effect of subtest, $F(8, 360) = 37.3$, $p < .001$, with Dot Coordinate correct responses taking significantly longer to produce compared to all other subtests. Mental Rotation (4) and Possible/Impossible Structures (5) correct RTs were significantly faster compared to all other subtests.

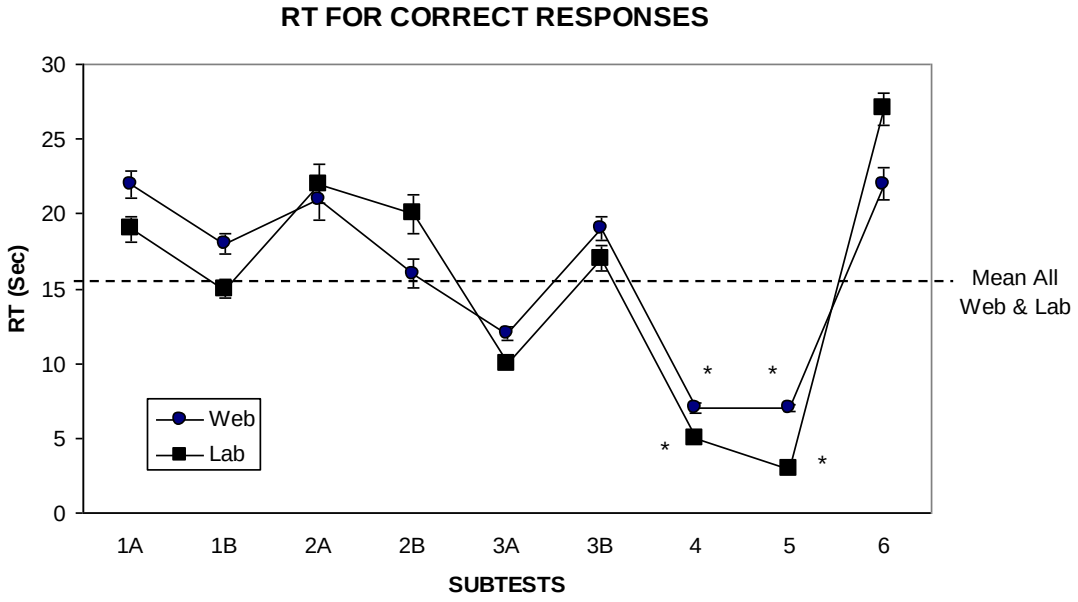


Figure 2. The comparison of mean correct response time and standard error per subtest for the Web and Lab samples. In some instances, the standard error is very small and is hidden by symbols used for Web and Lab RT. Significant difference ($p < .05$).

The pattern of results across subtests in Figure 2 is a mirror image of the pattern across subtests shown in Figure 1, suggesting that participants took longer to correctly answer harder subtests (lower percentage of correct responses) but took less time to correctly answer easier subtests (higher percentage of correct responses). This was further examined by correlating the percentage correct means with mean correct RT for each subtest, as shown in Table 3. Significant, positive and moderate to strong correlations were found for each subtest and the total for the web sample indicating that the harder the items in a subtest the longer participants took to produce a correct answer. This relationship was reflected to some extent in the lab-based sample, however, four of the correlations did not reach significance.

Table 3
Correlations Between Percentage Correct Mean Scores and Mean Response Times for Correct Answers for Web-based and Lab-based Samples.

Sample	1A	1B	2A	2B	3A	3B	4	5	6	Total
Web	.46**	.49**	.79**	.69**	.52**	.55**	.36*	.41*	.94**	.74**
Lab	.24	.14	.54**	.56**	.26	.42**	.25	.38*	.83**	.37*

Note. *p is significant at the .05 level. ** p is significant at the .01 level.

One thing this analysis did reveal was that the names given to the subtests were awkward, cumbersome and generally not very informative. A better labelling system emerged later in the development process and is referred to in the following chapters.

Discussion

Results for accuracy and RT were compared across both the lab and web studies to demonstrate that similar patterns emerge. The comparison was not expected to show that one methodology was superior to the other since the format of the two tests was very similar except for the input method (response pad versus mouse clicks).

There were no problems encountered in running the laboratory study, likely because of the controlled environment and the opportunity for the writer to address participant's concerns. Participants performed at a high level considering they did not have prior learning experiences in technical drawing, most likely because of the high academic achievement required to enter the Psychology program at the University of Newcastle. At this institution, students are admitted based on their University Admissions Index (UAI) and psychology students who participated in this study had a UAI of 89.1 or better. To allow some comparison, about 15% of school leavers who graduate after a senior high school education have a UAI of 89.1 or better.

Several issues were encountered in implementing the web study. The need to match the web and laboratory studies sometimes limited the writer's ability to fully exploit the benefits of the web interface. ColdFusion has the capability of allowing participants to click on the actual image that best represented their answer, which could have reduced the effort required to click on a number (web version) or press a response key (laboratory version) associated with the image. Other implementation issues included how to accommodate a variety of end-user connections and making allowances for timing differences due to network bandwidth. Since the web study was not carried out in a laboratory setting, the participant's computer and internet connection quality were outside the control of the study. This needed to be addressed by taking a lowest common denominator approach. While it was possible that users on fast PCs and fast internet connections could load the test data for each screen almost immediately, it was necessary to allow a few seconds for data to load onto slower machines before displaying any data to participants. This had no impact on the measurement of RT but may have resulted in some frustration for many participants.

Calibration issues raised by Krantz (2001) may account for some differences in results between the two studies, and the question of equivalence between sample characteristics (Buchanan, 2001) may offer further explanation. Despite the differences, results are similar providing confidence in reliability and validity between the studies.

The main technical difficulty with web delivery was caused by some participants discovering they could use the back button in the web browser, which resulted in data compilation problems. Participants were not explicitly told they could not use the back button and many may have considered this to be a reasonable practice to engage in. Of the 50 participants eliminated from the final analysis after entering the test phase, 18 were excluded because of the back button problem. Reips and Stieger (2004), point to log file analysis tools such as Scientific LogAnalyzer (http://psych-wextor.unizh.ch/loganalyzer/Analyzer5//) to mine data and analyse log files produced by web servers. In many cases, these tools help detect problems associated with data collection over the web. Although they help identify unusable data, they cannot eliminate the problems, only detect them. Hence, any future web implementations of 3DAT using ColdFusion will disable the back button by adding javascript to the ColdFusion code.

The remaining 32 participants eliminated from the final analysis were excluded because their test data were incomplete. For the 89 possible responses, excluded participants provided between two and 82 responses. Some participants may have failed to complete because they were not informed how many questions remained to be answered during testing. This information was included in later versions of the test to minimize the dropout rate. Note that none of the remaining data sets were eliminated on the basis of patterns in responses and RTs. Screening on this basis may be necessary for web-based data, particularly if a pattern of rapid responding indicates “clicking through”, however, there was no evidence of this pattern in the data collected.

The interface and response requirements of 3DAT were more complex than most other web-based tests that the writer is aware of. The web study was also more time consuming than most web-based tests and required dedication from participants to work through detailed instructions and complete practice trials before testing began. In contrast to the laboratory study, participants worked in isolation with no supervisory support. Collectively, these factors may have discouraged many potential participants. Of the 80 participants who made a start to the testing phase, some did not finish and others were excluded from the sample because of the data compilation problems caused by the browser back button problem.

The web study was developed to parallel the laboratory study so that reliability and validity could be compared. The results support previous findings that web-based studies can produce reliable and valid data (Birnbaum, 2004; Steyvers & Malmberg, 2003), and also extend these

findings to a more complex and demanding design than has been used in many previous web-based studies. The web study may also have more validity in terms of generalisation to the population because it was less student based. However, dropout rates are a concern. Reips (2002b) draws attention to the importance of examining dropout data and using this to improve online studies. Dropout numbers may indicate that a simpler and shorter test, and perhaps greater rewards for participation, will be required to obtain a larger sample. As well, improvements in design, such as simplifying instructions and changing the interface layout to take better advantage of the software tools are also likely to be helpful, together with improved and more widespread promotion strategies.

An essential standard in psychometrics is to evaluate the test items used in a test to decide if they are good or bad items in measurement terms. The *Classical Test Theory* (CTT) is one method that can be used to achieve this assessment, and it was applied to this study as a pilot though unplanned initiative after findings were known. In reality, the sample size was too small, even when the two studies (web and lab) were combined (n = 71), for any assessment to be totally convincing. However, the CTT procedure did identify underperforming test items that needed to be discarded despite some reservations. If nothing else, it was a starting point to improving the instrument for the two studies that would follow. The procedure reduced the 89 item test to 45 items and corrected an inconsistency in the number of test items allocated to each subtest. Some subjectivity was needed in choosing the 45 test items to account for the low sample size, but the process was mainly objective and helped achieve confidence in the procedure as a forerunner to what would follow. No detailed results are reported here because of the preliminary nature of the initiative. However, the CTT method and the psychometric properties it addresses are reported in far greater detail in chapter 4 where the sample size was more appropriate and the demographics were better suited to developing the 3DAT as an instrument for designers. Chapter 4 also draws attention to another method known as *Item Response Theory* (IRT) which was used to verify findings from the CTT method. IRT is introduced in chapter 3 as part of study 6.

The objectives of this study were concerned with investigating the feasibility of the web as a suitable platform for testing spatial ability and deciding about the psychometric qualities of initial subtests and test items. The study established the web as being suitable but some modifications were needed to improve its acceptance. Statistical analyses identified overlaps between a number of subtests suggesting that they would not all be needed in any final version of the 3DAT. Furthermore, 44 test items failed to reach the required psychometric standards to survive any scrutiny that would normally apply to a new ability test. In view of these outcomes, the objectives that guided this study were largely achieved.

Study 2 Investigating Validity

This study reports on research conducted by two Psychology Honours students (Day, 2006; Pollock, 2006) who were supervised by the writer. Their studies primarily investigated *convergent* and *discriminant* validity (divergent validity) for test items that were under consideration for the 3DAT. To a lesser extent, *validity with known groups* was also investigated. The two student researchers worked collaboratively but on different research questions. Essentially their studies compared the 3DAT with tests thought to measure similar abilities, and other tests thought to measure dissimilar abilities. In essence, if the purpose of the 3DAT is to assess spatial ability, it should show a strong relationship with tests of overlapping abilities and a weak or no relationship with tests of nonoverlapping abilities. This research also compared the performance of nondesigners (unskilled) and designers (skilled) with the expectation that designers would outperform nondesigners. This should be the case if the purpose of the 3DAT is to assess the spatial abilities required by designers. The objectives of Study 2 were to:

- investigate several types of validity using a range of established tests, and
- compare the performance of a design group and a nondesign group.

The 3DAT used in this study consisted of the same subtests used in study 1. However, after the item analysis was conducted, the number of test items were reduced to 45 as previously reported. The three subtests classified as two-way subtests contained 10 test items each, while the remaining three subtests contained five test items each. In simple terms, this equated to nine subtests with five test items in each that collectively made up the 45 test items. As a reminder to the reader, examples of these test items are shown in Appendix A (A1 to A6). While the 3DAT was computer-based, the other tests used to assess its psychometric properties were paper-based. These tests are established ability tests but some needed to be reduced in length to accommodate the constraints imposed by the nature of Honours projects. The paper tests are described below, and examples and sources are shown in Appendix A (Figures A7 to A11):

Space Relations: This test requires the mental manipulation of 3D objects in space. Each test item presents an unfolded view of an object and drawings of four optional 3D objects for participants to choose from. Participants match the unfolded view with one 3D object.

Minnesota Test: This test measures mental manipulation of 2D shapes. Participants are required to visualise how a number of different sections will combine to form a single shape. There are five options to choose from.

Mechanical Reasoning: This test presents pictorial mechanical situations linked to a simply worded question with several options for participants to choose from. The test items are typically concerned with rotation, pulleys, levers and loads.

Verbal Reasoning: This test measures the ability to understand whether the conclusion drawn from certain statements is correct or incorrect. Although the statements are really nonsense and not necessarily logical, the given test items measure a form of verbal reasoning. Of note is that the final score is equal to the number of answers marked correctly minus the number of answers marked incorrectly. Therefore, it is possible to receive a negative final result.

Numerical Reasoning: In this test, a set of numerical problems need to be solved based on reasoning. The test is not the same as numerical ability test. Test items are simple statements that require mathematical logic rather than mathematical ability. Most can be done mentally though a calculator can be used. Participants choose from four options.

The first three listed tests (*space*, *Minnesota* and *mechanical*) were considered to assess overlapping abilities consistent with the 3DAT. The *space relations* test is certainly seen as a measure of one form of spatial skill, however, both the *Minnesota and mechanical reasoning* tests are not strictly considered to be measures of spatial ability. Rather, they are viewed as nonverbal ability tests, though they still assess certain spatial concepts. These tests are often used for aptitude testing as part of batteries of tests for IQ measurement. Whilst the 3DAT measures a wider range of abilities, it was hypothesized that the 3DAT would positively correlate highly with these tests. It was also expected that correlations would be higher for participants with prior learning in spatial ability (designers). On the other hand, the remaining two paper tests (*verbal* and *numerical reasoning*) measure different abilities to spatial and were not hypothesized to correlate with the 3DAT for either group.

In simple terms, the validity of an instrument is the extent to which it measures the quality it claims to measure. Validity can be complex (e.g., construct validity) and there are various types that report different perspectives. Gregory (2004) asserts that validity is the most fundamental characteristic of a test and that its importance has been acknowledged by psychometricians for some time. As previously mentioned, this study was concerned with convergent validity which is revealed when a new test correlates with tests of similar abilities. This study also examined discriminant validity which is demonstrated when a new test does not correlate positively with tests that measure dissimilar abilities. On a smaller scale, this study also investigated validity with known groups which is evident when groups expected to do well perform better than those groups not expected to do well. Whilst the ultimate objective is to provide evidence of construct validity for a new test, the minor validities nevertheless contribute to this. Gregory makes the point that many psychometric theorists see construct validity as accumulative evidence provided by all forms of validity. In reality, this comes down to the other validities collectively contributing to the bigger and more complex measure of construct validity.

Methodology

Participants in this research were undergraduate university students recruited from both design and nondesign disciplines. The criteria for placement into either group was based on prior learning experience in a technical drawing environment. Those participants with prior learning experience were placed in the design group, and they came from the disciplines of engineering, architecture and construction management. There were 13 participants (9 males and 4 females) in this group. Participants with no prior learning experience were placed in the nondesign group and there were 18 participants (8 males and 10 females) in total. These participants came from the discipline of psychology. The design group received a small financial reimbursement to offset their incidental expenses, and the nondesign group received course credit for their participation.

Both groups completed the 3DAT in a spatial cognition research lab and the test was delivered using SuperLab Pro 2.01 under similar conditions to those reported in study 1. However, participants were able to complete the 3DAT is less time because of the reduced number of test items from the original 89 down to 45. With regard to the paper tests, participants in both groups received precise instructions to ensure they understood was what was required of them, and all tests included practice items to establish familiarity. The number of items in each paper test and the time allowed for completion are shown in Table 4.

Table 4
Number of Items and Time Allocated to Paper Tests

Test	Number Of Items	Time Allocated (Mins)
Space Relations	60	35
Minnesota Paper	64	20
Mechanical Reasoning	35	20
Verbal Reasoning	15	4
Numerical Reasoning	6	3

Total testing including the 3DAT and practice trials but excluding breaks took about two hours to complete. Tight supervision and the strict monitoring of time occurred to ensure participants were assessed under the same conditions.

Results

Correlation coefficients (*r*) were calculated for the 3DAT and all paper tests for both the design and nondesign groups to measure the level of convergent and discriminant validity that existed. Correlation coefficients for the design group are shown in Table 5, and for the nondesign group, they are shown in Table 6. Ideally, a high positive and significant correlation should exist for any combination of the spatial tests (*3DAT, mechanical, space and Minnesota*), but not between any of the spatial tests and the nonspatial tests (*verbal and numerical*).

Table 5
Correlation Coefficients for Design Group Across all Tests

	3DAT	Verbal	Numerical	Mechanical	Space	Minnesota
3DAT	1	.36	.53	.36	.63*	.53
Verbal	.36	1	.28	-.30	.08	.28
Numerical	.53	.28	1	.52	.60*	.75**
Mechanical	.36	-.30	.52	1	.79**	.47
Space	.63*	.08	.60*	.79**	1	.43
Minnesota	.53	.28	.75**	.47	.43	1

Note. * p is significant at the .05 level. ** p is significant at the .01 level.

For the design group, correlation is high between the *3DAT* and the *space* test, and also between the *space* and *mechanical* tests. Surprisingly, there is also a high correlation between the *numerical* test and both the *space* and *Minnesota* tests. Otherwise, the *numerical* test does not correlate significantly with any of the remaining tests. Noteworthy is that the *verbal* test does not correlate with any test used in the study.

Table 6
Correlation Coefficients for NonDesign Group Across all Tests

	3DAT	Verbal	Numerical	Mechanical	Space	Minnesota
3DAT	1	.13	-.13	.55*	.79**	.72**
Verbal	.13	1	.26	.26	.07	.34
Numerical	-.13	.26	1	.13	-.01	.07
Mechanical	.55*	.26	.13	1	.66**	.58*
Space	.79**	.07	-.01	.66**	1	.80**
Minnesota	.72**	.34	.07	.58*	.80**	1

Note. **p* is significant at the .05 level. ** *p* is significant at the .01 level.

For the nondesign group, correlation was high between the 3DAT and the mechanical, space and Minnesota tests. There was also a high correlation between the *space, mechanical and Minnesota* tests. However, there was no significant correlation for either the *numerical or verbal* tests with any other test used in the study.

Concentrating on convergent validity first of all, the 3DAT would ideally show a significant positive correlation with the *space, mechanical and Minnesota* tests. However, a high correlation with the *space* test is the most important because it is a benchmark for spatial ability, whereas the remaining two tests are classified as nonverbal tests. With these points in mind, this study provides evidence of convergent validity for both the design and nondesign groups although the evidence is stronger for the nondesign group. In other words, there is evidence that the 3DAT measures overlapping abilities with an established spatial ability test and two nonverbal tests that measure spatial concepts. Even though the evidence is stronger for the nondesign group, it is nevertheless still convincing for the design group because it is the *space* test that matters most of all. In terms of *effect size*, r = .10 is considered to be low, r = .30 is considered to be medium and r = .50 is considered to be high (Cohen, 1992). Using these

standards, the *r* values shown in Table 5 confirm that the effect size is high for the correlation between the 3DAT and space test. More convincingly, the *r* values shown in Table 6 confirm that the effect size is high for the correlation between the 3DAT and the *mechanical, space and Minnesota* tests. Importantly, effect size is also high between the three spatial tests (*space, mechanical and Minnesota*) for the nondesign group which reinforces that they are measuring the same ability. There is less evidence of this relationship for the design group, although it is partly established between the *space and mechanical* tests. However, this favourable finding is negated to some extent by the high correlation that exists between the *space and numerical* tests which is not ideal. Effect size is defined earlier in the *Terms and Definitions* section of this thesis, but essentially it is a measure of *practical significance* which helps put comparisons into perspective.

Turning now to discriminant validity, the ideal outcome is for the 3DAT not to show a significant correlation with the nonspatial tests which in this study are the verbal and numerical paper tests. As well, there would not be a significant correlation between the nonspatial tests and any of the spatial paper tests. With this in mind, the correlation between the 3DAT and both the verbal and numerical tests did not reach significance for the design or the nondesign group. Furthermore, the verbal test did not show a significant correlation with the mechanical, space or Minnesota spatial tests for either of the two groups. However, providing some inconsistency, a high correlation was found for the numerical test with the space and Minnesota tests within the design group, although not for the nondesign group. Overall, these results suggest that the 3DAT is not measuring overlapping abilities with the nonspatial tasks.

If the 3DAT is to be regarded as a good instrument for measuring the spatial ability of designers, then test takers from design disciplines such as engineering should perform significantly better than test takers from unrelated disciplines such as psychology. On the other hand, there should not be a significant difference in performance between the two groups on the nonspatial tasks. Conducting such investigations is essentially determining one form of validity described as *validity with known groups*, although this is referred to by Gregory (2004) as *theory-consistent group differences*. Gregory considers that this form of validity is evident when those people thought to be high on the construct being measured by the test achieve better results than those people thought to have low ability on the construct. Table 7 shows the results of an independent *t* test conducted for the design group and the nondesign group on the 3DAT and the spatial and nonspatial paper tests.

Table 7
Mean Differences Between Groups for the 3DAT and Paper Tests

	Design		NonDesign				
Test	M	SD	M	SD	df	t	p
3DAT	87.18	10.49	68.15	13.61	29	4.21	< .001
Verbal	30.77	34.16	10.37	31.87	29	1.71	.09
Numerical	80.77	22.41	70.37	34.09	29	.96	.35
Mechanical	86.59	10.84	78.73	13.59	29	1.73	.09
Space	84.87	11.35	72.69	15.24	29	2.43	.02
Minnesota	87.14	10.47	72.22	14.38	29	3.18	.01

Note. Design (n = 13), NonDesign (n = 18). Negative scores are possible for the Verbal test.

Considering the 3DAT first of all, Table 7 shows that the difference in performance between the design and nondesign groups is statistically significant (t (29) = 4.21, p < .001), and that the design group performed better than the nondesign group. Moving to the spatial paper tests, a significant mean difference is also shown in Table 7 for both the space and Minnesota tests favouring the performance of the design group (t (29) = 2.43, p = .02 and t (29) = 3.18, p = .01 respectively). For the third spatial test (mechanical), the mean difference is not significant though the mean is higher for the design group than the nondesign group. The significance value (p = .09) suggests, however, that the findings are trending towards significance. For *effect size*, where d is equal to *difference*, and where d = .20 is considered low, d = .50 is considered medium, and d = .80 is considered high (Cohen, 1992), the results can be shown to be more substantial. That is, the effect sizes are large for the 3DAT, space and Minnesota tests since d = 1.5, d = 0.9 and d = 1.2 respectively. These are large d values that help highlight the significant difference between the two groups on the 3DAT and spatial tests. The strength of these values, which are only considered when mean differences are significant, indicate that *validity with known groups* is established for this study even though the sample size was less than ideal.

To appreciate these results further, respective performances on the nonspatial paper tests (verbal and numerical) are deserving of mention. Table 7 shows that the mean differences between the design and the nondesign groups on both the verbal and numerical tests are not significant (t (29) = 1.71, p = .09 and t (29) = .96, p = .35 respectively). Noteworthy, the large SDs relative to the means for the verbal test suggest that skewness in the data breaks the assumption of normality required by the t test (histograms verified skewness). However, the application of a nonparametric statistical procedure (two independent sample *Mann-Whitney* test) confirmed the t test result (U = 74.50, Z = 1.71, p = .089, two-tailed).

Discussion

This study represented a start to addressing the important issue of validity and focused on convergent, discriminate and validity with known groups in particular. Convergent validity is

evident when a strong correlation exists between tests of similar abilities, and discriminant ability exists when there is no positive correlation between tests that measure dissimilar abilities. Validity with known groups is identified when test takers expected to do well outperform those who are not expected to do well. Each of these validities are indicators of a bigger and more complex validity known as construct validity. Study 2 is described as an indicative study because of some restraining factors. The sample size, for example, was less than ideal and should be considered when interpreting the findings. Moreover, some paper tests were modified or reduced in length to accommodate the limitations imposed by Honours projects. However, despite these setbacks, this study as an initial investigation provided some confidence in the process of validating the 3DAT.

Correlation coefficients and effect sizes mostly pointed in the right direction and therefore provided evidence of convergent validity for both groups. Gregory (2004), using IQ measures as examples, explains that two tests of similar abilities should have enough in common to produce a large correlation when jointly given to appropriate examinees. Four out of the six relevant correlation values in this study (see Tables 5 & 6) were greater than the criterion of $r = 0.5$ suggested by Gregory. On the other hand, the 3DAT did not correlate with the nonspatial paper tests as expected and hence provided good evidence of discriminant validity. Furthermore, there was strong evidence of validity with known groups because the design group performed better than the nondesign group on the 3DAT and all the spatial paper tests although the difference was not significant for the mechanical test. In support, the differences in performance between the two groups on the nonspatial tests were not significant as hypothesized.

The concept of effect size was introduced in this study together with standards advocated by Cohen (1992) because it is regarded as a measure of practical significance and a method for easy comparisons . Effect size is simple to calculate (see *Terms and Definitions*) and provides meaning and magnitude to changes and differences found in data produced from research. Kazis, Anderson, and Meenan (1989) recommend the use of effect size as a standard unit of measurement for benchmarking and argue effect sizes allow a clearer understanding of data variation. Although Kazis et al. applied the principles of effect size to health and medicine, their reasoning applies equally to other disciplines as well. The use of effect size is appropriate where correlation coefficients or mean differences are to be rated and compared.

The objectives for study 2 focused on elements of construct validity and investigating the difference in spatial ability between two categories of subjects divided into a design group and nondesign group. With the exception of some minor surprises, the results of this study favoured the psychometric properties considered desirable in the development of any ability test, and in this case, the 3DAT in particular.

Study 3 Gender Differences

This study reports on research conducted by a third Psychology Honours student (Laver, 2007) who was also supervised by the writer. Laver's studies focused entirely on gender differences in spatial cognition which the junior researcher investigated using the 3DAT and several spatial paper tests. A gender difference favouring males is constantly reported in the literature but discrepancies and variations in findings are evident. There is debate about the extent of gender differences, the spatial tests that best measure this, the influence of prior learning and whether the difference might be decreasing. Central to the debate is whether the difference is due to biological factors or whether sociological factors may be responsible. Sociological considerations include: the type of activities young people engage in as they approach maturity, perceived gender roles and stereotype priming that may exist in the home and community. As a reminder to the reader, these issues are dealt with in some detail in chapter 1. Study 3 provided a good opportunity to examine some of these issues and to see if the 3DAT would identify a gender difference still to be present in current younger generations. There was some expectation that this might be diminishing because gender roles are probably less separated today than they were for previous generations. In other words, many growing up activities are no longer labelled male activities as they might have been in the past since both genders were beginning to share similar educational and recreational experiences. Examining gender differences in spatial performance is worthwhile because it raises questions about equal opportunities, innateness, targeted training, the self-fulfilling prophesy and whether enough is being done to reduce the difference. These questions are important enough to be a main focus in any research on spatial cognition and it remains a concern that there is a tendency to accept the current situation as a *matter of fact* without giving it the real attention it deserves. This study compared the spatial performance of a male group and a female group who both had no prior learning experience in technical drawing. The objectives of Study 3 were to:

- conduct a preliminary investigation into gender differences,
- explore whether a link between spatial experience and spatial performance exists, and
- evaluate initial reliability of test items in the 3DAT.

The 3DAT used in this study was the same version reported for study 2. To assist the reader, it consisted of six subtests (three described as two-way) and a total of 45 test items divided among the subtests. The subtests were labelled: *2D-3D recognition, correct fold, true length, mental rotation, object decision and dot coordinate*. Note that the *object decision* subtest was previously named *possible/ impossible structures*. Examples of test items for each subtest are shown in Appendix A (A1 to A6). For the paper tests, there were spatial tests in common with study 2 as well, but for study 3 a baseline test was added. The common spatial tests were: *space relations, Minnesota form test and mechanical reasoning*. The baseline test was introduced as a

general ability test so that a comparison of academic standing between the two groups could be assessed before conclusions were drawn about group performances on the various tests. The use of a baseline test where the difference in performance is not significant gives some confidence that samples were drawn from the same population. The test used for this purpose (*raven's standard progressive matrices)* is described as a test of analytical intelligence that measures the educative ability to think clearly, shape insights and to identify relationships (Raven, Raven, & Court, 1998). The test is nonverbal and based on pattern recognition in the sense that a missing section has to be identified from eight alternatives to match a particular pattern. The patterns exhibit changes in two directions. Examples of each paper test including the general ability test are shown in Appendix A (A7 to A9, A12).

Part of this study included a questionnaire developed by the junior researcher (Laver, 2007) to explore any association between spatial performance and prior experience in spatially-oriented activities. It was anticipated that any significant difference found between the male and female groups in spatial performance could be explained in terms of spatial activities that participants engaged in during their early life experiences. The seven item questionnaire was named the s*patial activities survey* and consisted of broad questions representing sets of related activities thought to utilise some elements of spatial ability. Participants responded to each question by choosing one option from a Likert Scale of measurement. The questionnaire is shown at Appendix B.

Methodology

Participants who took part in this research were undergraduate university students who did not have any prior learning in formal design or graphical communication courses at the secondary or tertiary level. The idea was to recruit unskilled participants to allow a balanced comparison to be made between the two groups since the study aimed to gauge the level of gender difference in the general community. This would establish if the condition existed before any design-related training. Participants with prior learning would have introduced a confound and added some risk of distorting the research outcomes. 39 students (15 male, 24 female) chose to participate after considering a number of other research participation options advertised through an online research participation management system. The online system made it possible for researchers to stay at arm's length to the recruitment process and gave students the freedom to choose from many projects without feeling any form of coercion. Participants who volunteered for this study were first year psychology students who received course credit for their participation. University students were considered a good choice because both genders probably had equal opportunities and exposure to education, recreation and social activities during their developmental years. Participants were aware that the study intended to investigate possible gender differences in spatial cognition but were not told that this bias normally favoured males.

The intention was to avoid possible priming prior to testing and to avoid reinforcing a self-fulfilling prophesy if it existed among female participants. The age of the 15 males ranged from 18 to 25 (M = 21, SD = 6.42) and the age for the 24 females ranged from 18 to 40 (M = 24, SD = 6.42).

Both groups completed the 3DAT in a dedicated spatial cognition lab under conditions similar to those experienced by participants in study 1 and study 2. The s*patial activities survey* was given to participants first of all followed by the 3DAT, the spatial paper tests and the baseline test in that order. The administration of the paper tests including the completion time was in accordance with the standard instructions stipulated by test suppliers. All the paper tests were multiple choice in design but the number of items and the time allowed to complete them varied. It was not necessary to reduce the time allowed to complete the paper tests as it was for study 2. This removed a confound for study 3 which may have been an issue for study 2. Although study 3 was mainly about gender differences, it was also an opportunity to verify validity results reported for the 3DAT in study 2. The time taken to complete all tests including practice trials, instructions and breaks to safeguard against fatigue was approximately three hours.

Results

Independent samples *t* tests were used to examine means for statistical significance for the 3DAT and all paper tests across groups and results are shown in Table 8. For the spatial tests, significance was reached for the *3DAT and mechanical reasoning tests* and trended this way for the *space relations* test. Mean scores for these tests shown in Table 8 indicate a consistent gender difference favouring males. Effect size (*d*) was introduced in study 2 using a scale of d = .20 (small), d = .50 (medium) and d = .80 (large) and reported where mean differences were significant. Applying this standard to study 3, effect size was found to be high for both the *3DAT* and the *mechanical reasoning* tests (.88 and 1.29 respectively). The large effect sizes act to strengthen the evidence of gender differences. Interestingly, and contrary to expectations, females averaged higher scores than males on the *Minnesota* test though the difference was not significant. In two out of four cases, the spatial tests supported the hypothesis and showed a bias in spatial performance favouring males. A third test (*space relations*), approached significance and again in favour of males. The *ravens* test was included in the selection of paper tests to serve as a baseline measure of general academic ability so that researchers could gauge the level of homogeneity between the two samples. Ideally, the results should not show a significant difference between male and female performance. This was not the case and male performance was significantly better than females (t (37) = 2.51, p = .02). However, this is not as disappointing as it first seems because the *ravens* in retrospect may not have been a good choice for a baseline test. The *ravens* test belongs to a category of tests described in study 2 as

nonverbal which measure spatial concepts. Although the *ravens* test does not fit the description of an ideal spatial ability test because it does not include a 3D component, it does nevertheless measure some aspect of spatial ability. This being the case, the test chosen to benchmark general ability in this study may also be confirming a gender difference skewed towards males.

Table 8
Means and Standard Deviations for Gender Groups for the 3DAT and Paper Tests. Percent values are shown.

	Male		Female				
Test	M	SD	M	SD	df	t	p
3DAT	82.18	12.47	72.13	11.94	37	2.67	.01
Minnesota	74.38	12.37	76.63	12.58	36	.554	.59
Mech Reasoning	84.67	10.27	70.48	11.50	37	3.90	< .001
Space Relations	79.22	15.53	70.83	14.07	37	1.74	.09
Ravens	90.89	8.04	84.58	7.39	37	2.51	.02

Note. Male (n = 15), Female (n = 24). *p* significance is 2-tailed.

Though not a primary focus of this study, but worthy of mention nonetheless, is that study 3 also provided evidence of convergent validity for the 3DAT. Since the 3DAT and all of the paper tests (including *ravens*) purport to measure some aspect of spatial thinking, there should be sufficient spatial ability overlap between the tests to expect a significant level of correlation between the tests. This is mostly the case. A correlation analysis based on the Pearson standard showed a large effect size for the *3DAT* when paired with the *mechanical* (r = .62), *space relations* (r = .71) and *ravens* (r = .62) paper tests. Using the scale advocated by Cohen (1992), these values provided strong evidence of convergent validity for the 3DAT and consequently gave confidence that 3DAT development was moving in the right direction. However, once again, the *Minnesota* test produced surprising results. Because this test fits into a group of tests thought to measure spatial concepts, it too was expected to show a significant correlation with the 3DAT. However, this was not the case and it raises the question of whether the *Minnesota* would have been a better baseline test than the *ravens* for this study.

Focusing only on the 3DAT, Figure 3 shows the mean scores for male versus female performance on each of the subtests in the 3DAT. Overall difference in performance on the 3DAT was reported earlier as significant, however, Figure 3 indicates that significance was not evident in all subtests. In fact, it was really only achieved for the true length B and object decision subtests (t (37) = 2.13, p = .04 and t (37) = 2.93, p = .001 respectively). These two subtests then are likely to account for the significance shown for the 3DAT overall. At this point, it is timely to emphasize that study 3 was part of early investigations and part of this was concerned with identifying subtests and suitable test items. The better than expected performance on the remaining subtests may be explained in terms of results approaching ceiling and therefore test items failing to discriminate between the abilities of the two groups. This

observation does not apply to the dot coordinate subtest because mean scores are well below ceiling. There is also the possibility that results are providing some evidence that the gap in gender difference is narrowing when compared to earlier studies reported in the literature.

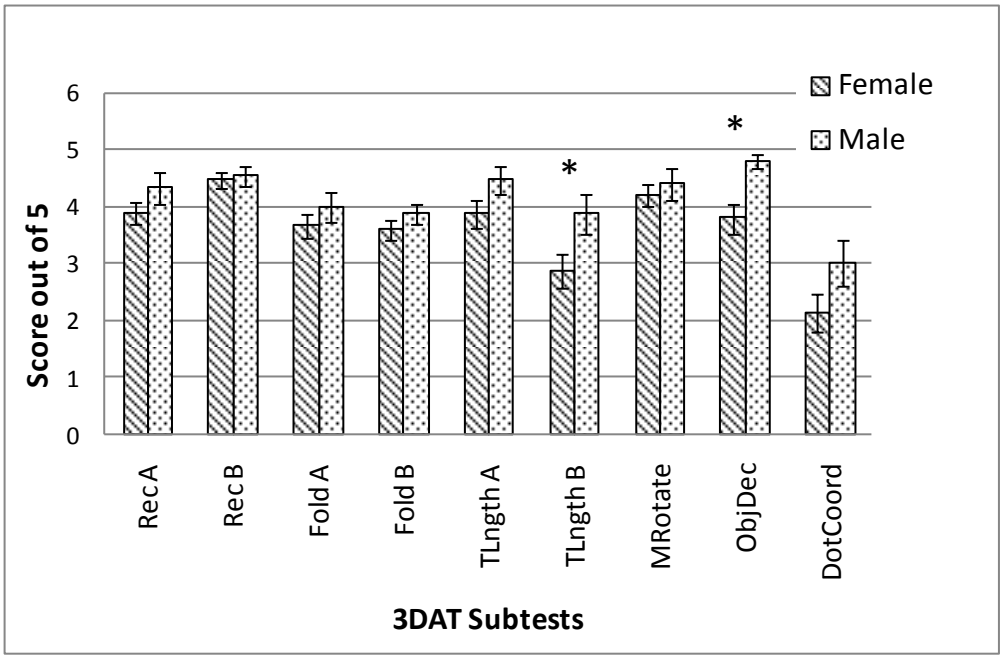


Figure 3. Mean correct responses and standard error for males and female groups for each subtest in the 3DAT. Scores are out of 5. ObjDec adjusted for unequal variance. *Significant difference ($p < .05$).

Table 9 shows the results of an independent samples *t* test conducted for each of the questions asked in the spatial activities survey and the survey overall. Dealing with the latter first, a significant difference in means was found between male and female participants based on the ratings they gave to each of the spatial questions asked in the survey. This difference favoured males and therefore indicated that they spent more time in their developmental years on activities that were thought to involve spatial skills. However, significance was not reached for all seven questions (refer Table 9) but it was reached for two questions (video games and map reading) and trended this way for a third question (object assembly). Considering the spatial skills embodied in the sets of activities represented by the seven questions, a case can be made that the three sets that reached or approached significance required a higher degree of spatial experience. For example, understanding 3D properties, visual perception and manual manipulation. Adding some weight to this, effect size (practical significance) was large for both video games and reading maps ($d = 1.12$ and $d = .67$ respectively). Although significance was not reach for all seven questions on the survey, the level of significance found for the two reported questions is substantial. Since the difference favoured males, the results imply that the better performance of males on the spatial tests might be explained in terms of the extra time they spent on spatial activities during their early years compared to their female counterparts.

Table 9
Results of Survey Overall and for each Survey Item. Object Assembly Adjusted for Unequal Variance.

	Male		Female				
Survey	Mean	SD	Mean	SD	df	t	p
Results Overall	3.28	.440	2.98	.371	37	2.29	.03
Drive Vehicle	4.40	.632	4.25	1.11	37	.475	.64
Active Sports	3.47	1.13	3.33	1.17	37	.352	.73
Video Games	3.27	.961	2.17	1.01	37	3.38	.002
Computers	4.07	.458	4.21	.721	37	.678	.50
Arts & Crafts	2.00	.756	2.33	1.13	37	1.01	.32
Map Reading	3.13	.834	2.67	.565	37	2.09	.04
Object Assembly	2.6	1.30	1.87	.741	19.78	1.97	.06

Note. Means are Based on a Likert Ranking of 1 to 5. Male (n = 15), Female (n = 24).

This study presented a further opportunity to investigate the internal consistency (reliability) of test items within each subtest. Because of the possibility that subtests were measuring different elements of spatial ability, it was logical to test for internal consistency within subtests rather than across all subtests. Aiken (1997) defines internal consistency as the degree to which all items *measure the same variable*, and Gregory (2004) describes it as a measure of *consistent interrelatedness*. Internal consistency is rated according to a coefficient index called Cronbach's Alpha (α) or coefficient alpha with a possible range of -1.0 to +1.0. The Cronbach-alpha formula is generally used to calculate this index. Table 10 lists the subtests that make up the 3DAT and shows coefficient alpha indexes for each of them. Reliability is covered in more detail in later sections since there are other forms of reliability besides coefficient alpha (e.g., *test retest* reliability). These are particularly addressed in chapter 4.

Table 10
Cronbach's Alpha Reliability Measures for 3DAT Subtest (n = 39).

Subtest	Alpha	Subtest	Alpha
2D 3D Recognition	.447	Mental Rotation	.548
Correct Fold	.158	Object Decision	.630
True Length	.722	Dot Coordinate	.695

Note. Alpha overall not reported for 3DAT. Refer Chapter 5 for explanation.

Benchmark alpha reliability coefficients were introduced in study 1 and α values greater than 0.7 were considered acceptable by psychometric standards while values greater than 0.8 were considered highly acceptable. Values approaching zero were seen to reflect poor internal consistency. Based on these standards, the alpha indexes shown in Table 10 mostly fall short of ideal measures and are not normally acceptable, although there are exceptions depending upon the purpose of the instrument. Guilford and Fruchter (1978) suggest that many standard tests with reliability coefficients less than .70 can be useful. In reality, the sample size was very low for testing reliability and the 3DAT was progressing through its initial stages of development.

However, the findings flagged an important concept if a good measure of spatial ability was going to be developed.

Discussion

The main objective of this study was to investigate whether the gender difference in spatial ability and reported in the literature as favouring males still existed, or whether there were signs that some change was occurring. Study 3 established that the gap between male and female spatial performance still remains, but there were encouraging signs that the gap was reducing. While the difference on the 3DAT was significant, it was only significant for two out of the six subtests if the two-way subtests are treated as single subtests. There was also an expectation that significance would be found on the three designated spatial paper tests but this only occurred for one of them. However, the baseline test which could be regarded as a form of spatial test did show a significant bias towards males although this was not its intended purpose. In contrast, a spatial paper test expected to show a gender difference favouring males did not produce significance. In fact, the mean scores were opposite to expectations in favour of females. Some caution is necessary in interpreting findings overall because sample sizes were low and the literature reports mixed results when the question of appropriate tests is raised. Voyer et al. (1995) for example contend that some tests are more applicable than others for identifying gender differences in spatial abilities. Furthermore, the authors suggest that the mental rotation task is the most likely to produce strong evidence of a gender difference. One subtest in the 3DAT (*MRotate*, Figure 3) was a mental rotation task based on the standard used by Metzler and Shepard (1988) and referred to earlier in this chapter. Interestingly, this study did not produce a significant difference for this subtest consistent with the findings of Voyer et al. although males did achieve a higher mean score than females.

A common explanation put forward by researchers for the gender difference in spatial performance is the difference in the sociological experiences (see *Terms and Definitions*) of males compared to females during their early developmental years (Rafi, 2006). Many younger people believe that they should be engaged in certain educational and recreational activities according to their gender, and this is often reinforced by the home and the expectations of their culture. By tradition, this means that males tend to adopt activities that are more spatially-oriented (e.g., object assembly) than females during childhood and adolescent years. Gender stereotyping generally denies females these shaping opportunities because they are seen as masculine endeavours and not relevant to females. Consequently, the spatial potential of females is often underestimated. The survey conducted by Laver (2007) and reported in this study showed some evidence of social impact. Sets of activities in the form of seven questions were surveyed with the participants and the two questions thought to represent the most relevant spatial activities (video games and map reading) produced significant gender differences in

favour of males. One of these (video games) is reported by Deno (1995) as the sole activity to positively correlate significantly with spatial ability for females. In other words, females may benefit from this activity alone if they engaged often enough in the various skills inherent in these games. Deno also adds that females appear to benefit more from visual activities than they do from physical activities. Examples of these might be video games and sports participation respectively. Of note is that a third activity (object assembly) likely to involve substantial spatial skills trended towards significance (t (19.78) = 1.97, p = .06). Although the Laver survey was not elaborate or extensive, it did nevertheless probe some specific areas thought to be relevant to spatial ability and it did reveal differences in activities where they appear to matter most.

The investigation of psychometric properties was also part of this study but not in a comprehensive way. Convergent validity was again identified as a positive for the 3DAT, and reliability in the form of internal consistency approached acceptable levels for some subtests, although not for others. In retrospect, the choice of the *ravens standard progressive matrices* as a baseline test was not the best choice possible since it is regarded as a nonverbal test that probably measures some aspect of spatial performance. Instead, a verbal reasoning, numerical reasoning or a nonsense syllogisms test would have been a better choice given that it was general academic ability that was being compared on this occasion. Results overall ran parallel with the mixed findings generally reported in the literature but they offered some optimism towards improving spatial ability for females. Study 3 largely achieved its objectives of exploring gender differences, investigating a link between spatial experience and performance and evaluating several psychometric properties of the 3DAT during an early stage of its development.

Chapter Summary

This chapter reported exploratory initiatives taken during the first stages in the development of the 3DAT. The scope varied but the main issues were psychometric properties, the viability of online testing, gender difference and the possible link between prior spatial learning and present spatial performance. The studies mostly involved participants who were nondesigners and from a university culture, but there were exceptions. One trial in one study was delivered online and it attracted participants from the wider community, and another study saw the introduction of a design group. Apart from the online trial, the studies were run under controlled laboratory conditions using software dedicated to the preparation and delivery of psychological experiments. Sample sizes were not large which prompted some caution in the interpretation of findings. Nondesign groups were chosen for most studies as a starting point to gauge the level of spatial understanding in the general community. However, a design group was introduced later to allow one form of validity to be assessed. Subsequent studies reported in later chapters focus entirely on participants with a design background.

From the outset, this writer believed that a correct measure of spatial ability would include a number of subtests based on a premise that spatial ability is likely to be a combination of several spatial factors, and different subtests were required to capture these. A central consideration was deciding about suitable subtests and the test items themselves. This required item analysis to evaluate psychometric properties such as item discrimination, item difficulty and item reliability. Some subtests and test items showed early promise, but on the other hand, some appeared to be unsuitable. Whilst sample sizes were less than ideal, it was possible to combine two trials (reported in study 1) to achieve a better (though not ideal) sample to at least conduct a provisional analysis so that very poor test items could be discarded. As a consequence, the 3DAT reduced from 89 test items to 45 after study 1 and the revised version was used in the studies that immediately followed. The concern about suitable subtests and test items was a constant throughout the initial development stage, but also for the other development stages reported in later chapters. For the most part, the evaluation of the 3DAT was based on choice accuracy, but reaction time (RT) was a factor for the lab and web trials conducted in study 1. Some care was needed with conclusions drawn from the RT data because strictly speaking, the 3DAT was not a speeded test and participants were advised that accuracy was more important than speed. Participants also knew that the test items would time out, and that the provisions were generous. Hence, there is every possibility that participants took their time in responding to the test items. However, RT was still reported because it is a measure of competency under certain conditions.

This initial stage of the 3DAT development essentially achieved its objectives and provided some valuable lessons. Online testing was considered feasible and worth pursuing further. There were issues to overcome such as the browser problem, a better programming platform and a more user-friendly interface. Also, a more effective naming system for the subtests was needed to improve verbal and visual communications. However, these realisations were seen as part of the normal process of test construction and development. In simple terms, this comes down to *testing the test* because many test items, though first of all seeming okay, turn out to be poor items serving no real purpose in a new test. Also clearly obvious was a need to substantially increase the sample size so item analysis could become a serious undertaking. As previously mentioned, the guide for meaningful item analysis is to have a sample equal to five test takers for every test item. Thus, for the 3DAT with 45 test items, this equates to a sample of 225 test takers. There was every possibility that the 3DAT would increase in the number of test items because of the likelihood of adding and evaluating new subtests. In this case, the sample would need to be greater than 225. In fact, this turned out to be the case, and details are reported in a later chapter. A large sample and a subsequent analysis were needed to have confidence in the psychometric properties of the 3DAT, but a large sample was also needed to produce standards

(norms) that could be offered to potential users of the 3DAT. The focus on gender during this early stage of development was meaningful because it established that the type of test can be a concern, and that any gender difference may relate to differences in prior sociological experiences. Effect size was introduced as a measure of practical significance and as an extension to statistical significance. This increased meaning because comparisons across datasets were given an added perspective. Also to emerge was a realisation to choose baseline tests more carefully, and to ensure that they are different to measures of spatial ability. The Laver questionnaire was simple and effective although not comprehensive, but it did serve to highlight differences in early sociological experiences between the genders. The outcomes from these studies paved the way for the eventual completion of the 3DAT. The process confirmed a number of concepts thought to be relevant to spatial ability and drew attention to others while flagging many things still needing to be done. These matters are addressed in the developmental phases that follow.

CHAPTER 3

TRANSITIONAL DEVELOPMENT

Overview

This chapter reports on a set of studies described as *transitional* because they moved the 3DAT from early investigation to final investigation. This phase explored further subtests, compared the degree of difficulty of test items belonging to the same subtest, and collated the opinions of subject matter experts (SMEs) about subtests and the importance of spatial ability to their disciplines. Part of the transitional phase included a special focus on content validity which had not been considered in previous studies, and a closer look at validity with known groups. The need for a simplified naming system for subtests became apparent after initial studies showed that the previous names were cumbersome to report in figures, tables and body text. Hence, new names were established during this phase that better described the focus of subtests and two letter acronyms (e.g., BR) were introduced to improve communications. The aim of the transitional phase was to further examine psychometric properties of subtests and test items and to consider the opinions of qualified designers from both industry and academia. This chapter presents procedures and outcomes divided into three studies named study 4, study 5 and study 6 respectively. Study 4 tested several subtests considered possibilities for the 3DAT and also compared different sets of test items within the same classification. Study 5 covered all issues associated with the SMEs while study 6 provided a more comprehensive comparison of a design group and a nondesign group to demonstrate that the 3DAT was targeting designers. Two of these studies (study 4 and study 6) were transitional from another perspective since they were the last incidents of the 3DAT based on using psychological experimental software (SuperLab) that was only suitable for a closed laboratory environment. After these studies, the 3DAT became an online test potentially available to any clientele in any environment in any location.

Study 4 Evaluating Subtests and Test Items

This study consisted of two trials and both were conducted with novice designers from the disciplines of architecture, construction management and technology education. The first part of the study evaluated the suitability of several new subtests, a process considered necessary because a number of previous subtests had proven unsatisfactory. The *object decision* subtest (previously named *possible/ impossible structures*) is a case in point since it was always too easy for any group in the earlier studies (refer Figure 1, 2 & 3). The new subtests were chosen because they appeared to measure the skills required by designers. There was no certainty about their final acceptance, but initially they appeared suitable and worthy of evaluation. These subtests collectively had elements of rotation, different viewing directions and 2D to 3D and 3D to 2D transformations. That is, all spatial attributes thought to be important to designers. Hence,

the first part of study 4 focused on new subtests. The second part of the study compared sets of test items from the same subtest classification to demonstrate that any variation in test item standards could impact on test performance even though they belonged to the same group. That is, an argument was being made that more detail should be provided about actual test items used in research. This suggests that some results reported in the literature are potentially misleading where comparisons are made between studies that used the same subtests but not necessarily the same test items. Not enough is said about the actual test items and the only mention is generally about the category of test items (i.e., subtests) but not the actual test items themselves. There is a case for reporting the psychometric properties of test items in every study where test performance is important. The objectives of study 4 were to:

- investigate the suitability of several subtests not previously considered for the 3DAT, and
- compare different standards of test items within the same subtests.

Table 11 lists all subtests used in Trial 1 and Trial 2, and new subtests and subtests common to both trials are indicated. The new labelling system for subtests was adopted from this point on and is included in Table 11.

Table 11
Subtests used in Trial 1 and Trail 2 with Status Shown.

Subtests	New Subtest	Subtests in Common
Dot Coordinate (DC)		Yes
Mental Rotation (MR)		
Mental Cutting (MC)	Yes	
Fold Unfold (FU)		Yes
True Length (TL)		Yes
Visualization (VZ)		Yes
Building Representations (BR)	Yes	
Engineering Drawing (ED)	Yes	
Transformation (TR)	Yes	
Recognition (RC)		

Subtests DC, MR, FU, TL and RC were previously used as part of the 3DAT development and were reported in chapter 2. Three subtests (FU, TL and RC) had slightly different names in previous studies, and one subtest used in previous studies (object decision) was discarded as reported earlier because it was found to be too easy. For each new subtest introduced, a description follows:

MENTAL CUTTING (MC)

A 3D view of an object intersected by a cutting plane is presented (USA University entrance examination as cited in Sorby & Baartmans 2000). The idea is to identify the resulting 2D shape of the surface when the top portion of the object is removed. Participants choose from 4 options (see Appendix A, Figure A13).

BUILDING REPRESENTATIONS (BR)

A 3D view of an object based on an arrangement of cubes is displayed with front and right views clearly labelled (Ben-Chaim et al., 1988). Participants are asked to identify the correct 2D back view of the object from 4 given options (see Appendix A, Figure A14).

ENGINEERING DRAWING (ED)

An object is presented as in isometric projection and 4 sets of related 2D views are given as possible answers (Alias, Black, & Gray, 2002). Each set contains a front view, a top view and an end view. Participants decide which set of 2D views match the 3D object (see Appendix A, Figure A15).

TRANSFORMATION (TR)

A top view of an object in 2D format is presented and a viewing direction is provided as a reference point. The object is an arrangement of cubes with numbers in strategic positions to indicate the 3D shape of the object (Olkun, 2003). Participants decide from four 3D options which one matches the given viewing direction (see Appendix A, Figure A16).

There were 6 test items in each subtest and each of these was developed by a 3D modeller under the direction, guidance, scrutiny and evaluation of the writer. Thus, the test items were not the same as those reported by other researchers in the literature so there was every possibility that their psychometric properties were different. In some cases, other researchers used a greater number than 6 test items. All objects depicted in the test items were novel in design and were intentionally not meant to represent everyday objects. From the outset, it was considered that test items based on real life objects may help interpretation and add an unnecessary confound to any of the studies conducted.

Methodology

Participants in the trials were first year university design students from the disciplines of architecture, construction management and technology education. 46 students formed the Trial 1 group (m = 30, f = 16), and 63 students formed the Trial 2 group (m = 39, f = 24). The trials were conducted as a class activity with Ethics approval to collect data because the activity aligned with course objectives and provided educational benefits to students. Participants received a summary of their overall result and for individual subtests. This was possible because a spreadsheet template that was purposely developed allowed participants to simply copy their data into the template and their results would display immediately after inbuilt functions automatically performed the necessary calculations. Participants were invited to consult with teaching staff for extra feedback and tuition if required. Importantly, participants were also able to nominate for their data not to be used for research purposes if they so decided. The 3DAT for these trials was created in a later version of SuperLab (4.06b) but was administered similarly to

the process reported in study 1 (chapter 2). However, there were some minor differences. For Trial 1, there were 36 test items divided equally into 6 subtests (DC, FU, MC, MR, TL & VZ), and for Trial 2, there were 48 test items divided equally into 8 subtests (DC, FU, TL, VZ, BR, ED, TR, & RC). These structures are reflected in Table 11. The trials were conducted in normal tutorial times in computer labs which could accommodate up to 30 participants at any one time. Another difference was that participants responded to the test items using the keyboard and not the response pad used in earlier studies. Supervisors guided participants through a set of practice trials also similar to the procedure reported in study 1 (chapter 2). There were some minor server difficulties and workstation problems in some labs that caused some frustration, but on average, the 3DAT and practice trials took about 40 minutes to complete. About 10% of participants did not provide consent for their data to be used for research purposes.

Results

Regarding the first part of this study, Figure 4 shows the results for the new subtests MC, BR, ED and TR subtests. 95% confidence intervals are shown for each subtest and chance scores are indicated as a separate plot. The criteria was simply that the confidence interval range should be above chance but below ceiling. Subtests that meet these criteria were considered definite possibilities for the 3DAT. For all four subtests, the confidence intervals were well above chance and satisfactorily below ceiling. However, two ranges (MC and TR) are approaching a ceiling effect. The results from this investigation provided sufficient evidence to pursue the suitability of these subtests further with results being reported in later studies.

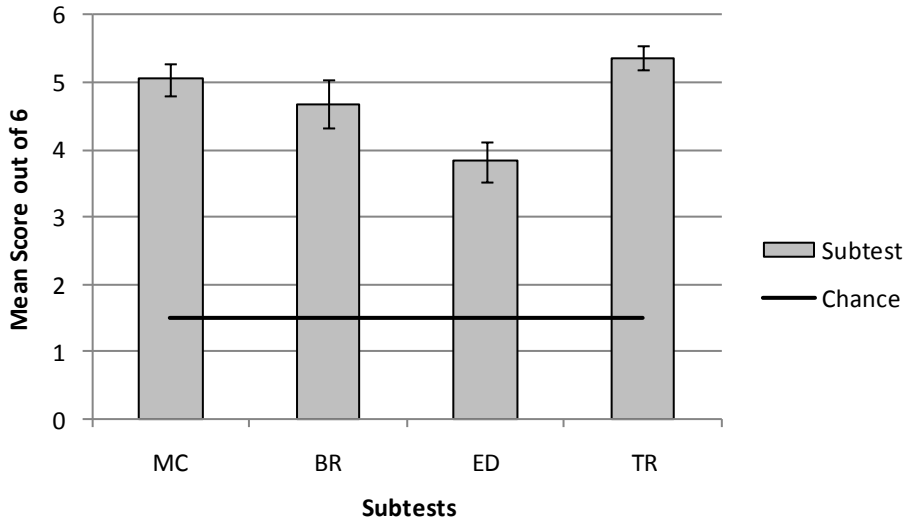


Figure 4. Mean correct and 95% Confidence Intervals for each new subtest under consideration. Means are scores out of 6. Chance scores are indicated for each subtest. Meaning of subtest abbreviations shown in Table 11. Sample (MC = 46, others = 63).

For the second part of this study, which was concerned with the different standards of test items from the same classification, comparisons were made between Trial 1 and Trial 2 for the

subtests in common (DC, TL, VZ & FU). Standards for test items were different for each of these subtests in each trial except for DC. That is, test items in the DC classification were the same for Trial 1 and for Trial 2. The DC subtest became the point of reference (control) so that benchmarking of the remaining three subtests could occur. For this part of study 4, *t* tests were conducted to see if there were significant differences between the different sets of test items used in the two trials. The results of the *t* tests are shown in Table 12 where it can be seen that differences in means are significant for all subtests except DC. This exception was expected.

Table 12
Means for 4 Subtests in common to Trial 1 and Trial 2 are shown. Different Standards of Test Items Apply Except for DC. Mean Differences and significance are also Shown.

	Trial 1		Trial 2		M diff			
Subtest	M	SD	M	SD	$T_1 - T_2$	df	t	p
DC	2.57	1.87	2.65	2.00	.08	107	.227	.82
TL	5.41	.686	2.68	1.12	2.73	107	15.74	< .001
VZ	4.46	1.33	1.54	1.05	2.92	107	12.83	< .001
FU	5.52	.584	4.27	1.29	1.45	107	7.89	< .001

Note. Trial 1 (n = 46), Trial 2 (n = 63). Maximum Score Possible = 6

The research question for this part of the study assumed no differences in sets of data, that is, the null hypothesis was (DC1 – DC2) = (VZ1 – VZ2) = (FU1 – FU2) = (TL1 – TL2). If any of these differences were not equal, it meant that at least one of the three subtests (VZ, FU or TL) changed between Trial 1 and Trial 2 more than DC did. Since the test items for DC were the same for both Trial 1 and Trial 2, then any difference between the results for trial 1 and trial 2 for VZ, FU or TL compared to DC would indicate that VZ, FU or TL had changed relative to DC, and this change would most likely be due to the test items not being the same. The only statistical test that is relevant to answering the research question is the interaction between subtests and trials which assesses the difference of the differences across four subtests. Hence the main effect for these was not relevant on this occasion.

A mixed RM ANOVA was conducted to test if the differences between Trial 1 and Trial 2 varied across subtests. The within subjects factor was *Subtests* (DC which was the control subtest and VZ, FU and TL were the treatment subtests), and the between subjects factor was *Group* (Trial 1 and Trial 2). The interaction *Subtest * Group* was significant using Greenhouse Geisser adjustment $F(2.53, 271.14) = 33.8$, $p < .001$. Hence, the null hypothesis was rejected indicating that the differences between Trial 1 and Trial 2 varied across the four subtests. This was then examined with three follow-up RM ANOVAs to compare the control subtest DC with each of the other three subtests. In each case the interaction terms were significant, FU $F(1, 107) = 17.1$, $p < .001$, VZ $F(1, 107) = 57.9$, $p < .001$ and TL $F(1, 107) = 53.0$, $p < .001$. This demonstrated that each of VZ, FU and TL had changed more than DC. A statistical summary of the RM ANOVAs is provided at Appendix C.

Discussion

The focus of study 4 was mostly on evaluating a number of new subtests regarded as possibilities for the 3DAT, and comparing sets of test items from the same classification to decide if the difference in their properties was significant. In effect, this study was fundamentally about item sampling and selection. Gregory (2004) nominates item selection as a potential source of measurement error and considers that particular sets of test items are not necessarily fair to all test takers. This alone justified the investigation of new subtests and having a close look at the impact of different sets of test items on performance. The plan was to come up with a number of viable subtests and to rely later on a more indepth psychometric analysis to identify the subtests best suited for the 3DAT. A good selection of diverse subtests was seen to be crucial if spatial factors were going to be identified. For test items, this study established that the difference in performance was significant where the standard of test items varied. As a form of review of main outcomes, the two aspects of study 4 (subtests and test items) are summarised below.

First of all, a range of subtests were needed so that optimum choices considered likely to identify spatial factors could be made. This meant that some preliminary evaluation beforehand was necessary short of item and factor analysis to allow initial identification to occur. A variety of screening methods are possible and some have been reported in other chapters, but for study 4, the criteria for accepting or rejecting a new subtest was simple and uncomplicated. Although straightforward, the assessment was nonetheless objective and in accordance with a statistical criteria. Study 4 was seen as a lead-in study to the bigger studies that would follow and was ideal for testing the potential of subtests not considered previously. With 95% confidence intervals for all new subtests being above chance and below ceiling, there was every good reason to take them to the next level of investigation.

Secondly, part two of this study established that a variation in the properties of test items did make a difference to performance. Three subtests, with the more difficult test items in Trial 2 and benchmarked against a subtest where the test items were the same for both trials, revealed a significance difference for the differences in the means. The attempt to vary the standard of test items for Trial 2 (i.e., make them more difficult) was clearly achieved for all three subtests with a value of $p < .001$ for all three comparisons and the possibility of achieving values of this order by chance on three occasions in the one study is very unlikely. The critical thing established here was that the actual test items themselves are important, yet their psychometric properties are rarely reported. As a consequence, there is not always certainty that different research projects report the same test items although the classifications (subtests) may be the same. This is particularly vital where comparisons are being made. This seems a weakness in the literature because it suggests that the actual test items are not critical, or that the actual test items reported

in comparative studies are assumed to be the same. This provides argument that more should be said about the actual test items in any research publication where spatial measurement is being reported. This is a reminder of the position taken by Gregory (2004) on item selection who also maintains that test developers have a duty to minimise unwanted confounds from poor item selection by carefully addressing the different stages of test development. His concern is that a test is only ever a sampling of a test taker's total knowledge, and is therefore prone to measurement error. Hence, the importance of making every effort to eliminate the various risk factors. Apart from the concern of measuring ability correctly, there is also the matter of researchers building on the research of others. Unless there is certainty about the actual test items in any measuring instrument, there should always be some doubt about research outcomes. Cohen, Swerdlik and Smith (1992) refer to the variation of test items between tests as *content sampling*, and use this to explain how test takers can achieve better results on one test compared to another although both claim to measure the same ability. Cohen et al. contend that this situation is possible simply because of the characteristics of the test items in the first test compared to the second test and add that this is one cause of error variance. In simple terms, researchers may be erroneously comparing spatial tests that use different test items although they are from the same classification with the same label (e.g., mental rotation).

Although not a comprehensive study, these trials nevertheless did provide a useful reminder of the importance of having a process in place to identify appropriate subtests and to evaluate test items under consideration. In so doing, the objectives of study 4 were largely achieved.

Study 5 Interviews with Subject Matter Experts

An important part of validating a new ability test is to consult with subject matter experts (SMEs) to seek their opinion on a range of related issues. In the case of the 3DAT, the thoughts of SMEs about possible subtests to measure spatial skills were crucial. The process of seeking and analysing the views of SMEs is another form of validity called *content validity*. Study 5 was entirely dedicated to collecting data and comments from SMEs about 25 subtests that were mostly identified in the scientific literature, and in many cases, also trialled in earlier studies reported in this thesis. One subtest (TL) was purpose-designed because it was thought to capture an important concept in technical drawing. Content validity in its broadest definition is focused on the content of a test and whether it elicits responses representative of the total mixture of skills that a test is thought to measure (Aiken, 1997). Adding to this, Aiken makes two very important points. First, he states that if SMEs agree that a test looks and performs like an instrument developed to measure what it is intended to measure, then the test is assumed to contain content validity. Aiken's second point is that the process of content validation should not be delayed until the development of the test has been completed. Instead, Aiken considers that expert opinion about the suitability of test items is required throughout the test construction

phase. Although the opinions of SMEs were not canvassed during the entire development of the 3DAT, they were sought and acted upon at a critical stage and well before development had been completed. Content validity is more than face validity (impression) and is a concept very relevant to achievement and ability tests. The 3DAT belongs to this group of tests.

Cohen et al. (1992) agree with Aiken (1997) that the content validity of a test is decided by how well it covers all aspects of the behaviour it is designed to measure. They discuss *blueprinting* as a procedure to ensure a test addresses all the important issues. Blueprinting might be described as pooled information from a variety of sources including SMEs, and from this, a test structure emerges. Cohen et al. outline one method of quantifying the degree of agreement among SMEs, and this produces a *content validity ratio* for an item being considered for the test. This method is based on a formula that takes into account the number of experts who agree that an item is essential to measure the skills required, and also the number of experts who disagree. A calculated benchmark value needs to be exceeded for this ratio to be regarded as significant. For example, where the number of experts is 10 and they are asked to make a judgment about whether or not an item is essential, that benchmark is calculated and found to be .62. Using this formula and where say 9 experts agreed an item was essential and one disagreed, the formula would return a content validity ratio of .80. Since .80 is greater than .62 and seen as significant, that item then would be regarded as having a sufficient degree of content validity. The explanation here of this method of quantification is intentionally brief to raise awareness that a statistical approach can be applied to evaluating content validity. Cohen et al. cite Lawshe (1975) as the advocate of this method and go into more detail in their coverage of this approach to quantifying content validity. For a construct such as spatial ability, a test would be regarded as being *content-valid* if there was a strong consensus among SMEs that it sampled spatial skills appropriate to designers.

Gregory (2004) is in agreement with both Aiken (1997) and Cohen et al. (1992) and points out that content validity is really no more than a sampling exercise and cites Bausell (1986) to support this position. Gregory explains that items that make up a test can be seen as a sample from a population of items that typify what a test developer wants to measure. The ideal solution is to capture examples representative of the full range of behaviours a test aims to measure and this is possible when a lot is known about what is being measured (e.g., arithmetic). However, for less tangible measures such as spatial ability, there is more reliance on informed opinion because of the impracticality of trying to sample the full spectrum of possibilities for something not as clearly defined as arithmetic. In such situations, Gregory is clear and suggests that content validity is the considered opinion of SMEs. In reality, Gregory maintains that a test developer essentially declares a panel of experts have reviewed the specifications and deemed the test to have content validity. Gregory also comments on the

quantification of content validity. He refers to the Lawshe (1986) model and similar approaches but regards these as specialised and not widely accepted. However, he considers they serve as a *commonsense* approach to collating expert agreement as a basis for content validity. Interestingly, Gregory outlines a quantification model not too dissimilar to Lawshe's model and it too is based on item ratings provided by SMEs. This method also uses a simple formula to calculate an index ratio called the *coefficient of content validity* and this is similar to the *content validity ratio* produced by the Lawshe formula. Importantly, Gregory advises his readers that the coefficient model is just one contribution of evidence in the assessment of a test and it alone is not enough to establish its overall validity. The *commonsense* approach to quantifying content validity reported here and advocated by various researchers has one distinct advantage of culling items identified as unsuitable by SMEs. However, one weakness is that it does not identify items that should be included in a test to make it more representative of the domain it targets (Gregory, 2004). This needs to be done by other means such as literature searches and the careful selection of ideas generated by practitioners in the workplace.

Content validity for the 3DAT was largely established from the opinions of SMEs after analysing the data they provided. However, there was not a total reliance on one panel of SMEs since individual SMEs not part of the panel also made contributions in what might be described as *lead-in participation*. The approach to quantifying content validity in this thesis (explained later in this chapter) was different to that described by Gregory (2004) and reported for Lawshe (1986). All methods, however, shared a common goal of implementing a procedure based on some form of statistical procedure to ensure objectivity. With respect to the method used in this thesis, one thing that was different to the reported methods was that a qualitative dimension was added to the process. As a consequence, the quest to establish content validity for the 3DAT took on two distinct characteristics. First, quantitative data were collected and used to establish a ranking of subtests being considered for the 3DAT. The second characteristic was the collection of qualitative data from SMEs that related to subtests and spatial ability in general. The quantitative data were generally more informative than the qualitative data. At this juncture, it is timely to point out that the coverage of content validity so far has focused on *item* as an alternative term for *question* or *task* that would be used in a test under development. For the 3DAT, the equivalent to *item* is *subtest*, and from this point on, subtest will substitute for item in any discussion about content validity since this is the focus.

What the authors who are reported in this section have in common is that they support the idea of SMEs helping to decide about the content validity of a test under development. They agreed that it was an important process and they were in favour of a statistical procedure to derive a criteria for accepting or rejecting subtests. Hence, the objectives of study 5 were to:

- identify and rank a range of subtests according to the opinions of SMEs, and

- collect and evaluate comments from SMEs about the composition of the 3DAT.

The careful management and systematic application of a scheme to elicit ratings and comments from SMEs was paramount to the success of developing the 3DAT. While content validity was not the only validity to be concerned with, it was nevertheless fundamental to the directions the 3DAT would take and the number of subtests that would finally be analysed. The protocol that was implemented to achieve the objectives was based largely on two procedures that could be described as *informed subjectivity* and *statistical analysis*. The approach was a mixture of both but the lead procedure was informed subjectivity closely supported by the statistical analysis. The application of these two procedures and their relationship are described in the paragraphs that follow. Note that many of the points raised are expanded upon in either the *Methodology* or *Results* sections of this study.

Methodology

25 possible subtests were presented to a panel of SMEs. Many subtests were identified largely from the literature, but key personnel with similar expertise to the panel members (e.g., this writer with industry experience, and a leading design academic with an interest in spatial ability) narrowed the selection to 25 subtests. As well, several of the subtests were tested in a number of previous studies and were reported earlier. The SMEs provided a score for each of the subtests using a five point Likert scale, and were invited to comment on each subtest and a number of spatial issues in general. The 25 subtests are listed in Appendix D and examples are shown in Appendix E. References for the subtests are also shown. Appendix F provides the questions asked for each of the Likert Scales. Both parametric and nonparametric tests were conducted on the data produced from the Likert scales, and comments were compiled and organised into categories. The 25 subtests were a starting point and the aim was to reduce this total to an effective number appropriate to the 3DAT at that stage of its development.

Because of the specialised nature of this study, participants who had professional experience in design were recruited from a range of disciplines such as engineering and architecture to achieve a representative sample. Participants were recruited from both academia and industry and were first identified through networks, affiliations and professional associations. An individual approach rather than a general call for volunteers was seen as more effective because of the need to recruit participants with a sound knowledge of spatial ability. Potential participants were contacted in writing initially and provided with an information pack and an invitation to participate. This was followed up with a personal phone call or an office visit. Recruitment was generally difficult because of poor availability of potential participants rather than any indifference towards the study. 15 participants (all male) took part in the study and the majority (80%) came from academia though most had professional experience outside of the

higher education sector. It was not possible to recruit female participants and this was both a disappointment and a limitation for this study. It would have been meaningful, for example, to have been able to assess whether females differed in opinion to males about critical tasks. The sample size was less than that targeted (30) but acceptable under the circumstances. The literature supports the use of a small sample in qualitative research where the purpose is to provide specialised information that can only come from a select group of skilled individuals (Willis, 2005). In these circumstances, a sample of between 5 and 15 participants is common (Willis, 2005). This view is supported by Kvale (1996) who explains that researchers can be guided by two principles when the purpose of the interview is information gathering. The two principles are: interview as many subjects as necessary to find out what you need to know, and conduct 15 interviews plus or minus 10. One strategy for overcoming the limitations of a small sample size was proposed by Spittaels and Bourdeaudhuij (2006). For multiphased research, they advocate using a larger sample in a subsequent phase to compensate for the smaller sample used in an exploratory stage. This offset occurred as part of this research in a number of studies, but in particular, a sample of 635 was achieved in a study reported in chapter 4. Finally, Willis (2005) and Kvale (1996) suggest that, should the collected information start to become repetitive after say 15 to 20 interviews, the interview process may cease at that point. In effect, this became a benchmark for this study and this repetition did start to occur around the 15th interview mark. Had this not been the case, recruitment, though difficult, would have continued.

The procedure was a structured interview process where each SME was consulted individually and asked questions about spatial ability and the 25 subtests in particular. The idea was to ascertain an industry perception of spatial ability relevant to design education and to identify spatial skills considered essential to the discipline. In addition to the standard demographic questions, the interview schedule consisted of two main sections. The first asked specific questions (7 in total) about spatial ability, while the second was concerned with scoring and commenting on the 25 subtests. SMEs recorded their score for each subtest on a Likert scale and this provided quantitative data to allow a ranking of the subtests to be determined. Comments about subtests and spatial ability in general were combined to produce the qualitative data. In essence, the interview schedule was really a detailed questionnaire which was treated as a script by the interviewer who maintained a strict protocol to ensure consistency for all SMEs. A copy of the questionnaire can be found at Appendix F. The reader is also reminded about Appendix E which contains examples for each of the 25 subtests. The interviews took approximately 60 minutes to complete and they were conducted in locations convenient to the SMEs. Notes were taken during the interviews and later transcribed to a database. Comments were also audio-recorded as a form of backup but also to clarify and expand on any note taking that took place. The qualitative data were manipulated and organised using Nvivo, a software package dedicated

to managing and categorising qualitative data. Likert scores were first recorded on paper and then converted to a computer data file so that a statistical analysis could take place. Both parametric testing (repeated measures ANOVA) and nonparametric testing (Friedman's test) occurred as *play safe* measures because of some doubt about normality in the data. This is possible when the sample size is low and as a consequence of using Likert scores since the data are often skewed and generally regarded as ordinal measurements. This approach of using parametric and nonparametric tests explains the method of quantifying content validity (degree of agreement) employed in this study which sets it apart from the quantification methods described by Gregory (2004) and proposed by Lawshe (1986).

Results

For the quantitative data, a parametric test (RM ANOVA) was applied to the 25 subtests to check for a significant difference in the rankings allocated by the SMEs. Significance was not confirmed by the RM ANOVA (main effect of subtests) using Greenhouse Geisser adjustment $F(6.18, 86.52) = 2.091$, $p = .06$. Because of there being some doubt about the normal distribution of the data, a nonparametric procedure (Friedman Test) was applied as a precautionary measure. If the data were normally distributed, then general agreement could be expected between the two methods of testing. The Friedman test is the nonparametric equivalent to the one-way within-subjects analysis of variance (RM ANOVA). This test was applied to the data and showed that the opinion of the SMEs varied significantly across the 25 subtests with the Friedman's statistic being returned as: $\chi^2 (24) = 52.7$, $p = .001$. The *monte carlo* statistic was also calculated and returned a value of $p < .001$. This added confidence to the asymptotic result since both p values were near the same. There would be concern, however, if the two values were substantially different. In such a case, the monte carlo would be given priority. In a sense, the difference in p values found for the parametric and nonparametric tests was a good thing because it highlighted the danger in using a parametric test only. This difference suggested that the data were not normally distributed and the preference would therefore be given to the nonparametric result. The nonparametric test, in showing a disagreement among SMEs about the ranking of the subtests was not ideal, but good in a way since it meant that some subtests could be eliminated because of their low ranking by SMEs. 25 subtests were always going to be too many for practical purposes, but the quandary was, how many should be retained and which ones should be discarded. While a RM ANOVA will show descriptive statistics for any analysis, a rank order is not displayed. To achieve rank order, a descriptive statistics procedure was conducted to produce the mean scores for each subtest and to place them in descending order. The nonparametric results were arranged in order of mean ranking provided by that procedure. The two methods of achieving a ranking are essentially different, but interestingly, the two rankings of the 25 subtests only differed marginally. Details of the statistical procedures

and the rankings of the subtests produced by both methods are shown at Appendix G. The next step was to decide on what subtests should be eliminated, and whether this would show an improvement in consensus among SMEs.

The SME ranking of the 25 subtests based on the mean rank scores estimated by the Friedman nonparametric test is shown in Figure 5.

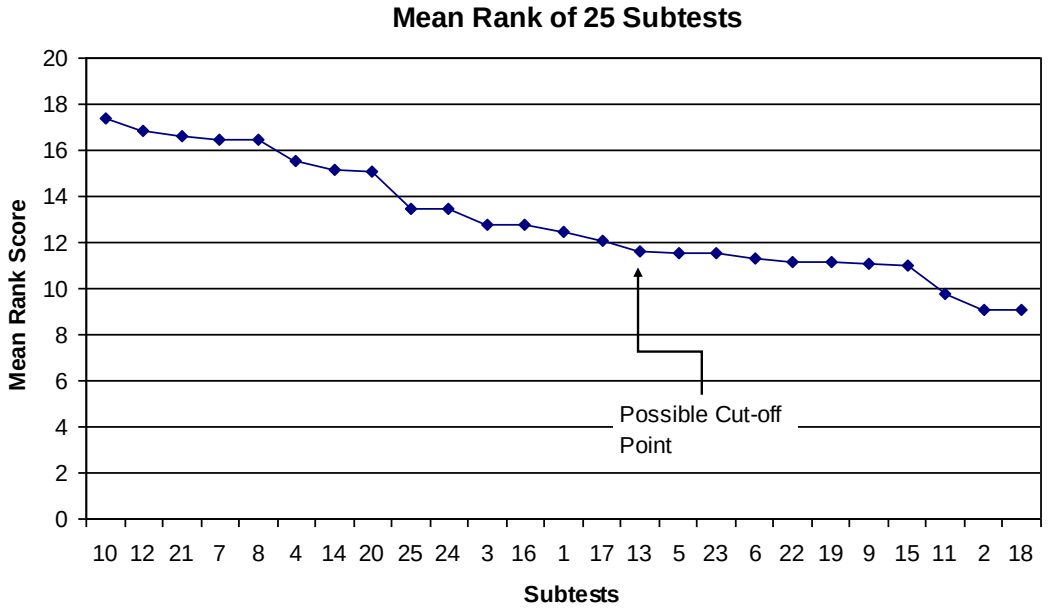


Figure 5. Mean rank scores for 25 subtests determined by SMEs and produced by Friedman nonparametric test. Key – 10: Building Representations 12: 2D to 3D Recognition 21: 2D to 3D Transformation 7: 3D Mental Rotation 8: Engineering Drawing 4: Mental Cutting 14: True Length Recognition 20: Surface Development 25: Mental Rotation 24: Mental Rotation 3: Surface Development 16: Dot Coordinate 1: Paper Folding 17: Mental Rotation 13: Correct Fold and Surface Development 5: Cube Construction 23: Building Representations 6: 2D Mental Rotation 22: Building Representations 19: 3D Mental Rotation 9: Space Relations Task 15: Possible/Impossible Structures 11:Water Level 2: Form Board 18: Cube Comparison. *Task* is synonymous with *subtest.*

Informed subjectivity was used to decide which subtests should be discarded. Subtest 13 shown in Figure 5 appeared to be a possible turning point in the graph since a flattening out of the plot begins, suggesting that subtests to the right of this point could be eliminated. As well, any subtest with a mean of less than 4.0 calculated from the Likert scores was deemed to have a mean too low to be accepted. The main effect for *subtest* produced from the RM ANOVA and represented graphically in Appendix G will make it more obvious to the reader why the value of 4.0 was chosen. Applying both criteria meant that 10 subtests would be eliminated and 15 would be retained for further consideration. One further step was required, and this was to decide if any one of the 15 subtests was measuring the same skill as any other subtest. A panel of two with expertise in spatial ability (including the writer) considered this question. As a consequence, a further three subtests were eliminated because of parallels in what they measured with other subtests. The three subtests eliminated were: task 1, task 3 and task 17 respectively. Task 1 (paper folding) was dropped in favour of task 13 (correct fold) because task

13 contained an obvious 3D element and was better suited to designers. Task 3 (surface development) was dropped in favour of task 20 (a second surface development task) because Task 20 was ranked higher by SMEs and it too was better suited to designers. Task 17 (mental rotation) was eliminated in favour of task 25 (a second mental rotation task) because Task 25 was ranked higher by SMEs and was more challenging than Task 17. Task 25 involved rotation about 3 axes, whereas task 17 involved rotation about one axis. To recap to ensure clarity: the rejection of three subtests was seen as highly desirable because of the similarity in the skills measured in 3 separate pairs of subtests (i.e., task 1 versus task 13, task 3 versus task 20 and task 17 versus task 25). The tasks eliminated were subtest 1, subtest 3 and subtest 17, noting that *task* is synonymous with *subtest*. The 3DAT now consisted of 12 subtests ready for further analysis.

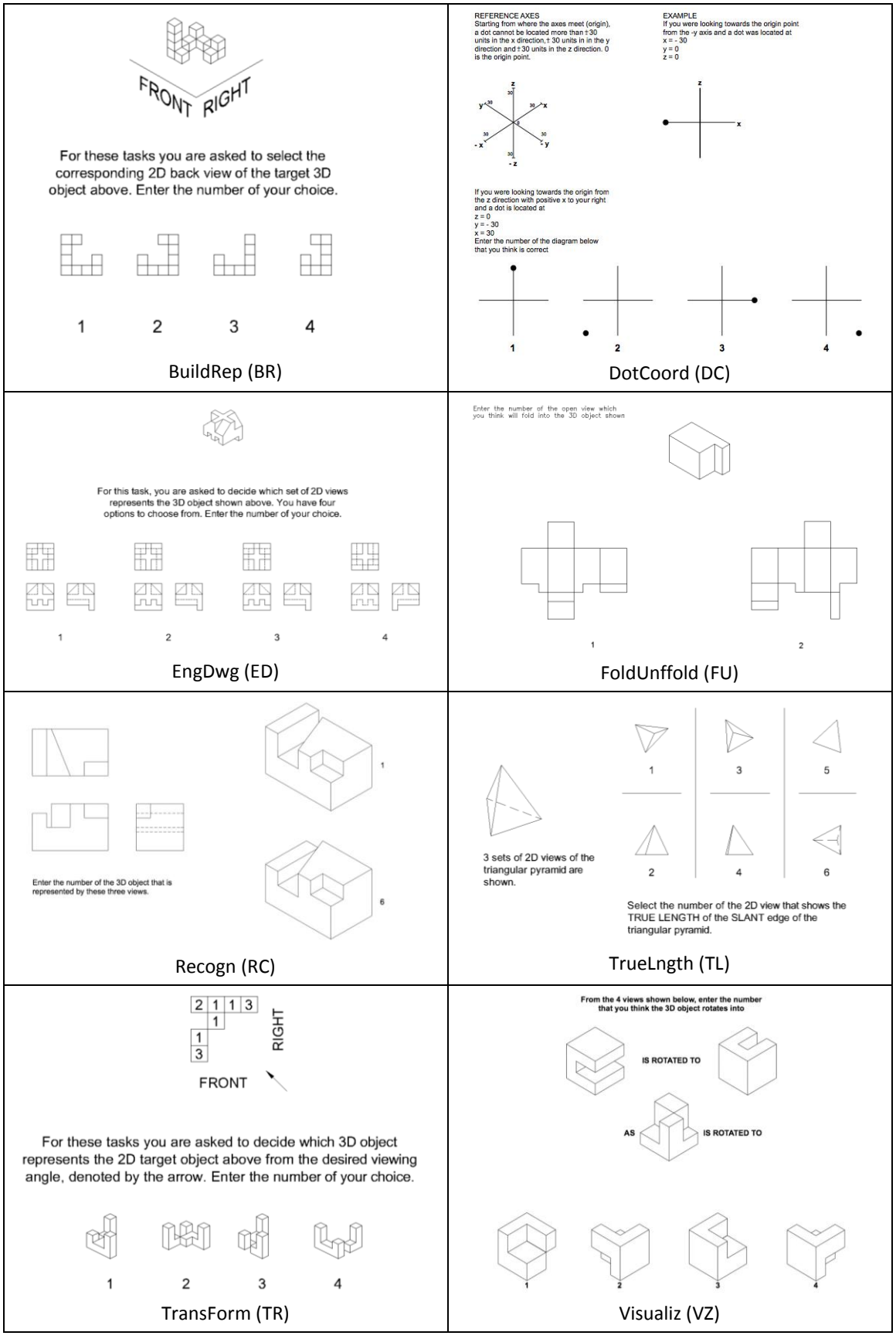
The statistical tests applied to the 25 subtests (parametric and nonparametric) were also applied to the 12 subtests. The idea was to see whether the tests would reveal a change in agreement among the SMEs about the 12 subtests compared to the 25 subtests, and to reassess the ranking of the subtests according to the Friedman test. The main effect of *subtests* from the RM ANOVA was not significant using Greenhouse Geisser adjustment $F(3.92, 54.84) = 1.14$, $p = .347$. Similarly, the Friedman test also showed that the opinion of the SMEs did not vary significantly across the 12 subtests with the Friedman statistic reported as: $\chi^2(11) = 13.3$, $p = .271$. The *monte carlo* statistic was also determined and it too was not significant with a value of $p = .272$. This again supported the asymptotic result since there was very little difference in the two p values. Hence, both parametric and nonparametric tests provided evidence of agreement among the SMEs about the ranking of the 12 subtests. In these circumstances, it is a good result if both tests failed to show a significant difference because the preferred outcome is agreement rather than disagreement. These findings were in contrast with those found for the 25 subtests, although that too was a desirable outcome on that occasion. Table 13 shows the ranking of the 12 subtests according to the Friedman method. However, both statistical methods produced a similar order of subtests with only slight differences in some positions. The second worksheet at Appendix G provides further details and the rank orders determined by both statistical tests.

Table 13
Rank Order of 12 Subtests According to the Friedman Nonparametric Statistical Test

Rank	Subtest	Code	Rank	Subtest	Code
1	10 Building Representations	BR	7	14 True Length Recognition	TL
2	12 2D to 3D Recognition	RC	8	20 Surface Development	SD
3	8 Engineering Drawing	ED	9	25 Mental Rotation	MR
4	21 2D to 3D Transformation	TR	10	24 Mental Rotation	VZ
5	7 3D Mental Rotation	RT	11	16 Dot Coordinate	DC
6	4 Mental Cutting	MC	12	13 FoldUnFold	FU

Note. Code = abbreviation for subtests.

To provide extra clarity, picture examples of the 12 subtests listed in Table 13 are shown below in Figure 6.



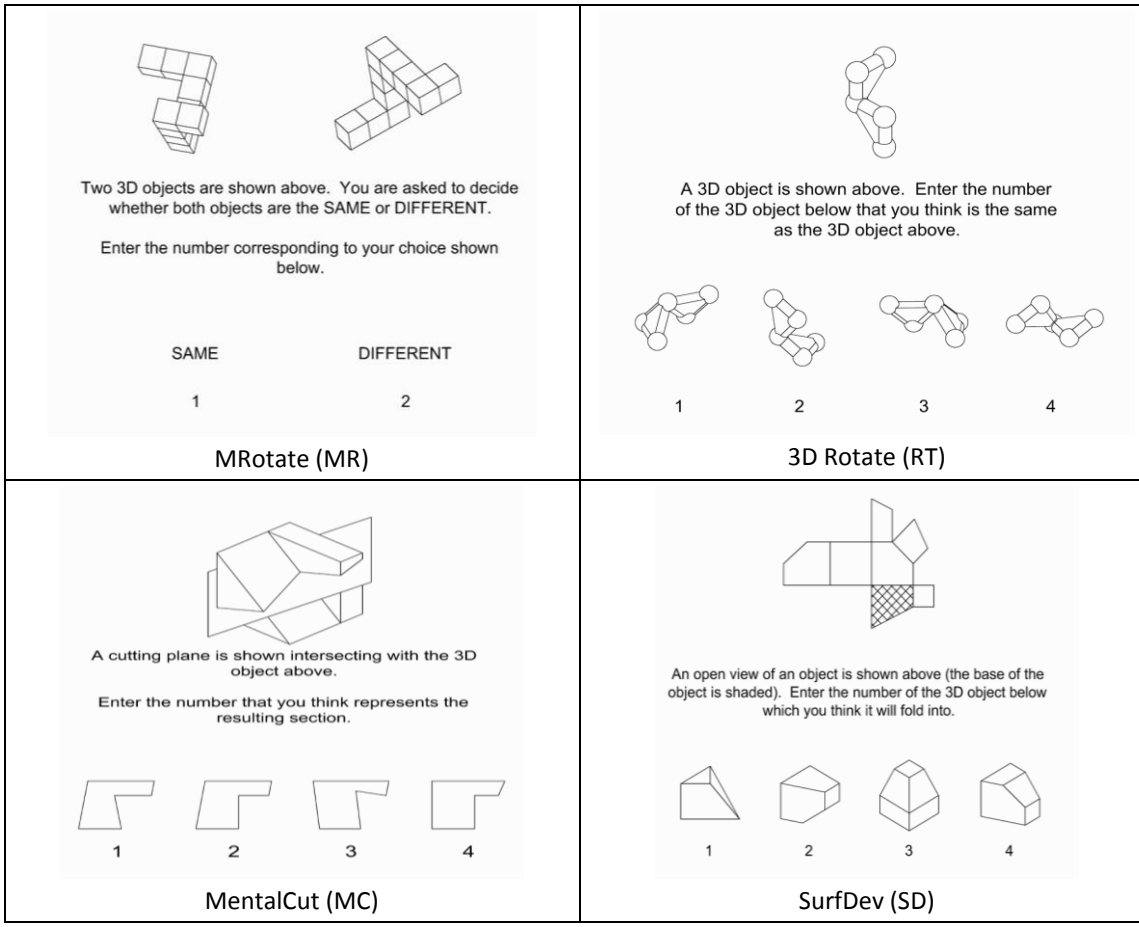


Figure 6. Examples of test items from 12 subtests that were identified and ranked by SMEs. The ranking given to the subtests is shown in Table 13.

The second part of this study was concerned with the collection of comments from SMEs about the 25 subtests and spatial ability in general. Their comments were directed towards specific areas and the interviewing style was standardised (scripted) to ensure the approach was the same for all SMEs. Appendix F provides details about the process followed for the SMEs. The SMEs were given every encouragement to make their thoughts known, and there was no limit on the number of comments they could make. This evoked comments outside of the targeted areas, but all comments were recorded, analysed and placed into appropriate categories. The process produced 1440 comments from 15 SMEs contained in a 27 page document and divided into nine categories and 22 subcategories overall. Some comments were placed into several categories because they matched the criteria of a number of searches that interrogated the Nvivo database. Some comments are not necessarily self-explanatory, but they are the raw comments recorded by the interviewer and they accurately reflect what SMEs had to say. The qualitative material was not expected to be as revealing as the quantitative data, but nevertheless, it was considered likely to add meaning to the quantitative findings reported earlier. All 1440 comments grouped into the various categories and subcategories are shown at Appendix H.

Table 14 provides a list of the categories and subcategories and the number of responses for each. The large number of responses shown (i.e., 109 to 224) are for categories that embraced

all 25 subtests, and therefore, the responses are associated with any or all of these subtests. Also, where there are no subcategories such as *agree* or *disagree*, this is an indication that the database searches did not identify comments that fitted these criteria. The details in Table 14 clearly show agreement that spatial ability is important, that comments on the subtests are generally positive, and that there is mostly agreement that spatial ability is lacking in the workplace as a skill. For the most part, there appears some consensus that spatial ability is viewed as an innate skill. Somewhat contradictory, Table 14 also indicates that a number of SMEs thought that spatial ability could be developed. Importantly, there is also some evidence that SMEs generally agreed that spatial ability is not specifically being taught in the higher education sector. The question of gender difference is interesting. Research on this issue is generally reported as showing a bias towards males. However, it seems like SMEs either think a gender difference does not exist, or they are uncertain about this. For the category focused on the 3DAT, the comments on a whole are quite satisfactory. That is, they are mostly supportive, or they offer ideas for improving some aspect of the subtests such as: labelling, distracter tasks, the number of answer options, degree of difficulty and about the general instructions that advise test takers what is required of them. In some cases, comments are very general and reflect the SME's own exposure to the subtests. Occasionally, the relevance of a subtest is questioned, and also whether or not a subtest is measuring a spatial skill or some other cognitive skill. For the one remaining category (assessment of spatial ability), the comments are similar to those just reported. That is, they are mostly supportive of the idea of assessing spatial ability, but there are other comments that could be described as *comments in passing*. One difference to the previously mentioned category is that comments in this category (assessment of spatial ability) focused specifically on individual subtests, and not the 3DAT as a whole. As a reminder to the reader, the comments for all categories and subcategories are shown at Appendix H.

Table 14

Categories, Subcategories and Number of Responses from SMEs.

Category	SubCat	Resp	Category	Subcat	Resp
Importance of Spatial Ability	Agree	77	Lack of Spatial Ability in the Workplace	Agree	24
Comments on Subtests	Positive	206		Disagree	3
	General	224		Unsure	9
	Suggestions	109	Spatial Ability is an Innate Skill	Agree	20
Comments on 3DAT as a 25 Subtest Model	General	223		Disagree	9
	Constructive	161		Suggestions	39
	Suggestions	109	Can be Developed	General	14
Assess. of Spatial Ability	General	160	Spatial Ability is being Taught	Agree	1
Gender Differences	General	7		Disagree	8
	Yes	6		Unsure	4
	No	10			
	Unsure	17			

Note. Total number of comments: 1440. Some comments qualified for several categories. The number of subcategories varies according to category. Resp = Responses.

A number of comments from SMEs had a direct impact on the shaping of the 3DAT, while others suggested ideas for future research. To provide examples, Table 15 shows a range of benefits to emerge from the survey, and a selection of the comments that helped define these.

Table 15

Examples of Benefits to the 3DAT and the Spatial Ability Research Area as a Consequence of SME Comments. Also Examples of Contributing Comments are Shown.

Benefit	Contributing Comment
Established the Importance of Spatial Ability	People lacking in spatial skills would not find employment in design fields
Indicated Value of Measuring Spatial Performance	If we had measurement tools like this we would be able to pick up problems early
Supported Relevance of 2D and 3D Relationship	Translating/correlating 2D to 3D is core
Mentioned of Other Important Spatial Skills	Translation of volume is an important issue and how plans relate to volumes is an issue
No Strong Evidence that Spatial Ability is being Taught	Spatial skills are required but not taught
Emphasized Uncertainty about a Gender Difference	It surprised me to learn of the finding that women have less spatial skills than men, I didn't realise it before
A Belief that Spatial Ability can be Developed	It can be learned/taught/improved/trained
Comments on the Number of Distracters	If you put in more distracters it would be harder
How to Improve Subtests	More credibility if a bit harder / too simple
Conceptual Issues highlighted	Some people can't see the true length of things, they don't understand what the concept means

The compilation of comments from SMEs was not without its difficulties, but the advantages outweighed the disadvantages.

Discussion

At the most basic level, the aim of study 5 was to identify the level of content validity that existed for the 3DAT at this stage of its development. To achieve this, it was necessary to measure the degree of agreement among a sample of professionals from the discipline of design. The content validity of a test is described as *the extent to which it covers all aspects of the construct it is designed to measure*. It was important to base the investigation of agreement on a statistical process to reduce subjectivity, and in this case, the approach involved both parametric and nonparametric procedures to assess the quantitative data collected. Although the methods are slightly different, it is a desirable outcome if both parametric and nonparametric outcomes basically agree. However, where there is some uncertainty, and normality is in doubt, it is better to trust nonparametric results. Essentially it is about using all the information available to make an informed decision. This became the quantification method for this study which is different to the methods advocated by the researchers reported earlier. However, all methods have in common the same ideals and objectives. Consistent with this, the idea was to produce a 3DAT containing a sample of items representative of the population of items that match the construct of spatial ability. The process of establishing content validity not only verified suitable subtests for the 3DAT, but it also made it obvious which subtests should be discarded.

In essence, the process of evaluating the opinions of SMEs produced a number of specific outcomes. First, it provided convincing evidence of the subtests that should be included in the 3DAT to ensure content validity. Second, it produced a ranking of subtests from say essential to nonessential. Third, it encouraged the use of statistical methods to objectively test agreement among a panel of experts. Fourth, it provided confidence in a 12 subtest model because of the endorsement received from those experts. Fifth, it revealed uncertainty among design professionals about gender differences, despite there being evidence from research of a gender difference favouring males. Supposedly, there are disadvantages for female novice designers if this difference is not recognised.

The ratings of SMEs varied significantly for the 25 subtests, but not for the 12 subtests. For the 25 subtests, the evidence of disagreement (i.e., a significant difference in opinion) was considered desirable for practical reasons because it meant that the number of subtests could be reduced. However, evidence of agreement (i.e., no significant difference in opinion) was hoped for at some stage since it would be less than ideal if SMEs could not agree on some number of subtests. Agreement occurred for the 12 subtest model. Naturally it was important to be able to move to the next stage of development with evidence of agreement about subtests from SMEs.

In a sense, the actual ranking of the subtests from 1 to 12 was academic since all subtests would be part of the next phase of evaluation. However, the ranking did reveal the order of importance of the subtests according to SMEs. What was more than academic, however, was the change from disagreement to agreement by SMEs when the 3DAT reduced to 12 subtests. Since significance was not found for this model, it meant that SMEs agreed about the choice and order of subtests. The ranking of subtests produced a few surprises. The top four (BR, RC, ED & TR) for example (refer Table 13) was not expected, but because of the 2D 3D relationship common to each of these subtests, this ranking was understandable. As an aside, the 2D 3D relationship rated highly in the comments from SMEs. On the other hand, two subtests clearly requiring some form of mental rotation (MR & VZ) did not rank as high as expected. Of note is mental rotation is probably a spatial skill required in most spatial manipulations. One other subtest (DC), considered to require a combination of spatial skills, did not rank highly either. Nevertheless, there were no real surprises with the actual 12 subtests chosen, only with the rankings they received. The 12 subtests embraced a mixture of visual skills, which meant there was a good chance that some would measure the spatial factors thought to collectively represent spatial ability.

For the qualitative data, the feedback about the 3DAT and spatial ability in general justified the survey undertaken with SMEs. With the benefit of hindsight, more questions could have been directed to particular concepts rather than allowing open comments intentionally encouraged by the interviewer. There should be provisions for open responses in any future studies, but not in an uncontrolled way since the compilation of qualitative data is more problematic than the compilation of quantitative data. Though there were 1440 comments, it was not feasible to group these into categories and subcategories other than those listed in Table 14. That is, most comments were stand-alone and could not be meaningfully classified into further categories or subcategories. However, considered individually, many comments reflect the thinking of individual SMEs, and it was therefore important to list all of them in Appendix H. Since the categorisation of comments was problematic, it made the statistical analysis and the reporting of findings more difficult than expected. However, the qualitative data collected for this research was additional to the quantitative data that would normally be collected when testing for content validity, and thus it provided information that would not be obvious from a totally quantitative approach. Importantly, comments from SMEs endorsed a conscious effort to measure spatial ability, and also spatial ability as a credible area of research. In view of these outcomes, the objectives of this study were achieved.

Study 6 Design and NonDesign Groups Compared

This study primarily investigated a further validity described as *validity with known groups* which was first introduced in study 2. To recap, a test designed to measure a specific construct

such as spatial ability should show higher scores for persons high in that ability when compared to persons considered low in that ability. For the 3DAT, this means that designers on average should perform better than groups who are likely to be low achievers on this test (e.g., health workers). Obviously, there would be a discomfort about a test developed for designers if the performance of the two groups was not significantly different and not biased towards the design group. As previously reported, Gregory (2004) refers to this validity as *theory-consistent group differences* and states that one way to increase the validity of a new instrument is to demonstrate that persons with diverse backgrounds and experiences on average will achieve theory-consistent scores on that test. Cohen et al. (1992) also acknowledge this form of validity but refer to it as *evidence from distinct groups,* though they also raise the alternative name of *the method of contrast groups*. The actual name given to this validity by different researchers is not as critical as the concept itself. However, a variety of descriptors does help convey what this validity represents and what judgment it makes. Cohen et al. are consistent with Gregory and explain this validity in a similar way. They state that for a test to be a valid measure of a particular construct (e.g., leadership), then people from groups assumed to differ on that construct will also have a correspondingly different performance on that instrument. Since this validity would be addressed a number of times in this study, and since there are several names that could be used, it made sense to settle on a particular name for convenience. For simplicity, this validity is referred to as *theory-consistent validity* for the most part of this chapter.

Study 6 also investigated three additional but lesser research questions. First, the performance of female participants was compared to the performance of males within two diverse groups. Gender was examined again to test for further evidence of bias but this time with a larger sample, better subtests and with two groups who were distinctly different in terms of their expected spatial understanding. Second, the standing of the test items underwent scrutiny using a method of item analysis called *item response theory* (IRT). As previously reported, this method is gaining in popularity among researchers and is an alternative to the traditional method known as *classical test theory* (Gregory, 2004). IRT was an important first step because it identified underperforming test items that were removed to ensure a more accurate evaluation of the 3DAT could occur. Third, the internal consistency (reliability) of the test items within each subtest was measured and reviewed. Although consistency is difficult to achieve under the best of circumstances because of potential measurement error, it was important nonetheless to gauge the level of reliability existing in the 3DAT during this stage of its development. Reliability can be seen as the repeatability and dependability of test scores and how well the test items correlate positively with each other. Reliability was first introduced in study 1 and later in study 3, and was further investigated in this study under better conditions. It is also covered more widely in chapter 4.

This study was implemented to examine a type of validity not previously applied to the 3DAT. It was also about revisiting gender differences and evaluating the psychometric properties of test items within the 3DAT. Hence, the objectives of study 6 were to:

- investigate theory-consistent validity for the 3DAT by comparing two diverse groups,
- examine gender issues in spatial performance across two diverse groups,
- conduct item analysis to evaluate test items, and
- assess the internal consistency of test items in each of the subtests.

This study centred on two dissimilar groups of participants classified in terms of their assumed spatial understanding and prior learning experience. One group was regarded as *skilled* because of their expected high level of spatial ability and they were called the *design* group. Participants in this group were novice designers. The second group of participants was regarded as *unskilled* because of their assumed low level of spatial ability. They were called the *nondesign* group and participants came from the humanities disciplines. Having two diverse groups meant it was possible to compare and contrast their performances on 12 subtests and potentially reveal something new about spatial skills. Included was the mental rotation subtest which is often reported as a task that females have the most difficulty with (e.g., Holliday-Darr et al., 2000). Both groups consisted of male and female participants.

The process of item analysis encompasses the measures of item discrimination, item difficulty, item reliability and item validity which are the tools of the psychometrician. The IRT method of item analysis is also referred to as *latent-trait theory*. The procedure is applied to each individual test item and produces an *item characteristic curve* (ICC). The ICC is a graphical display that represents item difficulty and item discrimination and is plotted on a graph where *ability* is measured on the horizontal axis and the *probability of a correct response* is measured on the vertical axis. The quality of the test item is indicated by the slope and position of the ICC graph. A good item is defined by a curve roughly shaped like an *S* that slopes upwards to the right with a lower asymptote at zero and an upper asymptote at one. A steep curve demonstrates high item discrimination while a flat curve demonstrates low item discrimination. The position of the curve along the horizontal axis is a measure of item difficulty. Curves positioned midway along the horizontal axis indicate the acceptable level of item difficulty. Curves to the left of centre reflect easier test items and curves to the right of centre reflect harder test items. The ability scale plotted on the horizontal axis is a distribution of the total scores for all test items in a particular test, not just the individual test item itself. An example of an ICC graph is shown at Appendix I along with source information. IRT is also part of study 9 and certain features of this method are covered in more detail there. The main point at this time is that IRT is different to

the traditional classical test theory (CTT) model and was used in this research as a complement to CTT rather than as a competitor.

Methodology

Participants in this study were university students mainly in their first year of enrolment. For the *design* group, Students with prior spatial learning experience were recruited by placing invitations on noticeboards throughout an engineering and built environment faculty. Recruitment information was also disseminated verbally by course coordinators and promoted using an online student and course management system. Other avenues such as student union magazines, web sites and international student bulletins were also utilised. Potential participants were provided with an information statement and possible times for participation. A signed consent form was required before any participation was possible. In effect, participants in the design group could be described as novice designers who came from disciplines such as engineering, architecture, mechatronics and construction management. For the *nondesign* group, psychology students with no prior spatial learning experience were sought through an online research participation management system. Potential participants were able to log onto the system and sign up for the study after sighting an information statement. Again, a signed consent form was required before participation in the study was possible. The design group consisted of 42 participants (30 male, 12 female) who received financial reimbursement for out of pocket expenses. The nondesign group was made up of 56 participants (24 male, 32 female) and they received course credit for their participation. Allowing for administration and practice trials, the 3DAT took about 75 minutes to complete. Study 6 also collected demographic information such as gender, discipline, experience and age group.

This study was the first implementation of the 12 subtest version of the 3DAT (72 test items) and it was conducted in a lab setting using Superlab software that was described in earlier studies. The 3DAT was also supervised and delivered under similar conditions reported in those studies. The two letter labelling system for subtests (e.g., BR) first mentioned at the start of chapter 3 continued into this study. A list of the abbreviations and what they stand for are shown in Table 13. As well, examples of each subtest are shown at Appendix J.

The analysis conducted included IRT which was used to reduce the number of test items in the 3DAT. Also included was a full factorial 12 * 2 * 2 (subtest * group * gender) mixed RM ANOVA design where *subtest* was a within-subjects factor and *group* and *gender* were between-subjects factors. Estimated marginal means were examined to test mean differences, and reliability was estimated using Cronbach alpha. Effect size was re-introduced to show the magnitude of mean differences for key research questions. The calculation of effect size was based on within population SDs and calculated either one of two ways, depending on whether

population SDs were similar or dissimilar. If SDs were similar, the method used was that advocated by Borenstein, Hedges, Higgins, and Rothstein (2009). Where the SDs differed greatly, then the method used was that recommended by Cohen (1988). The formulae used to calculate both methods are shown in the *Terms and Definitions* section of this thesis.

Results

The IRT method of item analysis was applied to the 72 test items across both groups with the intention of identifying poor performing items. Consistent with the principle of IRT, an ICC graph was produced for each test item with overall ability plotted on the horizontal axis and the probability of correctness plotted on the vertical axis (refer Appendix L). The idea was to discard the weak items and reduce the size of the 3DAT accordingly. Some subjectivity and compromise was necessary to satisfy a criteria of retaining the same number of test items in each of the subtests. From a research perspective this may not have been necessary, but from the experience of earlier studies, it was always easier to report data and communicate information about the 3DAT if there was a consistency about its structure and the number of test items in each subtest. After reviewing the ICC graphs, a decision was made to reduce the 3DAT from 72 to 48 test items and to conduct further analysis based on this number. In effect, this meant the 3DAT now consisted of 12 subtests with 4 test items in each. Regarding the ICC plots, it came down to retaining the test items with curves that best matched the ideal shape for an ICC graph. The reader is reminded that an example of an ICC graph that represents a good test item can be seen at Appendix I. Many of the ICC graphs were less than ideal, but the nearest to ideal were selected keeping in mind that 4 test items were required for each subtest. Appendix K lists all test items and indicates those that were rejected. Appendix L shows the ICC graphs produced for each of the test items which were created using JMP statistical software.

For the 48 test item 3DAT, a mixed RM ANOVA was applied to the data to measure the significance of main effects and interactions. The main effect for *subtest* was significant using Greenhouse Geisser adjustment $F(9.10, 855.69) = 45.89$, $p < .001$, thus indicating that the type of subtest had an impact on performance.

With respect to theory-consistent validity, the interaction between *subtest* and *group* was significant using Greenhouse Geisser adjustment $F(9.10, 855.69) = 2.57$, $p = .006$. This indicated that the differences between the design group and the nondesign group varied significantly across the subtests. Estimated marginal means were examined to explore the interaction *subtest* by *group* for each of the 12 subtests. Results are shown in Table 16. Note that full names for subtests are shown in Table 13.

Table 16
Estimated Marginal Means for Design and NonDesign Groups for 12 Subtests.
Differences of the Means, SE of the Difference of the Means and Significance are also Shown.

	Design		NonDesign		M diff	SE	
Subtest	M	SE	M	SE	D – N	D – N	*p*
BR	3.41	.188	2.70	.149	0.71	.24	.004
DC	3.62	.190	2.42	.150	1.19	.24	< .001
ED	2.42	.191	1.70	.151	0.71	.24	.004
FU	3.08	.186	2.17	.147	0.92	.24	< .001
MC	2.53	.198	1.87	.157	0.66	.25	.011
MR	3.27	.167	2.89	.132	0.38	.21	.077
RC	3.77	.134	3.03	.106	0.74	.17	< .001
RT	2.56	.185	2.22	.146	0.34	.24	.152
SD	3.71	.115	3.29	.091	0.42	.18	.005
TL	1.33	.161	1.28	.127	0.04	.21	.831
TR	3.58	.157	3.27	.124	0.31	.20	.125
VZ	2.92	.173	2.47	.137	0.45	.22	.045

Note. Design (n = 42), NonDesign (n = 56). Max. Score Possible = 4.

Details in Table 16 indicate a significant difference in performance between the design group and nondesign group for 8 out of the 12 subtests. Further, the main effect for group was significant $F(1, 94) = 23.62$, $p < .001$, thus reflecting an overall difference in performance between the two groups where the means were design = 3.01 and nondesign = 2.44. Vital to this study, all outcomes reported here were highly favourable for establishing the existence of theory-consistent validity for the 3DAT because the design group consistently scored better than the nondesign group.

Turning to gender results, the mixed RM ANOVA did not show a significant interaction for *subtest* and *gender* using Greenhouse Geisser adjustment $F(9.10, 855.69) = 1.24$, $p = .265$. This indicated that the differences between males and females did not vary significantly across the subtests. However, this result virtually mandated a check of the main effect for *gender* produced by the mixed ANOVA. A significant main effect for *gender* was found $F(1, 94) = 9.74$, $p = .002$, thus highlighting a difference in performance between the two genders. Interestingly, this significance did not apply to both the design and nondesign groups because of the significant *group* by *gender* interaction $F(1, 94) = 7.02$, $p = .009$. That is, for the design group, the difference in the means (male = 3.04, female = 2.99) was not significant ($p = .764$). However, for the nondesign group, the difference in the means (male = 2.78, female = 2.10) was significant ($p < .001$). This comparison suggests that the significant main effect for gender came from the nondesign group only. These results can be seen in Appendix M.

Aspects of reliability were described in earlier studies and concepts such as definitions, scales and specific characteristics were addressed. In this study, Cronbach alpha coefficients (α) were

calculated to show reliability for each of the 12 subtests used in this version of the 3DAT. Results are shown in Table 17.

Table 17
Cronbach Alpha Reliability Coefficients for 3DAT Subtests

Subtest	Alpha	Subtest	Alpha	Subtest	Alpha
BR	.71	MC	.53	SD	.25
DC	.77	MR	.37	TL	.31
ED	.47	RC	.49	TR	.55
FU	.55	RT	.45	VZ	.47

Note. Alpha overall not reported for 3DAT. Refer Chapter 5 for explanation.

As previously reported, benchmark standards for Cronbach alpha coefficients are defined as acceptable where α values are greater than 0.7, and very acceptable where values are greater than 0.8. Test items with α values approaching zero are seen to have very poor reliability qualities. Based on these standards, the α values in Table 17 are not ideal with only two subtests exceeding the 0.7 benchmark, and several falling into a low category of under 0.4. The study was below a sample size generally considered suitable for testing reliability, but still large enough to provide a reminder that this important attribute needed to be considered further. Internal consistency is treated more critically in study 9 where a sample of 635 was achieved.

One final thing to report is *effect size*. Though effect size can be applied to any two conditions where the difference in means is found to be significant, it was only applied to the 3DAT in four key areas to demonstrate the extent of the differences. Details are reported in Table 18.

Table 18
Paired Samples for the 3DAT Where Effect Size was Calculated. Means and SDs determined from Independent *t* tests.

	Sample 1		Sample 2		M diff		
Paired Samples	M	SD	M	SD	$M_1 - M_2$	*p*	*d*
Design – NonDesign	36.31	4.82	28.71	8.45	7.60	< .001	1.10
Male – Female (overall)	35.11	6.08	28.11	8.51	7.0	< .001	.95
Male – Female (design)	36.50	4.80	35.83	5.04	.667	.690	.14
Male – Female (nondesign)	33.38	7.11	25.22	7.73	8.16	< .001	1.09

Note. Effect size calculated for 3DAT only. Max. score possible = 48. *p* = probability, *d* = Cohen's d

Results from a *t* test do not include a measure of the effect but this can be established by calculating the effect size (Cohen's *d*). This provides a measure of the degree of difference in the two means in terms of standard deviations (Brace, Kemp, & Snelgar, 2009). Applying the criteria for effect size of $d = .20$ (small), $d = .50$ (medium) and $d = .80$ (large) introduced in earlier studies, values in Table 18 reveal large effect sizes for three out of four of the paired samples. Considering each in turn, this could be expressed as a large effect size was found for the *design – nondesign* pair in favour of the design sample, a large effect size was found for the *male – female (overall)* pair in favour of the male sample, and a large effect size was found for

the *male – female (nondesign)* pair favouring the male sample. Strictly speaking, Cohen's *d* would not normally be calculated for the *male – female (design)* pair since $p > .05$. However, it is reported here to allow a comparison with the other key areas, and to show that the difference in performance between males and females in this group was very minimal. This contrasted quite considerably with the large performance difference found between males and females in the nondesign group.

Discussion

In brief, this study was multifocused and investigated several important concepts in psychometrics. IRT was implemented to test the strengths and weaknesses of individual test items even though the sample size was less than an ideal for item analysis. This study in particular tested *theory-consistent* validity for the 3DAT with two very different groups in terms of their assumed spatial knowledge. One group was defined as skilled, while the other was defined as unskilled. Psychometric theory suggests that the skilled group would perform significantly better on the 3DAT if it measured the spatial skills they were thought to have. Gender as an important issue returned in this study to reveal a number of findings deserving of further consideration and raised several research questions for the future. Reliability as a measure of internal consistency of the individual test items within the 12 subtests was also assessed. Coefficient alpha indexes were produced to gauge the strength of this internal consistency which is essentially a measure of how well the items positively correlate with each other. The inferences and implications from these procedures are described in detail below.

The IRT procedure produced *easy to understand* ICC graphs (see Appendix L) where it was generally quite obvious which test items should be rejected. However, quite often it was difficult to decide between several items because of varying evidence of poor psychometric properties. To probe further, IRT was also applied separately to each group (design and nondesign) as an added investigation to help decide which items should be rejected. This turned out to complicate things because redefining the sample generally produced different shaped ICC graphs. Consequently, some subjectivity was necessary in the final decision making. Although IRT improved the 3DAT by reducing the number of test items, the subjectivity required tended to foster a preference for the CTT method of item analysis for this test developer. In a later study where the sample size was better suited to item analysis, a more substantial application of IRT was possible.

Paramount in this study was an impetus to gauge the level of theory-consistent validity evident in the 3DAT. That is, do those expected to do well on a purpose-developed test on average do better than those not expected to do well. Statistics reported in the *Results* section certainly supported this. Not only did a main effect for *group* reveal a significant difference on the 3DAT

with a bias towards the design group (those expected to do well), but 8 out of the 12 subtests in all cases showed a significant difference favouring the design group. To add weight to these findings, the effect size reported in Table 18 highlighted the extent of that difference. Thus, for design versus nondesign, the mean difference between samples was 7.60 and the 95% confidence interval for the population mean difference was between 4.91 and 10.28. The effect size was d = 1.10. Viewed against the benchmark indexes where d = .80 is large, this result provided good evidence of *theory-consistent* validity for the 3DAT. Importantly, all appropriate evidence produced by this study supported this validity.

The results for gender were interesting. When the whole sample was considered simply as males versus females and ignoring for a moment the groups they belonged to, a significant difference in performance on the 3DAT and a bias towards males was found. The mean difference between the genders was 7.0 and the 95% confidence interval for the population mean difference was between 3.96 and 10.04. Table 18 shows that the effect size was d = .95. Again, effect size was able to demonstrate the strength of the mean difference, and in view of the benchmark indexes, this result clearly indicates that the difference in gender was large. However, when the sample is divided into separate groups (design and nondesign), and only considering effect size on this occasion, a significant difference (male vs female) was found for the nondesign group (d = 1.09) but not for the design group (d = .14). Refer Table 18. This raises some interesting research questions. For example, is the often reported gender difference based on samples where there is a difference in sociological and early adolescent experiences, and if so, is this a fair comparison? On the other hand, are these results showing that females can improve given sufficient training since the design group would have received training in technical drawing and graphical communication courses during their undergraduate programs. In other words, achieving similar scores across genders for spatial ability may be possible, suggesting that spatial ability may not necessarily be an innate ability only. Another consideration is *self-selection*. That is, do both males and females with spatial aptitude tend to gravitate to careers where spatial understanding is an attribute? If so, then it is possible that no difference in spatial performance will be found between the genders. Further, self-selection samples are unlikely to be representative samples of the population.

Coefficient alpha indexes were disappointing for the most part indicating that many items should be reworked, however, some caution should be exercised in view of the sample size. Though reasonable for many studies, a sample of 98 falls short of an estimated sample of 240 needed for reliability testing of an instrument consisting of 48 test items. That is, five test takers per test item using Tabachnick and Fidell (1996) as a guide. A further consideration is that reliability is one of several attributes, and an otherwise good subtest should not necessarily be rejected where coefficient alpha is less than ideal. In essence, it comes down to the purpose of

the test and how critical the results might be. Kaplan (1987) argues that a test developer must decide if an increase in reliability is worth the extra time, effort and expense. He considers that it may only be worthwhile where it affects personnel decisions and it would therefore be risky to not try and improve reliability. In applications where the test is used for noncritical purposes, Kaplan suggests that the gains may not justify the costs involved. This implies that a test developer should not be too hasty in discarding subtests that fall short of a particular index number. However, this is not to dismiss the importance of reliability. A test that is not reliable is also not that valid in the broad meaning of the word.

The objectives of this study focused mostly on assessing several psychometric properties of the 3DAT as it stood at a particular stage in its development. The study was ambitious in a sense because it investigated a mixture of research questions with essentially a diverse group of participants. However, the undertaking was regarded as successful. An extra validity was established for the 3DAT, test items were evaluated and gender issues were identified. In view of these outcomes, study 6 was regarded as having achieved its objectives.

Chapter Summary

This chapter reported on the transitional phase in the development of the 3DAT and was particularly concerned about a range of psychometric properties. Studies in this phase investigated two forms of validity, test item characteristics and reliability as a measure of internal consistency between test items in each of the subtests. There was also an opportunity to test for gender difference in one study, although this was not its purpose, but made possible because of the experimental design. Participant demographics varied across the studies but for the most part, participants had some experience in spatial ability and could be classified as designers in some capacity. One study involved subject matter experts while another included a nondesign group made up of participants who were deemed unskilled in spatial understanding. Except for the subject matter experts, the samples included male and female participants. Achievements resulting from these studies included the establishment of content and theory-consistent validities for the 3DAT, and the application of the IRT method to identify weak test items that could be discarded. In addition, measures of coefficient alpha showed that some subtests met benchmark standards for reliability while others fell short of these. The testing of gender issues expanded on earlier studies in this thesis and agreed to some extent with the literature, but also challenged it other aspects. A case in point is the mental rotation task which is often found to be a task that females have trouble with. However, this study did not find consistent evidence to support this. The difference in gender performance in the design group reported in study 6 is intriguing since it showed no significant difference between male and female participants. This was not the case for the nondesign group where a significant difference was found. The literature generally suggests that there will be a difference in favour

of males, and sometimes questions the potential to improve spatial performance (especially for females) because of a perception that spatial ability may be an innate ability. Results for the design group certainly suggest that females can perform as well as males and that improvement is possible. Otherwise, the significant gender difference found for the nondesign group should also be evident in the design group. However, to be objective from a research perspective, the reason for not finding a significant gender difference in the design group could be the result of students self-selecting into academic courses.

There were several broad issues to emerge from these studies that are very relevant to psychometric test development. From the outset it was clear that many test items would need to be trialled with the expectation that many would be rejected before a viable number of test items of sufficient standard would be apparent. Gregory (2004) explains that it is common practice to prepare a surplus of test items perhaps double that anticipated with an expectation that this number will greatly reduce. This was the case for the 3DAT during these studies, and it was likely that it would be repeated in a later stage of development. A concern for any test developer is *error in measurement*. The issue here is that it is very difficult if not impossible to derive a true score of ability from any form of psychometric test. Gregory maintains that the best that can be expected is a good estimate and the prediction that the true score exists within a certain range. There are many sources of measurement error but according to Gregory, the four most likely sources are: *item selection, test administration, test scoring and systematic errors of measurement*. The aim then was to minimize the impact of each of these to the highest degree possible. Two of these sources were addressed in earlier studies and continued into this development phase. *Item selection* is the first of these and the attention given to item analysis and item sampling were the best safeguards against this source of error. The second source was *test scoring* and it was largely counteracted by the design of the 3DAT itself. Measurement error in test scores generally occur when some level of subjectivity is required to make a judgment about an answer given by a test taker. The 3DAT test items are not in this category and require no more than a correct/incorrect response. For the lab version of the 3DAT that operated during this phase of development, the response required was simply a single stroke of the keyboard. Furthermore, the responses are computer-managed where results are stored as electronic data files which removes another form of measurement error. Attempts to reduce measurement errors need to be ongoing throughout test development, and these two sources are revisited in chapter 4. Importantly, the remaining two sources identified by Gregory and the total issue of measurement error are overviewed in chapter 5. One side effect that a test maker must consider is the impact of measurement error on reliability. Tests that are flawed will return inconsistent results which is opposite to what reliability is all about. More is said about reliability in later

chapters. The main intention here is to raise this important consideration known as *error in measurement* and the difficulty it poses for achieving a true measure of ability.

This *transitional* phase also hosted a mixture of special outcomes that were important to the testing and evaluation that would occur in the final studies. It first of all trialled a simplified labelling system for subtests and test items that improved verbal and visual reporting of information, data and results. This system was adopted and remained part of the 3DAT from this point on. This transitional phase was also the last time the 3DAT was delivered on a platform dedicated to stimulus presentations suited only to laboratory setups. After this, the 3DAT was converted to an online version and tested in an open environment where global access was possible, though login permission was controlled by the test developer. The superlab studies were complete and the lessons learnt assisted in the design and implementation of the online version. IRT procedures reduced the 3DAT to 48 test items leaving a residual of 24 test items that could be reworked to improve their psychometric properties if necessary. These were reworked and went forward to the next set of studies. A final issue concerned the true length (TL) subtest. Any testing of the 3DAT prior to study 6 included a set of TL test items that attempted to measure this elusive concept. Almost all of the testing produced better results than expected, possibly because participants were able to derive the correct answers using a process of deduction without really understanding the concept of true length. Study 6 saw the introduction of a set of *very difficult* TL test items to counter this, but they turned out to be too difficult for participants from any persuasion. Interestingly, the TL subtest was ranked 7 out of 25 by the SMEs referred to in study 5 which is a measure of the importance they gave to this concept. It was felt that testing of this subtest should continue into several future studies before any reworking was contemplated. The measurement of TL seemed appropriate for a test of spatial ability for designers, but the best items for this test were yet to be identified.

CHAPTER 4

FINAL DEVELOPMENT

Overview

The studies reported in this chapter were classified as *final* because they were concerned with the final assessment of the psychometric properties of the 3DAT based on lessons learnt and as a consequence of having a good sample size. In many respects, these studies represented the culmination of studies reported in previous chapters, and in effect, they stood as the final evaluation of the 3DAT. Part of this included the investigation of one remaining though minor validity study to complete the overall assessment of validity. This final development was also the point where the 3DAT first operated as a fully functional online test although some *testing of the test* still needed to be done. This position was reached after many modifications of the interface design and reviews of the programming options. The 3DAT was almost fully established in a psychometric sense, but what still needed to be done was to decide on the final form of the instrument. That is, how many subtests were required to make up the 3DAT, what spatial skills would they measure, how many test items were necessary and exactly which ones should they be. The 3DAT online was developed using Adobe Flex which is a software development kit for the preparation and deployment of cross-platform rich internet applications based on the Adobe Flash platform. Adobe Flex is used to build web applications suited to all major browsers, desktops and operating systems.

This chapter consists of three studies labelled study 7, study 8 and study 9 respectively. In brief, study 7 primarily investigated *test retest reliability,* but it also produced a benchmark index able to be used to help calculate actual learning beyond the *learning effect* that accompanies practice. Study 8 reports face validity based on quantitative and qualitative feedback received from design students who participated in one study. Study 9 was particularly important given that item and factor analyses were possible because of the large number of participants who took part in this study. Study 9 was also important for other reasons. It consisted of more objectives than the other two studies and these were intentionally diverse to investigate a mixture of research questions. To achieve these objectives, different procedures for item analysis were implemented, gender issues were investigated again, RT was revisited with a hope of increasing knowledge about spatial factors, and general academic ability was considered as a factor in spatial performance. The two main aims of chapter 4 were to decide what the final composition of the 3DAT should be in terms of subtests and test items, and essentially to bring this research to completion.

Study 7 Test Retest Reliability

Study 7 primarily investigated a different form of reliability than those covered in earlier studies known as *temporal* reliability, and this reliability was measured using the test retest method. In practical terms, this came down to giving the 3DAT to an appropriate selection of test takers a first and second time with a reasonable interval in between. To determine this form of reliability, correlation procedures are carried out and the value of the correlation coefficient index (*r*) indicates the strength of the linear relationship between the two test results. The correlation coefficient in some contexts can also be treated as a *reliability coefficient* and used as a measure of consistency of test scores (Gregory, 2004). Essentially this means that if the scores on one test are highly consistent with the scores on the same test given to the same group a second time, then a strong positive correlation will exist between the tests, perhaps approaching the upper limit of +1.0. This concept introduces an important approach to assessing the reliability of an ability test under development. In reality, it was expected that the results for the second application of the 3DAT would show improved scores because of practice. This is generally referred to as the *learning* or *practice effect* and is a common phenomenon that occurs when the same test is given a second time to the same recipients. This is particularly so for ability or achievement tests such as the 3DAT. However, this is not generally a problem because an improved performance on the second attempt does not impact on the reliability of the instrument provided both attempts are strongly correlated. In other words, if an ability test is ideally reliable, then the scores on the second administration of the test should be highly predictable from the scores on the first administration of the test (Gregory, 2004).

Another approach to testing temporal reliability is known as the *alternate forms* method and it involves the administration of two forms of the same test that were developed to satisfy the same specifications. Both forms cover the same content, and test items are comparable in terms of item difficulty and other psychometric properties. When given to the same group of test takers and administered using a counterbalanced design, statistics such as means and standard deviations turn out to be quite similar (Gregory, 2004). One downside to this method, however, is being confident about test item equivalence otherwise described as *item sampling*. Item sampling differences are inherent in the alternate forms method of testing reliability. Cohen et al. (1992) make the point that test takers may perform well, or possibly not so well on a particular version of a test, although not because of their ability, but because of the actual test items selected for the test. This can be one source of measurement error. Of note is that item sampling is not a concern for the test retest method since the same test items are used in both administrations of the test. One further disadvantage of the alternate method is cost. It is simply too time consuming and labour intensive to develop two parallel tests because of the tedious *testing of the test* procedures that must take place. However, one advantage over the test retest

method is that the practice or learning affect is reduced, although this is not generally an issue unless for some reason achieving similar means on the two tests is important.

Because of its *gold standard* status (Gregory, 2004), and in view of the disadvantages of using the alternate forms methods for temporal reliability, the test retest method was selected for this study. Hence, the objectives of study 7 were to:

- assess the reliability of the 3DAT using the test retest method, and
- establish a standard for practice effect that should be discounted when determining actual learning that has occurred in a learning environment.

This second objective is fundamentally about producing an index from the test retest methodology that could be used as a benchmark for the improvement that normally occurs with the second administration of a test given to the same recipients. The intention is for this index to be treated as a *norm* (i.e., a standard) by design educators who would subtract this from any overall evidence of learning to derive the actual learning that has occurred. In the view of this writer, educators quite often mistake learning from practice to be actual learning as a consequence of some formal training, strategy or treatment they have introduced into the classroom. Quite often, these interventions have not in themselves brought about any real improvement in performance. Finally, as points of clarification, the approach to determining reliability reported here is termed the *test retest method,* and the reliability it measures is known as *test retest reliability*. The latter is in the form of an index (0 to +1.0) which serves as an estimate of reliability for the test.

Methodology

Participants in this study were design students from a variety of undergraduate programs. A sample size of 104 (77 male, 27 female) was achieved by combining two groups who satisfied the criteria appropriate to the test retest method. The first group of 54 participants (43 male, 11 female) were participants in a study intentionally developed to measure test retest reliability. The second group of 50 participants (34 male, 16 female) were the control group in a project that investigated the effectiveness of learning tasks to improve 3D understanding. Both groups had very similar backgrounds and the experimental conditions that operated for both groups were identical. Further, the control group did not experience any intervention between pretesting and posttesting in the project they participated in. From this point on, the two groups are treated as one group and are primarily referred to as *participants* or *test takers*.

The 3DAT consisted of 12 subtests and 72 test items and it was delivered as fully functioning online test. It was given to participants on two separate occasions with an interval of one week between administrations. The 3DAT was identical for both administrations. Participation occurred in a research lab in small groups of one to six, depending upon the availability of

participants. Recruitment was similar to that reported in study 6 for the design group, but essentially it came down to advertising through electronic networks and online noticeboards. A study supervisor with appropriate experience was appointed to conduct all sessions and to manage all aspects of study delivery. Practice trials were given to participants to allow them an opportunity to become familiar with the requirements of each subtest. Participants were able to ask clarifying questions but they were not given any strategies to help determine correct answers for any of the test items. Within the 3DAT, each subtest was preceded with an instructional screen that also explained what was required to complete the test items within each subtest. These screens did not time out and remained on display until a participant was ready to move on. Time on task was approximately two hours for each administration of the 3DAT but this included the practice trials and some management time. Participants received financial reimbursement to offset any inconvenience or out of pocket expenses. They were generally well motivated and appeared to see participation as a challenge, and also to appreciate the educational benefit because they recognised the relevance of spatial ability to their discipline studies. The 3DAT generated a performance summary for each individual which they received after completion, and they seemed interested in the results. For the most part, they appeared keen to participate a second time to see if they could improve on their first performance. Participants were supervised for the full period of participation.

The analysis included an inspection of scatter plots to test for a linear relationship and histograms to check for normality and homogeneity in both datasets. These processes helped decide what statistical procedures would be implemented. A *Pearson* correlation analysis followed to measure the strength of the relationship between the two administrations of the 3DAT, and therefore the reliability coefficient in this instance. A *Spearman* nonparametric correlation procedure was also carried out as a precaution. To examine the practice effect, a *paired samples t test* was applied to the two datasets to determine if the mean difference was significant. An equivalent analysis was also conducted at the item level (correct/ incorrect) rather than at the 3DAT level (total scores) to verify findings. That is, contingency tables were generated to ascertain percent agreement and again to determine the practice effect. The *McNemar test for repeated measures* was used to establish whether the practice effect was significant.

Results

A scatter plot is shown in Figure 7 (*test 1* versus *test 2*) and it visually indicates a linear relationship between the two datasets of reasonably high positive strength. Of note is that the scatter plot provides evidence of some outliers at the lower end of the scale. Outliers in this distribution are not strictly outliers in the true meaning of the word, but are treated accordingly because the sample is not evenly spread with only a few test takers represented at the bottom

end of performance. The question is, how representative are they of the population? Notwithstanding, there is a good estimate at the top end indicated by the cluster of data points.

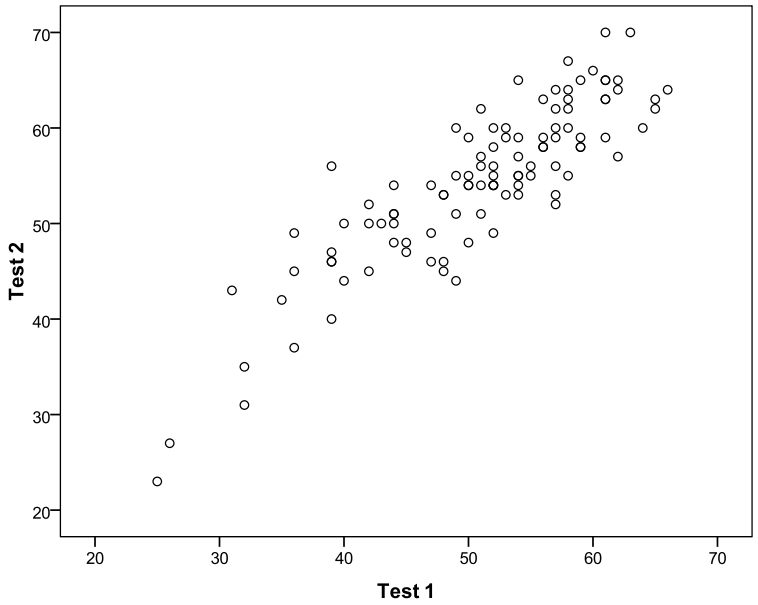


Figure 7. Scatter plot showing relationship between two administrations of the 3DAT to the same test takers with a 7 day separation. Some evidence of outliers are shown in the bottom quarter of the graph.

Histograms produced for both administrations of the 3DAT did not reveal any violations of normality or homogeneity in the data except for some minor skewness to the right for test 1. Descriptive statistics indicate similar distributions for test 1 (M = 50.79, SD = 8.83) and test 2 (M = 54.28, SD = 8.49). To clearly identify the strength of the relationship between test 1 and test 2, correlation procedures were carried out. Though a parametric procedure (*Pearson*) seemed adequate, a nonparametric procedure (*Spearman*) referred to earlier was still carried out to increase confidence in the findings. To take this further, the same procedures were also applied to the datasets with outliers removed (i.e., two test takers). After considering the scatter plot, the criterion for removal was scores less than or equal to 30 on either test. Results are shown in Table 19.

Table 19
Correlation Between both Administrations of the 3DAT and thus
Reliability Coefficients based on the Test Retest Method for Determining Reliability

Procedure	Outliers IN	Sig (2 tailed)	Outliers OUT	Sig (2 tailed)
Pearson	.874**	< .001	.848**	< .001
Spearman	.853**	< .001	.844**	< .001

Note. ** *p* is significant at the .01 level. Outliers IN (n = 104), Outliers OUT (n =102)

Reliability coefficients shown in Table 19 are high in all instances, though slightly less for *Outliers OUT*. Noteworthy is that the indexes came closer together for Pearson and Spearman when outliers were removed. Even though the methods of calculation are different, this result was not unexpected since Pearson is more sensitive to variability than Spearman. With outliers

removed, this sensitivity is reduced. As reported in earlier chapters, reliability coefficients greater than 0.8 are considered very acceptable in psychometric terms.

The paired samples *t* test that was applied to the two datasets revealed that the mean difference between the datasets was statistically significant. Results are shown in Table 20.

Table 20
Means for both administrations of the 3DAT used in the test retest method are shown. Mean difference and significance are also indicated.

Test 1		Test 2		M diff			
M	SD	M	SD	$T_1 - T_2$	df	t	p
50.79	8.83	54.28	8.49	3.49	103	8.17	< .001

Note. Sample (n = 104). Maximum Score Possible = 72

In percentage terms, the mean for test 1 shown in Table 20 is equal to 70.5% (50.79/72*100), and the mean for test 2 is equal to 75.4% (54.28/72*100). Thus, the percentage improvement and therefore the mean difference is equal to 4.9%. These values are important and will act as benchmarks for the 3DAT to help evaluate actual learning in a classroom. Statistical details for the procedures up to this point are shown at Appendix N.

Contingency tables were produced using the cross tabulations procedure in SPSS to determine the level of agreement across the datasets and to investigate whether improved performance overall was significantly different to any reduced performance overall. For this analysis, the results of the 3DAT are expressed as *correct* or *incorrect* (yes/ no) answers for each of the 72 test items. Since the sample size was 104, this meant that 7488 (72 x 104) responses were considered. Table 21 shows the outcome generated from the procedure.

Table 21
Contingency Tables for Test Retest of the 3DAT Showing Different Counts and Totals in Percentage Terms

			Test 2		
			No	Yes	Total
Test 1	No	Count	1313	893	2206
		% of Total	17.5%	11.9%	29.5%
	Yes	Count	530	4752	5282
		% of Total	7.1%	63.5%	70.5%
Total		Count	1843	5645	7488
		% of Total	24.6%	75.4%	100%

Note. McNemar Test of Exact Sig (2 tailed) < .001. No of Valid Cases = 7488

The results shown in Table 21 indicate that 17.5% of responses were incorrect for both tests, while 63.5% responses were correct for both tests. Interestingly, 7.1% of responses indicate a reduced performance on the second test by test takers, however, 11.9% of the responses indicate an improved performance on the second test. Results in Table 21 show percent agreement between the two administrations of the 3DAT to be 81% (17.5% + 63.5%). Furthermore, the McNemar test using binomial distribution indicates a significant difference in the number of

reduced performances compared to the number of improved performances between test 1 and test 2 in favour of the second test (n = 7488, exact p < .001). Details outlining the analytical procedures are shown at Appendix O.

To evaluate the assumption that practice effect could be used to determine actual learning, a full factorial 2 * 2 (*Tests* * *Group*) mixed RM ANOVA was conducted to test if the differences across test 1 and test 2 for two groups varied significantly. The results from the test retest participants were compared with the results from an engineering graphics class that pretested and posttested students using the same version of the 3DAT. The class consisted of 57 design students who had similar backgrounds to the test retest participants (university design students enrolled in the same courses in the same Faculty over the same period). Spatial skills and concepts were an integral part of the curriculum for this class. For the ANOVA, the within subjects factor was *Tests* (test 1 and test 2) and the between subjects factor was *Group* (Norm and Geng). *Norm* represented the test retest sample, and *Geng* represented the students in the engineering graphics class. (The label *Geng* is a course code that stands for *general engineering*.) The main effect for *Tests* was significant using Greenhouse Geisser adjustment F(1, 159) = 123.45, p < .001, thus confirming a performance difference on the two administrations of the 3DAT. However, the statistical test most relevant to the research question is the interaction between *Tests* and *Group* which is a measure of the difference of the differences between test 1 and test 2 across groups. The interaction *Tests* * *Group* was significant using Greenhouse Geisser adjustment F(1, 159) = 14.28, p < .001, thus indicating that improvement beyond practice was statistically evident. A graphical representation of the interaction is shown in Figure 8.

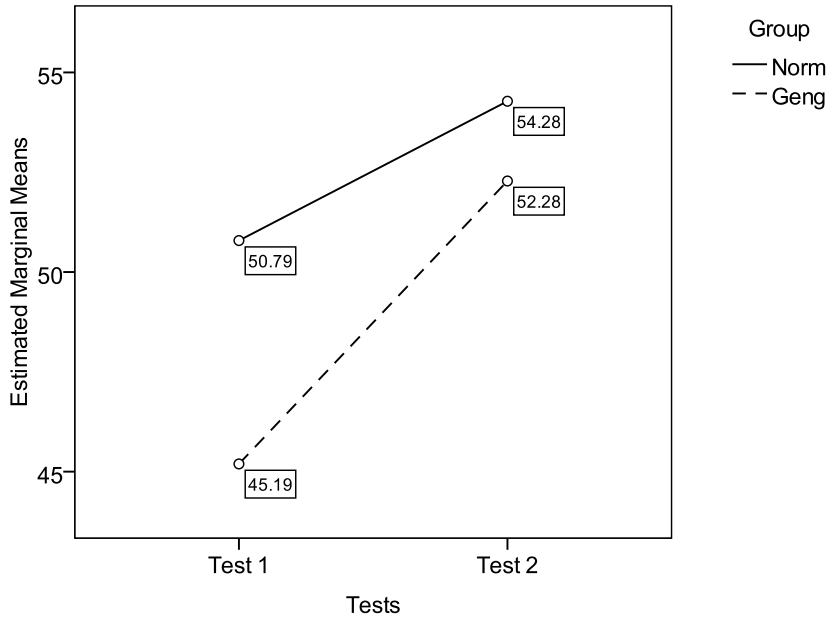


Figure 8. Interaction between *Test* and *Group* showing means for the *Norm* group and the *Geng* group on Test 1 and Test 2. The interaction is significant (p < .001).

The important thing established here is that a statistical process can be used to objectively accept or refute the existence of any actual (real) learning in a classroom environment. Other information required to perform such a calculation is shown in Table 22. Noteworthy is that the differences in the means for both groups (Norm and Geng) are statistically significant.

Table 22
Means and SDs for Groups *Norm* and *Geng* are Shown. Mean Differences and Significance Determined from Paired Samples *t* tests.

	Test 1		Test 2		M diff			
Group	M	SD	M	SD	$T_1 - T_2$	df	t	p
Norm	50.79	8.83	54.28	8.49	3.49	103	8.17	< .001
Geng	45.19	9.84	52.28	8.69	7.09	56	6.92	< .001

Note. Norm (n = 104), Geng (n = 57). Maximum Score Possible = 72

The application of this information to demonstrate how actual learning can be calculated is reported in the *Discussion* section of this study. The results of the full factorial ANOVA are shown at Appendix P. One note of caution in interpreting the interaction analysis reported here is the unequal performance at baseline for the two groups. This potentially limits the conclusions that can be drawn from the interaction.

Discussion

The test retest method applied to the 3DAT produced high test retest reliability coefficients for both Pearson and Spearman procedures based on datasets that included and excluded outliers. In all cases, the indexes were greater than 0.84 providing confidence in the stability of the 3DAT. One difficulty is to be certain about what is an acceptable reliability index, though there appears to be general agreement that reliability indexes greater than 0.8 are very acceptable. However, Gregory (2004) states, that although many authors suggest reliability should be higher than this, there is really no hard and fast rule about an index value. To emphasize this, and to return to a point made in study 3, Guilford and Fruchter (1978) maintain that many tests with reliability indexes as low as .70 prove to be meaningful, and even tests with values less than this can have applications in research. Aiken (1997) writes that for most achievement tests, the reliability indexes are in the .80s and .90s, though he is suggests the meaning of these values depends on the method of attaining the reliability index. Aiken suggests that the test retest and Cronbach alpha methods are likely to yield higher indexes compared to the alternate forms method. This contrasts somewhat with Gregory's viewpoint touched on earlier that the test retest method is the *gold standard* approach to determining reliability. For the test retest method, what must also be considered is the impact of measurement error which pervades any testing of reliability. The results of the second test are suspect because of the potential of *intervening events*, for example, lack of motivation, fatigue, memory and a participant's emotional state. Such conditions may affect the results of the second test and therefore impact on the calculation of a reliability coefficient. Thus, the interval between testings is particularly a factor in this regard. Cohen et al.

(1992) report a median reliability index of .88 for the test retest method with an interval between tests of seven days (n = 113), and a reliability index of .66 with an interval between tests of 12 months (n = 182). Though it is no guarantee, the interval of seven days between tests for the 3DAT appears to be optimum and intervening events should have been kept to a minimum since conditions were stable and consistent at the time of testing. One further consideration is that reliability of a test can generally be improved by adding extra test items to that test (Aiken, 1997). However, there is one qualifier, and that is, the extra test items need to have the same psychometric properties as existing test items. Though this is a concept more applicable to alpha reliability, it should also apply to test retest reliability because of the positive relationship that is likely to exist between alpha and test retest reliability. In essence, the point here is that the reliability reported for the 3DAT in Table 19 could be improved by simply increasing the length of the test by adding suitable test items. More is said about this potential in study 9.

With the reliability coefficient consistently presenting as greater than .84, the reliability of the 3DAT determined by the test retest method appears to satisfy any criteria of *very acceptable*. Considering some improvement may be possible with greater regard for intervening events, and by increasing the number of test items, the main outcomes of this study were encouraging. Also having conducted the test retest with an interval of seven days increases the significance of these outcomes. Though not raised in earlier sections of this thesis, deciding on what that interval should be was a difficult decision. The decision of seven days now seems justified. One final thing from Cohen et al. (1992) is that, although the worth of a test is often gauged by the reliability index it reports, any such index can only be fully appreciated along with the unique circumstances surrounding its application.

Results shown in Table 20 indicate that the mean for test 2 (54.28) was greater than the mean for test 1 (50.79) which demonstrates that test takers performed better on the second administration of the 3DAT. Importantly, also reported in Table 20 is a significant difference between the two means (t (103) = 8.17, $p < .001$). Consequently, these statistics are reporting the *practice effect* which is a concept explained earlier in this study and generally occurs with the second administration of a test. In other words, these results demonstrate that significant learning in all probability has taken place and the results therefore could be factored into a statistical procedure designed to measure actual (real) learning. In many respects, moving the focus at this point away from psychometrics to classroom learning may seem like a digression, but the idea is to show one real benefit of being able to quantify practice effect. The results reported earlier for the mixed RM ANOVA and illustrated in Figure 8 indicated that the difference of the differences (*Norm* versus *Geng*) was significant. Effectively, this means that actual learning could be determined after deducting the practice effect, and a formula for this

procedure could be expressed as: *total learning – practice = actual learning*. Using raw values (mean differences) shown in Table 22, the example in this study (*Norm* vs *Geng*) could be written as *Total* (7.09) – *Practice* (3.49) = *Actual* (3.60). Converting this to a percentage, actual learning could then be shown to be 5% (3.60/72*100) noting that there were 72 test items in this version of the 3DAT. Expressed another way, students in the *Geng* group improved their spatial performance in real terms by 5% over the period of their course. The interval between the two administrations of the 3DAT for the *Geng* group was 12 weeks which implies this estimate is conservative. This is suggested because the practice effect was determined from test retest results based on an interval of seven days. If the test retest had occurred with an interval of 12 weeks, the difference in means was likely to be less because of intervening events such as reduced recall memory. In this case, the amount to be deducted from total learning would be less and hence the measure of actual learning would be higher. Thus, using a difference based on a seven day interval provides a conservative estimate. The main issue here is not to be judgmental about the outcome, but simply to demonstrate that an evidence-based statistical procedure can be used to objectively determine if actual learning has taken place. In many respects, this measure of practice effect can be treated as a defacto norm and used with any pre/post application of the 3DAT with a direct benefit to educators and researchers. The term *defacto* is relevant because, to call it a norm *proper*, the reported result would need to be repeated across several evaluations to be confident of promoting it as a universal standard. This *digression* led to the creation of a template that can be used to measure learning differences and to report if any of that learning is actual. The calculation is based on an independent samples *t* test using summary output of two paired *t* tests as inputs. A completed version of the template based on the *Norm* and *Geng* data is shown at Appendix Q. More is said about the template in chapter 5.

A second approach to analysing the test retest data based on contingency tables was conducted and results are shown in Table 21. The first approach is characterised as the *Pearson correlation and paired samples t test model* while the second is characterised as the *contingency tables* model. From this point, these models are referred to as *model 1* and *model 2* respectively. The models involved different statistical procedures and the purpose of model 2 was mainly to confirm the findings from model 1 as a confidence initiative. Two important outcomes from model 2 and copied from Table 21 are listed as follows:

- no / no agreement 17.5%
- yes / yes agreement 63.5%

The 17.5% indicates the percentage of responses that were *incorrect* for both attempts, and the 63.5% indicates the percentage of responses that were *correct* for both attempts. Combining these two percentages indicates that 81% of responses were in agreement across the two

administrations of the 3DAT. In other words, this is a measure of repeatability (reliability) and can be expressed in index form as 0.81. Importantly, this index is similar to the reliability coefficient found for both Pearson and Spearman correlations procedures (r = .84) carried out for model 1. Thus, two statistical estimates based on different procedures are in agreement and indicate very acceptable reliability for the 3DAT.

Two further outcomes from model 2 and again copied from Table 21 are listed as follows:

- yes / no performance 7.1%
- no / yes performance 11.9%

The 7.1% indicates a reduced performance for test takers moving from test 1 to test 2, while the 11.9% shows an improved performance moving from test 1 to test 2. The McNemar test verified that the difference between the two results was significant ($p < .001$) in favour of the improved result. This outcome is parallel to the significance level found from the t test conducted in model 1 ($p < .001$). In simple terms, more test takers improved than those who did not, and the difference was significant. This then is confirmation of the practice effect.

Acceptable reliability for the 3DAT was established primarily using the test retest method (Pearson correlation and t test) and confirmed by a second statistical procedure (contingency tables). Furthermore, a practice effect in quantifiable terms was identified and also confirmed by a second procedure. As a consequence, the objectives for study 7 were considered achieved.

Study 8 Face Validity

One validity that does not fit the normal understanding of validity is called *face validity*. In essence, face validity is really concerned about appearance, or a perception that a test seems appropriate as opposed to an indepth statistical perspective that would be expected for say content or theory consistent validities. In every respect, face validity is a consideration from the test taker's point of view rather than that of a test examiner or a test developer. Though not important from a technical viewpoint, face validity is nevertheless important. A test may satisfy all other requirements for a good test in psychometric terms, but if it is not seen as relevant by the test taker, then there is some doubt about how successful the test can really be (Gregory, 2004). Consequently, there is some risk of nonserious attempts, indifference, poor motivation or a lack of confidence in the test itself (Cohen et al., 1992). To take this concern one step further, two tests may measure the same ability, but the test that appears to be more relevant to the test taker will carry the most meaning. In other words, it is a better test if test takers recognise its appropriateness. Of greater necessity is to not overstate the importance of face validity. Whilst face validity is seen as important, an instrument has little value if it does not produce meaningful data. However, considering the potential harm in ignoring face validity, there is justification in reporting the status of face validity for a test under construction. Consequently,

study 8, although not a large study, was concerned with the face validity of the 3DAT. The investigation was carried out as a secondary study in a larger study where it was possible to survey the opinions of test takers about spatial issues. Face validity was considered important enough to deserve its own individual focus; accordingly, the objectives of study 8 were to:

- assess the face validity of the 3DAT based on quantitative data collected from research participants,
- evaluate the ratings given to several spatial topics by research participants, and
- review comments from research participants about spatial ability.

The larger study that hosted the face validity survey was the test retest reliability study reported in study 7. The survey was conducted with one group of undergraduate design students who participated in that study.

Methodology

54 undergraduate students (43 male, 11 female) from six design disciplines completed the face validity survey. The disciplines and the number of students from each are shown in Table 23.

Table 23
Number of Participants who Completed the Face Validity Survey and their Respective Design Disciplines

Discipline	Num	Discipline	Num
Architecture	7	Engineering	36
Design & Tech Teaching	4	Industrial Design	2
Graphic Design	2	Construction Management	3

Note. Engineering Discipline Consists of Several Specialisations. Sample (n = 54)

The survey was made up of four sections. The first section was concerned with definitions, instructions and some basic demographic information such as discipline. The second section consisted of 10 questions all related to spatial topics such as gender performance, innate ability and the importance of spatial ability to respective disciplines. The third section was concerned with the 12 subtests and simply asked the same question about each subtest; that is, *was the subtest relevant to the participant's degree*. Both the second and third sections produced quantitative data and were scored using a five point Likert Scale. The five categories were: *strongly disagree, disagree, neither agree or disagree, agree, strongly agree*. The fourth section produced qualitative data and invited comment about many aspects of spatial ability and the 3DAT in general. There was no limit to the number of responses possible, and these were later grouped into a number of categories. The survey was given to participants in paper form after they completed the second administration of the 3DAT as part of the test retest study they participated in. The survey took about 20 minutes to complete, and a copy is provided at Appendix R.

Results

The results for the 12 subtests are reported first of all. The graph in Figure 9 shows the means and 95% confidence intervals for each of the subtests based on the Likert scores. The range of possible means was between 1 to 5 and Figure 9 indicates the lowest mean was 3.15 for the TL subtest and the highest was 4.17 for the ED subtest. With all means being above 3.0, it suggests that all subtests on average were regarded as relevant by the participants.

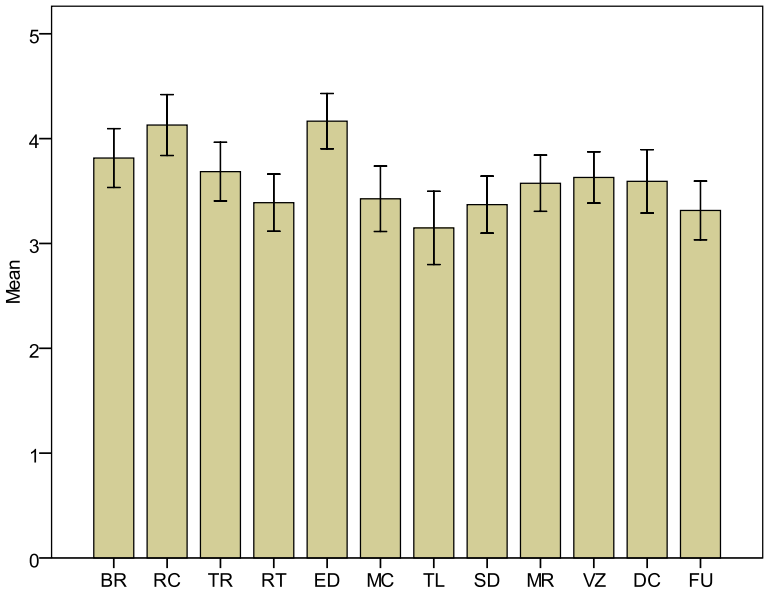


Figure 9. Bar Graph showing means and 95% confidence intervals for each subtest based on Likert scores produced from the face validity survey. Participants were asked the same question for all subtests, that is, did they think the subtest was relevant to their degree. Clearer identifiers for subtests are shown in Table 24. Sample (n = 54).

A better appreciation of how participants rated the subtests is shown in Table 24. The data are presented a little different to that normally done for Likert scales. What is presented instead is the percentage of participants who clearly agreed that a subtest was relevant to their degree, and as a comparison, the percentage of participants who clearly disagreed. The *agree* category is a sum of the strongly agree and agree scores, and the *disagree* category is a sum of strongly disagree and disagree scores. In other words, *strongly agree* and *agree* were collapsed into one category, and the *strongly disagree* and *disagree* were also collapsed into one category. This approach deliberately disregards those participants who were uncertain about the merit or otherwise of a subtest. The main idea was to be more informative and present a clearer picture of how the subtests were rated by the participants.

Table 24
Agree and Disagree Summaries for all 12 Subtests based on Likert Scores.

Subtest	Agree	Disagree	Subtest	Agree	Disagree
BuildRep (BR)	67	15	TrueLngth (TL)	43	43
Recogn (RC)	80	15	SurvDev (SD)	50	22
TransForm (TR)	63	15	MRotate (MR)	56	17
3D Rotate (RT)	54	26	Visualiz (VZ)	57	11
EngDwg (ED)	80	9	DotCoord (DC)	63	20
MentalCut (MC)	54	26	FoldUnfold (FU)	48	26

Note. Agree and Disagree Values are in Percentage Terms. Sample (n = 54)

The detail provided in Table 24 is deserving of comment. First of all, the subtests BR, RC and ED received the highest ratings. Of note is that these subtests are very much about 2D to 3D and 3D to 2D transformations. Most, if not all experts in this field would agree that these skills are the most fundamental of all requirements to communicate graphically in a design environment. Thus, it is encouraging that participants (novice designers) recognised these skills and rated them accordingly. The RT, MR and VZ subtests all received a medium to low rating in comparison, and all three involve mental rotation in one form or another. It is curious therefore, that these subtests were not perceived as highly relevant since the ability to mentally rotate is commonly regarded as an essential skill in spatial ability. The SD and FU subtests both rated low, and the skills they tested were similar. The TL subtest rated the lowest, which is understandable, but at the same time is also surprising. This subtest was never well understood in the format presented in this study by any group who experienced it, which probably explains its very low rating. On the other hand, the concept of true length as opposed to apparent length seems a fundamental differentiation that designers should be able to make. It is thus a small concern that this subtest was not rated more highly by participants. Somewhat appropriate, these outcomes serve to remind design educators that learners should not have the final say in all curriculum decisions.

Moving to the second set of results, Figure 10 shows the means and 95% confidence intervals for each of the *specific spatial topics* questioned in the survey according to the Likert scores. The range of possible means was between 1 to 5 and Figure 10 shows the lowest mean to be 1.98 (spatial ability is an innate ability that cannot be improved), and the highest mean to be 4.31 (spatial ability can improve with proper training).

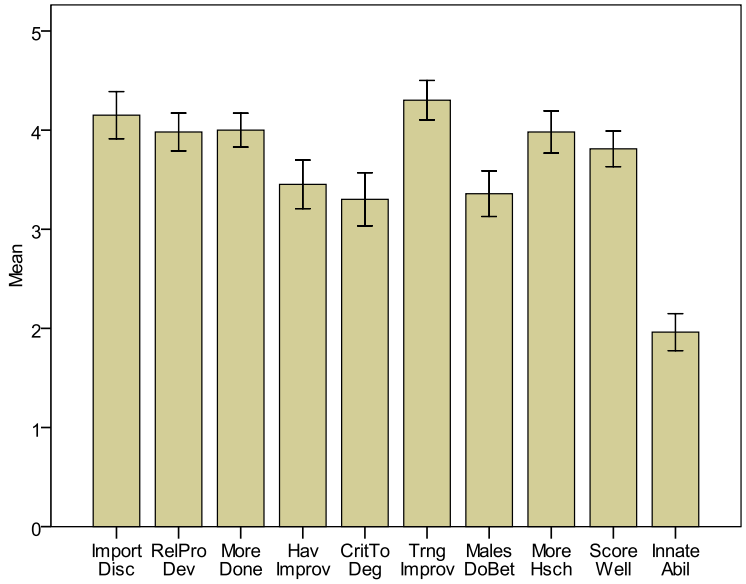


Figure 10. Bar graph showing mean and 95% confidence intervals for 10 questions concerned with spatial topics. Horizontal labels are abbreviated descriptors of questions asked. Clearer descriptors are shown in Table 25. Sample (n =54).

Based on the same rationale, and using the same approach applied to the 12 subtests just reported, Table 25 shows the percentage of participants who agreed with the 10 spatial questions asked in the survey, and also the percentage who disagreed.

Table 25
Agree and Disagree Summaries for 10 Spatial Questions based on Likert Scores. Survey Question Numbers are Shown in Parentheses.

Question	Agree	Disagree	Question	Agree	Disagree
SA is important to my discipline *(1)*	85	7	SA can improve with proper training *(6)*	89	2
Subtests are relevant to prof development *(2)*	80	2	Males do better than females on SA tasks *(7)*	33	7
More could be done at university *(3)*	81	0	More can be done at high school *(8)*	78	4
SA has improved since starting university *(4)*	46	13	I expect to do well on the SA test *(9)*	80	6
SA is critical to success in my degree *(5)*	41	17	SA is an innate skill that cannot be improved *(10)*	2	81

Note. Agree and Disagree Values are in Percentage Terms. SA = Spatial Ability. Sample (n =54)

With reference to Table 25, the importance of spatial ability is apparent from the ratings given to questions *(1)* and *(2),* and to a lesser extent, questions *(3)* and *(8).* However, these are somewhat in contradiction with question *(5)*. Encouragingly, participants believed spatial ability could be improved according to the ratings given to question *(6),* and this is very well supported by the strong disagreement registered for question *(10)*. From a training viewpoint, questions *(3)* and *(8)* indicate a large proportion of participants felt more could be done to improve spatial understanding. Interestingly, question *(7)* suggests that there is not a very large consensus for the idea that males perform better than females on spatial tasks, which is somewhat contrary to

what the literature says. One particular point from question *(7)* is that a large percentage of participants (60%) were undecided about this question. From question *(9)*, it seems confidence was high among participants since 80% of them thought they would do well on the 3DAT. One final point, and again from a training perspective, question *(3)* suggests that 81% of the participants felt more could be done at university to improve spatial ability, which is emphasized markedly with nobody shown to disagree with this view.

The final set of results can be described as qualitative because they represent the comments received from participants in response to the survey. The comments are not extensive, and for the most part they suffered from similar disadvantages as for the qualitative data collected in study 5 (SME interviews). For example, categorising the data was problematic because only open comments were requested, and as a result, it was difficult to fit all comments into meaningful categories. As well, some comments were simply a repeat of what was reported earlier for the quantitative data. However, it was possible to derive four broad classifications, although one of these related specifically to future research and another to the feedback built into the 3DAT. Hence, just two classifications are reported here (*spatial ability is important*, and *training leads to improvement*) because of their special significance to the 3DAT. To ensure meaning, a simplified synopsis of the qualitative data is shown in Table 26. What is provided are the two classifications, the different subcategories and some examples of comments from various participants. Finally, the reporting is necessarily brief because the focus was primarily on face validity and some comments tended to move away from this. The full set of comments for the two classification can be found at Appendix S.

Table 26
Classifications, Subcategories, Number of Reponses and Selected Comments from Test Takers are Shown

Classification	SubCat	Selected Comments
Spatial Ability is Important	Agree *(23)*	Designers need to be able to conceptualise a draftsman's drawings. In my faculty you need this every day. Engineers need to be able to think visually whilst designing. Imperative.
	Disagree *(2)*	It has very little consequence to my degree. For the majority of time, no but sometimes yes, i.e., teaching engineering studies and tech drawing.
	General *(9)*	Without visualisation, creativity can only be one dimensional. I am really surprised that training like this has not been incorporated in high school and even primary school. I do really see a beneficial side to spatial thinking.
Training Leads to Improvement	Agree *(21)*	Revision of spatial concepts can always improve spatial ability. Everyone begins with an innate talent for spatial ability however it can be improved through practice. All can be improved, how much improvement is the question
	Disagree *(4)*	With some people it can be difficult to improve. I think some people are naturally gifted while others will never get it.
	General *(19)*	So far I have only done first year. It would have been nice to incorporate some course material that could enhance our spatial ability right from the start. So far there has been none. I believe this is a life skill and more should be done. More focus on this core skill.

Note. Values in Parentheses indicate Number of Responses

At first glance, the comments shown for each of the subcategories in Table 26 appear to come from one participant. However, this is not the case because different comments have been merged to form one paragraph. Of note is that there is very little negativity reflected in these comments, and overall, they endorse the importance of spatial ability and strongly support the idea that spatial ability can be improved with training.

Discussion

First and foremost, the presence or otherwise of face validity is really a judgment call by test takers or by test users to some extent. Aiken (1997) describes face validity as the degree to which an instrument appears to measure the characteristic it is meant to measure. Strictly speaking, the evidence of face validity for the 3DAT was demonstrated in Figure 9 and supported by Table 24 where only two of the 12 subtests failed to rate above 50% *agree*. One of these was the true length (TL) subtest and a plausible explanation was given for this. This study, however, moved beyond simply rating the subtests and sought the opinion of test takers about specific spatial topics and spatial ability in general. The aim was to demonstrate the existence of face validity from two additional perspectives. Figure 10 and Table 25 report the first

perspective (*specific spatial topics*), and Table 25 in particular indicates that five out of the six most applicable spatial topics were rated at least 80% *agree* by the test takers. The comments from test takers listed in Table 26 report the second perspective (*importance of spatial ability and training leads to improvement*). Collectively these two perspectives support the quantitative results because they add an extended view on how test takers regard spatial issues. Thus, the findings from three different probes (subtests, spatial topics and general comments) strongly suggest that test takers agreed about the relevance of the 3DAT, which in effect is the same thing as saying the 3DAT has face validity.

It is possible that the difference between ratings given to each of the survey questions may have something to do with the range of disciplines represented in this study. For example, novice architects could feel quite different to novice engineers about the relevance of certain subtests and spatial topics to their discipline. Further, while some *% agree* values listed in Table 24 and 25 may appear to be low or medium in strength, they could in fact be stronger in reality than they seem. That is, data distributions for any survey do not generally show extreme skewness in either direction since many respondents play it safe and choose the neutral rating. A skewness would be required to get *% agree* values to be greater than most of those shown in Table 24 and 25, and this would only happen if test takers chose other than neutral. However, what it comes down to is that a test only needs to look and feel right to possess face validity, and the *% agree* values are merely a measure of the extent of this agreement. The level of acceptance is not generally questioned, and as Gregory (2004) points out, a test has face validity if it looks valid to test takers, test administrators and test developers alike. In view of the level of face validity reported for the 3DAT, study 8 is considered to have achieved its objectives.

Study 9 Item and Factor Analysis

Many of the procedures reported in this study were applied in earlier studies (e.g., item analysis), but the 3DAT was now at a stage where a more robust application of the procedures could take place. A good sample size had been achieved and the 3DAT was now reduced to a very plausible set of subtests and test items ready for final evaluation. The research had reached a point where finalisation was possible and conclusions about a number of aspects could be drawn with confidence. Study 9 was diverse and comprehensive and covered a wide range of objectives and particularly built on the experiences of earlier studies. The diversity embedded in this study also meant that a substantial contribution could be made to construct validity since evidence from many sources is required to demonstrate this form of validity (Gregory, 2004). Construct validity is complex and was first introduced in study 2.

What is now outlined is the range of investigations that were carried out in study 9. First of all, item analysis was conducted using the *classical test theory* (CTT) method which was also used in study 1 and identified as the traditional approach to item analysis. CTT is fundamentally

concerned with testing for item difficulty, item reliability and item discrimination. A second method was also used and it was first introduced in study 1 as well and later covered in greater detail in study 6. It is known as *item response theory* (IRT) and there are three variations to this model based on the number of parameters it investigates. A one parameter model estimates item difficulty, a two parameter model estimates item difficulty and discrimination, and the three parameter model estimates item difficulty, discrimination and guessing (Aiken, 1997). The one parameter model is also known as the *Rasch Model*, and it was named after Georg Rasch, a Danish mathematician and statistician who pioneered work on IRT. Because of its popular use and it being the default of the JMP software, the two parameter model was decided upon for this study. The IRT method is technically different to that of the CTT method and was used in this study to gain experience, for comparative purposes and as a confirmation measure. The reason then was threefold. Importantly, although both CTT and IRT methods were applied to the data, CTT was given priority and test items were rejected according to this method. Following item analysis was a return to testing reliability as a measure of internal consistency using coefficient alpha. The intention was to identify other test items that could be eliminated after item analysis had been completed. Collectively, item analysis and coefficient alpha were expected to identify underperforming test items that should be removed from the 3DAT. Later, the internal consistency of the surviving test items within each subtest would also be calculated and reported.

Moving on, it was extremely important (especially since this was the final study) to see if spatial factors could be identified. Hence, exploratory factor analysis was used to investigate the factorial composition of the 3DAT. The objective was to determine how many spatial factors may exist to collectively define the spatial ability construct. Initially, factor analysis was applied to the choice accuracy dataset since it was seen as the likely source for identifying factors, but the procedure was later applied to the RT data with surprising results. Further, in an attempt to present a final account of gender differences, a number of statistical procedures were applied to the dataset to compare the spatial performance of male and female participants. This was a good opportunity because, although there was an imbalance in numbers, the large dataset did mean that more females than usual would be represented. Importantly, the question of gender differences did require some closure after such a long investigation. As a special focus, the relationship between general academic ability and the 3DAT was also investigated. There is some disagreement about the strength of this relationship with some suggestions that spatial ability may correlate with academic ability, but on the other hand, there is also a thought that spatial ability may not be closely related at all. It seemed meaningful, therefore, to include an investigation that could probe this question. Thus, procedures were put in place to examine the relationship between *university admission indexes* (UAI) and the 3DAT since UAI is seen as a

reasonable measure of general academic ability (see *Terms and Definitions*). Because of the likelihood of a gender difference favouring males in the dataset, the relationship of UAI versus 3DAT was adjusted for gender to try and give a clearer insight into the strength of this relationship. Of potential relevance, Deary, Strand, Smith, and Fernandes (2007) in their study of intelligence and academic achievement found a correlation of r = .81 between intelligence and educational achievement scores in national examinations for adolescents. Although this current study investigated the relationship between UAI and the 3DAT and not UAI and IQ, the findings of Deary et al. raise the interesting question of how closely is spatial ability related to general intelligence. This question is outside the focus of this study, but it is well-suited to post PhD research.

Since there were many issues planned for investigation in this study, it meant that the objectives were varied and high in number. Consequently, the objectives of study 9 were to:

- conduct item analysis to evaluate test items using CTT methods,
- confirm item analysis using IRT methods,
- measure internal consistency of test items within relevant subtests,
- conduct exploratory factor analysis to identify subtests for the 3DAT,
- test for gender differences in spatial performance,
- investigate the relationship between academic ability and the 3DAT, and
- verify the presence of spatial factors based on reaction time (RT) data using exploratory factor analysis.

The 3DAT was in its final stages of development and at a point where very definite claims could be made about what its composition should be, what skills it particularly measured and what psychometric properties could be accorded to its subtests and test items. The investigation started with the 12 subtest version of the 3DAT (72 test items) but the configuration changed progressively according to the analysis being undertaken at different times.

Methodology

The sample of 635 (male = 535, female = 100) were university students from 15 design disciplines and the number of students in each are indicated in Table 27.

Table 27

Disciplines and Numbers of Participants in Study 9.

Discipline	Num	Discipline	Num	Discipline	Num
Const Management	93	Mining Engineering	33	Graphic Arts	2
Mech Engineering	143	Mechatronics	55	Science & Other	7
Civil Engineering	95	Technology Teaching	37	Architecture	88
Chem Engineering	54	Inform Technology	1	Matls Engineering	10
Elect Engineering	3	Industrial Design	2	Environ Engineering	12

Note. Num = Numbers. Sample (n =635)

This good sample size was achieved because of the interest in spatial ability from two universities across several design schools that represented a mixture of design disciplines. The full sample was not used for every analysis because some categories of UAIs (e.g., *other*) could not be utilised. Also, for some procedures, something about participant data in several categories (e.g., RT) appeared odd and also needed to be rejected. The sample of 100 females came from 11 of the 15 disciplines shown in Table 27, but the larger numbers of females were from architecture (29), chemical engineering (15), technology teaching (15), construction management (13) and mechanical engineering (9).

The 3DAT given to participants was the 12 subtest online version which was able to be delivered to any classroom in any location where internet access was available. The 3DAT operated from a server maintained by the host university and participants were tested in their normal computer laboratories. The test was run as a class activity with Ethics approval to collect data for research purposes because the activity was seen to align with course objectives. Importantly, participants were able to elect for their data not to be used for research purposes if they were inclined this way. Where this occurred, data collected from these participants were filtered from the primary dataset before any analysis took place. About 9% of participants did not provide consent. The subtests were presented in random order and the 6 test items within each subtest were randomised as well. Instructional videos to explain the requirements of the 3DAT and practice trials were available to participants, but academic staff responsible for the running of the 3DAT at each location decided not to use these. Instead, their unanimous choice was to utilise the instructional screen that preceded each of the 12 subtests. The instructional screens explained what participants needed to do to complete the test items within each subtest. Participants could take their time in studying these since they did not time out and therefore remained onscreen until participants elected to move on. After completion of the 3DAT, participants received an overall summary of their results displayed onscreen in graphical form and a breakdown of their performance in each of the subtests. There was also an option for participants to review any test item they scored incorrectly. If participants required a record of their effort, they were able to send themselves an email containing an automatically generated

text summary of their results. These few features alone helped convince the various Ethics committees of the educational value in running the 3DAT as a class activity. Guidance, administration, supervision and organisation provided to participants were similar to procedures reported for study 7 (test retest study). However, time on task for this study was a little less and averaged about 75 minutes but this included management time. A separate CSV file with a unique filename was generated for each participant and it recorded demographic information, choice accuracy and reaction time. All CSV files regardless of where the participation took place were initially saved and stored on the server at the host university. All CSV files were then merged into one large file using a simple DOS command and later converted to an XLS file in readiness for export to both the SPSS and JMP software packages. As a reminder to the reader, the names and abbreviations for each subtest are shown in Table 13, and picture examples are available at Appendix J.

In many respects, the analytical procedures selected for this study can be identified from the study objectives listed earlier. For example, to decide about the properties of test items, the CTT method of item analysis was implemented and the IRT method was conducted as a form of verification. There was no real requirement to run both methods, but in view of research opinion being somewhat divided, there seemed to be learning value in comparing the two methods. To measure internal consistency, Cronbach alpha was applied to the dataset. This was done first of all to help decide what test items could be rejected, and later to provide measures of internal consistency within each of the final subtests. In an attempt to identify common underlying variables within the 3DAT, exploratory factor analysis (FA) was conducted with the idea that the 12 subtests may be summarised into a fewer number of factors. Although the 3DAT consisted of 12 subtests with each potentially measuring a different spatial skill, FA would possibly reveal a smaller set of dimensions that explain the 3DAT. To clarify this important concept, Gregory (2004) offers the example of the decathlon athletic event. Whilst the event itself consists of 10 separate track and field activities (e.g., javelin, hurdle and long jump), it can, however, be reduced to four critical attributes, namely: speed, strength, coordination and endurance. Similarly, the aim of applying FA to the 3DAT was to reduce it to several critical skills as well. To explore gender differences, independent samples *t* tests were conducted to determine significance in performance on all subtests and the 3DAT overall. To increase the meaning of the results, *effect size* was again introduced to provide practical significance. Finally, to examine the relationship between UAI and the 3DAT, a regression was performed using the SPSS *general linear model* (GLM) including an analysis of covariance (ANCOVA) to partial out the influence of *gender* because of its potential as a confounding variable. Controlling for gender effect was decided upon because males are consistently reported to perform better than females on spatial tasks. By removing gender, a clearer insight into the

relationship between UAI and the 3DAT was possible. Four increments of 10 UAI units were used; that is, 61 – 70, 71 – 80, 81 – 90 and 91 – 100. Also instigated was a linear regression manipulation to obtain the correlation coefficient for UAI versus 3DAT after adjustment for gender.

The final analytical procedures were a series of progressive steps that reduced the 3DAT from 12 subtests (72 items) to 9 subtests, then to 7 subtests and finally to a 5 subtest model consisting of 20 test items. The closing procedures that tested gender, UAI, RT choice accuracy, internal consistency and reliability were all based on the final composition of the 3DAT.

Results

CTT ANALYSIS

The CTT analysis considers item difficulty, item reliability and item discrimination. For each, an index number is produced within the range of 0 to +1. The index is a measure of each of these properties. These measures are defined in the *Terms and Definitions* section of this thesis, but they are partly repeated here for convenience with one or two points added. Item difficulty is a measure of test item difficulty and is calculated according to the number of test takers getting the item correct. The lower the index number, the more difficult the test item is. Index numbers approaching either end of the range indicate underperforming items. That is, low index numbers represent test items that may be too difficult, and high index numbers represent test items that may be too easy. The chance factor has to be adjusted for, however, the formula used to calculate item difficulty takes this into account. For item reliability in the CTT model, it is really another measure of internal consistency although there are specific considerations. Test items tend to be dichotomous scores such as yes/ no or correct/ incorrect responses. On the other hand, the total test score will be a continuous variable. In simple terms, to calculate the correlation between the scores on the test item with the scores on the total test, a particular statistic termed the *point biserial correlation coefficient* is used (Gregory, 2004). What must also be considered is the variability in the test item score in terms of its standard deviation. In order to produce an index for item reliability within the CTT model, the product of the test item standard deviation and the point biserial correlation is calculated (Cohen et al., 1992). Importantly, the higher the index value, the greater is the reliability of the test item. Item discrimination is a little less complicated. It is a measure of how well a test item separates the high scorers on the test as a whole from the low scorers on the test as a whole. In theory, the test item is a good discriminator if all the test takers in the upper group get the item correct, while those test takers in the lower group get the item incorrect. In other words, a test item is not discriminating ideally if high achievers on the test as a whole score poorly on the test item. The index generated from the statistical procedure can range from -1.0 to +1.0, and the closer the index is to +1.0, the more effective the item is in discriminating between high and low

performers. However, indexes of this magnitude are rarely achieved (Aiken, 1997). Items on the other end of the scale (-1.0) should be discarded immediately. After all, this indicates that all test takers in the lower group scored the item correctly while all the test takers in the upper group scored the item incorrectly. This is the opposite to how it should be, and essentially, items with positive indexes can be acceptable, but items with negative indexes at least should be reworked. Distracter options also play a significant part. For test takers who do not know the correct answer, each distracter should be approximately equal in appeal, and the number of test takers therefore choosing each of the distracters should be about the same.

To achieve the CTT item analysis, a sizable spreadsheet file was established to calculate the indexes for all three factors (item difficulty, item reliability and item discrimination) outlined above. This file is shown at Appendix T. Ideally, any test item failing to satisfy any three index ranges would be rejected, but this was too severe on this occasion because only a few items would have survived the cut. Consequently, some subjectivity was introduced to adjust for this. As a result, a criteria was decided upon where any test item that failed to satisfy two out of the three index ranges would be rejected (see Appendix T for statistical details). There was also a desire to maintain an equal number of test items in each of the subtests. As reported in study 6, earlier experience had shown that it was generally easier to convey information about the 3DAT if there was a consistency about its structure including the number of test items in each of the subtests. Based on preliminary analysis, there was every indication that 4 test items for every subtest would satisfy mostly an objective analysis with only minor subjectivity being necessary. Using the criteria decided for the CTT analysis, 25 of the 72 test items were rejected. Of these, a concentration occurred for the TL, RC and SD subtests such that 6/6, 5/6 and 4/6 test items respectively were rejected. This effectively meant that these three subtests were rejected by the CTT procedure. The remaining 10 rejected test items were spread almost evenly across the nine remaining subtests. These results and some that are soon to be reported are shown at Appendix U. The actual test items rejected are also indicated. Of note is that the reliability criterion was the most standout in eliminating test items followed by the difficulty criterion.

IRT ANALYSIS

Moving now to IRT item analysis which was introduced into this study (among other reasons) to confirm the results from the CTT method. As previously stated, the two parameter model which investigates item difficulty and item discrimination was used for this IRT analysis. A detailed description of the IRT method was provided in study 6, and the reader is reminded that an example of an ideal *item characteristic curve* (ICC) is shown at Appendix I. Also provided are the ICC graphs that were generated for each of the 72 test items, and these are shown at Appendix V. Also shown are parameter estimates for item difficulty and item discrimination.

The CTT analysis clearly indicated that the TL, RC and SD subtests should be rejected, and the IRT procedure agreed with this analysis. IRT showed the test items in TL to be very poor since there was virtually no discrimination occurring, and the difficulty level was far too high. For RC, both analytical methods agreed that the test items were less than satisfactory, but interestingly, not for the same reason. In fact, they were somewhat contradictory. CTT rejected most of the RC test items because they were outside the discrimination and reliability criteria, though, item difficulty was shown to be satisfactory. In contrast, IRT rejected the same test items based on difficulty not being satisfactory (items too easy). There was also disagreement about item discrimination between the two analytical methods. CTT showed most RC test items to have a low discrimination index, though, IRT indicated discrimination to be satisfactory. For the SD subtest, there was agreement between CTT and IRT about most items in this subtest not being satisfactory, and for both methods, a mixture of reasons were given. That is, some items were too easy, some were too hard and there were variations in reliability measures. The TL, RC and SD subtests accounted for 15 of the test items rejected by CTT and IRT analyses, and for the remaining 10 rejected items, there was almost total agreement between the two methods, but with one or two notable differences. An example is the test item labelled *FU4*. The CTT procedure rejected this item because it fell outside the index range for both item difficulty and item reliability. On the other hand, the IRT showed this item to be almost ideal for item difficulty, and at least very satisfactory for item discrimination. Thus, IRT in fact provided argument why this item should be accepted rather than rejected as indicated by the CTT analysis. Looking further, CTT and IRT agreed that the item MR3 should be rejected, but nevertheless, were in conflict. CTT identified item difficulty as acceptable, but not item discrimination. In contrast, IRT identified item difficulty as not acceptable, but reported item discrimination to be acceptable. That is, the two analytical methods were in agreement and rejected MR3, but not for the same reason. For the remaining eight test items which were spread across the seven subtests not mentioned thus far, there was agreement between CTT and IRT that the items should be rejected. Summing up CTT versus IRT, both methods agreed in all but one case that the same test items should be rejected, but curiously, the reasons were sometimes different, and in one case, the analyses came to opposite conclusions. In percentage terms, this can be expressed as 72% agreement to reject (same reason), 24% agreement to reject (different reason) and 4% disagreement (CTT: reject, IRT: retain).

INTERNAL CONSISTENCY

As an adjunct to the CTT analysis to help decide which test items should be rejected, internal consistency was examined using Cronbach alpha. Among other things, this procedure produces alpha coefficients that would result if any one test item were to be excluded from the dataset. These coefficients indicate either an improvement in internal consistency, or a decrease in internal consistency. Thus, this analysis identified two subtests (BR and TR) where alpha would

improve if particular test items (BR1 and TR6) were removed from the respective subtests. Only a small number of rejections by this method were expected because the CTT procedure itself measures reliability, and many items were rejected earlier given that they fell short of the required criteria. The final alpha coefficients for the subtests are reported later in this study.

EXPLORATORY FACTOR ANALYSIS

The starting point for FA was to decide whether the *orthogonal* or *oblique* rotation method of analysis was the most appropriate to be used with the *accuracy* dataset. If orthogonal, the *varimax* rotation technique would be chosen, or if oblique, the choice would be the *direct oblimin* technique. The two methods are different because, for orthogonal, the components (factors) are assumed not to be correlated. On the other hand, for oblique, the components will be correlated. To decide between the two methods, the oblimin procedure was run using a *principal component* extraction analysis to produce a *component correlation matrix* table since this output would provide guidance about which method should be adopted. The varimax procedure does not produce a similar matrix because there is no expectation of correlation between components. The component correlation matrix is shown in Table 28.

Table 28
Component Correlation Matrix Produced from Factor Analysis using the Direct Oblimin Rotation Method.

Component Correlation Matrix					
Component	1	2	3	4	5
1	1.000	-.213	-.220	-.333	.219
2	-.213	1.000	.193	.222	-.109
3	-.220	.193	1.000	.196	-.151
4	-.333	.222	.196	1.000	-.201
5	.219	-.109	-.151	-.201	1.000

Note. Relates to Choice Accuracy Data

Tabachnick and Fidell (1996) provide a criteria to assist in the selection of a rotation method. Their position is that correlation coefficients of .32 and above should serve as a determinant for using oblique rotation, and by implication, coefficients less than .32 suggest that orthogonal rotation is better suited to the dataset. Table 28 clearly shows that, for the 10 possible pairs, only one pair (component 1 and 4) produced a coefficient of .32 or higher, and in this case, the value of .33 was barely above this measure. Since nine out of 10 combinations were less than the criteria suggested by Tabachnick and Fidell, a FA was conducted using a principal component extraction with varimax rotation. An eigenvalue of greater than 1.0 indicated a possible seven components and this was supported by the scree plot. Variance explained was 55.56%. After a number of trials using different configurations of suppressed loadings and numbers of components, the best loading was found to occur for a five component model. The trials also showed that the test items for two subtests (FU and ED) did not correlate with any component

regardless of any configuration when loadings ≤ .3 were suppressed. Consequently, these subtests were removed from the 3DAT. Results of a five component FA are shown in Table 29.

Table 29
Factor Analysis based on the Principal Component Extraction Method and Varimax Rotation. Loadings onto 5 Components are Shown

	Component				
	1	2	3	4	5
MR2	.517				
MR4	.551				
MR5	.514				
MR6					
RT1	.434				
RT4	.543				
RT5	.372				
RT6	.399				
VZ2	.338				.314
VZ3	.451				
VZ4	.420			.362	
VZ5	.379				
DC1				.690	
DC3				.716	
DC4				.564	
DC5				.694	
TR2		.772			
TR3		.735			
TR4		.716			
TR5		.747			
BR2			.661		
BR3			.770		
BR4			.673		
BR6			.780		
MC1					.533
MC2					.683
MC4					.670
MC5					.456

Note. Relates to Choice Accuracy Data (7 Subtests). Loadings ≤ .3 Suppressed.

Table 29 revealed that correlations were generally in the medium to high range keeping in mind that loadings of ≤ 0.3 were suppressed. Test items for three subtests (MR, RT and VZ) all loaded onto the same component except for one test item (MR6) which did not load on any component. This common loading meant that the three subtests were pointing to the same underlying factor, and therefore two could be removed from the 3DAT. Or, as an alternative, the best four test items across the three subtests could be chosen. The decision was to retain the MR

subtest and drop the RT and VZ subtests because of its frequent use in research, and hence its benchmark potential. No other common loading occurred with the remaining subtests DC, TR, BR and MC all loading onto separate components. Some minor cross loading occurred for test items VZ2 and VZ4, although this did not matter since the intention was to discard the VZ subtest. As a confirmation measure, the data were also analysed using a principal component analysis with direct oblimin rotation. Interestingly, the results (shown at Appendix W) agreed almost completely with the results presented in Table 29 thus supporting the choice of varimax.

Before moving to the final analyses of the 3DAT, it is appropriate to clarify the state of the 3DAT at this point considering there were 12 subtests at the start of study 9. Subtests TL, RC and SD were eliminated mainly from the CTT analysis, but with some contribution from alpha procedures. Subtests FU and ED were removed by preliminary FA, and subtests RT and VZ were rejected because they loaded onto the same component as MR. This meant that the final 3DAT consisted of the five subtests MR, DC, TR, BR and MC and 20 test items divided evenly between the subtests. The statistical methods that follow are based on this model.

GENDER DIFFERENCES

Before investigating the main issue of differences in gender performance across the 3DAT and each of its subtests, a number of preliminary investigations were conducted to gauge the standing of the gender data overall. An independent samples *t* test was conducted to determine if there was a significant difference in gender UAI scores. Not all UAI bands were included in the analysis and the *other* category for example was excluded and bands reported lower than what the institutions normally accept were also excluded. For the latter, it was thought that students from other educational systems did not convert their UAI equivalent scores correctly to UAI scores appropriate to the institutions they were studying at. Five increments of 10 UAI scores were used; that is, 51 – 60, 61 – 70, 71 – 80, 81 – 90 and 91 – 100. Hence, for the UAI comparison only, the sample was reduced to 538 (male = 459, female = 79) and the results of the *t* test showed that the difference in means for males and females for UAI was not significant t (536) = 1.90, $p = .058$. The mean score for males was 6.41 and for females it was 6.65 on a scale where *6* equalled the UAI band of *71 to 80* and *7* equalled the UAI band of *81 to 90*.

Also conducted was a 5 * 2 (subtest * gender) mixed RM ANOVA design where *subtest* was a within-subjects factor and *gender* was a between-subjects factor was applied to the dataset. The main effect for subtest was significant using Greenhouse Geisser adjustment F (3.90, 2466.59) = 27.44, $p < .001$, thus indicating that the type of subtest had an influence on overall performance. A more critical research question was answered by the *subtest * gender* interaction. The interaction was not significant using Greenhouse Geisser adjustment F(3.90, 2466.59) = 1.48, $p = .207$. This indicated that the differences between males and females did not vary substantially across the five subtests. Preliminary investigations therefore did not reveal any serious concerns.

Turning now to differences in performance on the 3DAT and across the five subtests, the results of independent samples *t* tests are shown in Table 30.

Table 30
Means and Standard Deviations for Gender Groups for Subtests and the 3DAT. Effect Sizes Shown. BR, MR and TR Adjusted for Unequal Variance.

	Male		Female					
Subtest	M	SD	M	SD	df	*t*	*p*	*d*
BR	3.15	1.201	2.81	1.37	129.08	2.33	0.021	0.28
DC	2.47	1.38	1.97	1.43	633	3.29	0.001	0.36
MC	2.60	1.18	2.39	1.1	633	1.62	0.106	0.18
MR	3.02	1.05	2.53	1.1	135.08	4.15	< .001	0.46
TR	2.99	1.31	2.79	1.45	130	1.29	0.198	0.15
3DAT	14.23	3.71	12.49	4.06	633	4.23	< .001	0.45

Note. Male (n = 535), Female (n = 100). Maximum Score Possible (subtest = 4, 3DAT = 20)

Table 30 indicates that the difference in performance between males and females was significant for three out of five subtests (BR, DC & MR) and for the 3DAT overall. To appreciate the differences further, effect sizes were small to medium using Cohen's (1992) scale reported earlier where d = .2 (small) and d = .5 (medium). The most substantial results are for the subtests DC (d = .36) and MR (d = .46), and the 3DAT overall (d = .45). Of special note is that males outperformed females in every case, although not all differences were significant.

RELATIONSHIP BETWEEN ACADEMIC ABILITY AND THE 3DAT

For this investigation, four UAI bands were used (61 – 70, 71 – 80, 81 – 90 and 91 – 100). The UAI (51 – 60) used in the gender investigation was not included because of the low sample size (17) in comparison to the other UAI bands (81, 175, 177 and 88 respectively). As a consequence, the overall sample for this investigation was 521 (male = 445, female = 76). An ANCOVA was performed to test the effect of UAI after partialling out the influence of gender. After adjusting for gender, there was a significant effect of the between subjects factor *UAI* in that $F(1, 518) = 23.639$, $p < .001$. To check the assumption of homogeneity of regression for the two levels within the gender variable, the *gender* * *UAI* interaction was tested. The GLM procedure showed that the data did not violate the assumption of homogeneity of regression slopes, $F(1, 517) = .735$, $p = .392$. This meant that effects for UAI did not depend on gender.

The coefficient for effect of UAI on the 3DAT was .812. Since the range of 3DAT scores was 18 points (min = 2, max = 20), this is equivalent to a 4.5% improvement for every 10 units of UAI (.812/18*100 = 4.5%). Viewing this from another perspective, this result means that on average, it can be expected that UAI will account for a difference in performance of 13.5% between the lowest UAI increment reported (61 to 70), and the highest increment reported (91 to 100) since 3 increments * 4.5 = 13.5%. In more precise terms, that is, working from the midpoint of each increment (65.5 and 95.5 respectively), this can be expressed as a difference of

30 UAI units between the lowest and the highest UAI increment. For the gender effect, the parameter estimate was 2.06 which means that males on average scored 2.06 points better than females on the 3DAT after adjusting for UAI. This is equivalent to a 11.4% better performance by males than females on the 3DAT regardless of UAI score (2.06/18*100 = 11.4%). The linear regression confirmed estimates of .812 and 2.06 respectively, and after adjusting for gender, this procedure indicated that the partial correlation coefficient for UAI and 3DAT was .209.

RT FACTOR ANALYSIS

Preliminary FA was applied to the RT data to confirm the presence of spatial factors, and the procedure produced surprising but better than expected results. RT data typically is skewed (thus breaking normality assumption), and also exhibits nonconstant variance with higher RTs having a greater variability (thus violating the assumption of constant variance). Log transformations usually correct most of these problems. As a precaution against skewness and nonconstant variance influencing FA, a *log transformation* of the data was carried out and FA was applied to the log RT data. The results of FA for both varimax and direct oblimin rotations based on the log RT data are shown in Appendix W. Where results differ substantially between actual RT data and log RT data, a decision is made about which data should be reported. Actual RT data will provide the more complex view, while the log RT data will provide a simplified view. Where there is little difference between the two analyses, FA for actual RT data would be reported. In comparing the results of FA for the two datasets (actual RT data and log RT data), very few differences were found. Subtests loaded uniquely onto components, no cross loading occurred, and differences in loadings were marginal. Since there were no real difference in pattern structures, a FA based on actual RT data is reported here. The reader is advised that the methods and procedures covered in this section run parallel to those reported earlier for FA based on the accuracy data. Accordingly, some brevity will be apparent. The starting point for this analysis was also to decide whether the orthogonal or oblique rotation method should be used with the RT data. The direct oblimin procedure produced a component correlation matrix which is shown in Table 31.

Table 31
Component Correlation Matrix Produced from Factor Analysis using the Direct Oblimin Rotation Method.

	Component Correlation Matrix				
Component	1	2	3	4	5
1	1.000	.147	.218	-.206	.198
2	.147	1.000	.130	-.113	.157
3	.218	.130	1.000	-.178	.178
4	-.206	-.113	-.178	1.000	-.207
5	.198	.157	.178	-.207	1.000

Note. Relates to Reaction Times (RT) Data

Applying the criteria advocated by Tabachnick and Fidell (1996), the matrix showed no correlation coefficient of .32 or above. Based on this criteria, a FA using principal component extraction with varimax rotation was conducted and produced the results shown in Table 32.

Table 32
Factor Analysis based on the Principal Component Extraction Method and the Varimax Rotation Method. Loadings onto 5 Components are Shown.

	Component				
	1	2	3	4	5
BR2					.615
BR3					.581
BR4					.597
BR6					.573
MR2		.689			
MR4		.613			
MR5		.709			
MR6		.697			
DC1				.564	
DC3				.635	
DC4				.579	
DC5				.618	
TR2			.515		
TR3			.711		
TR4			.592		
TR5			.668		
MC1	.725				
MC2	.611				
MC4	.681				
MC5	.724				

Note. Relates to Reaction Times (RT) Data. Loadings ≤ .3 Suppressed.

An eigenvalue of greater than one indicated a possible five components which was supported by a scree plot also produced. Variance explained was 46.03%. The results in Table 32 indicate correlation coefficients mostly tending towards the high range with all test items being accounted for and no cross loading occurring. Loadings ≤ 0.3 were again excluded from the procedure. Loadings are clearly unambiguous with all five subtests loading onto separate components. Again, as for the accuracy data, FA using direct oblimin rotation was applied to the data for comparative purposes and results were near identical to those for varimax rotation with the five subtests each loading onto separate components. Loadings were inclined towards the high end of the scale with all test items within each subtest pointing to the same component with no cross loading evident. The results of the two rotation methods are shown at Appendix W.

FINAL STATEMENT OF FACTOR ANALYSIS AND RELIABILITY FOR THE 3DAT

To make a final statement about the psychometric properties of the 3DAT, three critical analytical procedures based on final composition of the 3DAT (five subtests) are reported. The first of these is a FA using choice accuracy data, the second is a measure of internal consistency according to Cronbach alpha, and the third is test retest reliability.

In an earlier section, Table 29 provided the results of FA based on a seven subtest version of the 3DAT before it was reduced to the final five subtest model. The results of FA for the five subtests using a principal component extraction with varimax rotation are shown in Table 33.

Table 33
Factor Analysis based on the Principal Component Extraction Method and Varimax Rotation. Loadings onto 5 Components are Shown.

	Component				
	1	2	3	4	5
MR2	.667				
MR4	.655				
MR5	.660				
MR6	.329				
DC1				.695	
DC3				.750	
DC4				.572	
DC5				.718	
TR2		.778			
TR3		.729			
TR4		.727			
TR5		.756			
BR2			.660		
BR3			.776		
BR4			.672		
BR6			.793		
MC1					.527
MC2					.728
MC4					.694
MC5					.489

Note. Relates to Accuracy Data (5 Subtests). Loadings ≤ 0.3 Suppressed.

An eigenvalue of greater than 1.0 supported a possible five components and the variance explained was 49%. Noteworthy is that all test items within each subtest pointed to the same unique component with no cross loading occurring for any of the 20 test items. Correlations tended mostly towards the high range with two exceptions (MR6 and MC5) falling in the low to medium range. However, this was an improvement for MR6 which failed to register under the seven subtest model, and a slight improvement for MC5 (refer Table 29). To be consistent with earlier procedures, FA using direct oblimin rotation produced very similar results and again

justified the focus on varimax. A comparison of the two rotation methods can be seen at Appendix W.

Cronbach alpha coefficients (α) for accuracy scores were produced as a measure of internal consistency for the five subtests and results are shown in Table 34.

Table 34
Cronbach Alpha Reliability Coefficients for 5 Subtest 3DAT Model

Subtest	Alpha	Subtest	Alpha	Subtest	Alpha
BR	.726	DC	.677	TR	.759
MR	.453	MC	.485		

Note. Test items in each subtest = 4.

Based on the standards previously introduced (i.e., α = .7 acceptable, α >= .8 very acceptable), results shown in Table 34 represent mixed standards. Alpha coefficients for subtests BR, DC and TR are in the acceptable range, but subtests MR and MC fall short of this. One test item from each of these subtests (MR6 and MC5 respectively) are shown in Table 33 to have low loadings which suggests their removal from the subtests may improve alpha values. The *Item Total Statistics* table produced from the Cronbach alpha formula indicates a slight improvement when MR6 is removed (.46), but no improvement when MC5 is removed.

Correlation coefficients for test retest reliability for the 12 subtest 3DAT model were provided in Table 19 (Study 7 Test Retest Reliability) and both Pearson and Spearman procedures were reported. The same procedures were conducted for the final five subtest 3DAT model and results are shown in Table 35.

Table 35
Correlation Between both Administrations of the 3DAT Showing Reliability Coefficients based on Test Retest for the 5 Subtest Model.

Procedure	Coefficient	Sig (2 tailed)
Pearson	.828**	< .001
Spearman	.812**	<. 001

Note. ** *p* is significant at the .01 level. Sample (n = 104).
Test retest method reported in Study 7.

Using the criteria of coefficients greater than .8 being very acceptable in psychometric terms, the reliability coefficients shown in Table 35 fit with this criteria. Again, Pearson versus Spearman coefficients are very similar, and compared with coefficients listed in Table 19, results are also very similar though slightly less for the five subtest model. For convenience, Pearson (12 subtests) = .848 and Pearson (5 subtests) = .828, and likewise, Spearman (12 subtest) = .844 and Spearman (5 subtest) = .812. In all cases, $p < .001$. Considering there were fewer items in the five subtest model which normally reduces measures of reliability, the similarities are statistically important and reflect favourably on the final version of the 3DAT.

Discussion

Study 9 was ambitious because it investigated a large number of diverse issues guided by a larger than usual set of objectives. To be effective with both item analysis and FA in particular, a large sample was required. Tabachnick and Fiddell (1996) provide a useful guide to what this sample should be and suggest at least five participants for every test item or variable being developed. Considering there were 72 test items at the start of study 9, this study then exceeded the sample recommended by Tabachnick and Fiddell (72 x 5 = 360) since it achieved a sample of 635 participants. This provided good statistical power and therefore confidence in the results. Adding to the strength of the sample size is that it was spread over a wide variety of design disciplines and it captured a good number of female participants (n =100) which is much larger than is normally possible.

In one respect, the 3DAT with 20 test items could be considered to have a low number of test items and that it may be a better test if it contained a greater number of test items. However, to achieve this, an increase in the level of subjectivity would have been required to overrule statistical outcomes. While some minor subjectivity did occur in developing the 3DAT, judgments about the suitability of test items were mostly objective because of the analytical procedures that were applied. Any further subjectivity was seen to weaken any argument that would be made in support of the 3DAT. Somewhat related, item selection was difficult where three subtests (MR, RT and VZ) loaded onto the one component indicated by FA. In many respects, simply choosing the four best test items based on their correlation measure with the component was the easiest and perhaps the best solution. However, preference was given to the four MR test items because of the prominence of the mental rotation (MR) task in the spatial cognition literature. There was also a preference for having a uniform number of test items.

The alpha coefficients shown in Table 34 for the subtests MR and MC are satisfactory but less than ideal. A contributing factor is the low number of test items (4) in each subtest. Increasing the length of a test is one way of improving coefficient alpha (Aiken, 1997). Aiken provides a formula that can be rephrased as *new* α *= (increase length factor x orig* α*) / (1 + orig* α *(increase length factor – 1)).* Applying this formula to the α value for the MR task shown in Table 34 (.453) and based on a proposal to increase the length of the subtest by a factor of three (i.e., 3 x 4 = 12 test items), the *new* α value will be .713 since *(3 x .453) / (1 + .453(3 – 1)) = .713*. The qualification, however, is that the new test items would need to have the same psychometric properties as the original test items. Increasing the number of test items also improves the probing of the *spatial factor* that the subtest is designed to measure. In essence, it has the same effect as increasing the sample size for a particular study. That is, by increasing the number of test items or sample size, a measure closer to the *true* ability is achieved.

There was an inclination to disregard the RT data in study 9 because the 3DAT was not strictly a speeded test and therefore the variation in reaction times may have been too great for meaningful analysis. However, this would have been a mistake. After some deliberation, the RT data were deemed to be meaningful because they would reflect natural response times from participants and thus serve a real purpose since different sets of RTs in this instance were regarded as measures of different skills. The analysis of RT data turned out to be a lot less complicated to assess than expected, and shaped up to be at least as good an indicator of spatial factors as the choice accuracy data. In fact, several approaches to FA consistently produced good indicators of spatial skills to the point where results were far better than expected. Knowing that RT data are normally skewed and also to have nonconstant variance, a play safe alternative was implemented. The data were *log transformed* which is known to improve the normality of the data and reduce the amount of nonconstant variance. FA was repeated with the log RT data and produced very similar results to the actual RT data which gave confidence that the violation of assumptions did not influence the FA results. In some respects, this approach is similar to using parametric (Pearson) and nonparametric (Spearman) correlation procedures as a precaution when some uncertainty about the data exists. When the difference between the methods is large, it is a warning that there is something unusual about the data and the analysis may then become more complex. Log RT provides a similar note of caution.

In summary of study 9 objectives, test items were assessed and reduced in number using the CTT method with agreement from IRT about the same items for the most part. Measures of item interrelatedness using Cronbach alpha contributed to this reduction although not to a large extent because item reliability was addressed as part of CTT. A gender difference favouring males was found for all subtests though not significant for the MC and TR subtests with the mental rotation task (MR) proving to be the strongest indicator of this difference. These findings were consistent with the literature reported earlier (e.g., Voyer et al., 1995), and unfortunately, there is no convincing evidence of improvement for female designers. However, effect sizes were not large and this perhaps is an encouraging indicator. FA for both choice accuracy and RT data strongly indicated five spatial factors for the spatial ability construct. There was an expectation that the direct oblimin rotation method would be the most appropriate for the data because of an expected correlation between the components. This anticipation was largely based on the disagreement in the literature (refer chapter 1) about the number of factors, or in some cases, the belief that spatial ability is unidimensional. However, the coefficients reported in the *component correlation matrix* tables did not support direct oblimin rotation. In essence, this was a good result because it supported the concept of discrete spatial factors for the 3DAT. If the *component correlation matrix* table showed otherwise, it would indicate cross loading of test items and thus would weaken the argument for unique spatial factors. Lastly, although

analytical procedures did reveal a significant correlation between UAI and the 3DAT, its effect size was below medium according to the Cohen (1992) criteria. On this occasion, UAI was taken as a reasonable measure of the general academic ability of students entering university. With these outcomes in mind, the objectives for study 9 were considered to have been achieved.

Chapter Summary

This section brings together a summary of the main points from three studies where one of the studies in particular was complex because of the diverse range of research questions it investigated. The collective intention of these studies was to bring closure to the development of the 3DAT by reducing it to a final number of subtests and test items after the scrutiny of a concluding set of psychometric procedures.

For study 7 (test retest reliability), the *test retest* method was chosen from two possibilities to test the reliability of the 3DAT. This method was chosen because of its reputation among researchers as the psychometric standard and since it avoided the disadvantages of the alternative method. There was the likelihood of an improved performance on the second administration of the test, but this was not really a concern because it was the measure of correlation between the two tests that mattered most of all. The reliability index for the 3DAT was produced from this procedure. Improvement on the second test did occur and a significant difference between the two results was found and hence confirmed a practice effect. The test retest method was the primary procedure for determining reliability, but a second procedure (contingency tables) based on a different statistical approach was performed to also demonstrate the reliability of the 3DAT and that a significant practice effect existed. In reality, the conduct of two procedures was not strictly required, but nevertheless both were worth doing because the agreement between the two procedures provided confirmation for a very critical issue (reliability) in test development. In effect, the second procedure proved to be an advantage rather than any real disadvantage. Always a concern in any form of testing is potential measurement error. However, the conditions under which the 3DAT was tested for reliability were sound, and a short interval of seven days between testing helped in reducing confounds of this nature. Consequently, testing conditions were well controlled providing confidence that external influences were kept to a minimum. Study 7 also identified a statistical procedure that could be implemented to measure actual learning in a classroom setting after allowing for a practice effect. Some elaboration of this and a description of a resulting template is held over until chapter five.

For study 8 (face validity), the focus was on a validity somewhat less important in psychometric terms than other forms of validity, but important nevertheless. This validity is measured by the opinion of test takers who need to see the relevance of a test they undertake, otherwise there is some likelihood of increased measurement error. In other words, there is a potential risk of the

test not being taken seriously because of a perception of irrelevance. In this study, participants endorsed all subtests as appropriate measures of spatial skills except the TL subtest where the *agree and disagree* scores shown in Table 24 were equal. The poor rating to TL is understandable because the psychometric investigation revealed that the test items in this subtest were not well handled, and this therefore was likely to have contributed to the rating this subtest received. This is an example of where a concept thought to be important to designers is not always recognised or understood. Noteworthy is that some results identify strong agreement (e.g., RC & ED), some identify medium agreement (e.g., TR & DC) and two identified low agreement (SD & FU). Both of the latter subtests were about 3D objects in the unfolded state (surface development) and participants having to recognise what the developments looked like when folded into a solid object. Perhaps participants were making a judgment about a practical skill they thought they did not need rather than about a spatial concept the subtests were designed to measure. Quite correctly, the subtests do not measure a practical skill that designers normally require in the workplace, but they do measure one form of mental manipulation which is a cognitive skill considered important to designers. On issues of spatial ability, Table 25 clearly established the relevance of spatial ability with strong agreement for nine out of the 10 survey questions presented. The remaining question (*spatial ability is an innate skill that cannot be improved*) was an understandable exception since it was a reverse question in essence such that *disagree* was the preferred choice from a research perspective. Importantly, it rated highly.

The survey conducted was detailed and it investigated face validity at three levels and hence went further than most studies normally do. Because there was strong support across all three levels, there was compelling evidence that face validity existed for the 3DAT.

For study 9 (Item and Factor Analysis) which was a complex study because it dealt with many objectives, the focus was essentially on the final developmental stages of the 3DAT. This meant that conclusions were now required and any final assessments needed to be completed. For this to be possible, a selection of statistical procedures were conducted which could be summarised as measuring, identifying and confirming subtests, test items and spatial factors.

A feature of this study was the comparison conducted between the CTT and IRT item analysis methods. The procedures are technically different since the CTT method makes no assumptions about the distribution of data and results are derived objectively without the need for any subjective decision making, or at least in theory. In comparison, the IRT method is based on probability and assumptions and the belief that the test is unidimensional (Cohen et al., 1992). Cohen et al. (1992) also argue that the ability being assessed is not directly measured, and instead, is based on estimated scores and the visual appraisal of the ICC plot. Another difference between the two methods is that CTT considers three parameters (difficulty, discrimination and reliability) while the IRT method used in this study considers two parameters (difficulty and

discrimination). However, despite the differences, both procedures were almost always in agreement about the test items which no doubt would have presented a difficult dilemma had it been otherwise. As previously mentioned, IRT is gaining in popularity among test developers, but there is nevertheless a lasting uncertainty in the minds of some researchers because of the unidimensionality that is assumed in this procedure (Cohen et al., 1992).

Another statistical procedure that was vital to this study was FA. To achieve the final analysis, it was necessary to experiment with a number of settings and combinations in what really amounted to an educated round of trial and error manipulations. Tabachnick and Fidell (1996) advocate an investigative approach to FA and encourage experimentation using a range of options and configurations within the software including different extraction methods. They state that the researcher will ultimately settle on a combination that produces the best scientific solution to their research question. In fact, subjectivity plays a bigger part in FA than researchers generally acknowledge, and this subjectivity continues after FA has produced a satisfactory result. FA helps identify factors, but what those factors are, and what names and descriptions are given to them are decided by the researcher. Gregory (2004) reinforces this view and asserts that a researcher moves from the objective phase provided by statistics to an informed subjective phase to derive a number of final decisions. Gregory further points out that it is not surprising that diligent researchers may come up with different conclusions even though they based their analyses on the same dataset. The acknowledgment of the role of subjectivity in FA is reassuring because it was necessary to introduce some subjectivity into this study to bring about several satisfactory conclusions.

The results for the UAI section of this study were interesting. The range of UAIs (4 x 10 unit bands) was reasonably large for a university standard. Accepting UAI as a reasonable measure of general academic ability, and in view of UAI being the main entry requirement for design programs, the effect size (.209) found for the relationship between UAI and the 3DAT was weak to moderate in strength. This is a concern considering spatial ability is regarded as a critical aspect of graphical understanding in a design context. Since the UAI versus 3DAT relationship was not strong, it suggests that UAI should not be the only criteria for entry into design programs. It also suggests that spatial ability may not be well represented in general academic ability. The study showed that a 13.5% improvement in performance on the 3DAT could be expected across a range of 30 UAI units between a UAI of 65.5 and 95.5 units. Whether this is a substantial or an insignificant amount is outside the scope of this thesis, and this may be a question that is best answered by researchers from the discipline of education and training. However, superficially, it seems like design educators can expect students at the high end of UAI to achieve better results on the 3DAT than those on the low end, but not in keeping with the usual expectations linked to UAI scores because of the weak correlation between UAI

and the 3DAT. The study also revealed that design educators can expect female students to do nearly 12% less well on average than male students on a measure of spatial ability. This differential implies that it would be very difficult for female students on average to receive the highest grade possible (*high distinction*) because these grades generally apply to the 85% to 100% range, that is, over a spread of 15%. The point here is that there is not big difference between 12% and 15%. Taking this one step further, and considering the male and female distributions across the four UAI bands (61 – 70 to 91 – 100) where the sample was 521 (445 male, 76 female), the actual number of *high distinctions* equivalent for the 3DAT was found to be 145 (33%) for males and 12 (16%) for females. Interestingly, there was a significant difference between UAI means based on gender that actually favoured females. To be pedantic, and also to explain one important concept, the 12% value is strictly an *observed* score based on a calculation using a range of 18 which was reported in the *Results* section of this study. This value, however, can be converted to an *absolute* scale using a value of 20 which is the potential range of scores on the 20 test item version of the 3DAT. Based on this, the difference between genders on average would be 10.3% (2.06/20*100). The 2.06 value used in this calculation is the parameter estimate for the gender effect which was also reported in the *Results* section of this study. The 10.3% value has more practical significance as a benchmark figure and could be applied to any assessment where the 3DAT is used.

In retrospect, the UAI data may have been better from a research perspective if it had been collected as individual scores (e.g., 73, 77) rather than in bands such as 71 – 80. This would have produced better information and improved outcomes because more detail would have been available for the statistical analysis. Consequently, this would have increased confidence in the results and allowed better predictions to be made.

A good level of confidence existed in the results produced by study 9 because a sequential and detailed analytical approach was applied to the datasets which was possible because of the large sample that was achieved. The analyses produced convincing evidence of five distinct spatial factors and a trust in five subtests to measure these factors. The only remaining task to be completed at this final stage was to name these factors and to assign good descriptors to define their uniqueness. This was done and is reported in chapter five.

CHAPTER 5

DISCUSSION AND CONCLUSIONS

Overview

The intent of chapter 5 is to bring this thesis to a close and to address several aspects that have not been fully addressed elsewhere in this thesis. The latter is essentially a finishing off process to tidy up some important aspects deserving of formal consideration. These aspects are pertinent to the project overall and help ensure a sense of completion about the thesis. This chapter is intentionally brief, and the format is summary in style and no new findings or statistical analyses are reported. Issues covered are grouped under several main headings that follow.

3DAT Developmental Phases

The 3DAT progressed through a number of modifications before reaching its final configuration of five subtests and 20 test items, and each developmental phase was described in detail in previous chapters. For convenience, a summary of the different versions of the 3DAT is provided in Table 36. Related information and brief outcomes are also shown.

Table 36
Nine Developmental Phases of the 3DAT and Related Information

Phase	Study	Num of Subtests	Num of Test Items	Platform	Changes and Outcomes After Implementation
1	1	9	89	SuperLab & ColdFusion	Reduced to 45 items after CTT analysis. Some subtests in doubt due overlap.
2	2	9	45	SuperLab	No changes required. Study mostly tested a range of validities.
3	3	9	45	SuperLab	Object decision subtest discarded, new subtests and test items required.
4	4	10	60	SuperLab	New labelling system introduced, assembled subtests for SME evaluation.
5	5	25	NA	NA	SMEs ranked 25 subtests. 25 subtests reduced to 15 then to 12.
6	6	12	72	SuperLab	IRT reduced 3DAT to 48 test items prior to further statistical analyses.
7	7	12	72	Adobe Flex	24 discarded test Items from Phase 6 reworked. No other changes required.
8	9	12	72	Adobe Flex	CTT and EFA reduced 12 subtests to 9, then to 7 and then finally to 5 subtests.
9	NA	5	20	Adobe Flex	Final version of the 3DAT consisting of 5 subtests and 4 test items in each.

Note. CTT and IRT: item analysis procedures, Superlab: lab software, ColdFusion and Adobe Flex: online software, EFA: exploratory factor analysis, Num: number, NA: not applicable

Table 36 makes it clear that the 3DAT did not simply reduce in size from a large number of subtests and test items down to a lower number of subtests and test items in a sequence of uniform reductions. Instead, the number of subtests and test items varied either up or down for

each version of the 3DAT according to the findings of the analysis that followed each implementation of the 3DAT. That is, each version of the 3DAT was a result of outcomes determined from the previous version. Of note is that 119 test items were first considered in pilot studies carried out before Phase 1 was initiated.

Spatial Factors Identified

Chapter 4 explained the statistical processes that identified five spatial factors that collectively deliver a measure of the spatial ability construct. Figure 11 illustrates these factors and the spatial ability construct relationship in graphical form. Importantly, names have been assigned to each of the five factors.

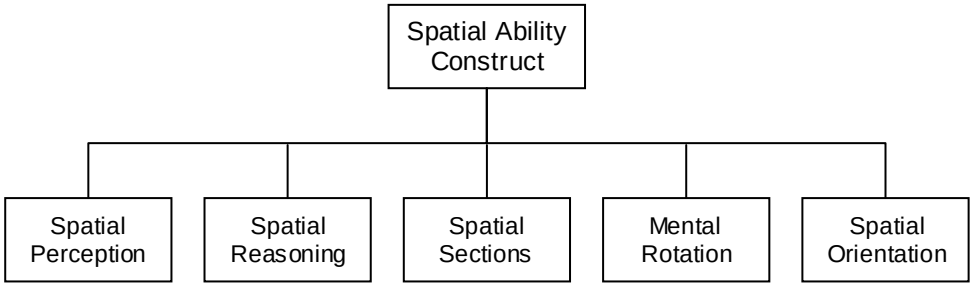


Figure 11. Structural diagram indicating the five spatial factors considered to collectively assess the spatial ability construct. Both choice accuracy and RT data were used to identify five factors (n = 635).

To add clarity, Table 37 itemises the spatial factors with two letter abbreviations shown, and also the subtests within the 3DAT that measure each of these factors. Any future reference to the 3DAT will see a renaming of these subtests to better match the factors they measure, and the two letter abbreviations will be used when required for simplified reporting.

Table 37
Spatial Factors and Subtests that Measure each Factor. Abbreviations Shown in Parentheses.

Spatial Factor	Subtest that Measures Factor
Spatial Perception (SP)	Building Representation (BR)
Spatial Reasoning (SR)	Transformation (TR)
Spatial Sections (SS)	Mental Cutting (MC)
Mental Rotation (MR)	Mental Rotation (MR)
Spatial Orientation (SO)	Dot Coordinate (DC)

Note. Newly Established two Letter Abbreviations for Spatial Factors are Indicated.

The five spatial factors have been given names that approximately describe the skill they measure. Some explanation of why the names were decided upon is thought necessary, and to help appreciate the descriptions that follow, Appendix J contains examples of test items from the subtests that measure each of these factors. First of all, *spatial perception (SP)* requires the skill to visualise the 2D shape of a 3D object when viewed from the opposite direction to a given view. A correct perception or understanding of the 2D 3D relationship is fundamental to this skill. For *spatial reasoning (SR)*, the skill necessary also involves an understanding of the

2D 3D relationship, but some reasoning is also required because the starting point is a plan view (2D) and the profile of the object has to be reasoned from labels in number form that indicate the height of different parts of the object. That is, test takers start with a 2D view and need to reason a correct 3D solution from a set of numbers. The *spatial sections (SS)* factor describes the task of visualizing the shape of a 2D surface that is produced when a 3D object is intersected by a plane at any given angle and after the top portion of the object has been removed. The *mental rotation (MR)* factor is straight forward and is simply a case of being able to mentally rotate a 3D object into another 3D object to decide whether both objects are the same or different (mirror image). The task can be made more difficult by adding complexity to the shape of the object and by changing the angles of inclination, however, the task still remains a mental rotation skill. For *spatial orientation (SO),* the skill measured overlaps skills demonstrated in other factors to some extent, but essentially it is one of orientation. The test taker has to imagine being in different viewing directions to conclude a solution for the task. Being able to mentally place self in sequentially different positions and the use of working memory are fundamental to solving this task. The names given to the spatial factors were based on objective assessment, but it is certainly possible that other names equally descriptive could have been chosen. However, the given names are distinctive and reasonably represent the spatial skill they claim to measure.

Construct Validity

Of all the validities, construct validity is the most difficult to assess, and perhaps for no other reason than there not being a single procedure to measure this validity. Gregory (2004) describes construct validity as the most intangible of the validities which cannot be determined on the basis of a few elementary investigations. Gregory further points out that construct validity relates to psychological tests that measure complicated, multiple and theoretical traits and he provides examples such as psychopathy, leadership and intelligence. Importantly, Gregory adds that no single criterion entirely exists to define construct validity, and that a diverse set of research procedures are required to identify this complex measure. Gregory further argues that a researcher must accumulate a range of evidence from many sources to demonstrate construct validity. Aiken (1997) is of a similar view to Gregory, and maintains that construct validity is not measured by a single method, but instead is measured by a network of studies devised to assess a test that claims to measure a particular psychological construct. Cohen et al. (1992) also agree with Gregory and Aiken and suggest that construct validity is increasingly seen as an amalgamating concept for all forms of validity evidence. There is little doubt that spatial ability is a complex and multifaceted construct, and from the comments cited here, construct validity for the 3DAT should be reported based on collective evidence from a variety of sources. This evidence certainly exists for the 3DAT.

To validate the 3DAT, a number of investigations were carried out based on the theory of spatial ability and the standards of psychometric test development. Gregory (2004), Aiken (1997) and Cohen et al. (1992) all provide similar choices for determining construct validity, and the quest to develop the 3DAT exceeded their minimum recommendations. The validation procedures applied to the 3DAT are listed in Table 38. Also shown are the sections where they are reported and brief evidence of outcomes.

Table 38
Validities that Contribute to Construct Validity for the 3DAT. Locations within the Thesis are also Indicated.

Validation Procedure	Chapter	Study	Evidence
Convergent	2	2	Appropriate positive correlations and effect sizes
Divergent	2	2	No positive correlations with nonspatial tests
Content	3	5	Established from collective opinions of SMEs
Theory Consistent	3	6	Designers significantly outperformed nondesigners
Face	4	8	Agreement among test takers about relevance of 3DAT
Factor Analysis	4	9	Appropriate loadings for accuracy and RT data

Note. Main Sources of Construct Validation are Shown. Minor Evidence Established in Other Sections.

Essentially, establishing construct validity for a test is based on a range of diverse investigations that mostly focus on validation and factor analysis. There are also other possibilities such as the mental processes that occur in responding to test items (Aiken, 1997), or an increase or decrease in test scores as a factor of age according to some theoretical prediction (Cohen et al., 1992). However, these options were not applicable and therefore were not applied to the 3DAT. In real terms, evidence of construct validity can be identified in most practical studies that assess test scores produced by appropriate test takers (Gregory, 2004). Because of the many sources identified in Table 38, evidence of construct validity for the 3DAT is convincing, and this is particularly the case for some sources reported.

Why the 3DAT is Different

The ideas for the 3DAT largely came from spatial tasks sighted in the literature. The tasks acted as idea generators and each test item for the 3DAT was *invented* using the tasks as a starting point. The test items used up to study 3 were developed by the writer using CAD software, and test items developed after this were created by a 3D modeller commissioned by this writer who provided examples and the following instructions:

- produce a consistency in layout and design,

- answer options to be near equally plausible,
- create variation in the difficulty of test items,
- new test items subject to evaluation by the writer before acceptance,
- some reworking of test items may be necessary after evaluation,
- vary the position of the correct answer among possible options, and
- produce more test items than required to allow a choice.

As a consequence, sets of test items were established and each set became a subtest within the 3DAT. The *invention* of exclusive test items was the first distinguishing feature of the 3DAT. The importance of the test items themselves was often referred to in this thesis, and accordingly, they were reworked and critically assessed using item analysis to achieve acceptable psychometric standards.

There are other features as well that also set the 3DAT apart from other measures of spatial ability. Most importantly, the 3DAT was established as an online test that could be used simply as a spatial diagnostic tool to assist learners, or it could be used to collect data for research purposes. In both capacities, the 3DAT provides a summary of performance and a capability to review test items scored incorrectly. Perhaps a special feature of the 3DAT is that it was developed using both choice accuracy and RT data to identify the spatial skills it would measure, and that there was convincing agreement between the datasets. Another feature is the *real learning* template that emerged from this research that could be used by design educators to estimate real learning in a classroom after adjusting for practice effect. Further, unlike other spatial ability tests, the 3DAT measures five specific spatial skills which were all identified from exploratory factor analysis. In many respects, this summary supports the rationale and hypotheses for this thesis.

3DAT and Problems with Current Tests

Under a subheading titled *Problems with Current Tests* listed in chapter 1, the writer pointed to deficiencies with existing spatial tests. In brief, the main problems were:

- restricted in what they reveal about spatial understanding,
- unable to identify more than one spatial skill,
- do not accommodate different solution approaches,
- uncertainty about the number of spatial factors,
- many are nonverbal ability tests that measure 2D skills only, and
- psychometric properties are not generally known.

Fortunately, the 3DAT overcomes these problems as a result of psychometric procedures. For example, the 3DAT consists of multiple subtests that measure a range of spatial skills, and

consequently reveal significant information about the spatial performance of test takers. Because of the mixture of subtests and the variety of skills they measure, the 3DAT is also likely to account for different solution strategies. Importantly, each subtest has a 3D component and therefore this critical consideration is measured by each of the subtests. Because the rigour of psychometrics was applied to the development of the 3DAT, the procedures ensured that the selected test items satisfied psychometric requirements. On the other hand, the 3DAT fell short of one or two desired outcomes. The final version consisted of 20 test items and this renders some risk to reliability and increased practice effect if used too often say as a spatial diagnostic test. Also, some subtests considered very appropriate to designers such as TL did not survive psychometric testing. Further, some psychometric properties such as the internal consistency of test items within some subtests could have been a little more significant.

Learning Issues

This research revealed several learning issues deserving of attention from design educators. The most notable relates to gender from two perspectives. First of all, a gender difference in varying degrees and favouring males was consistently found in relevant studies, and scores were not particularly high for females on the more difficult subtests such as the DC subtest. There was also evidence that females found considerable difficulty with the mental rotation tasks, and these are generally reported in the literature as showing the strongest gender difference. The second perspective draws attention to an imbalance in the genders studying design at tertiary level. For the major sample reported in this research (n = 635) the proportion was in the order of six males for every one female (male = 535, female = 100). If it were not for the popularity of architecture among females, and also chemical engineering to a lesser extent, the ratio would be much greater than this. Engineering across all disciplines, except for say chemical engineering, is most conspicuous because of its inability to attract female students. Collectively, the two perspectives are making a statement about female participation in engineering programs. In what appears to be a general worldwide shortage of engineers suspected by the writer, recruitment measures so far have not been effective and there remains widespread concern that more females are not being attracted to engineering disciplines.

Also to emerge from this research was a difference in results between the five factors of spatial ability measured by the 3DAT. This made it obvious that the 3DAT could identify specific skills that test takers were deficient in, and consequently in need of some form of learning assistance. Using the 3DAT for spatial diagnostic purposes does mean that particular difficulties will be identified, which in turn means educators can develop learning tasks that target these difficulties. A case in point is the DC subtest. In concise terms, this subtest requires the ability to work from 3D to 2D, some working memory and for the test taker to view an object from a number of viewing directions. An appropriate learning task could be devised as a three stage

activity that aligns with each of these skills and allows the learner to improve one skill before moving to the next. Thus, the diagnostic potential of the 3DAT is that it may recognise specific abilities that require some form of remediation. Without this potential, design educators are at risk of turning to unproven tasks with an unrealistic expectation of bringing about improvement. The 3DAT with its ability to measure five separate spatial factors will provide data to assist in the development of purpose designed 3D learning tasks.

The *real learning* template referred to earlier also fits nicely into this section. It offers educators an objective method based on statistical procedures to determine whether any real learning beyond the practice effect has actually occurred in the classroom. As a reminder to the reader, a copy of the template is shown at Appendix Q. All that is required is for an educator to copy and paste numerical data from a 3DAT pretest and posttest into dedicated columns within the template. After entering the data, functions built into the template calculate total learning, subtract practice effect and report real learning in percentage terms. The template will also provide the *t* statistic and the statistical probability (*p*) and the correct APA format for reporting the result of a statistical test. Other calculations occur in the template (e.g., degrees of freedom, mean differences), but the educator does not have to interpret these to appreciate the essential calculations. The template is derived from the 12 subtest 72 test item version of the 3DAT but it can easily be adjusted to suit any version including the final five subtest 20 test item model. The re-calibration of practice effect is also possible. The calculation of *real learning* is conservative and is likely to be an underestimate of real learning, and reasons for this are also detailed in the *Discussion* section of study 7. The results produced by the template remove subjectivity and provide good evidence of whether or not a classroom activity designed to improve spatial performance has been successful. In so doing, the results may prompt the educator that more has to be done to bring about improvement, but importantly, it will safeguard against misconceptions about perceived spatial learning. Interestingly, the same principles could be applied to any learning environment where the effectiveness of a learning intervention is wanting to be known.

Future Research

There is potential to extend research that focuses on the 3DAT and spatial ability in general. A great deal of this would be linked to extra test items, additional subtests, improved online features and the development of 3D learning tasks to improve spatial understanding.

Perhaps the first opportunity to consider is revisiting subtests used in earlier versions of the 3DAT which were rejected by one of the psychometric procedures carried out. The TL, ED and RC subtests immediately come to mind. All three appear to be relevant to designers because one subtest (TL) investigates the important concept of true length, and all three subtests probe the understanding of the 2D 3D relationship. For the ED subtest in particular, it closely resembles a

skill that designers would use on a regular basis when graphical communication is important. Of note is that ED is an abbreviation for *engineering drawing*. The only reason the three subtests were rejected was because the test items within each did not satisfy psychometric standards. In other words, they were not rejected because they were conceptually wrong. For the TL subtest, the design of the test items appeared difficult for test takers to interpret, however, the question is whether this was a fault of the test items themselves, or was it an indication that the concept of true length was not well taught or not well understood by test takers. Even though there is a rote learning procedure that test takers can follow for this subtest, there was no evidence that this had occurred. Being able to identify this deficiency alone would make the inclusion of the TL subtest worthwhile. For the ED subtest, the visual information for the most part was convoluted and cluttered with too many hidden detail lines thus making it a difficult task to comprehend. For RC, it was clearly a subtest characterised by the 2D 3D relationship that is important to designers, but the test items were almost always rejected because they failed to meet the minimum standards for reliability. In view of their likely contribution to spatial ability, all three subtests should be reworked and tested again since they appear to tap into important spatial skills required by designers. Just to strengthen this argument further, all three subtests were rated highly by subject matter experts (SMEs) and two out of the three were rated highly by the test takers who reported face validity. The reintroduction of these subtests in a later release of the 3DAT may complement the current subtests, or in fact, they may reveal other spatial skills not yet identified by the 3DAT.

There is also other potential for research. For example, the existing four test items per subtest for the 3DAT appears to be the minimum number to achieve acceptable reliability. Hence, one area that should be investigated is increasing the number of test items to say six or eight to improve reliability. The *Discussion* section of study 9 showed that reliability could be improved by adding new test items provided they had similar psychometric properties to existing test items. Thus, a research project that develops additional test items would be worthwhile. Another area to consider is investigating whether further information can be extracted from the RT data generated by the 3DAT. In particular, perhaps it can be shown that RT is also a measure of spatial competency in addition to the number of test items scored correctly which is currently the only measure used. A starting point is the work done by Roberts and Stankov (1999) who found a relationship between intelligence and processing speed. They point out, however, that the relationship appears dependent upon manipulations of task difficulty, and importantly, that a number of models that purport to explain the link between intelligence and processing speed are unsustainable. A further research option is the development of spatial ability tests for other disciplines using the 3DAT as a starting point. Possible examples include medical imaging, aviation (e.g., air traffic control) and x-ray image analysing (prohibitive cargo). Another

research opportunity is interface design for online tests. The 3DAT was a complex development that required over 500 images to produce the final version. It is possible that lessons learnt from the 3DAT experience could lead to enhancements that improve the readability and useability of interfaces with the aim of further reducing measurement error, increasing levels of participation and enabling more speedy delivery.

One final research possibility relates to the development and testing of 3D learning tasks to improve spatial understanding. In many respects, this is the next big step. Spatial diagnostics and measurement should be followed by efforts to improve spatial ability, and perhaps particularly so for females. Some preliminary research and investigation conducted by this writer suggests that, to be effective, learning tasks should be interactive, allow object manipulation controlled by the learner, provide feedback and have different levels of competency built in. This approach is best described as *active exploration*. Ideally, the learning tasks would be online and should at least develop mental rotation skills, visualisation skills, and 2D to 3D and 3D to 2D understanding. The advent of modern and comprehensive software has provided new opportunities to develop spatial skills. This technology should be explored.

Closing Comments

As the very last step in bringing this thesis to a close, there remains several issues to highlight that did not fit with other sections in this chapter, and it seemed remiss not to give them a proper mention. Each of these are addressed in separate paragraphs below.

Unidimensionality is a term applied to tests on the assumption that each test item is measuring the same underlying factor. However, this is not always the case. The 3DAT for example should not be seen as a one-dimensional test since it was developed specifically to measure a number of spatial factors, or in other words, a number of dimensions. The Cronbach alpha index for the 3DAT (across all 20 test items) is .761, and this reasonably high index appears to be an anomaly because there are five discrete factors (dimensions) measured by the 3DAT and the correlation coefficients between them are low (refer Table 28). However, there is a phenomenon where internal consistency for a test can be shown to be high although a test such as the 3DAT is clearly not one-dimensional. Gregory (2004) points out that traditionally, Cronbach alpha has been thought of as a test of unidimensionality, or stated another way, a test of one dimension. Gregory clarifies this and adds that it is possible for a test to appraise more than one factor, but at the same time still show a very strong alpha index. Cohen et al. (1992) supports this view and refers to a measure like the 3DAT as *heterogeneous* (i.e., designed to measure more than one factor) as opposed to *homogeneous* (i.e., designed to measure one factor). A researcher should be aware that it is almost automatic that alpha will be high for a test where alpha is also high for some subtests that measure different factors within that test. The explanation is that these values contribute to alpha overall. This occurrence is something that should not be taken out of context,

and nor should too much be made of it. A researcher is wise not conclude that a high alpha for a test overall is an indicator of unidimensionality and is therefore measuring one factor only. The assumption of unidimensionality may also be an issue for IRT analysis, but this was addressed in the summary section of chapter 4.

Various sources of measurement error were covered in earlier parts of this thesis, but the bigger mention occurred in the summary section of chapter 3. This section particularly drew attention to four main sources of measurement error identified by Gregory (2004), and two of these (item selection and test scoring) were treated in some detail. The remaining two sources (systematic errors and test administration) are discussed here to finish off the overall coverage of measurement error. Systematic errors of measurement occur when a test consistently measures some unknown dimension besides the intended dimension. An example might be a test designed to measure spatial ability, but at the same time, also unintentionally measures nonverbal reasoning. The difficulty for the test developer is that this form of error is not always obvious or easily identified. The best safeguard against systematic errors, although there is no guarantee, is to produce a test in strict accordance with recognised psychometric test development standards (Gregory, 2004). In the case of the 3DAT, the potential conflict between 3D ability, 2D ability and nonverbal reasoning was always considered. Moving to test administration as a source of measurement error, this may occur when a test developer cannot be totally confident about test conditions (e.g., environmental comfort), the test taker's mental state (e.g., anxiety) and the test supervisor's manner (e.g., inconsistency). To a large extent, the delivery of the 3DAT manages to avoid most of the more serious measurement errors induced by test administration. The 3DAT is delivered online and requires only a low level of management and control. It is also self-contained in that it includes instruction screens, instructional videos and practice trials. In addition, the breaks built into the 3DAT which are controlled by the test taker help reduce fatigue. Another advantage of the 3DAT being online is that it can be delivered in a computer laboratory or in the home where the test environment (e.g., sound, air-conditioning & lighting) can mainly be controlled to advantage the test taker. The test user and the test developer in particular are concerned about all sources of measurement error simply because of the impact they have on test delivery and reliability. Measurement errors reduce the repeatability of a test and therefore raise concerns about results obtained from any application of the test (Gregory, 2004). All this serves as a reminder that a psychological measure is only ever an estimate, and that a true measure of ability is probably never achieved.

The development of the 3DAT was a progressive undertaking that occurred across a sequence of many studies that were less revealing at first, but most revealing towards the end. Fundamental to the research methodology were the stated hypotheses and the study objectives that evolved from these. Strict compliance was foremost during every step, and as a consequence, the final

outcomes of this research supported the hypotheses and study objectives. The 3DAT will continue to be reviewed since any psychological test should undergo continual scrutiny even after the comprehensive evaluation it underwent during development. In reality, the key concepts in any test development are *reliability* and *validity* and both should be treated as *dynamic* and subject to regular review.

The 3DAT can be seen and experienced by visiting a special website set up for this purpose, and to do this, the reader is directed to: http://psych.newcastle.edu.au/SpatDiag/3DAT/ and the password to access the site is: "3DAT". Spatial ability is an attribute that is not readily understood, but its importance to the world of design, graphical communications and product manufacture is widely acknowledged. Measuring and improving spatial ability are likely to become even more critical in the coming years with advances in technology, greater demands for design initiatives, changes in building and engineering structures, and with the expectation of faster delivery by consumers. The development of the 3DAT has been a very satisfying experience for this writer, and it is now becoming a reward in itself since the 3DAT has been used, and continues to be used by a number of universities in Australia. Different applications of the 3DAT will continue to be reported in various publications, and at different conferences with a view to encouraging others to participate in spatial research. At the time of writing, a large government agency was showing a keen interest in adapting the 3DAT to assist in the selection and training of personnel in positions very dependent upon spatial ability. This writer has been approached to assist in that development.

REFERENCES

Adanez, G., & Velasco, A. (2002). Predicting academic success of engineering students in technical drawing from visualization test scores. *Journal of Geometry and Graphics, 6*(1), 99 - 109.

Aiken, L. R. (1997). *Psychological testing and assessment*. Boston: Allyn and Bacon.

Akasah, Z., & Alias, M. (2006). *Bridging the spatial visualisation skills gap through engineering drawing using the whole-to-parts approach.* Paper presented at the Australasian Association of Engineering Education Conference, Auckland University of Technology, New Zealand.

Alias, M., Black, T. R., & Gray, D. E. (2002). Effect of instructions on spatial visualisation ability in civil engineering students. *International Education Journal (http://iej.cjb.net), 3*(1), 1-12.

Allahyar, M., & Hunt, E. (2003). The assessment of spatial orientation using virtual reality techniques. *International Journal of Testing, 3*(3), 263-275.

Ben-Chaim, D., Lappan, G., & Houang, R. (1988). The effect of instruction on spatial visualization skills of middle school boys and girls. *American Educational Research Journal, 25*(1), 51-71.

Bertoline, G. R., & Miller, D. C. (1990). A visualization and orthographic drawing test using the macintosh computer. *Engineering Design and Graphics Journal, 54*(1), 1-7.

Birnbaum, M. (2004). Human research and data collection via the internet. *Annual Review of Psychology, 55*, 803-832.

Blasko, D., Holliday-Darr, K., Mace, D., & Blasko-Drabik, H. (2004). VIZ: The visualization assessment and training Web site. *Behavior Research Methods, Instruments, & Computers, 36*(2), 256-260.

Boersma, N., Hamlin, A., & Sorby, S. (2004). *Work in Progress - Impact of a remedial 3D visualization course on student peformance and retention.* Paper presented at the 34th ASEE/IEEE Frontiers in Education Conference, Savannah, GA.

Bore, M., & Munro, D. (2002). Mental Agility Test. *Personal qualities assessment, Newcastle, Australia: TUNRA.*

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis.* Great Britain: John Wiley and Sons Ltd.

Brace, N., Kemp, R., & Snelgar, R. (2009). *SPSS for Psychologists* (4th ed.). Great Britain: Palgrave Macmillan.

Branoff, T. (1998). The effects of adding coordinate axes to mental rotations task in measuring spatial visualization ability in introductory undergraduate technical graphics courses. *Engineering Design and Graphics Journal, 62*(2), 16-34.

Caplan, P., MacPherson, G., & Tobin, P. (1985). Do Sex-Related Differences in Spatial Abilities Exist? *American Psychologist, 40*(7), 786-799.

Carpenter, P. A., & Just, M. A. (1986). Spatial Ability: An Information Processing Approach to Psychometrics. . In R. J. Sternberg (Ed.), *Advances in the Psychology of Human Intelligence* (Vol. 3, pp. 221-253). Hillsdale: NJ: Erlbaum.

Carroll, J. B. (1992). Cognitive Abilities: The State of the Art. *Psychological Science, 3*(5), 266-270.

Carroll, J. B. (1993). *Human Cognitive Abilities: A Survey of Factor-analytic Studies* (1st ed.). USA: Cambridge University Press.

Cohen. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

Cohen, Swerdlik, M. E., & Smith, D. K. (1992). *Psychological Testing and Assessment*. Mountain View, California: Mayfield Publishing Company.

Cohen, J. (1988). *Statistical Power for the Behavioural Sciences* (2nd Edition ed.). New Jersey: Lawrence Erbaum Associates, New Jersey.

Coleman, S., L, & Gotch, A., J. (1998). Spatial perception skills of chemistry students. *Journal of Chemical Education, 75*(2), 206.

Contero, M., & Naya, F. (2006). Learning support tools for developing spatial abilities in engineering design. *International Journal of Engineering Education, 22*(3), 470-477.

Contero, M., Naya, F., Company, P., Saorin, J., & Conesa, J. (2005). Improving visualization skills in engineering education. *Computer Graphics in Education*(September/October), 24-31.

Cooper, L. A. (1990). Mental representation of three-dimensional objects in visual problem solving and recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition, 16*(6), 1097-1106.

D'Oliveira, T. (2004). Dynamic spatial ability: an exploratory analysis and a confirmatory study. *The International Journal of Aviation Psychology, 14*(1), 19-38.

Day, J. (2006). *Validating a measurement instrument of 3D understanding with unskilled groups.* Unpublished Degree of Bachelor of Psychology Honours Thesis, School of Psychology, The University of Newcastle, Callaghan NSW Australia.

De Lisi, R., & Wolford, J. (2002). Improving children's mental rotation accuracy with computer game playing. *The Journal of Genetic Psychology, 163*(3), 272-282.

Deary, I., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35*, 13-21.

Deno, J. (1995). The relationship of previous experiences to spatial visualization ability. *Engineering Design Graphics Journal, 59*(3), 5-17.

Duesbury, R. T., & O'Neil, H. F. (1996). Effect of type of practice in a computer-aided design environment in visualizing three-dimensional objects from two-dimensional orthographic projections. *Journal of Applied Psychology, 81*(3), 249-260.

Fitts, P. M. (1964). Perceptual-motor skill learning. In A. W. Melton (Ed.), *Categories of human learning* (pp. 243-285): Academic Press.

Garg, A., Norman, G., & Sperotable, L. (2001). How medical students learn spatial anatomy. *The Lancet, 357*(February 3), 363-364.

Gluck, J., & Fitting, S. (2003). Spatial strategy selection: interesting incremental information. *International Journal of Testing, 3*(3), 293-308.

Gregory, R. J. (2004). *Psychological Testing: History, Principles and Applications* (4th ed.). Boston: Pearson Education Group Inc.

Guilford, J., & Fruchter, B. (1978). *Fundamental Statistics in Psychology and Education* (6th ed.). New York: McGraw-Hill.

Hartman, N., & Bertoline, G. (2005). *Spatial visualization tests and their relationships to contemporary CAD tools: advocating more than just mental rotations tests*. Paper presented at the American Society for Engineering Education Annual Conference & Exposition 2005.

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law appealed: The case for an exponential law of practice. *Psychonomic Bulletin and Review, 7*, 185-207.

Holliday-Darr, K., Blasko, D., & Dwyer, C. (2000). Improving cognitive visualization with a web based interactive assessment and training program. *Engineering Design and Graphics Journal, 64*(1), 4-9.

Hsi, S., Linn, M., & Bell, J. (1997). The role of spatial reasoning in engineering and the design of spatial instruction. *Journal of Engineering Education, Winter*, 151-158.

James, K. H., Humphrey, G. K., & Goodale, M. A. (2001). Manipulating and recognizing virtual objects: where the action is. *Canadian Journal of Experimental Psychology, 55*(2).

Kaplan, R. M. (1987). *Basic Statistics for the Behavioral Sciences*. Boston: Allyn and Bacon, Inc.

Kass, S. J., Ahlers, R. H., & Dugger, M. (1998). Eliminating gender differences through practice in an applied visual spatial task. *Human Performance, 11*(4), 337-349.

Kazis, L. E., Anderson, J. J., & Meenan, R. F. (1989). Effect sizes for interpreting changes in health studies. *Medical Care, 27*(3, Supplement), 178-189.

Krantz, J. H. (2001). Stimulus delivery on the web: What can be presented when calibration isn't possible.

. In I. U.-D. R. M. Bosnjak (Ed.), *Dimensions of Internet Science* (pp. 113-130). Lengerich, Germany: Pabst.

Kvale, S. (1996). *Interviews: an introduction to qualitative research interviewing*. Thousand Oaks, California: Sage Publications.

Kyllonen, P. C., & Lajoie, S. P. (2003). Reassessing aptitude: introduction to a special issue in honor of Richard E. Snow. *Educational Psychologist, 38*(2), 79-83.

Laver, M. (2007). *Gender bias in spatial cognition performance.* Unpublished Bachelor of Psychology Honours Thesis, School of Psychology, The University of Newcastle

Callaghan NSW Australia.

Leopold, C., Gorska, R. A., & Sorby, S. A. (2001). International experiences in developing the spatial visualization abilities of engineering students. *Journal for Geometry and Graphics, 5*(1), 81-91.

Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development, 56*(6), 1479-1498.

Maier, P. H. (1998). Spatial geometry and spatial ability: how to make solid gemetry solid? In E. Cohors-Fresenborg, K. Reiss, G. Toener & H. G. Weigand (Eds.), *Selected papers from Annual Conference of Didactics of Mathematics 1996* (pp. 69-81): Osnabreck.

McGee, M. G. (1979). Human Spatial Abilities: Psychometric Studies and Environmental, Genetic, Hormonal, and Neurological Influences. *Psychological Bulletin, 86*(889-918).

McGraw, K., O, Tew, M., D, & Williams, J., E. (2000). The integrity of web-delivered experiments: can you trust the Data? *Psychological Science, 11*(6, November 2000), 502-506.

Medina, A. C., Gerson, H. B. P., & Sorby, S. A. (1998). *Identifying gender differences in the 3D visualization skills of engineering students in Brazil and in the United States.* Paper presented at the International Conference of Engineering Education, Rio de Janeiro.

Metzler, D., & Shepard, S. (1988). Mental rotation: effects of dimensionality of objects and type of task. *Journal of Experimental Psychology: Human Perception and Performance, 14*(1), 3-11.

Miller, C. L. (1992). Enhancing visual literacy of engineering students through the use of real and computer generated models. *Engineering Design Graphics Journal, 56*(No 1 Winter, 1992), 27-38.

Miller, C. L., & Bertoline, G. R. (1991). Spatial visualization research and theories: their importance in the development of an engineering and technical design graphics curriculum model. *Engineering Design Graphics Journal, 55*(3), 5-14.

Olkun, S. (2003). Making connections: improving spatial abilities with engineering drawing activities. *International Journal of Mathematics Teaching and Learning*(April 2003), 1-10.

Osborn, J. R., & Agogino, A. M. (1992). An interface of interactive spatial reasoning and visualization. *ACM 0-89791-513-5/92/0005-0075*(May 3-7 1992), 75-82.

Paivio, A. (1986). *Mental Representations: A Dual Coding Approach*. New York: Oxford University Press.

Pellegrino, J. W., Alderton, D. L., & Shute, V. J. (1984). Understanding spatial ability. *Educational Psychologist, 19*(3), 239-253.

Pollock, R. (2006). *Validating an instrument of 3D understanding with skilled groups.* Unpublished Bachelor of Psychology Honours Thesis, School of Psychology, The University of Newcastle, Callaghan NSW Australia.

Potter, C., Kaufman, W., Delacour, J., Mokone, M., van der Merwe, E., & Fridjhon, P. (2009). Three dimensional spatial perception and academic performance in engineering graphics: a longitudinal investigation. *South African Journal of Psychology, 39*(1), 109-121.

Potter, C., & van der Merwe, E. (2001). *Spatial ability, visual imagery and academic performance in engineering graphics.* Paper presented at the International Conference on Engineering Education, Oslo, Norway.

Quaiser-Pohl, C., & Lehmann, W. (2002). Girls' spatial abilities: charting the contributions of experiences and attitudes in different academic groups. *British Journal of Educational Psychology, 72*, 245-260.

Rafi, A. (2006). On improving spatial ability through computer-mediated engineering drawing instruction. *Educational Technology & Society, 9*(3), 149-159.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Standard Progressive Matrices Raven Manual: Section 3* (2000 ed.). San Antonio, Texas 78259 USA: Harcourt Assessment Inc.

Reips, U.-D. (2002a). Standards for internet-based experimenting. *Experimental Psychology, 49*, 243-256.

Reips, U.-D. (2002b). Internet-based psychological experimentation. *Social Science Computer Review, 20*(3), 241-249.

Reips, U.-D., & Stieger, S. (2004). Scientific LogAnalyzer: a web-based tool for analyses of server log files in psychological research. *Behavior Research Methods, 36*(2), 304-311.

Roberts, R., & Stankov, L. (1999). Individual differences in speed of mental processing and human cognitive abilities: toward a taxonomic model. *Learning and Individual Differences, 11*(1), 1-120.

Salthouse, T. A. (1991). Age and experience effects on the implementation of orthographic drawings of three-dimensional objects. *Psychology and Aging, 6*(3), 426-433.

Schacter, D. L., & Cooper, L. A. (1990). Implicit memory for unfamiliar objects depends on access to structural descriptions. *Journal of Experimental Psychology: General, 119*, 5-24.

Sexton, T. J. (1992). Effect on spatial visualization: Introducing basic engineering graphic concepts using 3D CAD technology. *Engineering Design Graphics Journal, 56*(3), 36-43.

Sorby, S. (1999). Developing 3-D spatial visualization skills. *Engineering Design Graphics Journal, 63*(2), 21-32.

Sorby, S. (2000). Spatial abilities and their relationship to effective learning of 3-D solid modelling software. *Engineering Design Graphics Journal, 64*(30-35).

Sorby, S. (2007). Sustained efforts in developing 3-D spatial skills for engineering students. *Publication details not known.*

Sorby, S., & Baartmans, B. J. (1996). A course for the development of 3-D spatial visualization skills. *Engineering Design Graphics Journal, 60*(1), 13-20.

Sorby, S., & Baartmans, B. J. (2000). The development and assessment of a course for enhancing the 3-D spatial visualization skills of first year engineering students. *Journal of Engineering Education, 89*(3), 301-307.

Sorby, S., Drummer, T., Hungwe, K., & Charlesworth, P. (2005). *Developing 3D spatial visualization skills for non-engineering students.* Paper presented at the American Society for Engineering Education Annual Conference & Exposition.

Speelman, C., & Kirsner, K. (2005). *Beyond the Learning Curve*. New York: Oxford University Press.

Spittaels, H., & Bourdeaudhuij, I. D. (2006). Implementation of an online tailored physical activity intervention for adults in Belgium. . *Health Promotion International, 21*(4), 311-319.

Steyvers, M., & Malmberg, K. (2003). The effect of normative context variability on recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 29*(5), 760-766.

Strong, S., & Smith, R. (2001). Spatial visualization: fundamentals and trends in engineering graphics. *Journal of Industrial Technology, 18*(1), 2-6.

Sutton, K., Heathcote, A., & Bore, M. (2005). *Implementing a web-based measurement of 3D understanding.* Paper presented at the Australian Computer-Human Interaction Special Interest Group Conference, Canberra, Australia.

Sutton, K., Heathcote, A., & Bore, M. (2007). Measuring 3D understanding on the web and in the laboratory. *Behavior Research Methods, 39*(4), 926-939.

Sutton, K., & Williams, A. (2007). *Spatial cognition and its implications for design.* Paper presented at the International Association of Societies of Design Research 2007, The Hong Kong Polytechnic University School of Design.

Tabachnick, B., & Fidell, L. (1996). *Using Multivariate Statistics* (3rd ed.): HarperCollins College Publishers.

Tartre, L. A. (1993). Spatial Skills, Gender and Mathematics. In E. H. Fennema & G. C. Leder (Eds.), *Mathematics and Gender* (pp. 27 - 59). St Lucia: University of Queensland Press.

Thurstone, L. (1950). Some Primary Abilities in Visual Thinking. *Proceedings of the American Philosophical Society, 94*(6), 517-521.

Ullman, K. M., & Sorby, S. (1995). Enhancing the visualization skills of engineering students through computer modeling. *Computer Applications in Engineering Education, 3*(4), 251-258.

Voyer, D., Voyer, S., & Bryden, M. (1995). Magnitude of sex differences in spatial abilities: a meta-analysis and consideration of critical variables. *Psychological Bulletin, 117*(2), 250-270.

Wanzel, K. R., Hamstra, S. J., Anastakis, D. J., Matsumoto, E. D., & Cusimano, M. D. (2002). Effect of visual-spatial ability on learning of spatially-complex surgical skills. *The Lancet, 359*(January 19, 2002), 230-231.

Willis, G. B. (2005). *Cognitive interviewing: a tool for improving questionnaire design*. Thousand Oaks, California: Sage Publications.

Workman, J. E., Caldwell, L. F., & Kallal, M. J. (1999). Development of a test to measure spatial abilities associated with apparel design and product development. *Clothing and Textiles Research Journal, 17*(3), 128-133.