DEVELOPMENT OF THE MULTIPLE APTITUDE TEST – FORM D (VISUAL-SPATIAL) FOR CAREER GUIDANCE IN A PRIVATE UNIVERSITY

SABRENA GABRIELLA AROSH

FACULTY OF EDUCATION
UNIVERSITY OF MALAYA
KUALA LUMPUR

2019

DEVELOPMENT OF THE MULTIPLE APTITUDE TEST – FORM D (VISUAL-SPATIAL) FOR CAREER GUIDANCE IN A PRIVATE UNIVERSITY

SABRENA GABRIELLA AROSH

DISSERTATION SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF EDUCATION (MEASUREMENT AND EVALUATION)

FACULTY OF EDUCATION
UNIVERSITY OF MALAYA
KUALA LUMPUR

UNIVERSITI MALAYA

ORIGINAL LITERARY WORK DECLARATION

Name of Candidate: Sabrena Gabriella Arosh

Registration/Matric No: PMB160001

Name of Degree: Master of Education

Title of Project Paper/Research Report/Dissertation/Thesis ("this Work"):

Development of the Multiple Aptitude Test – Form D (Visual-Spatial) for Career Guidance in a Private University

Field of Study: Education (Measurement and Evaluation)

I do solemnly and sincerely declare that:

(1) I am the sole author/writer of this Work;
(2) This Work is original;
(3) Any use of any work in which copyright exists was done by way of fair dealing and for permitted purposes and any excerpt or extract from, or reference to or reproduction of any copyright work has been disclosed expressly and sufficiently and the title of the Work and its authorship have been acknowledged in this Work;
(4) I do not have any actual knowledge nor do I ought reasonably to know that the making of this work constitutes an infringement of any copyright work;
(5) I hereby assign all and every rights in the copyright to this Work to the University of Malaya ("UM"), who henceforth shall be owner of the copyright in this Work and that any reproduction or use in any form or by any means whatsoever is prohibited without the written consent of UM having been first had and obtained;
(6) I am fully aware that if in the course of making this Work I have infringed any copyright whether intentionally or otherwise, I may be subject to legal action or any other action as may be determined by UM.

Candidate's Signature Date

Subscribed and solemnly declared before,

Witness's Signature Date

Name:
Designation:

ABSTRACT

This study presents the development of items for the visual-spatial component of a Multiple Aptitude Test currently being developed in a local private university in Malaysia. Despite the benefits of aptitude testing in career guidance and counselling, it is often overlooked in favour of interest and personality testing. Within Malaysia, aptitude testing has been implemented in primary schools, but it is not placed within the career guidance and counselling process to help guide students in making education and career decisions. To meet the need for holistic contextualised assessment tools for career counselling, the career guidance unit of a local private university developed a test battery known as the HELP Career Readiness Evaluation System (HELP CaRES) which included assessments of career readiness, employability skills, personality, interests, and aptitude suitable for students aged 14-25. However, the first three versions of the Multiple Aptitude Test proved too easy to accurately measure the aptitude of students aged 17-25. Due to a lack of large visual-spatial item pool, the current study aims to develop visual-spatial aptitude items that are well matched to this age group. An initial 30-item instrument measuring skills of figure-ground perception, object assembly, progressive series, spatial orientation, spatial visualisation, and visual discrimination was piloted with 149 first-year undergraduates. Pilot results of the confirmatory factor analysis showed good model fit and the instrument had acceptable levels of person and item reliability according to Rasch analysis. 15 items were removed to be replaced with new items, 13 items retained with no changes, and two items were retained with changes. After amendments were undertaken, the scale was distributed to 203 first-year undergraduates of a local private university. It was found that the scale exhibited improved model fit from the pilot study based on the confirmatory factor analysis with better reliability indices in terms of item

and person reliability based on the Rasch analysis. A total of nine items need to be replaced, 19 items are retained with no changes and two items require distractor amendments based on the findings of the distractor analysis. A comparison of results between the pilot and the actual study was presented and discussed. The findings of the current study show empirical support for Carroll's (1993) three-factor theory and extend its applicability to an Asian context. The scale under development also presents an empirically-validated career assessment tool that is well-grounded in literature and contextualised to an Asian setting. Future directions of the study include further revisions of the items to ensure validity and reliability of the scale, establishing norms using standard setting techniques, and application of two- and three-parameter item response theory models to refine item characteristics.

Keywords: Visual-spatial aptitude, assessment, test development, Rasch analysis, confirmatory factor analysis

PEMBINAAN UJIAN PELBAGAI APTITUD – JENIS D (SPATIAL VISUAL) UNTUK PEMBIMBINGAN KERJAYA DALAM UNIVERSITI SWASTA

ABSTRAK

Kajian ini menghuraikan pembinaan item untuk komponen spatial visual sebuah Ujian Pelbagai Aptitud yang sedang dibina di sebuah universiti tempatan swasta dalam Malaysia. Walaupun penggunaan ujian aptitude membawa banyak manfaat kepada pengguna, ia sering diabaikan berbanding penggunaan ujian minat dan personaliti. Dalam Malaysia, pengujian aptitude telah dimulakan pada peringkat sekolah rendah tetapi dibuat di luar konteks program bimbingan dan kaunseling kerjaya yang bertujuan membimbing pelajar dalam membuat keputusan berkaitan dengan pendidikan dan kerjaya. Untuk memenuhi keperluan instrumen penilaian yang holistik dan berkonteks tempatan, pusat bimbingan kerjaya sebuah universiti tempatan swasta membina bateri ujian yang dikenali sebagai Sistem Penilaian Kesediaan Kerjaya HELP (HELP CaRES) yang mengandungi ujian kesediaan kerjaya, kemahiran pekerjaan, personaliti, minat, dan aptitud yang sesuai untuk pelajar berumur 14-25 tahun. Walau bagaimanapun, ketiga-tiga versi awal Ujian Pelbagai Aptitud terlalu mudah untuk mengukur aptitud pelajar berumur 17-25 tahun dengan tepat. Oleh kerana bank item untuk item spatial visual tidak mencukupi, kajian ini bertujuan membina item aptitud spatial visual yang sesuai digunakan untuk julat umur ini. Sebuah instrumen 30-item awal yang mengukur kemahiran persepsi bentuk-latar, pemasangan objek, siri progresif, orientasi spatial, visualisasi spatial, dan diskriminasi visual digunakan dalam kajian rintis dengan 149 mahasiswa tahun pertama. Hasil kajian rintis dengan analisis pengesahan faktor menunjukkan fit model yang baik dan instrumen juga menunjukkan nilai keboleh percayaan item dan orang yang boleh diterima menurut hasil analisis Rasch. 15 item dikeluarkan untuk diganti dengan item baru, 13

item dikekalkan tanpa perubahan, dan dua item dikekalkan dengan perubahan. Setelah perubahan dibuat, instrument diedarkan semula ke 203 mahasiswa tahun pertama di universiti tempatan swasta. Hasil analisis pengesahan faktor menunjukkan penambah baikan fit model dari kajian rintis dengan indeks keboleh percayaan yang lebih baik dari segi keboleh percayaan item dan orang berdasarkan keputusan analisis Rasch. Sejumlah sembilan item perlu diganti, 19 item dikekalkan tanpa perubahan, dan dua item memerlukan perubahan distraktor berdasarkan keputusan analisis distraktor. Perbandingan keputusan antara kajian rintis dan kajian sebenar dibentangkan dan dibincangkan. Hasil kajian ini menyokong teori tiga faktor Carroll (1993) secara empirikal dan mengembangkan keboleh gunaan teori ini dalam konteks Asia. Instrumen yang sedang dibina juga merupakan alat untuk penilaian kerjaya yang disahkan secara empirik, mempunyai asas yang kukuh berdasarkan literatur, dan sesuai dalam konteks Asia. Langkah seterusnya untuk mengembangkan hasil kajian ini termasuk pembetulan item selanjutnya untuk memastikan keboleh percayaan dan kesahan instrumen, mewujudkan norma dengan menggunakan kaedah menetap piawai, dan aplikasi model teori respon item dua- dan tiga-parameter untuk menyempurnakan ciri-ciri item.

Kata kunci: Aptitud spatial visual, penilaian, pembinaan ujian, analisis Rasch, analisis pengesahan faktor

ACKNOWLEDGEMENTS

I would like to extend my thanks and regards to my supervisor Dr Shahrir Jamaluddin for his guidance in completing this dissertation. I am also grateful to Dr Maria Felicitas (Marife) Mamauag and CAREERsense of HELP University for allowing me to collect data under their test development project. Ultimately, my work here would not have been possible without the support of my parents who financed my studies, and my husband, Isaac Ho, who kept me sane in difficult times. My coursemates, Nadia and Sumaia, were invaluable in their guidance and support of the dissertation process and I am eternally grateful to Ross Stephenson without whom I may have never been able to print this successfully. Finally, my deepest appreciation to all my friends for their patience when I was struggling to complete this. The journey begins anew.

TABLE OF CONTENTS

Chapter 1: Introduction

Chapter 2: Literature Review

Chapter 3: Methodology

Chapter 4: Findings

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

CF	:	Closure flexibility
CFA	:	Confirmatory factor analysis
CFI	:	Comparative fit index
ETS	:	Educational Testing Service
FGP	:	Figure-ground perception
GIMP	:	GNU Image Manipulation Program
INFIT MNSQ	:	Infit mean squared
IRT	:	Item response theory
MAT	:	Multiple Aptitude Test
MAT-D	:	Multiple Aptitude Test – Form D
ML	:	Maximum likelihood
OA	:	Object assembly
OUTFIT MNSQ	:	Outfit mean squared
PS	:	Progressive series
RMSEA	:	Root mean square error of approximation
SO	:	Spatial orientation
SR	:	Spatial relations
SV	:	Spatial visualisation
VD	:	Visual discrimination
VS	:	Visual-spatial
WLS	:	Weighted least squares
WLSMV	:	Weighted least squares means and variance adjusted

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

Career counselling or career guidance programs refer to comprehensive and developmental programs that aim to help people make informed choices regarding educational pathways and occupational decisions (Office of Career, Technical, and Adult Education, 2014). The experience of a career counselling program has been shown to be related with many benefits in terms of educational outcomes, social facets, and economic consequences. Higher graduation rates, better educational achievement, greater job satisfaction, and increased employee productivity are just some of the numerous positive outcomes of undergoing career counselling programs (Gillie & Isenhour, 2003). As such, any institutes of education across primary, secondary, and tertiary levels seek to incorporate career development as a necessary adjunct to standard curriculum.

Generally, the career counselling or guidance process encompasses activities such as providing career education which is information about what a career is and how to develop one, giving insight into the world of work with regards to employment trends and job requirements, providing assessment for individuals to better understand themselves so they can make informed choices about career pathways that are aligned to their personality and capabilities, and provide information to prepare individuals for the career search and recruitment process. Within the Asian culture, career guidance and counselling has developed differently as compared to Western contexts (Arulmani et al., 2011). Arulmani et al. (2011) noted that the collectivistic nature of Asian societies created unique pressures within the career decision-making process of Asian citizens. Despite the need for culturally sensitive instruments that would reflect the nuances of Asian culture, it has been seen that many theories and tests used arose from adaptation of those from the West (Arulmani et al., 2011). The following section will discuss the development of career

guidance and counselling in Malaysia while an overview of career guidance development in other Asian countries will be detailed in chapter two.

1.2 Career Guidance in Malaysia

In Malaysia, the history of career counselling has been tied very closely to that of school counselling as both began in schools (Lloyd, 1987, as cited in Pope, Musa, Singaravelu, Bringaze, & Russell, 2002). Pope et al. (2002) stated that career counselling in Malaysia had a unique twist to it compared to the American practices on which it was based. They noted that counsellors had to account for the collectivist orientation of individuals, the concept of filial piety and dependence on the family structure which would impact the counselling environment (Pope et al., 2002). The use of vocational assessments was also noted as needing further research due to the lack of local norms for majority of the tests being used and only a select few being translated and renormed (Pope et al., 2002). Most studies from within Malaysia focused on interest testing as seen in the work of Amla Mohd. Salleh (2010). Drawing from the pioneering work of Awang (1976, as cited in Amla H. M. Salleh, 1984) that translated Holland's Vocational Preference Inventory to Malay for administration among Malay secondary school students, Amla H. M. Salleh (1984) conducted a similar translation on Holland's Self Directed Search and expanded the validation of its psychometric properties across different groups in Malaysia (Amla Mohd. Salleh, 2010).

Pope et al. (2002) suggested that career counselling began at the time that Malaysia gained independence from the British in 1975, but Quek (2001) puts it as early as 1939. However, it was formally introduced in 1967 when the Malaysian Ministry of Education published a circular calling for all schools at both primary and secondary levels to establish guidance teachers (Quek, 2001). However, it only came fully into focus during the 1980s with the rise of drug abuse among Malaysian adolescents (See & Ng, 2010). In the following years from 1982 to 1984, the role was named (Career Guidance

Teacher) and subsequent circulars concerned the establishment of counselling facilities for the teachers' use (Quek, 2001). Pope et al. (2002) suggested that this catalysed an increase in the number of faculty members at the tertiary level continuing their doctoral studies in America.

However, despite the Ministry's best intentions, there existed multiple problems with the implementation of the Career Guidance Teacher role. Firstly, there was a shortage of adequately trained staff to deal with the need among the schools (Quek, 2001). In addition, the lack of certification led to variations in the level of service provided due to the variations in counsellor qualifications, ethics adhered to, and types of services provided among other things (Quek, 2001). In response to this, the Guidance and Counselling Unit nested under the School Division of the Ministry was established in 1984, with the purpose of establishing guidelines of practice for the guidance teachers (Quek, 2001). In addition, the role was renamed as School Counsellor thus leading to the overlap of the counsellor's area of expertise in both career counselling as well as emotional counselling (Quek, 2001).

Most school counsellors were usually teachers who worked for an extra certification beyond their teaching degree (Pope et al., 2002; Quek, 2001). School counsellors usually held multiple duties; in addition to their teaching load, they conducted counselling for social, emotional, and academic issues on top of assisting students with career decision-making (Pope et al., 2002; Quek, 2001). The Ministry of Education put an end to the excessive load of the teacher-counsellor by turning the counsellor position to that of a full-timer which saw at least one counsellor working full-time in every secondary school by 2000 (See & Ng, 2010). In contrast, elementary schools still lack counsellors with most of the work still being borne by teachers (See & Ng, 2010). Career counselling in corporate settings has seen less focus than that of school counselling. It

was put into place in the 1990s but dealt with more of human resource issues in addition to career placement and within-company transfers (Pope et al., 2002, Quek, 2001).

See and Ng (2010) have identified the Malaysian counselling profession as currently being at its adolescence stage. They suggested three areas of improvement in terms of school counselling: firstly, to develop a comprehensive theoretical orientation around which to frame practice (See & Ng, 2010). In line with this, the use of tests and assessments needs to be further clarified (See & Ng, 2010), an area that has been noted as deficient since 1987 by Scorzelli (as cited in Pope et al., 2002). See and Ng (2010) further suggested that the area requires more research to truly define the field and guide it towards a unified framework of practice. Lastly, school counsellors currently tend to learn most of their practice on the job, which impedes efficiency (See & Ng, 2010). Proper research-based practice needs to be put into place not only to improve the quality of Malaysian counsellors, but standardize the quality of care provided across the board.

Concurrently, Amla Mohd. Salleh (2010) also paralleled these issues adding that other challenges in the field included a lack of support from school administration, a lack of resources for counsellors, and the overemphasis solely on academic performance over career exploration and development that hindered the students' initiative in seeking out career guidance. Although the establishment of full-time career counsellors prompted the Ministry of Education to develop a guidance curriculum that aimed to develop students academically by aligning them to study streams based on their learning abilities and achievements, promote psychosocial and mental health, and provide career development through self-exploration and career planning supported by the use of inventories and assessments, Amla Mohd. Salleh (2010) found the curriculum to be too wide in scope for the small number of available school counsellors to effectively impart to students.

1.3 Career Assessments and Testing

Focusing specifically on career assessments, career assessments have a long-standing use in aiding people to find careers that suit them. Not only can career assessments guide initial career choice, it plays a role in life-long career development. Krumboltz and Vidalakis (2000) have noted the use of career assessments in being able to help existing employees identify areas of weakness in terms of skills to be improved, interest areas to develop, existing beliefs that need change, habits to be broken, and personality characteristics that explain work behaviours to be modified. The earliest model of career assessment in career counselling followed a 'matching model' which looked at classifying jobs according to shared features or characteristics, and 'matching' a person's aptitudes and capabilities to a certain class or family of jobs (Bellows, 1941).

The matching model typically looked at matching individuals to vocations based on three criteria. First was by abilities, seen as early as World War II when soldiers were assigned duties according to their aptitudes (Super, 1983). Interest measures were originally conducted separately from aptitude measures before being used in tandem in later matching models (Super, 1983). Finally, values were an additional area of matching that failed to generate as much research as interests and aptitudes but were subsequently included in matching models (Super, 1983). The classical assessment process consisted of five stages: Initial assessment, in-depth testing of aptitudes, interests, and values, assessment of data gathered, counselling and planning, and a final follow-up (Super, 1983). Although this process based on the matching model was empirically supported and widely accepted, Super (1983) believed that it failed to account for an important factor, namely, the individual's readiness for career exploration.

As such, Super (1983) proposed a holistic developmental model of career assessment. His four-stage model closely mimics that of the matching model with a few modifications. He maintained stage I as the preview or initial assessment (Super, 1983).

Stage II took a more in-depth view of assessing the client's readiness (Super, 1983). His assessment included measures of work salience (work motivation or awareness of need to work), career maturity (career readiness in terms of availability of information, decision-making skills, and reality-orientation), aptitude, and interests (Super, 1983). Stage III revolved around review of data and planning following communication such as with the client or with the family involved (Super, 1983). Finally, Stage IV included the overall review and discussion (Super, 1983). At this stage, the client could go over the assessment results with the counsellor, revise or accept the results, and use them in further career planning (Super, 1983). The follow-up session was still nested under Stage IV. An understanding of the client's personality was also advised by Super (1983) as he believed personality greatly shaped other factors of career readiness. Given the developmental and holistic nature of the model he proposed, Super (1983) was keen on the use of multiple instruments in a test battery to measure all the different areas.

However, Gottfredson (2003) found that aptitude testing had slowly been fading from the career counselling and assessment process. This was extremely puzzling despite its longstanding history of use in career assessment in conjunction with interest testing (Gottfredson, 2003). She soon discovered that the main reason for its lack of use was the counsellor's reluctance to inform clients of their shortcomings in certain abilities (Gottfredson, 2003). It is believed that low levels of ability somehow translate to a greater struggle in achieving success, and this led counsellors to be very wary of providing aptitude results (Gottfredson, 2003). Gottfredson (2003) argues that aptitude testing is still vital in the career assessment and counselling process given its informational value in predicting later job performance. She did, however, add that the tests in existence were mostly developed in the United States of America and, thus, were mostly psychometrically valid only in the contexts of their development, namely "native-born, English-speaking Americans" (Gottfredson, 2003, p. 121). In addition, the training of

skills related to different aptitudes could be largely influenced by curriculum; therefore, differences may be found according to the use of different curricula where some aptitudes may be emphasised more than others (Gottfredson, 2003). This was supported by the early position of Carroll (1993) that the proliferation of school curricula was a vital opportunity to study the differential effects of curricula on cognitive test outcomes.

Within Malaysia, aptitude testing has slowly become a point of focus with the implementation of school-based assessment (*Pentaksiran Berasaskan Sekolah*, PBS). Originally proposed in 2007, PBS included additional forms of assessment aside from the typical central examinations which included psychometric tests to assess students' in their abilities and interests as well as their readiness for learning (Ong, 2010). According to the proposed system, these assessments would be put in place up to the lower secondary level, but the upper secondary level would solely focus on centralised examinations (Ong, 2010). These plans finally came to fruition this year as Primary Year 6 students who sat for the 2017 Primary School Assessment Examination (*Ujian Penilaian Sekolah Rendah*, UPSR) received the Primary School Assessment Report which included assessment results for their classroom, psychometric, and sports, physical activity, and curriculum components of PBS (Azura Abas & Hashini Kavishtri Kannan, 2017). However, the psychometric assessment encompassed aptitudes in topics such as music, language, mathematical logic, kinaesthetic intelligence, and naturalistic intelligence (Azura Abas & Hashini Kavishtri Kannan, 2017). It was also unclear how these assessment results would be used in career exploration or guidance for the students.

The inclusion of these components under PBS was set to be continued under the Malaysia Education Blueprint for the period of 2013-2025 (Ministry of Education, 2013). However, although the blueprint stated that aptitude testing would focus on measuring thinking and problem-solving skills (Ministry of Education, 2013), it is also unclear if this will be used to inform the education or career counselling process. In addition, Ong

(2010) has outlined the challenge faced by the Ministry in moving towards holistic assessment including the development of psychometric tests after a longstanding historical focus on achievement testing, and proper communication and delivery of results. Further to that, not much literature can be found regarding the development of such instruments by the Ministry. As such, although steps are being taken to inculcate the assessment of aptitudes in our local education system, it is unclear as to what extent the assessments being developed have been empirically tested in terms of their psychometric properties and to what extent they provide use within the education and career guidance process of students.

1.4 Problem Statement

Given the historical background and development of career guidance and counselling in Asia, it can be seen that career counselling and assessment has for the most part been seen by governments and institutions of different countries across Asia as a means to grow a competent, skilled, and productive workforce. However, there has been a lack of test development to suit the unique cultural context of Asia within which Western models and theories have found questionable validity (Tien & Wang, 2016; Zhou et al., 2016). Many countries within the Asian region have called for an increase in evidence-based practise and the need for nuanced, effective tools to aid the career guidance process (Jin, 2017; See & Ng, 2010; Tien & Wang, 2016; Wong, 2017; Yeo et al., 2012; Yuen et al., 2014; Zhou et al., 2016).

In Malaysia itself, career counselling is still moderately young with a lack of research-based practice in place to aid counsellors in their work (See & Ng, 2010). There is a lack of appropriate career assessment tools where the only two known localised adaptations of tests are focused in the realm of interest testing, leaving much to be explored for effective career development to take place (Amla H. M. Salleh, 1984; Amla Mohd. Salleh, 2010). Additionally, the use of aptitude testing has lagged behind interest

testing despite its importance in effective career guidance (Gottfredson, 2003) and there is a lack of information regarding the development, psychometric properties, and usefulness of local aptitude assessments that have been implemented at the school level.

In response to this need for a contextualised assessment tool that adequately measures various aspects of career readiness as recommended by Super's developmental model of career assessment, the career counselling unit of a local private university has been undertaking an ongoing test development project known as the HELP Career Readiness Evaluation System (HELP CaRES) (Chong et al., 2016; Darwishah et al., 2016; Mamauag, Tai, Arosh, Teo, & Ooi, 2016). The test battery includes various measures of aptitude, interest, personality, employability skills, and career readiness in accordance with Super's recommendations to provide a holistic measure of clients such that the career counselling process is sufficiently informed to assist clients in more effective career decision-making (Chong et al., 2016; Darwishah et al., 2016; Mamauag et al., 2016). The tests were aimed to be grounded within the Malaysian context as a contextualised tool for career counselling in Malaysians within the age range of 14-25, from lower secondary to the tertiary level (Chong et al., 2016; Darwishah et al., 2016; Mamauag et al., 2016).

The aptitude measure developed is known as the Multiple Aptitude Test (MAT) which measures aptitudes such as verbal reasoning and visual-spatial reasoning. It consists of three parts: verbal analogies, number-letter series, and visual-spatial aptitudes (Mamauag et al., 2016). The visual-spatial skills measured encompassed seven constructs: figure-ground perception, object assembly, progressive series, surface development, visualisation and orientation, visual discrimination, and topology (Mamauag et al., 2016). A more in-depth discussion of these constructs will be given in chapter two. Statistical methods used to evaluate the psychometric properties of the original three forms of the MAT showed that although they were valid and reliable, the

difficulty of the items was not fairly matched to individuals of the tertiary age group (17-25) as the items were too easy (Mamauag et al., 2016). In line with the recommendations from Mamauag et al. (2016), more difficult items from each type were identified for a further round of testing. However, although the verbal analogy and number-letter series consisted of larger item pools from which other items could be drawn and tested, the visual-spatial items were lacking in excess items.

1.5 Aim

The aim of the current study is to develop an aptitude scale that measures higher levels of visual-spatial aptitude suitable for tertiary students aged 17-25.

1.6 Research Objectives

The research objectives are as follows:

1. To determine the psychometric properties of the Multiple Aptitude Test – Form D (Visual-Spatial) [MAT-D(VS)].
    1.1 To establish the construct validity of the MAT-D(VS) using confirmatory factor analysis.
    1.2 To determine the reliability of the MAT-D(VS) using Rasch modelling.
    1.3 To determine the item fit of the MAT-D(VS) using Rasch modelling.
2. To examine the quality of the item distractors in the MAT-D(VS) using distractor analysis.

1.7 Research Questions

Given the aforementioned research objectives, the following research questions were outlined:

1. What are the psychometric properties of the MAT-D(VS)?
    1.1 To what extent does the MAT-D(VS) exhibit construct validity, such that the items load sufficiently on the underlying factors?

1.2 What is the reliability of the MAT-D(VS) in terms of person and item characteristics?

1.3 What is the item fit of the MAT-D(VS) in terms of person and item characteristics?

2. What are the qualities of the distractors found in the items of the MAT-D(VS)?

1.8 Significance of the Study

Development of this contextualised tool will help produce an instrument that is well-tailored to measure the abilities of Malaysian undergraduate students who fall within the intended age range. This will allow for more effective career counselling to take place as the students would be better informed about their strengths and weaknesses so they can make better career decisions.

In addition, it will help career counsellors in the counselling process as they can be assured that the tools they use are giving them accurate depictions of the ability of the students thus assisting in the development of a more standardized and empirically-supported practice. Furthermore, development of the test battery in accordance with Super's framework is in line with See and Ng's (2010) recommendations for the development of theoretical frameworks of practice which help define the use of assessments for the ease of the counselling process. Lastly, it adds to the body of literature about context-specific career counselling within the Malaysian culture that accounts for local social and cultural influences on the career decision-making process.

1.9 Limitations

A few limitations have been identified within the scope of this study. These limitations are mostly due to the private financial sponsorship of the test development process.

Firstly, the instruments are all developed in English as it is the language of instruction at the private tertiary educational institution within which the test is being

developed. Even though the visual-spatial scale is meant to measure non-verbal ability, instructions are mostly given in English and thus, may be misunderstood by students with poorer English proficiency.

In addition, restrictions exist in terms of the sample that can be obtained at this point of development. As the instrument is still in the development stage of establishing validity and reliability, it is only allowed to be distributed with the educational institution that it is being developed. It is not allowed to be offered to external samples due to the private nature of its financial sponsorship given that the test development process is being funded by the institution. In turn, the generalizability of the instrument is thus limited to the university sample within which it was developed.

1.10 Delimitations

The first delimitation is that one construct under visual-spatial aptitude was removed from the scale due to certain limitations of the researcher's understanding and the constraints of the development schedule. This will be further discussed in chapter two. Additionally, although testing can only be conducted within the boundaries of the sponsoring educational institution, access to other faculties may be restricted due to differences in semester scheduling and the time constraint placed by the schedule of test development.

1.11 Definition of Terms

The following definitions will apply to the terms used in this study. Further details regarding the derivation of these definitions will be provided in chapter three:

Aptitude : An attribute of individuals revealed by differences in the levels of task difficulty, on a defined class of tasks, that individuals perform successfully when they have

		the opportunity and motivation to do well (Gottfredson, 2003, p.117)
Figure-ground perception	:	The ability to recognise and utilise figure-ground cues to discern figural boundaries from their respective backgrounds. Examples of such cues include but are not limited to: proximity, similarity, symmetry, parallelism, good continuation, and closure.
Object assembly	:	The ability to figure out how separate pieces of disassembled images come together to form one whole, two-dimensional image, much like a jigsaw puzzle. The pieces of disassembled images can be rotated, displaced, or both, before they are re-assembled to form a coherent image.
Progressive series	:	The ability to uncover abstract relations or rules underlying a series of geometric progressions, and infer the missing entries from the series based on these abstract relations.
Spatial orientation	:	The ability to comprehend and identify spatial relationships that exist in regard to an object, such as understanding the shape of an object as it is viewed from different angles, and identifying the position of the object in space in relation to other objects or oneself.
Spatial visualisation	:	The ability to mentally manipulate an object with regard to changes in perception of its physical shape, such as the folding and unfolding of a flat surface to a

		3-dimensional object, active object rotation around a focal point, twisting of the object, and inversion/mirroring of the visual.
Visual discrimination	:	The ability to detect if two objects or images are duplicates, or similar, and identify the similarities or differences between the visuals.

1.12 Summary

A review of the current context of career counselling within the Asian region showed that career counselling is a relatively young field of practice, particularly in Malaysia. Although much guidance has been taken from Western models and practices, limited applicability of Western-developed theories and assessment tools have been found in the Asian cultural context. In Malaysia, the practice of career counselling requires much research-based evidence to be found to properly establish frameworks of practice and use of contextualised assessment tools. Most models of career assessment focus on matching individuals to careers based on limited information. A more holistic, developmental model by Super (1983) was used as the groundwork for development of a career assessment test battery by a local private university which included measures of aptitude, interest, personality, employability skills, and career readiness. However, initial forms of the aptitude test were found to be too easy at the tertiary level, hence, the current study aims to develop a visual-spatial aptitude scale as a fair measure of ability for students at this level of education. The subsequent chapter will present a critical review of related literature pertaining to the framework of aptitude measurement, relevant skills, and the use of statistical methods to establish validity and reliability.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter will present an overview of the development of career guidance and counselling in an Asian context as well as discuss the relevant literature pertaining to aptitude testing and the constructs under study. The theoretical and conceptual frameworks are presented and discussed. Finally, a review of the statistical methods that will be used in this study will be given.

2.2 Career Guidance and Counselling in Asia

Across a variety of Asian countries, it has been seen that career guidance and counselling has demonstrated certain parallels in terms of its development. Firstly, most career guidance programs were initiated based on a government-centred impetus to accurately and efficiently place manpower within the workforce. Over time, this evolved to become more focused on developing the latent potential of individuals within their respective career paths. Secondly, assessments within the field were largely based on new developments in Western career guidance theories and tests. The move towards indigenization of tests has slowly picked up in recent years but there remains a focus on the use of translated tests without adaptation to local cultures and nuances. The following sections will provide detailed descriptions of the development of career guidance and counselling in various Asian countries with regard to these two highlighted aspects.

2.2.1 Career guidance in Shanghai, China

Zhou, Li, and Gao (2016) provide an extensive review of developments in career guidance and counselling in Shanghai since its inception in 1977 till 2015. The authors identified four phases of development beginning with government-mandated job allocation to the current phase of career counselling. After the Cultural Revolution,

career guidance emerged within the community in the form of government regulations known as job allocation (1977-1992; Zhou et al., 2016). The government took it upon themselves to allocate jobs to graduates of colleges and vocational schools and the graduates had no choice in the matter as it was intended to establish a socialist society (Zhang, Hu, & Pope, 2002). There was a keen emphasis on the interests of the country and society without accounting for personal interests and abilities as even at the school level, teachers would emphasize how important it was for the students to prioritize the country's needs over their personal preferences (Zhang et al., 2002). This system arose to fulfil the shortage of labour created by the Cultural Revolution (Zhou et al., 2016). Many problems emerged with mandatory allocations when the personal needs and preferences of the graduates conflicted with work expectations and environments (Zhou et al., 2016).

As the economy progressed, mandatory job allocation evolved into job allocation with demand-supply meetings (Zhou et al., 2016). Colleges pushed for meetings between employers and educational institutions to negotiate allocation plans to suit the needs of the changing employment landscape (Zhou et al., 2016). Toward the end of the job allocation phase, the number of graduates far outweighed the demand in state-owned institutions and small-to-medium enterprises and foreign-owned companies were looking for trained professionals (Zhou et al., 2016). The shift allowed for legislation to be made that gave greater autonomy to employers to turn away less-qualified candidates. In addition, vocational education started being implemented at the secondary school level with the influence of models from Western nations (Zhou et al., 2016). However, a major issue was still the lack of knowledge and skills related to career decision-making.

Vocational guidance came to the forefront between the years of 1993 to 1999. At this point, focus was shifted to providing advice to college graduates. The changeover from allocation meant that graduates now had to learn about the job market and how to

develop their own careers, and service providers were no longer government middlemen and had to assist in developing graduates (Zhou et al., 2016). At the beginning of this phase, the Shanghai Graduates Vocational Guidance Centre (SGVGC) was established as a non-profit, non-governmental institution to aid graduates and senior secondary school students by providing them information about employment trends and market information (Zhou et al., 2016). Within colleges and universities, emphasis was placed on equipping graduates with knowledge about labour policies and providing employment information; this was done by the vocational guidance centres within the colleges (Zhou et al., 2016).

By the year 2000, the third wave of change shifted the national focus to career education where all college students had to take up formal career courses to help them in designing and developing their individualized career paths (Zhou et al., 2016). This was partly due to the rapid and continuous growth of higher education which was now open to all layers of society (Wang, 2010, as cited in Zhou et al., 2016). However, the increase in graduates created a glut in the employable workforce, particularly in already saturated developed cities (Zhou et al., 2016). The government allied with institutions like the SGVGC to create a network of information about career education, services, and guidance which was meant to help bridge gaps between employers, and university and school graduates (Zhou et al., 2016). In addition, certification of career guidance professionals came to the forefront with China taking the lead of countries such as the United States, Japan, and Hong Kong among others (Zhou et al., 2016). Career education was further implemented at all levels of college and universities with certified practitioners providing guidance services (Zhou et al., 2016).

China has only recently moved into the fourth stage of development which shifted towards career counselling (2012-2015). The main difference was the new onus by practitioners to assist people in understanding themselves better such that they are able to navigate the competitive Chinese job market (Zhou et al., 2016). Thus, a great need for

qualified practitioners created opportunities for further training and development of professionals (Zhou et al., 2016). Changes were also seen in the Chinese National College Entrance Examination (NCEE) which was implemented in 2014 where the original system was entirely dependent on test scores for college entrance and college choice (Zhou et al., 2016). However, the new system allowed for students to make subject choices according to their interests, and college admission would take into account not only test scores, but also personal qualities, secondary school academic profiles, and performance on career tests (Zhou et al, 2016).

Zhou et al. (2016) noted that although the development of career guidance and counselling has come very far since 1977, much improvement was still necessary in areas of accessibility by underprivileged groups, adaptation and indigenization of assessment theories and instruments, and establishment of professional associations for career counsellors. They also emphasized the need for integration of career education, assessment, counselling and other career-related activities in providing holistic services to their stakeholders (Zhou et al., 2016).

In 2017, a new roadmap for increasing employment was presented by the state council (Jin, 2017). Specific groups such as graduates, ethnic minorities, and retirees were targeted as a point of focus to increase employability (Jin, 2017). In addition, due to identified stressors that complicated the career finding process of college graduates, the government proposed more effective implementation of career guidance and counselling to aid college graduates in obtaining high-quality employment (Jin, 2017). Stricter guidelines of practise and an emphasis on credentialing professionals were some of the steps implemented by the government to promote an improvement in the standard of practice among career guidance professionals in Shanghai with a view to expand the accessibility of services to more graduates within mainland China (Jin, 2017).

2.2.2 Career guidance in Taiwan

Career guidance and counselling in Taiwan began approximately 50 years ago with the implementation of career education at the middle school level for individuals who did not intend to pursue further studies (Tien & Wang, 2016). It has evolved from aiming to help individuals find employment to defining career as the process of education and employment over the course of a lifetime (Tien & Wang, 2016). Taiwan has approached career guidance differently from mainland China in that from the beginning, they had already taken much lead from Western theories and models particularly in the areas of interests, career barriers, decision-making difficulties, and career adaptability (Tien, 1997, 2009; Tien & Wang, 2016). They had also adopted a highly scientific approach to career guidance wherein it was believed that programs for individual career development should be grounded in theory and efficacy verified through research (Tien & Wang, 2016).

In terms of applicability of Western theories within the Taiwanese context, it was generally found that most theories could be verified; some with modifications such as Holland's hexagonal interest typology or some with partial support such as Lent's social cognitive career theory (Tien, 1997, 2009; Tien & Wang, 2016). Certain psychological instruments such as the Career Barriers Inventory and Career Decision-Making Difficulties Questionnaire were successfully adapted for use within the Taiwanese culture (Tien & Wang, 2016). However, the authors identified certain problems that existed within the counselling landscape.

Firstly, the structure of the education system provided a fairly linear focus that mostly emphasized academic excellence above all else. Career education was ignored at the elementary school level as it was believed that the focus of students should be on academic achievement to be able to enter junior high school (Tien & Wang, 2016). Again, at the junior high level, students who perform well academically are nurtured to

progress to senior high whereas others plan for a vocational technological school (Tien & Wang, 2016). Despite the streaming system in place, this system does not account for individual skills, interests, and desires. There is a lack of information for students to undergo self-exploration in terms of personal strengths or work requirements before they embark on their careers (Tien & Wang, 2016).

Students who progress to senior high school face even greater pressure to make career decisions without adequate information (Tien & Wang, 2016). They are expected to make choices as soon as the end of first year, and the choices are largely predetermined by how well they perform on the College Entrance Examination (Tien & Wang, 2016). Once they have entered college, job-seeking skills became the primary point of focus to prepare graduates for the world of work (Tien & Wan, 2016). A new policy implemented in 2014 incorporated a compulsory career education module for high school students (Tien & Wang, 2016). This was done to help students understand themselves aside from providing them with more information about careers, work, and possible college majors (Tien & Wang, 2016).

In addition, the authors outlined a lack of trained professionals available to provide counselling services despite the government's investment in counselling facilities (Tien & Wang, 2016). A lack of counselling services for adults experiencing career transitions, women, and retired soldiers has also been identified as an area for improvement (Tien & Wang, 2016). Finally, the authors noted that most companies and corporate organizations fail to properly provide assistance in employee career development (Tien & Wang, 2016). Overall, the authors recommend that steps should be taken to improve career counselling services for adults, and provide more guidance for school students (Tien & Wang, 2016). More research should also be done to improve evidence-based practice within the field (Tien & Wang, 2016).

2.2.3 Career guidance in Hong Kong

In a manner similar to Malaysia, Hong Kong was previously colonised by the British until its recent independence in 1997. Career guidance was implemented relatively early compared to other Asian countries, and first began with the appointment of teachers as career masters who would help students be ready for the world of tertiary education and careers (Yuen, 2006, as cited in Yuen, Leung, & Chan, 2014). However, early implementation of career guidance was rudimentary, comprising of career talks and site visits (Wong, 2017). With subsequent policy developments, schools have focused on career development as a key area in education with career teachers designing and carrying out career education programs in their respective schools in addition to their existing teaching workload (Ho, 2008). Specific curriculum materials were even developed by the Hong Kong Association of Careers Masters and Guidance Masters to strengthen the content of the school syllabus (Ho, 2008).

University level career guidance was focused on bridging the gaps between employers and graduates by providing recruitment services, providing graduates with job market information, linking graduates with internship opportunities, and tracking post-graduation employment (Yuen et al., 2014). The responsibility to plan careers falls on the shoulders of graduates as the universities offer services such as online portfolio systems to help students assess themselves and track their career development in addition to the provision of internship opportunities and the promotion of compulsory work-integrated education (Fung & Wong, 2012). Aside from obtaining assistance from the Labour Department, adults are able to reach out to non-governmental bodies such as the Hong Kong Employment Development Service to meet professional career counsellors who provide aid with individualised career planning, conduct career assessments, and counsel clients through career transitions and work-related stress (Yuen et al., 2014). In terms of areas of improvement, Yuen et al. (2014) have called for constant evaluation of practice,

and the need for research-based techniques to be continuously studied such that efficacy of interventions and practices can be determined.

Wong (2017) posits that many problems are still enduring in the practice of career guidance and counselling. Though government initiatives have, to a certain extent, improved accessibility and quality of services rendered, Wong (2017) finds that most teachers still lack the necessary training for effective career guidance and counselling to take place. In addition, the stress of compliance to government mandates has resulted in intensified workloads on the part of teachers who provide career counselling (Wong, 2017). As a result, few students reap the benefits of understanding and increased self-awareness following the guidance sessions (Wong, 2017). Overall, Wong (2017) finds that for the practice to improve, more administrative support and development of focused curricula to enhance employability is needed.

2.2.4 Career guidance in Singapore

The closest neighbour to Malaysia, Singapore has seen rapid economic growth since its independence from the Federation of Malaya in 1965. Much emphasis has been placed on developing an efficient and productive workforce through quality education (Yeo, Tan, & Neihart, 2012). The Singapore Ministry of Education (MOE) has been spearheading school counselling programs with cooperation from the National Institute of Education (NIE) with the original intention of providing aid to students from low-income families (Yeo et al., 2012). Counselling for social and emotional problems was typically referred to external agencies until the Students Care Service was established in the 1970s (Yeo et al., 2012, Yeo & Lee, 2014). Around this time, the NIE set up the Guidance Unit to bring awareness regarding the need for student guidance and to train teachers in counselling practices but it was not impactful enough leading to its dissolution in 1977 (Yeo et al., 2012).

1987 brought the resurgence of the Pastoral Care and Career Guidance branch by the MOE which pioneered the systematic establishment of student counselling services in schools (Yeo et al., 2012; Yeo & Lee, 2014). An increased focus on developing school counsellors and ensuring their credentials were up to standard helped the government in placing school counsellors in every school within the short span of eight years (Yeo & Lee, 2014). As school counselling developed, career counselling and guidance subsequently gained traction in educational institutions. The early 1990s saw the release of a computer-assisted software for career guidance known as the Job Orientation Backup System (JOBS) aimed at students of the secondary level (Yeo et al., 2012).

This was eventually the basis for the development of the Orientation System for Careers (OSCAR) which took the instrument online making it more accessible for students who were able to get information quicker as well as undergo career assessment thus giving teachers more time to provide personal guidance to students (Yeo et al., 2012). OSCAR was discontinued in 2009 and was replaced by eCareers.sg which was an online interactive portal for career guidance developed conjointly by the MOE and University of Wisconsin-Madison (Yeo et al., 2012). As with other Asian nations, Singapore has become concerned with the applicability of Western theoretical models and interventions and are looking to expend research into studying the validity of theories as well as efficacy of practices (Yeo et al., 2012).

In summary, it is seen that many Asian countries face similar issues within the realm of testing for career guidance and counselling. This is mainly due to the problems that occur regarding the degree of applicability of Western career theories and tests. In moving forward with more accurate assessment to assist the guidance and counselling process, more nations are turning to developing their own assessment tools as well as investing more research into establishing validity of theories and practices.

2.3 Definition of Aptitudes

Typically, aptitudes are seen as the accumulation of informal learning and experiences a person gathers over the course of their life (Cohen & Swerdlik, 2009). Unlike achievement tests, the focus of aptitude tests is to measure an ability for the purpose of making predictions rather than assess structured knowledge that has been taught (Cohen & Swerdlik, 2009). Generally, aptitude tests draw on broader categories of abilities compared to general achievement tests (Cohen & Swerdlik, 2009).

Gottfredson (2003) defines ability as "an attribute of individuals revealed by differences in the levels of task difficulty, on a defined class of tasks, that individuals perform successfully when they have the opportunity and motivation to do well" (p. 117). In this manner, abilities can be highly individualised and differentiated from one another. Cognitive abilities are those that require the manipulation of facts, knowledge or ideas within the mind (Gottfredson, 2003). Not to be confused with factual or procedural knowledge, ability refers to capability or proficiency to be able to gather knowledge (Gottfredson, 2003).

In general, cognitive abilities can be arranged in a hierarchy depending on the level of specificity. One example is Vernon (1960, as cited in Anastasi & Urbina, 1997) where factors under general intelligence (commonly recognised as Spearman's *g* factor) can be divided to broad families of abilities and then further subdivided to narrow categories of abilities finally ending in very specific abilities. The hierarchical models of aptitude and abilities have become more widely accepted as they are seen as useful from both a theoretical and practical standpoint (Anastasi & Urbina, 1997). From a theoretical standpoint, the models are considered superior as they effectively bridge the two main schools of thought between the single general factor of intelligence and multiple factors of intelligence (Anastasi & Urbina, 1997). It has been found that hierarchical models are mathematically equivalent to multiple factor models from a methodological standpoint

(Anastasi & Urbina, 1997). In addition, they show practical usefulness in the flexibility of use given that scores can be given as a sum of clusters or per cluster, allowing for more precise identification of underlying strengths and weaknesses (Anastasi & Urbina, 1997). The model that will be focused on in this paper is Carroll's three-stratum theory.

2.4 Theoretical Framework: Carroll's Three-Stratum Theory

Carroll developed the three-stratum theory of intelligence in 1993 by analysing over 400 existing factor analytic studies of cognitive ability (Gottfredson, 2003). The theory posits that cognitive abilities exist in a hierarchy comprising three levels or strata (Carroll, 2003). Stratum I, a lower-order level, consisted of narrow abilities that Carroll believed to be linearly independent such that they were possibly correlated but separate factors (Carroll, 2003). This stratum comprised highly specific abilities that were usually the skills measured by psychological instruments such as standardized tests of mental ability (Gottfredson, 2003). Stratum I abilities were reflective of individual experience, learning, and the use of performance strategies due to its specificity (Johnson & Bouchard, 2005). There are no set abilities that are listed as stratum I abilities as studies that usually aim to test the model use the scales that exist in the assessment tool used, as in Johnson and Bouchard (2005) but Carroll (2003) estimates there to be around 60-70 stratum I abilities.

Stratum II comprised of general abilities that existed in broad behavioural domains (Johnson & Bouchard, 2005). Most cite Carroll as proposing eight factors (Gottfredson, 2003) but Carroll (2003) has mentioned that there could be between eight to ten factors. Stratum II factors were similar to stratum I in terms of being linearly independent (Carroll, 2003) but would have substantial correlations in diverse samples given that they are expected to measure the common factor, stratum III. Stratum II abilities are broader than stratum I but more specialised than stratum III where the categories of abilities can be identified (Johnson & Bouchard, 2005). Stratum III consists

of a single factor and represents general intellectual ability (Carroll, 2003). This parallels to the general mental ability identified by Spearman also known as *g*-factor or simply *g* (Carroll, 2003). Figure 2.1 shows the model for Carroll's three-stratum theory.

Carroll's theory derived from a close examination of different ability tests through the use of factor analysis (Carroll, 1993). As such, his theory posits that the location of a construct at each stratum emerges from the significant loadings of variables on the respective factor (Carroll, 1993). Generally, classifications were not meant to be overly rigid as the strata were mostly indicators of the degree of generality of an ability (Bickley, Keith, & Wolfle, 1995; Carroll, 1993). In addition, it is assumed that the factors correspond to phenomena in reality that form the basis of cognitive performance (Carroll, 1993). In other words, the theory assumes that factors indicate real clusters of mental ability that are exhibited through performance on tests.

Carroll (1993) concluded that data from various longitudinal studies taken at different periods over a span of several years demonstrated that stability exists for specific abilities. This was supported by Gottfredson (2003) who noted that the more general an ability was, the more stable and heritable it seemed to be. For example, with stratum II abilities, stability and heritability were seen to increase with age as both aspects levelled out to fairly high levels in late adolescence (Gottfredson, 2003).

In addition, the narrower an ability, the more it was subject to environmental influences and could be further trained (Gottfredson, 2003). Thus, stratum I abilities were most trainable. Carroll (1993) supported this in addition to positing that training at stratum I would eventually lead to improvements in strata II and III. Furthermore, it was identified by Carroll (1993) that changes in the differences between spatial and verbal abilities could be tracked to differences in exposure during work and educational experiences. He found that individuals who were given more exposure to verbal-dominated curricula showed greater development of their verbal aptitudes whereas those who were taught curricula

that emphasized technical subjects had greater aptitude for spatial and technical constructs (Carroll, 1993).

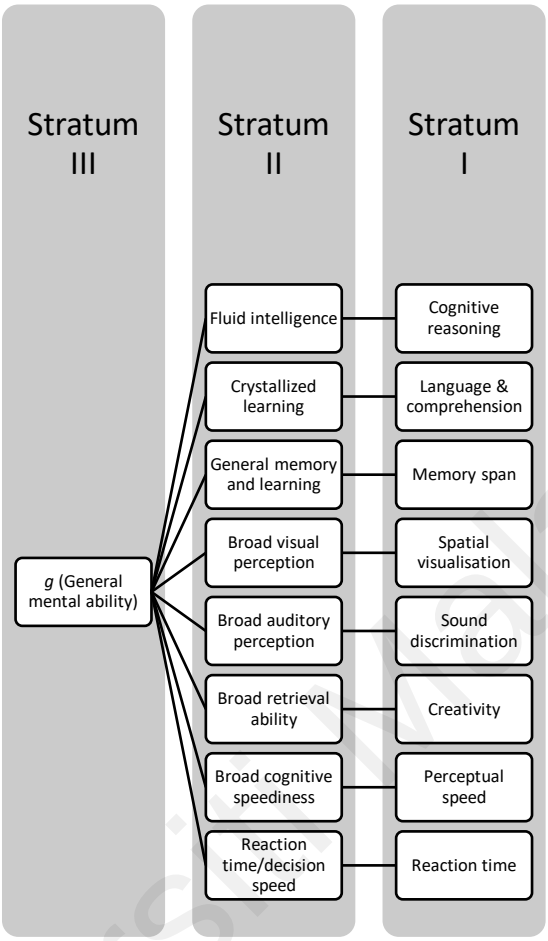

Figure 2.1. Theoretical framework: Carroll's three-stratum theory with sample stratum I abilities

On the whole, the most general abilities were seen to be the best predictors of later performance (Gottfredson, 2003). Stratum III or *g* was seen to predict performance to some degree across all job types but the predictive power of stratum II was dependent on the extent it reflected *g* (Gottfredson, 2003). It was also found that correlations between stratum II abilities tended to be lower in higher IQ populations, indicating that the shape of the ability profile may be better predictors of job performance as compared to simply examining the levels for those at a higher IQ range (Gottfredson, 2003). Additionally, cognitive abilities were generally seen to be independent of interests and personality (Gottfredson, 2003) as seen by the low correlations between these constructs as

demonstrated in studies such as Ackerman and Heggestad (1997). There was seen to be certain overlap in some abilities against interest and personality but the cases were highly specific (Ackerman & Heggestad, 1997; Gottfredson, 2003).

In terms of the impact of aptitude testing on actual performance, it was generally seen that tests of mental abilities, in addition to predicting some degree of performance across all jobs, was the best predictor of performance in jobs that had high cognitive complexity when performance was measured in an objective manner and performance was related to the main technical job duties (Gottfredson, 2003). Non-cognitive traits such as interests and personality contributed little to predictions of job performance as compared to mental ability (Gottfredson, 2003).

It was also seen that work experience and psychomotor ability was a good predictor of performance in jobs that required little mental ability, the complete opposite of that as predicted by *g*, and at higher average levels of experience, the advantage provided by higher levels of *g* was still maintained (Gottfredson, 2003). This is important to note as the core functional duties of a job were mostly distinguished by the level of cognitive complexity involved followed by the content of the work (Gottfredson, 2003). Knowledge of *g* can also help individuals understand the degree to which they will be able to master a profession as stratum III was seen to be related to mastery of work but stratum II was related to suitability for a field of work regardless of level (Gottfredson, 2003). In this manner, assessing the level of stratum II abilities can aid in career decision-making similar to having knowledge of interests and personality such that the right career can be chosen according to the known aptitude profile.

Particularly focusing on college and university students, it was seen that college students typically fell in the upper ranges of the intelligence distribution as a minimum of average intelligence would have been required to reach tertiary education (Humphreys & Yao, 2002, as cited in Gottfredson, 2003). This was supported by the earlier findings

of Balke-Aurell (1982, as cited in Carroll, 1993) which found a relationship between educational level and intelligence (*g*) factor such that the higher the level of *g* factor, the higher the level of education attained. Three stratum II abilities were identified as being important predictors of degree choice, namely, mathematical reasoning, verbal aptitude, and spatial aptitude which are usually measured in college-level entrance examinations such as the Graduate Record Examination (GRE) (Gottfredson, 2003). Furthermore, it was determined by Kell, Lubinski, and Benbow (2013) that individuals who had high levels of these abilities went on to complete professional degrees in business, law, medicine, information technology and other STEM (Science, Technology, Engineering, and Mathematics), attain prestigious occupations in their late 30s, and work for highly impressive organizations.

When it comes to the delineation between cognitive and non-cognitive constructs, it was interesting to note that ability profiles showed a certain degree of overlap with interest profiles (Gottfredson, 2003). It was seen that students tended to have interests in areas that they had high aptitudes for, such as those high in spatial abilities favoured hard sciences and this was reflected in their choice of degree (Gottfredson, 2003). Kell et al. (2013) supported this idea with the finding that even gifted students chose fields of study and work environments that were highly suited to their respective strengths. On the other hand, when levels of ability were high across most abilities, interest or another non-cognitive factor would override ability such that students would choose courses according to what they liked (Gottfredson, 2003).

2.5 Which Aptitudes should be Assessed?

At the very minimum, Gottfredson (2003) recommended that verbal abilities, spatial-mechanical abilities, mathematical reasoning, and clerical speed should be assessed because these aptitudes were relevant to large groups of professions. Hegarty and Waller (2005) have stated that visual-spatial abilities, in particular are important in

many daily activities; this was paralleled by Newcombe and Shipley (2015) who stated that spatial ability was not only crucial for navigating the world around us but also that higher spatial ability was seen as necessary for creativity, design of new tools and new living spaces.

Many studies have also noted the link between spatial ability and academic performance in the fields of mathematics and sciences (Hegarty & Waller, 2005; Newcombe & Shipley, 2015; Yilmaz, 2009). This was idea was identified as early as in the work of Vernon (1969; as cited in Anastasi & Urbina, 1997) who examined interrelations and cross-contributions between factors in his hierarchical model in relation to both educational and vocational achievement. It was seen that spatial abilities tended to be linked to scientific and technical abilities in addition to mathematical abilities when examined in tandem with number abilities (Vernon, 1969; as cited in Anastasi & Urbina, 1997). This was further supported by various authors cited in Yilmaz (2009) who identified that spatial skills contributed to understanding advanced mathematical topics, understanding of concepts in both mathematics and science, and that spatial skills correlated highly with success in science subjects. More recent evidence was provided by Zhang and Lin (2015) who found that visual-spatial skills predicted various arithmetic outcomes which encompassed written arithmetic, non-symbolic arithmetic, and arithmetic problems represented in words.

Mathewson (1999) also stressed the importance of understanding visual-spatial skills in the sciences as well as the arts due to the information-rich quality of graphical representations of concepts in both fields. He identified various aspects of visualisation as being embedded within the context of science education including aspects such as location, ordering, and Gestalt among many others (Mathewson, 1999). Similarly, Ivie and Embretson (2010) stated that multiple intelligence tests included scales for spatial ability. High levels of spatial ability have also been seen as related to high levels of

creativity in the fields of art, science, and mathematics (Ivie & Embretson, 2010). A longitudinal study by Humphreys, Lubinski, and Yao (1993, as cited in Ivie & Embretson, 2010) showed that spatial ability emerged as a predictor of career paths particularly for the fields related to engineering, scientific research, and the arts. Due to this, the authors suggested that entrance tests for tertiary education should incorporate at least one spatial ability measure.

Though the benefits of assessing spatial ability are numerous, there remains confusion and inconsistency among researchers in defining the construct and naming its sub-factors (McGee, 1979, as cited in Yilmaz, 2009). This has contributed to unclear definitions of visual-spatial ability and its components. One of the most recent definitions of visual-spatial ability has been provided by Newcombe & Shipley (2015) that stated it concerns the mental representation of "shapes, locations, paths, relations among entities and relations between entities and frames of reference" (p. 180) which can undergo transformations to help the individual navigate, construct, and manipulate the physical environment. This definition expanded that of Lohman (1979, as cited in Carroll, 1993) where spatial ability was seen as "the ability to generate, retain, and manipulate abstract visual images" (p. 305). Carroll (1993) saw visual-spatial ability as "individuals' abilities in searching the visual field, apprehending the forms, shapes, and positions of objects as visually perceived, forming mental representations of those forms, shapes, and positions, and manipulating such representations "mentally"" (p. 304) which guided his exploration of this ability using factor analysis.

Regarding the structure of the construct, Hegarty & Waller (2005) elaborated that visual-spatial ability shouldn't be approached as a singular, undifferentiated construct; rather, it is made up of separate abilities. This was seconded by Linn and Peterson (1985, as cited in Yilmaz, 2009) which suggested that spatial ability comprised of many sub-skills including but not limited to the ability to use maps, the ability to solve geometry

problems, as well as being able to recognise a two-dimensional representation of an originally three-dimensional object. This proposition built upon the hierarchical theory of Carroll (1993) and other theorists such as Lohman and Thurstone. Most prominently, Carroll conducted a factor analysis of over 140 different datasets across different tests and measures to isolate five major clusters of spatial abilities; namely, spatial visualisation, spatial relations, closure speed, flexibility of closure, and perceptual speed (Carroll, 1993). Consequently, Yilmaz (2009) argued for the inclusion of spatial orientation, spatiomeporal ability, and environmental ability as subfactors of visual-spatial ability as measures of testing have expanded beyond the original paper-and-pen or block format commonly used in Carroll's time.

In facilitating the development of aptitude test items, Magno (2009) created a taxonomy of test items to guide item writers in delineating between various constructs. Such a taxonomy was useful in outlining the various skills that can be measured within a test (Magno, 2007). Magno (2007) categorised nine non-verbal schemes of classifying test items. Eight of these nine were used in creating the three original Multiple Aptitude Test (MAT) forms. Literature pertaining to the eight constructs will be presented subsequently and the links to Carroll's subfactors will be discussed. In particular, the various operational definitions will be presented upon which the current operational definitions of constructs in this study are based.

2.5.1 Figure-ground perception

Figure-ground perception is considered a fundamental adaptive skill in humans due to its importance in object perception and visuomotor behaviour (Kimchi & Peterson, 2008). Certain studies have looked into the properties which delineate figures from the background such as size of the region, symmetrical or convex properties of regions, area covered, and familiarity of objects among other aspects (Kimchi & Peterson, 2008). Kimchi and Peterson (2008) found that figure-ground perception was a process that could

happen even when individuals aren't paying attention to the stimuli. Here, figure-ground perception was defined as "the process by which the visual system organizes a visual scene into figures and their backgrounds" (Kimchi & Peterson, 2008, p. 660). Peterson (2015) further explained this process as one of determining if an object has a defined shape that is delineated by the presence of a shared border. The image or region beyond the shared border will be perceived as continuing regardless of the presence of the object thus establishing its role as the 'ground' (Peterson, 2015).

This idea of a perceived border was mirrored in Ghose and Palmer (2016) who identified that extremal edges (illuminated borders along the edge of a curved surface) provided bias in helping individuals identify figures, and, together with gradient cuts (illumination gradients that are "blocked" by edges) create bias in identifying the background. This was mostly seen in three-dimensional images with various light and shadow variations (Ghose & Palmer, 2016). This was also concurrent in the review by Wagemans et al. (2012) where the authors suggested that any two adjoining regions within the visual field usually lead the perceiver to organize the image in a figure-to-ground manner such that the shared border is seen as the occluding or 'blocking' edge of one region over another. Some principles that governed perception of figure versus ground included convexity, degree of symmetry, and size of area, with newer research pointing to cues such as texture (as seen in the extremal edges) and perceived motion (Wagemans et al., 2012). A study by Grossberg (2016) explains the neurological mechanism in the visual cortex that allow for three-dimensional perception of objects that occlude (figure) and occluded objects (ground) based on two-dimensional stimuli which are due to properties such as identification of boundary ownership, Gestalt rules, and selectivity in determining the side of a figure among others.

Mapping back to Carroll's (1993) factors of spatial ability, figure-ground perception can be considered akin to closure flexibility (CF) which encompassed

"finding, apprehending, and identifying a visual pattern, knowing in advance what is to be apprehended, when the pattern is disguised or obscured in some way" (p. 363). A simpler understanding was presented by Yilmaz (2009) that presented CF as the ability to find patterns or figures that have been hidden in larger, more complex patterns after the examinees have been told what they need to look for. One example of a measure of CF were Gottschaldt Figures, where simple figures were embedded in more complex ones (Carroll, 1993). This construct has been identified measured in tests by Thurstone, Educational Testing Service (ETS) and French (Carroll, 1993). Carroll (1993) went on to identify characteristics of test items which tested for CF including the given stimulus must be obscured in some way, the stimulus picture must be a geometric design, and the examinee is pre-informed of the simple figure (stimulus) before being asked to search for it within the more complex pattern; it can be seen these criteria match the more recent ideas of Wagemans et al. (2012).

In summary, figure-ground perception is a visual ability that relates to perceiving the distinction between a figure and its respective background by delineating the borders of the figure that separate it from the background. Based on the preceding studies, figure-ground perception is defined in this study as the ability to recognise and utilise figure-ground cues to discern figural boundaries from their respective backgrounds. An example of a task measuring figure-ground perception is illustrated in figure 2.2.

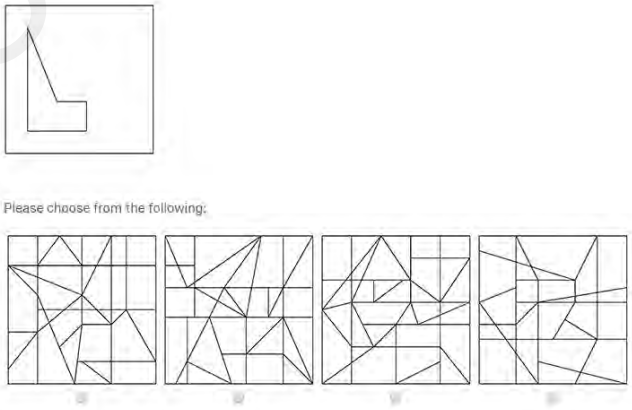

Figure 2.2. Example of a figure-ground perception task. Taken from unused MAT-D item pool

2.5.2 Object assembly

Object assembly has often been seen as a subset of spatial visualisation (Ivie & Embretson, 2010). Lohman (1998, as cited in Yilmaz, 2009) examined visualisation in relation to folding of complex figures, thus contributing to the idea that object assembly would be subsumed under spatial visualisation. This was supported by Carroll (1993) which suggested that items pertaining to object assembly loaded heavily on factors of visualisation. However, Carroll (1993) maintained that, though these two concepts were similar, object assembly tasks were more related to his concept of spatial relations (SR) based on Lohman's conception of the construct, which had a greater focus on problem-solving rather than visualisation with no outcome.

Object assembly requires the respondent to "mentally assemble a two-dimensional object that has been separated into multiple pieces that may have been displaced or rotated or both" (Ivie & Embretson, 2010, p. 324). Generally, the individual is then presented a series of choices to which they must match the mentally assembled image (Ivie & Embretson, 2010). Magno (2009) suggested a more linear presentation of the image where respondents would have to select the order of arrangement in segments of an image to determine how parts should be placed in order to create the whole object. He also stated that object assembly skills are useful in tasks that require putting together or creation of objects as they combine visual analytical skills, visual synthesis skills, and construction skills (Magno, 2009). In this manner, object assembly can be seen as measuring more than mere visualisation skills as higher order thinking is required to attain the right construction of the object. As such, it may be delineated from the subfactor of spatial visualisation.

Object assembly tasks have often been found in military testing settings. In development of items for the American Army's Project A test, Busciglio, Palmer, King, and Walker (1994) stated that object assembly has been evidenced to be a useful measure

of complex, *g*-loaded problem-solving skills and of general spatial ability. In addition, they found it integral in improving the validity of the Armed Services Vocational Aptitude Battery (ASVAB) across various military occupations (Busciglio et al., 1994). It is defined in the ASVAB as the "ability to figure out how an object will look when its parts are put together" (Personnel Testing Division, Defense Manpower Data Center, 2008, p. 1). Several characteristics of test items were identified by Busciglio and colleagues (1994) such as whether items would have marked connections between the pieces or whether the items would constitute a puzzle form. Certain recommendations were made by the authors regarding test item design such as not to include mirror images in correct answer options due to sex differences in ability to perform mental rotations, no distractors would be incorrect on the basis of including a mirrored image, and not more than one distractor should contain the wrong number of pieces (Busciglio et al., 1994).

To sum up, object assembly is a visualisation task with a problem-solving focus. Beyond mere visualisation, object assembly requires the candidate to analyse the placement of each component within the overall image in order to accurately complete the task. In this study, object assembly has been defined as the ability to figure out how separate pieces of disassembled images come together to form one whole, two-dimensional image. An example of a task is shown in figure 2.3.

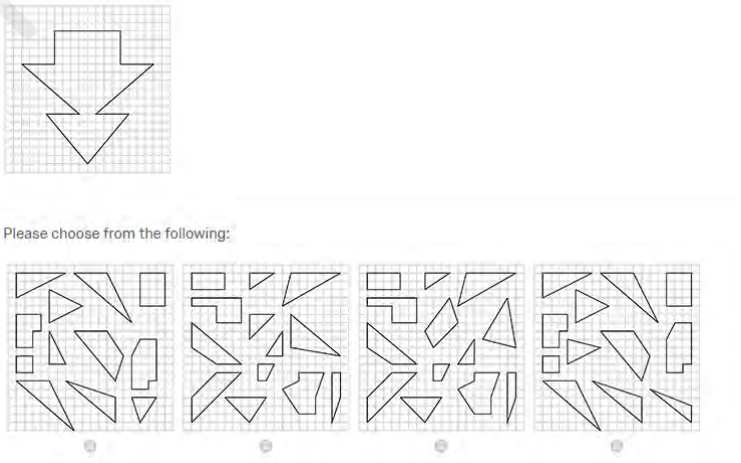

Figure 2.3. Example of an object assembly task. Taken from unused MAT-D item pool

2.5.3 Progressive series

Progressive series are seen as a figural representation of analogical reasoning (Blum, Holling, Galibert, & Forthmann, 2016). Analogical reasoning is a process whereby the individual makes inferences regarding certain unknown entities which are based on available knowledge that has been collected from previously encountered similar and familiar entities (Blum et al., 2016). In this manner, analogical reasoning is a process of prediction based on what is previously known about relationships between objects and information. Beyond mere spatial ability, progressive series examines logical thought and analytical skills in tandem with visualisation of the relationships between objects. This was evidenced in Carroll (1993) that proposed some spatial visualisation tasks loaded heavily on measures on reasoning ability, thus linking the two concepts.

A well-known figural measure of analogical reasoning would be Raven's Progressive Matrices. These tests consist of a set of designs or sequential diagrams with one missing component (Raven, 2000). The test-takers are required to choose the correct answer option which completes the design or series based on a number of given options (Raven, 2000). The progressive matrices are designed to measure two factors of cognitive ability; eductive ability which relates to the ability to find sense and meaning out of chaos and confusion or form mental knowledge structures which aid in unravelling complexity, and reproductive ability which refers to the ability to obtain, remember, and reproduce certain knowledge or information that has explicitly been communicated between individuals (Raven, 2000). Carroll (2014, as cited in O'Hare & McGuinness, 2015) stated that the progressive matrices have been empirically supported to be a measure that correlates the highest with fluid intelligence. They are generally used to test individuals' ability to understand abstract, novel relations and use that understanding in solving problems, independent of formal education or experience (Shamama-tus-Sabah, Gilani, & Iftikhar, 2012).

Multiple definitions have developed from the use of Raven's progressive matrices. An analysis of Raven's progressive matrices by Carpenter, Just, and Shell (1990) define progressive series as the ability to figure out abstract relationships between objects while managing a number of problem-solving goals within the working memory. Another paper presented by Kunda, McGreggor, and Goel (2012) define the progressive matrices as a series of "geometric analogy problems in which a matrix of geometric figures is presented with one entry missing, and the correct missing entry must be selected from a set of answer choices" (p. 1828). It can be seen that the shared elements of each operational definition can be related back to the original skills that Raven set out to measure. Kunda et al. (2012) also proposed a computational model for problem-solving of the progressive matrices stating that multiple cognitive factors play a role in solving the matrices, and that their findings could suggest the interplay between visual representation and cognitive reasoning.

To summarise, progressive series present an interesting form of visual-spatial reasoning such that the tasks involve elements of analytical reasoning in tandem with visualisation skills. Candidates must understand and manipulate the abstract relationships between images in a visual space to arrive at the correct answer. Based on the literature, progressive series is defined as the ability to uncover abstract relations or rules underlying a series of geometric progressions, and infer the missing entries from the series based on these abstract relations. An example of this task is shown in figure 2.4.

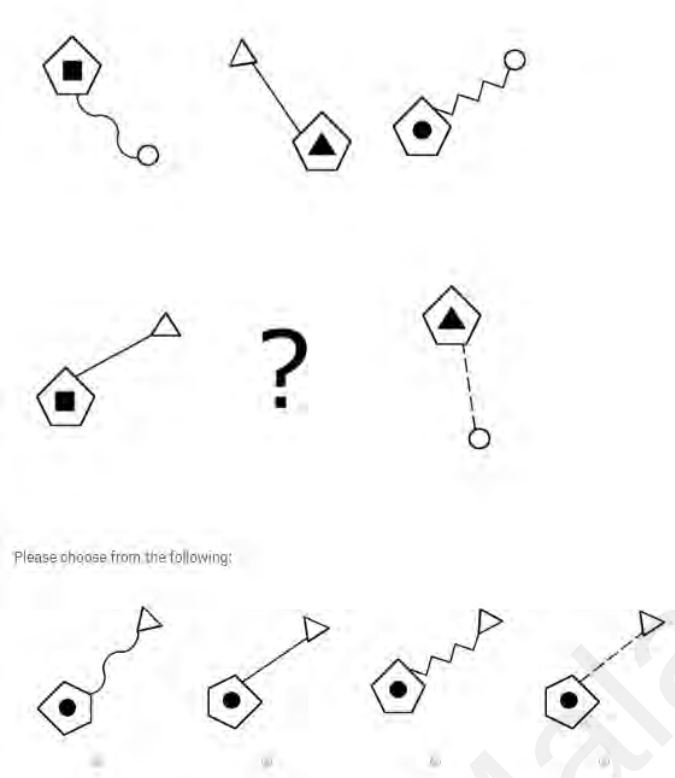

Figure 2.4. Example of a progressive series task. Taken from unused MAT-D item pool

2.5.4 Surface development

In surface development tasks, the respondent is presented with a two-dimensional net or template that can be folded or assembled into a three-dimensional object (Magno, 2009). Under this skill, the respondent has to imagine how the object will look like as it develops from its two-dimensional framework. Typically, these skills have been seen as necessary in areas of drafting, physics, as well as courses with mechanical and analytical components (Magno, 2009). Surface development is often seen as a type of task under the skill of spatial visualisation. Like spatial visualisation, there has been a marked difference in performance based on gender that favours males (Stumpf, Mills, Brody, & Baxley, 2013). Harris, Newcombe, and Hirsh-Pasek (2013) found that development of surface development skills was highly correlated with age in young children, beginning around the age of 5 years.

However, few studies discuss surface development as a construct of its own. For example, in Olkun (2003), surface development is discussed as a type of test item to measure spatial visualisation. This was further supported by Carroll (1993) who noted

that surface development items of a metal folding test (that he considered similar to the actual surface development items of an ETS test kit) loaded within the factor of spatial visualisation. Hence, in the current study, surface development is nested as a component of spatial visualisation. To illustrate, a surface development item is shown in figure 2.5.

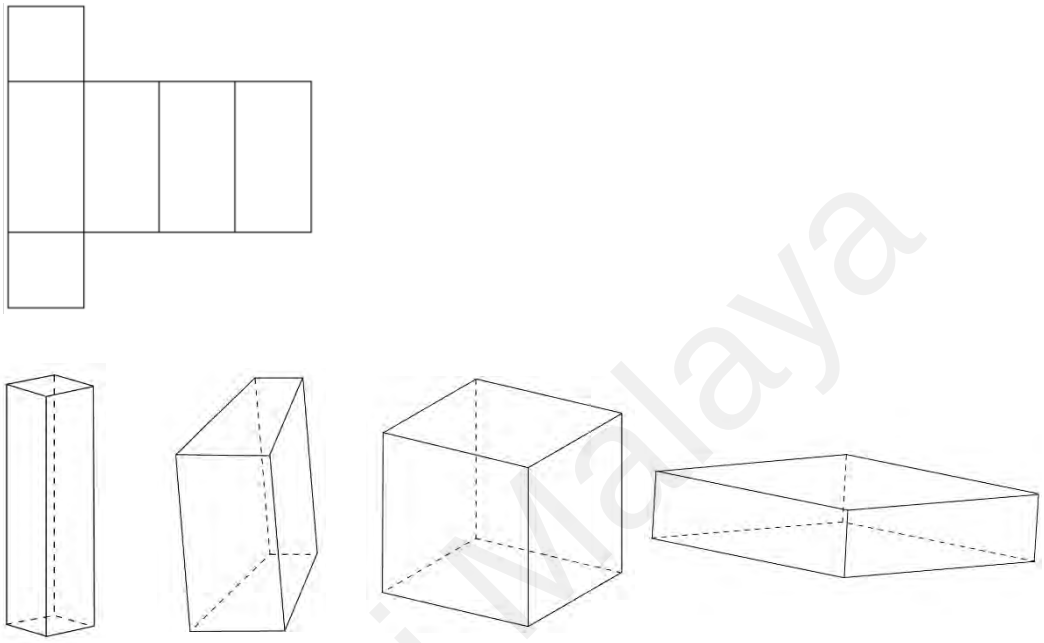

Figure 2.5. Example of a surface development task. Taken from MAT demo item pool. Candidates choose the three-dimensional object that correctly represents the two-dimensional net

2.5.5 Visualisation and orientation

Spatial visualisation and orientation is one of the most researched constructs within the scope of visual-spatial aptitude. It is considered to be highly representative of visual-spatial aptitude as a whole. Generally, it is assumed that males have higher levels of visualisation and orientation skills as compared to females. A study by Quaiser-Pohl, Neuburger, Heil, Jansen, and Schmelter (2014) confirmed this slight advantage in males compared to females as early as age 10. The authors argued that the existence of this difference could be due to gender stereotypes that are learnt within the primary school years or hormonal changes particularly an increase in testosterone for males that increases the asymmetry in cortical function at this age (Quaiser-Pohl et al., 2014). However, a

more recent study by Lin and Chen (2016) found that these differences could be minimised with training.

In general, much literature delineated between spatial visualisation and spatial orientation as two separate constructs. The differences can be seen in the operational definitions of both constructs. In Katsuoloudis, Jovanović, and Jones (2014), spatial visualisation is the ability to mentally picture the rotation of an object, the process of folding or unfolding flat images into various positions, and the changes in position of objects within a represented space. The definition proposed by Yang and Chen (2010) shared many similar characteristics with that of Katsuoloudis and colleagues such as the mental manipulation of spatial information but did not include rotation of objects which they considered a separate process. Gorska and Sorby (2008, as cited in Katsuoloudis et al., 2014) agreed with the idea of mental manipulation of visual stimuli but included processes such as rotation, twisting, and pictorial inversion.

In contrast, spatial orientation was related to ability to picture oneself seeing an object from various angles but still recognising that the object was the same (Katsuoloudis et al., 2014). In addition, spatial orientation involved being able to picture movements, internal or external, of parts within an object or pattern, and mentally represent spatial relationships where one's own bodily and spatial orientation was included as a factor in the problem (Katsuoloudis et al., 2014). Lin, Chen, and Lou (2014) proposed that spatial orientation helped humans answer two important questions, namely where were they in space and where are objects situated around them. In this manner, spatial orientation referred to the ability to determine and understand one's position in the environment (Lin, Chen, & Lou, 2014).

The delineation of spatial visualisation and spatial orientation is further supported by the work of Lohman (1979, as cited in Carroll, 1993) where Lohman defined spatial orientation as imagining how a stimulus piece looks from a different perspective, but

visualisation covered a broader array of test forms. Carroll (1993) also separated these constructs but failed to find factor loadings to suggest the existence of spatial orientation as a factor of its own.

In summary, much overlap has been seen between spatial visualisation and spatial orientation, whereupon some researchers measure them as a singular construct whereas others choose to delineate the two. Based on the preceding studies, this study defines the two constructs separately with spatial orientation defined as the ability to comprehend and identify spatial relationships that exist in regard to an object, and spatial visualisation is defined as the ability to mentally manipulate an object with regard to changes in perception of its physical shape. Tasks measuring each of these constructs are shown in figure 2.6 and figure 2.7 respectively.

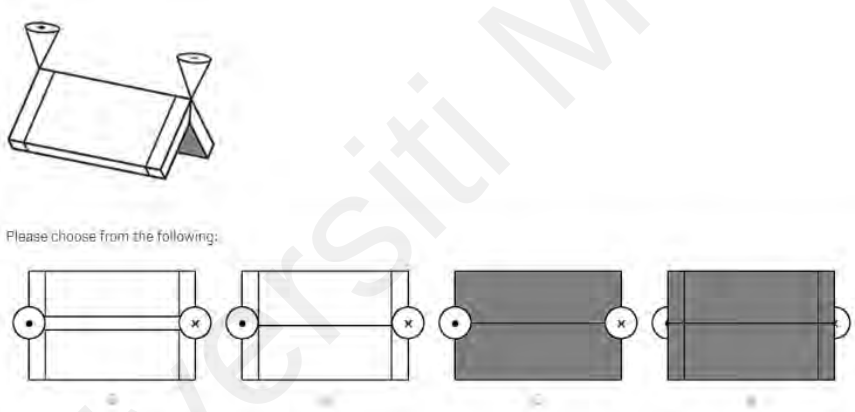

Figure 2.6. Example of a spatial orientation task. Taken from MAT-D

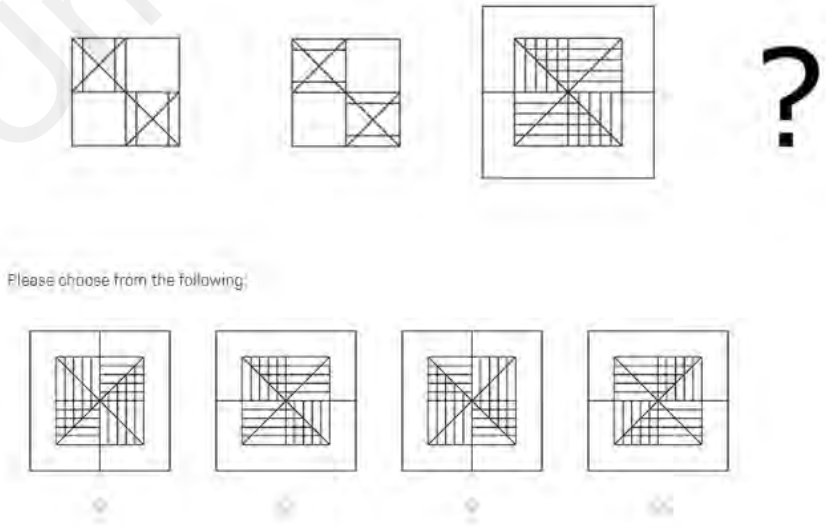

Figure 2.7. Example of a spatial visualisation task. Taken from unused MAT-D item pool

2.5.6 Visual discrimination

Visual discrimination is one of the simplest constructs under the umbrella of visual-spatial ability. At the simplest level, visual discrimination tasks involve determining the similarities and differences between figures in a given set of stimuli (Magno, 2009). Visual discrimination has been exhibited in other species such as sharks where the sharks displayed the ability to discriminate between different shapes (Fuss & Schluessel, 2015) and monkeys which was related to an active process to locate and obtain information from the environment, whereupon this process was learnt through experience (Roitman & Shadlen, 2002).

Visual discrimination skills are typically necessary in courses related to the social sciences. These skills are highly related to the development of reading ability particularly in the case of discrimination between different types of letters (Catts, Fey, Zhang, & Tomblin, 2001) and deficits in visual discrimination have been identified in individuals suffering from schizophrenia (Martinelli & Shergill, 2015). Interestingly, visual discrimination abilities can be affected by respondents' feelings of competence. Zacharapoulos, Binetti, Walsh, and Kanai (2014) found that as participant's self-efficacy increased, so did their scores on tasks of visual discrimination.

In Carroll's framework, visual discrimination was seen as characteristic of perceptual speed (Carroll, 1993). This in turn was based on the definition provided by French (1951, as cited in Carroll, 1993) which stated that perceptual speed was "characterized by the task of finding in a mass of distracting material a given configuration which is borne in mind during the search … ability to compare pairs of items or to locate a unique item in a group of identical items." (p. 345). One sample of test items that closely mirrors visual discrimination items was the Identical Pictures Test from an ETS factor kit from 1963 which required examinees to pick out a picture that matched the stimulus from a line-up of geometrical figures (Carroll, 1993). Though

Bunderson (1967, as cited in Carroll, 1993) was able to derive perceptual speed as a construct on its own, Carroll (1993) found that perceptual speed often overlapped with another construct known as spatial scanning, thus not being able to clearly establish it as a standalone subfactor.

To sum up, visual discrimination measures the candidate's ability to tell apart two objects and select the object that matches the given stimulus exactly. This study defines visual discrimination as the ability to detect if two objects or images are duplicates, or similar, and identify the similarities or differences between the visuals. A sample task is depicted in figure 2.8.

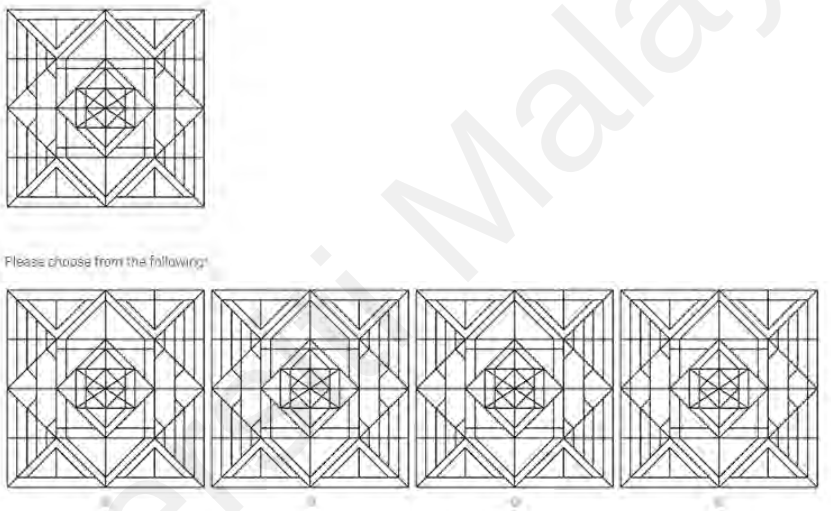

Figure 2.8. Example of a visual discrimination task. Taken from unused MAT-D item pool

2.5.7 Topology

Topology as a construct in itself proved to be extremely broad and complex to understand. The theoretical basis for understanding topology was seen to be highly mathematical and mostly dealt with three-dimensional imagery. A paper by Butner, Gagnon, Geuss, Lessard, and Story (2015) defined topology as a "graphical representation of differential equations, sometimes called a state space, phase space, vector field, or phase portrait" (p. 1) although it was typically used in representing elevation in geographical maps. They proposed that theories of change could be translated

into topographical maps which in turn could be mathematically tested. However, this was more related to a conceptual use of topology as a method of theory representation which was unable to inform the visual-spatial conception of topology.

Another study by de Freitas and McCarthy (2014) defined topology as the study of "properties of geometric objects that remain unchanged under bi-uniform and bi-continuous transformations" (p. 43). In this manner, topology was concerned with the motions of "bending, stretching, twisting or compressing elastic objects" (p. 43, de Freitas & McCarthy, 2014). Most of the examples given in this study were related to untangling knots which required physical manipulation and thus, unsuitable for the purposes of the current study.

The idea of unchanged properties was reflected in a study by Godoy and Rodríguez (2004) which sought to establish metric refinements of topological relations where topological relations were associated with the relative size of objects, degree of overlap between objects, and distance between objects. Although they framed it within the context of information search, the main idea was that in order to get the correct object, the system needs to be able to search available options for one that satisfies the limitations and constraints that are outlined by the spatial relationships as shown in the request (Godoy & Rodríguez, 2004). The most basic definition was found in Magno (2009) which stated that "topology is a measure of spatial relationship which is the ability to see two or more objects in relation to each other" (p. 43). However, the current researcher found the definition to be too simplistic to describe the construct at hand given the complexity of extant literature. Due to the time constraints of the test development process, the construct topology was removed from the current scale development.

2.6 Problems in Assessing Visual-Spatial Aptitude

From the preceding sections, it can be seen that spatial ability is a highly beneficial construct to be assessed within the context of education and vocational guidance. In

addition, it is a common construct that is included in various multiple aptitude batteries (D'Oliveira, 2004). However, it remains a source of confusion for researchers in being able to clearly define the underlying subfactors of spatial ability due to issues as detailed in D'Oliveira (2004). Firstly, it has been seen that there is a lack of consistency in definitions of visual-spatial ability. D'Oliveira articulates how certain abilities share similarities in the descriptions but carry different names, and some with identical names are defined differently.

In addition, there is great variation in the number of abilities identified as belonging to visual-spatial ability (D'Oliveira, 2004). Names of factors have failed to be standardized across the literature, and tests of ability add to the controversy given the wide variation in names and item content (D'Oliveira, 2004). The failure to reach standardization of terms within the literature is not a new problem as Carroll's (1993) seminal book presented a scientific and empirical effort to use factor analysis in clearly establishing and validating the different subfactors of various abilities that fit into his three-factor theory, though few researchers have undertaken such effort since. In this manner, providing accurate definitions for constructs within this study is crucial to clearly operationalise the chosen constructs.

2.7 Conceptual Framework

The conceptual framework of this study is formulated based on Carroll's (1993) three-stratum theory of cognitive abilities. The constructs measured are based on the idea of specific abilities at the level of stratum I. Items are developed to measure the specific constructs which are expected to show relationships to one another given that they will map onto the general ability of broad visual perception at stratum II. Based on the review of related literature, certain constructs from the first version of the MAT were adjusted in the MAT-D. This is also supported by the scale development guidelines of DeVellis (2003) where constructs must be clearly identified prior to item-writing. Further details

of the development process will be provided in chapter three. Based on the current framework, topology is no longer considered in the visual-spatial scale of the MAT-D, and visualisation and orientation is split into two separate constructs. In addition, surface development is subsumed within the measurement of spatial visualisation. Greater parallelism to Carroll's theory and Yilmaz's (2009) recommendations was achieved with the redefinition of constructs to be assessed. Figure 2.2 depicts the conceptual framework of visual-spatial aptitude as used in this study and figure 2.3 shows the changes in the visual-spatial constructs measured in the first three forms of the MAT to the MAT-D.

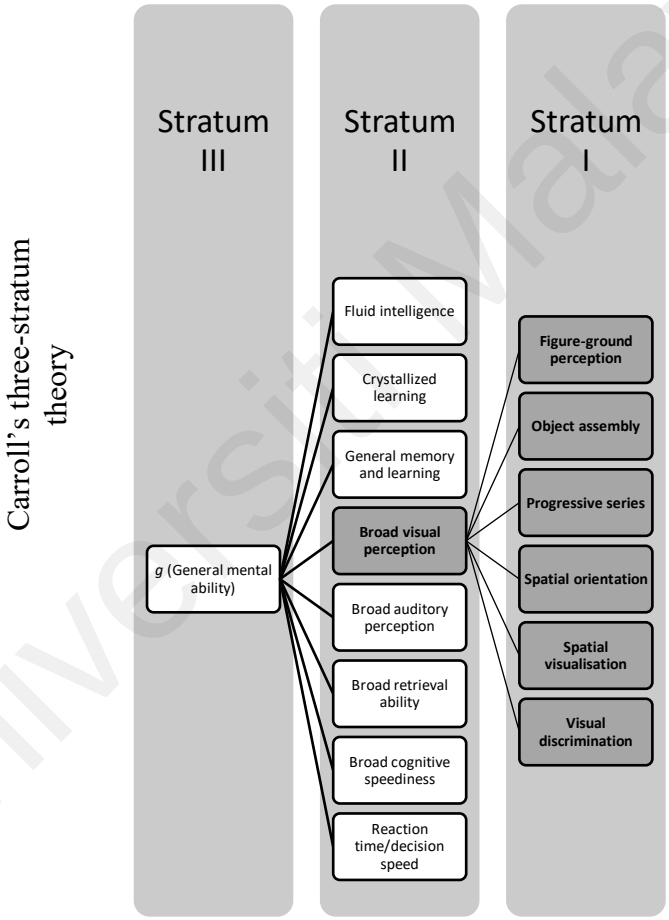

Figure 2.9. Conceptual framework of the development of the MAT-D(VS). Shaded constructs depict the constructs measured in the current study

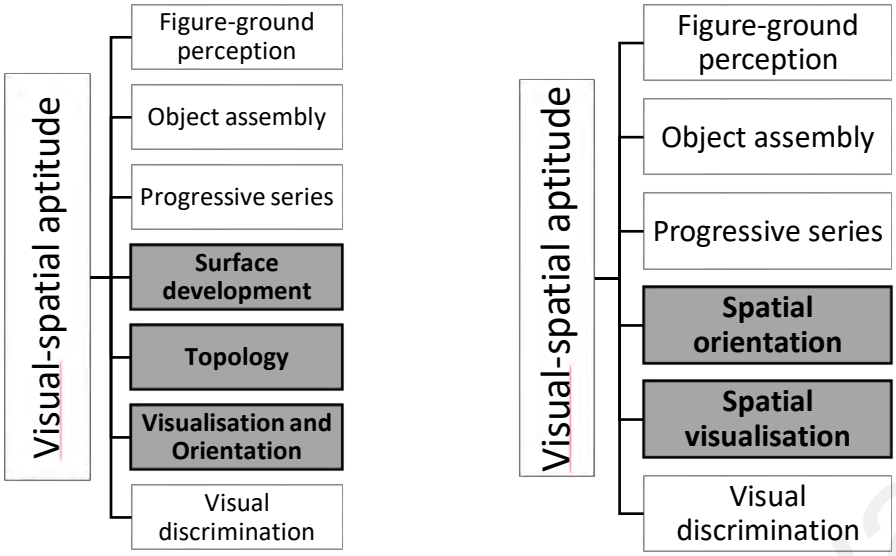

Figure 2.10. Comparison of constructs in original MAT [left] and MAT-D(VS) [right]. Changes to constructs are shown in shaded constructs

2.8 Summary

This chapter detailed the development of career guidance and counselling in various Asian nations and discussed the types of aptitudes that can be assessed according to theory followed by a brief overview of Carroll's three-stratum theory which was the framework upon which the current study was based. Literature pertaining to the constructs to be measured was given particularly to provide an understanding on the various operational definitions that are available followed by the presentation of the conceptual framework of the study. Problems relating to assessment of visual-spatial aptitude were also highlighted to provide context within the current study of the need for re-definition of constructs. The next chapter will describe the methodology of the study in terms of the design, sample and population, and procedure. The instrument under development is described and an overview of the test development process is given. The pilot study and subsequent results are also presented.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter provides a brief overview of the design of the study, the intended population, and the sample utilised. The MAT-D(VS) is described in detail before the pilot study and accompanying results are presented. A short description of the procedure is provided, and the data analysis techniques are discussed.

3.2 Design

This study used a quantitative research design utilising the cross-sectional survey. With the cross-sectional survey method, participants will be administered the scale only once and no retest of the scale will be given (Sedgwick, 2014). This method allows the researcher to gather a large set of data from the population within a short time period thereby reducing the chances of participant attrition (Sedgwick, 2014) and the data obtained would be representative of the population as the sample will include a range of ages, genders, and races thereby providing a 'cross-section' of the intended population (Fraenkel, Wallen, & Hyun, 2012).

The online mode of distribution was selected in this study compared to paper-and-pen administration for a few reasons. Firstly, the sample consists of students from a local private university who are well-acquainted with answering surveys and tests through online platforms. Secondly, it allows for greater convenience on the part of the respondents as they can answer at their own pace (Fraenkel, Wallen, & Hyun, 2012). It is of low cost to the researcher as there is no need for preparation and printing of test materials and allows for quicker turnaround time as there is no need to schedule predetermined test administration time slots (Fraenkel, Wallen, & Hyun, 2012). In addition, no data is lost as the platform hosting the scale can be adjusted to not allow for non-response (Qualtrics, 2017).

3.3 Population

The complete test assessment battery is intended to aid career guidance counselling for secondary and tertiary education students with an applicable age range of 14 to 25 years old. As the MAT-D(VS) was developed as a higher-ability alternative to the earlier MAT-A, -B, and -C, the intended population for this scale is foundation or sixth form and first-year undergraduate students of local public and private tertiary institutions. The expected age range is between 17 to 25 years of age.

3.4 Sample

The sample consisted of 203 first-year undergraduate students from the Department of Psychology of a local private university which fits the recommendations by Şahin and Anıl (2017) that proposed a minimum sample of 150 participants meets the requirements to run the Rasch model analysis regardless of test length based on a comprehensive study of sample size and test length for analyses of varying parameters. It was found that a sample size of 150 participants provided similar estimates across test lengths between 10 to 30 items using a simple one-parameter estimation (Şahin & Anıl, 2017) which can be attributed to the preciseness of logits within the Rasch measurement model thus enabling the full variability of participants to be reflected by samples of such size. The size also meets the sample size recommendations of Hair, Black, Babin, and Anderson (2010) which suggests a minimum of 20 participants for each construct being measured in order to run a confirmatory factor analysis.

The mean age of participants was 19.63 with a standard deviation of 1.352 and the overall ages ranged between 18 to 29 years old. Participants were recruited from first-year undergraduate career guidance classes where they were offered extra credit for their participation. Participants were briefed in class that the study sign-ups would be available through the university's own online platform for research participation known as iPsy. Once participants signed up, they were emailed the link to the study and the procedure of

the study was explained to them. Completion of the scale was monitored through the Qualtrics platform. Participants who successfully completed the scale were given extra credit which was added to their final scores for the career guidance subject class.

Convenience sampling was used in this study. This sampling method allows for greater access to the population under study based on the time constraints of the test development schedule and student availability according to the semester system of their courses (Bordens & Abbott, 2011). Although one criticism of convenience sampling is that opportunities for sampling are limited to as far as the researcher has easy access to (Bordens & Abbott, 2011), this method is suitable as the online link can be easily distributed to students via the online student portals for news and announcements. As the sampling was carried out in accordance with other rounds of testing (for development of other scales within the battery) under the career development department of the university, it allowed for access directly to the population of study and, thus, the sample will be highly representative of the population. Therefore, although arguably limited in generalizability (Bordens & Abbott, 2011), convenience sampling is suitable for the purpose of the development of the MAT-D(VS).

3.5 Instrument

The MAT-D(VS) is a 30-item scale to measure visual-spatial aptitude consisting of six subscales. Each subscale consists of five items, and two types of scores are obtained; scores per subscale are the sum of all the correctly answered items and a total visual-spatial score which is the sum of all subscale scores. Each item provides four response options with one correct answer. The presentation of items in each subscale is preceded by an information sheet that briefs the participants regarding the nature of the subscale and a sample item. Please refer to appendix A for samples of the information sheets. Due to the private nature of the development of the MAT-D(VS), the full scale will not be included in the appendices.

3.5.1 Initial development

The process of development was adapted from the guidelines recommended by DeVellis (2003). DeVellis (2003) outlined eight steps in his guidelines that were fulfilled in a linear manner. A graphical representation of DeVellis' guidelines can be seen in figure 3.1 on the subsequent page. Determination of constructs to be measured was related to review extant literature in the field to have a clear idea of the construct under measure (DeVellis, 2003). This was related to the step of identification of constructs in the current study. Item generation encompassed steps two to four of the DeVellis guidelines as detailed in the following sections. No items were necessary for validation purposes; hence, the step was removed. Administration to a developmental sample was done in two phases, the pilot, and the post-amendments stage. Unlike DeVellis' (2003) guidelines, the current development process could loop back to amendments if evaluation of items deems it necessary. Optimization of scale length was also unnecessary as the scale is relatively short and length was informed by preceding versions.

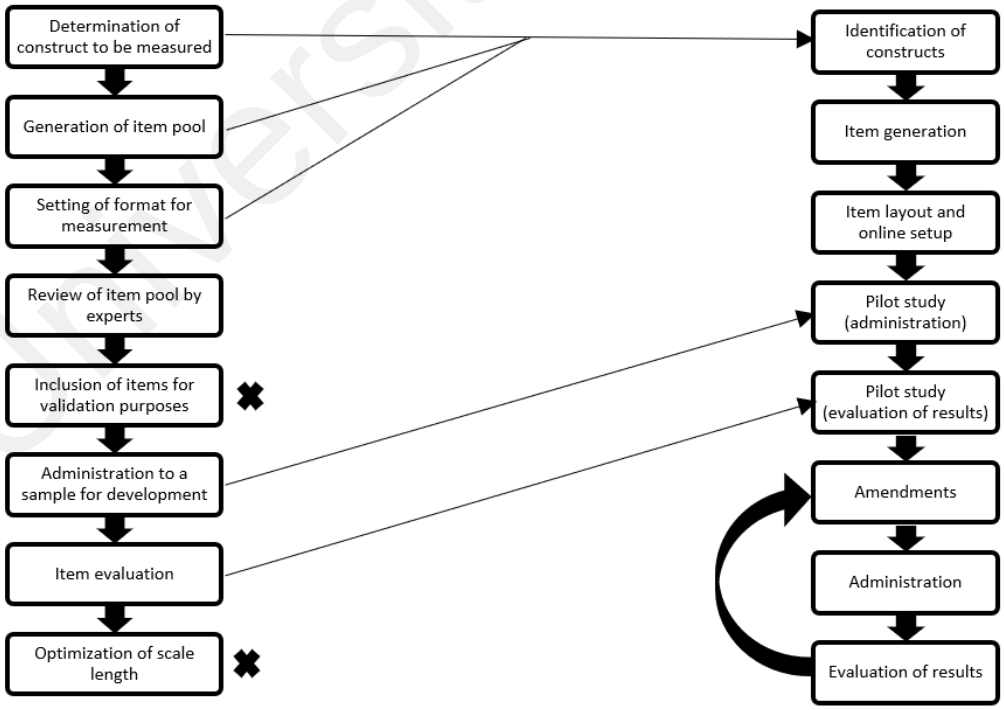

Figure 3.1. Guidelines for scale development based on DeVellis (2003) [left] and development process of the current study [right]

3.5.1.1 Identification of constructs

The MAT-D(VS) is intended to be a measure of visual-spatial aptitude at higher ability levels as compared to the MAT-A, -B, and -C. It roughly follows the same template as the initial versions of the MAT with slight changes to the underlying operational definitions and constructs measured. In the first stage of development, a review of literature was conducted to clearly define the constructs under study. The relevant literature discussed can be seen in chapter two.

From the review, the operational definitions that were presented in chapter one were developed. Although most constructs remained similar to those in the earlier MAT, the original construct of visualisation and orientation was delineated to spatial visualisation and spatial orientation separately. This is due to ambiguity in the literature presented regarding the unification of the construct. In addition, surface development was seen as a form of spatial visualisation as there seemed to be overlap between the cognitive processes measured by the two constructs. As such, surface development was nested as a part of spatial visualisation. Lastly, topology was identified as heavily mathematical in nature and comprised of a multitude of theories. In this manner, it proved too difficult to properly define and was beyond the scope of the researcher's ability to measure. Hence, topology was excluded from the MAT-D(VS). The list of constructs with their respective definitions can be seen in table 3.1.

Table 3.1

Operational Definitions of Constructs in MAT-D(VS)

Construct	Operational Definition
Figure-ground perception (FGP)	The ability to recognise and utilise figure-ground cues to discern figural boundaries from their respective backgrounds. Examples of such cues include but are not limited to: proximity, similarity, symmetry, parallelism, good continuation, and closure.
Object assembly (OA)	The ability to figure out how separate pieces of disassembled images come together to form one whole, two-dimensional image, much like a jigsaw

	puzzle. The pieces of disassembled images can be rotated, displaced, or both, before they are re-assembled to form a coherent image.
Progressive series (PS)	The ability to uncover abstract relations or rules underlying a series of geometric progressions, and infer the missing entries from the series based on these abstract relations.
Spatial orientation (SO)	The ability to comprehend and identify spatial relationships that exist in regard to an object, such as understanding the shape of an object as it is viewed from different angles, and identifying the position of the object in space in relation to other objects or oneself.
Spatial visualisation (SV)	The ability to mentally manipulate an object with regard to changes in perception of its physical shape, such as the folding and unfolding of a flat surface to a 3-dimensional object, active object rotation around a focal point, twisting of the object, and inversion/mirroring of the visual.
Visual discrimination (VD)	The ability to detect if two objects or images are duplicates, or similar, and identify the similarities or differences between the visuals.

3.5.1.2 Item generation

Based on the operational definitions, the items were generated by two separate researchers. The six constructs were divided between the two researchers; one researcher handled item generation for figure-ground perception, object assembly, and progressive series while the other researcher generated items for spatial orientation, spatial visualisation, and visual discrimination. The items were designed by referring to the operational definitions and creating problems that fit within the described parameters of the definitions.

Items were hand-drawn on graph paper before undergoing digitization using version 2.8 of the GNU Image Manipulation Program (GIMP; The Gimp Team, 2017). 3D images were digitized using FreeCAD version 0.17 (FreeCADweb.org, 2017). Due to time constraints, only five items were generated per subscale. Each item was followed by four response options following the format of preceding versions. Items were reviewed by the remaining members of the research team. At face value, items seemed to be

representative of the operational definitions on which they were based. As per recommendations from DeVellis (2003), the items were also reviewed by an assessment consultant from a local private university who agreed that the items showed content validity as they were related to specific constructs under the umbrella of visual-spatial aptitude. This was also in line with the process outlined by Sumintono and Widhiarso (2013) to obtain expert evaluation of the items as a means of determining content validity prior to pilot testing.

3.5.1.3 Item layout and online setup

The online survey delivery platform, Qualtrics, was used to host the presentation of the items. Once digitized, items were resized for presentation on Qualtrics. Each subscale was preceded by the subscale information page. Five items were presented per page. Each stimulus item was placed in a row above the four answer options. For the list of options longer than the width of the page, a scroll bar was available below for the participants to move in order to view the item in full. The survey options were set such that participants could only select one answer per item, participants had to provide a response to all items before proceeding to the next page, and participants were not allowed to navigate to previous subscales once they moved to the following subscale, similar to previous MAT versions.

Only participants who were distributed the link would be able to access the survey. The first draft of the survey was distributed to three members of the research team to check for spelling errors and obtain feedback for improvement. This was in line with the recommendations by Sumintono and Widhiarso (2013) to check for readability of items in a mini-pilot. Due to the image-heavy nature of the scale, an additional instruction was added to each subscale information page for participants to wait for the page to complete loading before responding. A scoring key was added to Qualtrics so a total score was obtained on the scale, but this score was not shown to respondents.

3.6 Procedure

The link to the scale was distributed to the participants via email. Within the email, participants were reminded to complete the scale in a quiet area with no distractions, and that the scale had to be completed in a single sitting. The estimated time taken for completion was between 45 to 60 minutes but there was no set time limit. Participants were presented the items one subscale at a time through the online delivery platform, Qualtrics (Qualtrics, 2017). On the first page, the informed consent document (Appendix E) was shown and participants indicated their consent to participate in the study by clicking 'Next'.

The first page presented was a general information sheet regarding the different sections within the scale. The next page was the information sheet for figure-ground perception. Participants were reminded on this sheet that they would not be able to move back to the previous page once they had submitted their responses. The following page was the figure-ground perception subscale. This subscale was followed by the object assembly information page, the object assembly subscale, and so on for the subscales of progressive series, spatial orientation, spatial visualisation, and visual discrimination. For each item, participants chose the answer option they believe to be correct based on the subscale being answered. The final section consisted of demographic information items.

After their full answers were recorded, participants were thanked for their time. Participants did not receive a score report due to the developmental nature of the study, though scores were submitted to the class lecturer to determine eligibility for extra credit. Any and all questions were communicated to the researcher via email.

3.7 Data Analysis

In this study, the psychometric properties of the newly-constructed MAT-D(VS) were tested statistically as per the recommendations of Furr (2011) and DeVellis (2003). Confirmatory factor analysis was conducted using Mplus (Version 7.0; Muthén &

Muthén, 2012) to determine construct validity. Furr (2011) finds construct validity to be one of the most important types of validity given that it is a reflection of the internal structure of the construct, is highly impactful on the accurate interpretation of the scores derived from the scale and signifies a degree of representation of the psychological processes underlying the individuals' responses. For these reasons, the contemporary viewpoint prioritises construct validity over the tripartite perspective of content, criterion, and construct validity (Furr, 2011). This is supported by Cattell (1978) who suggested construct validity, as determined through the use of factor analysis, is part of continuous, empirical test evaluation that is necessary to establish conceptual validity, and Anastasi and Urbina (1997) who referred to factor analysis as a relevant statistical technique in establishing construct validity due its refined nature in being able to determine interrelationships in behavioural data.

To determine the reliability of the MAT-D (VS), Quest (Version 2.1; Adams & Khoo, 1996) was used to run the Rasch model analysis that will provide values of item and person reliability. This software was chosen due to the features of the software that allowed for identification and removal of items and persons that do not fit to the measurement model (Adams & Khoo, 1996). This allowed for more accurate measurement of the psychometric properties of the scale. In measuring reliability of a test, Rasch models provide outcomes in terms of internal consistency, item reliability, and person reliability. Internal consistency is typically reported in the form of Cronbach's alpha (α) where the value relates to the extent to which items within a test are seen to measure the same construct (Tavakol & Dennick, 2011). Generally, acceptable values of Cronbach's alpha lie between 0.70 to 0.95 but some researchers caution against an alpha value higher than 0.90 which may indicate redundancy of items (Tavakol & Dennick, 2011). Person reliability in Rasch is similar to Cronbach's alpha as it is an indicator of the extent to which a person will have the degree of traits prescribed by their score on the

test (Linacre, n.d.). Item reliability, on the other hand, denotes the extent to which an item that is meant to measure a higher level of ability can really measure the higher level (Linacre, n.d.). On the whole, all three measures are typically examined to ensure that the test is reliable.

Quest was also used to run the distractor analysis once non-fitting cases are removed from the initial analysis as per the recommendation of Sumintono and Widhiarso (2013). Removal of participants who show high degrees of misfit according to the person fit statistics is necessary in avoiding the impact of person misfit on the psychometric properties of the scale (Sumintono & Widhiarso, 2013). This will allow for greater accuracy in examining the quality of the distractors during the distractor analysis.

The use of two measurement models (the factor analytic model and the Rasch model) in this study serves a complementary purpose in establishing validity and reliability of the scale at both the construct and item level. Where factor analysis is the method of choice in determining construct validity, the use of factor analysis is solely in determining the degree of relationship between the items and the underlying construct through empirical testing of its mathematical structure (Anastasi & Urbina, 1997) thereby allowing items that do not show significant relationships to the underlying construct to be removed at a preliminary stage of scale development. This presents an essential step in establishing conceptual validity for the latent construct (Cattell, 1978).

The use of the Rasch model, however, is primarily to establish scale reliability and conduct effective item analysis which a method of quality control at an item level. Given that the validity and reliability of a test is highly dependent on the quality of the items, item analysis plays an important part in the selection of high quality items and substitution or revision of poor items (Anastasi & Urbina, 1997). The Rasch model is particularly suitable for this purpose as the relationship between each item and the underlying construct is assessed in determining the quality of the item (DeVellis, 2003).

It also allows for increased precision as the difficulty can be fine-tuned at the item level and the items can be matched to the specific degrees of the underlying construct (DeVellis, 2003). Particularly for items with hierarchical levels of the underlying construct, IRT and the Rasch model are highly appropriate as they can discern the different levels of increasing difficulty in the items (DeVellis, 2003).

Though it is argued that the Rasch model can also provide evidence for validity, this is only through the meeting of the assumption of unidimensionality whereupon the scale is considered to measure the same underlying construct when the data fits the model (Bond & Fox, 2015). However, this still differs from construct validation and does not allow for the examination of the degree of relationship between item and construct as is done in factor analysis. Therefore, both factor analysis and Rasch analysis are suitable to be used in tandem to ensure that the scale consists of high quality items and has good psychometric properties of validity and reliability.

Raw data was downloaded from Qualtrics in the Excel format. Responses are recorded in the form of multiple choice answers (A, B, C, or D) to each respective item. Data was converted to .txt format to be used within Quest. In the Quest workspace, a scoring key can be input to the Quest control file in order for the program to recognise correct and wrong responses. This allows the dichotomous Rasch model to be applied where a correct answer indicates a higher level of the ability measured in the participants and a wrong answer indicates lower levels of ability (Bond & Fox, 2015). In Excel, a scoring key was applied to convert responses to binary output (1-0) to be used within Mplus before the resulting data was converted to the .csv format to be read by the program. Mplus was selected as the software of choice to conduct the confirmatory factor analysis as it is capable of running a factor analysis on binary (dichotomous) data which requires the use of estimators based on tetrachoric correlations as opposed to Pearson correlations (Muthén & Muthén, 2012).

Raw total scores obtained on the scale ranged between 5 to 29 out of 30 questions (M=16.94, SD=4.798). Normality testing was conducted on the total scores of the scale which were represented as an interval scale with a possible score range between 0 to 30. This was to determine if the total scores fell within a normal distribution such that all ranges of student performance as shown by the scores were able to be captured by the scale and to ensure that the participant responses showed no form of skew or bias that may affect interpretation of the results (Field, 2013). The distribution of scores obtained was approximately normal with skewness and kurtosis (Skewness=-0.254, SE=0.171; Kurtosis=-0.194, SE=0.340) within the acceptable range of ±2.00 as suggested by Field (2013). The Shapiro-Wilk test for normality was used as per Field's (2013) recommendation for sample sizes less than 2000 and results revealed that normality was assumed for the distribution of scores on the scale (W(203)=0.987, p=0.58). Therefore, the distribution of total participant scores on the scale approximated a normal distribution. Statistical output for descriptive statistics and normality testing of the sample can be found in appendix F.

3.7.1 Measurement model: Confirmatory factor analysis

Confirmatory factor analysis (CFA) is a statistical technique that is grounded in the principles of exploratory factor analysis. It is often used in confirming or establishing a known pattern of correlations and covariances between constructs based on predictions of theory or the results of previously run analyses (DeVellis, 2003). CFA has previously been established as one of the main techniques to determine the various categories under one form of aptitude (Magno, 2009) and has previously been used in establishing support for Carroll's three-stratum theory (Bickley et al., 1995).

The advantages of conducting confirmatory factor analysis are typically related to its flexibility such that correlated and uncorrelated factors can be mixed within the same model in accordance to the base theory (DeVellis, 2003). According to Furr (2011), CFA

is often the method of choice when the underlying structure (in terms of number of constructs measured by the items) is clear to the researchers. An understanding of the internal structure of the scale is crucial in informing the reliability of the scale, with regard to its internal consistency, and the validity of the scale as accurate interpretation of the results is highly dependent on the match between the structure of the scale and the underlying structure of the construct under measure (Furr, 2011). In this study, CFA is chosen as the model structure of the MAT-D(VS) has previously been determined in the initial versions of the MAT.

In relation to scale development, CFA allows the researcher to finetune the developed scale by assessing the measurement model based on its 'goodness-of-fit' where the degree of match between the hypothesized model and the data is evaluated, and by examining the factor loadings of each item onto the intended construct (Furr, 2011). Generally, model fit is considered 'good' when there is a high degree of consistency between the measurement model and the collected data; 'poor' fit arises when there is inconsistency between model and data (Furr, 2011). Indices and thresholds for determining 'goodness-of-fit' will be discussed below.

Factor loadings provide information regarding the extent to which each individual item contributes to or is related to the underlying construct (Furr, 2011). It is expected that variations in the examinees' responses to the items will be due to variations in their underlying levels of the construct assessed (Furr, 2011). Thus, if the item loads well onto the underlying factor, significantly large positive values are expected to be seen (Furr, 2011). Correlations between the underlying factors can also be examined to provide information regarding how closely the factors are linked (Furr, 2011).

3.7.1.1 Method of estimation

In terms of the type of estimation used in the analysis, binary data (variables with only two response categories) require the use of tetrachoric correlations as compared to the usual Pearson's product-moment correlation (Beauducel & Herzberg, 2006). Maximum likelihood (ML) estimation has been seen to be a highly popular basis for running CFA but this usually only follows for data which follow the normal distribution (Beauducel & Herzberg, 2006).

A previous estimation following the weighted least squares (WLS) method was originally suggested for categorical data but comparisons of both estimation methods showed large amounts of bias using the WLS estimates particularly when sample sizes were small (Hoogland & Boomsma, 1998, as cited in Beauducel & Herzberg, 2006). The weighted least squares means and variance adjusted (WLSMV) estimation was introduced by Muthén and colleagues as a revision of the WLS estimator which was expected to work with smaller samples while producing less bias (Beauducel & Herzberg, 2006). A comparison study by Beauducel and Herzburg (2006) found that the WLSMV was more appropriate than the ML estimation with variables that had two to three response options at a roughly equivalent sample size.

3.7.1.2 Fit indices

The main indices that are used to determine goodness-of-fit are the chi-square test of goodness-of-fit, the Steiger-Lind Root Mean Square Error of Approximation (RMSEA), and the Comparative Fit Index (CFI). The chi-square test presents one of the oldest methods of determining goodness of fit. It is a measure of the difference between sample covariance matrices and model-fitted matrices (Hu & Bentler, 1999) where smaller, non-significant values are taken as an indicator of a lesser degree of 'mis-fit' between the models thus supporting the measurement model (Furr, 2011). As such, good model fit is determined at $p>0.05$, where the difference between the matrices is not

significant (Hooper et al., 2008). Bentler and Bonett (1980, as cited in Mulaik, James, Van Alstine, Bennett, Lind, & Stilwell, 1989) proposed that the chi-square test was appropriate to be used with large samples as the statistical power to detect small differences between model and data would be increased. However, the test relies heavily on assumptions of normality, and deviations from normal can result in model rejection (McIntosh, 2006, as cited in Hooper et al., 2008).

The RMSEA is based on the chi-square test and works appropriately for large sample sizes (Kenny, 2015). Rigdon (1996) suggests that the RMSEA is suitable in confirmatory contexts as its behaviour is unpredictable with smaller samples. It is a highly popular measure of model fit due to the fact that it favours parsimony, and it has the ability for confidence intervals to be calculated using the RMSEA value (Hooper et al., 2008). Kenny (2015) recommends that the 90% confidence interval should provide a lower value that is close to 0.00 or not more than 0.05 and an upper limit not greater than 0.08. Good model fit is determined at RMSEA<0.06 (Hu & Bentler, 1999), though some authors have provided an upper threshold for fit at RMSEA<0.08 (Furr, 2011). PCLOSE provides a test of the null hypothesis for the RMSEA where p>.05 allows the conclusion that the model fit is 'close' (Kenny, 2015).

The CFI by Bentler (1990, as cited in Hooper et al., 2008) is based on a revision of the Normed Fit Index (NFI), which was one of the first incremental fit indices proposed by Bentler and Bonnett (1980, as cited in Kenny, 2015) but was problematic given that the index would continuously increase as more parameters were added (Kenny, 2015). The improved CFI has been seen as complementary to the RMSEA (Rigdon, 1996). It accommodates well for various sample sizes and shows good performance even in small samples as compared to the NFI which can underestimate model fit for small samples (Hooper et al., 2008). Initially, acceptable limits were set at CFI≥0.90 which was eventually raised to CFI≥0.95 to prevent acceptance of misspecified models (Hu &

Bentler, 1999; Rigdon, 1996). The CFI has emerged as a popular fit index to report due to its robustness against extreme variations in sample size (Fan et al., 1999, as cited in Hooper et al., 2008).

3.7.2 Measurement model: Item response theory

Item response theory (IRT) is a psychometric theory that is used in test development particularly in the assessment of latent or underlying traits or abilities (Baker, 2001). It is slowly becoming one of the most widely used techniques in developing educational assessments, psychological tests, and clinical assessments (An & Yung, 2014). It is assumed that an examinee or respondent on a test has a certain degree of the underlying ability that is meant to be measured by the test (Baker, 2001). As such, the degree of ability can be represented using a numerical score that is obtained through the test (Baker, 2001). At a certain score of ability, it is then assumed that there exists a probability that the examinee will provide a certain response (Baker, 2001). To illustrate, in a test of academic achievement, a student of high ability will likely be able to choose the correct answer to a test item.

These concepts align with principles of invariant measurement such that the ability of the individual as represented by their test score is the closest actual measure of the underlying ability and, thus, will not vary across different types of measurement tools (Englehard, 2013). Bond and Fox (2015) have supported this idea given that psychometricians strive to create invariant measures. It is hoped that values obtained by a measurement tool for an underlying construct should not vary greatly across all similar and suitable contexts (Bond & Fox, 2015). Similarly, it is expected that within the same context, all measurement tools that were created to measure the same construct should provide similar values (Bond & Fox, 2015).

3.7.2.1 Rasch model

Though many models exist under the umbrella group of IRT models, the Rasch model remains one of the most well-developed and used models. The creator of the Rasch model, Georg Rasch, developed the statistical model based on his ideas of specific objectivity which related to being able to make independent comparisons between test items as well as between test respondents (Englehard, 2013). Specific objectivity is concept that Englehard (2013) posits as being supportive of invariant measurement, an idea paralleled by Bond and Fox (2015) where both estimation of test item difficulty level and test respondent ability ought to be invariant regardless if the same sample was split into subgroups and re-analysed.

The Rasch model was considered unique in its ability to obtain invariant measures due to the separation between test items and test respondents at the conceptual level which allowed for the simultaneous mapping between items and persons (Englehard, 2013). It also relied on the use of total scores which provided leeway for the consistent analyses of varying response patterns, unlike the earlier model proposed by Birnbaum (Englehard, 2013). In addition, IRT models seek to find fit between the model and the real-world data (Yu, 2017). A statistical model that accurately reflects reality is often seen as overly messy and complicated; hence, Rasch models provide parsimony in the sense that data also undergo adjustment to attain model fit (Yu, 2017).

The model typically runs item calibration that is assumed to be person-invariant (the difficulty of the item is estimated independently of the respondents' data) and person proficiency estimation which is also assumed to be item-invariant (a measure of the individual's ability without accounting for attributes of the item), which adds to the objectivity of the model (Yu, 2017). Essentially, person-invariant measurement of test items involves the development of calibrated test items which are independent of the sample used for calibration and item-invariant measurement of persons allows for

estimation of an individual test respondent's level of ability independent of the test items used in the measurement process (Englehard, 2013). Overall, the statistical analysis runs in cycles where the estimations of item difficulty and person ability that emerge from data are used to fit the model, and then the model is used to project or predict further data, and model-data fit is achieved after a few cycles or iterations (Yu, 2017).

To run the iterative analysis, data is transformed to a unique scale within the constraints of the Rasch model. The model typically runs on a cumulative logistic distribution which requires the data to be transformed to log-odds units or logits (Englehard, 2013). Where data are usually obtained in the form of nominal or ordinal scaling, the logarithmic transformation within the Rasch model allows the data to be scaled along an interval scale (Bond & Fox, 2015). Linacre and Wright (1989) define a logit as "the distance along the line of the variable that increases the odds of observing the event specified in the measurement model by a factor of 2.718.., the value of "e", the base of "natural" or Napierian logarithms used for the calculation of "log-" odds" (para. 7). Thus, data is transformed to units of equal length within the logit scale hence providing the qualities of interval data (Linacre & Wright, 1989). Logits relating to a person's ability are derived from the natural log odds values for success on items that have been selected to represent the "zero" of the interval scale (Wright & Stone, 1979). Similarly, logits of an item's difficulty are seen in the natural log odds values that represent the failure of persons who have "zero" ability (Wright & Stone, 1979). These transformed values can fall within a range of $\pm\infty$ but more often have been seen to fall in the range of ± 5.00 (Englehard, 2013). Transformation to a uniform, interval-scaled logit allows for the independent comparisons of items and persons due to its independence from the pool of items and calibration sample used as the logit values are unaffected by variations from these two sources (Englehard, 2013).

Rasch models generally run estimation for one parameter, namely, the difficulty parameter which is sometimes known as the threshold parameter, up to a maximum of three parameters including estimations of item discrimination and the guessing factor (Yu, 2017). The one parameter model is most suitable for test construction to accurately match test difficulty against the range of students' ability (Yu, 2017). In most literature, the level of difficulty (item) or level of ability (person) is represented as theta (θ). Simply put, theta is the representation of the logit value for either item difficulty or person ability (Yu, 2017). Theta is usually mapped against a probability axis to denote the probability of a correct response given the level of difficulty or ability (Yu, 2017). This creates a sigmoid graph known as an item characteristic curve (ICC) where, at a theta coefficient of 0, there is a 50% probability of obtaining a correct answer (Yu, 2017).

Knowing that theta can be representative of either difficulty or ability allows researchers to map both these constructs along the same plane in an 'item-person' map, commonly known as a Wright map (Bond & Fox, 2015). Wright maps are aligned along the logit scale which is drawn down the middle of the figure and the placement of persons and items is done according to the corresponding theta values (estimated difficulty or estimated ability) (Bond & Fox, 2015). Due to the interval nature of the logit scale, distances between points on the scale are equal at 1 logit per interval (Bond & Fox, 2015). Generally, persons mapped at higher thetas have higher levels of ability and items mapped at higher thetas have greater levels of difficulty. Wright maps present a quick way to determine if the item difficulties are fairly matched to the examinees' abilities as fair tests would show similar-shaped distributions on either side of the logit scale.

In order to examine how well the data fits to the model, researchers often examine the degree of item misfit, which can be represented as the infit mean-square or the outfit mean-square (Yu, 2017). Infit mean-squares are seen as information-weighted as they are more sensitive to items that provide responses that are accurate and reflective of the

person, not that which is beyond the person's ability as seen in outfit mean-squares (Linacre, 2002). In other words, infit mean-squares place more weightage on performance of persons whose ability is more closely matched to the item's difficulty, thus providing insights that are more sensitive to the performance of the item itself (Bond & Fox, 2015).

Mean-square values are just a measure of randomness or distortion within the model with 1.0 as the expected value (Linacre, 2002). Mean-squares that are reported in the statistical output tend to be unstandardized; hence, they are often interpreted as the overall degree of item misfit (Bond & Fox, 2015). Standardized fit statistics, however, are an indicator of the likelihood of the amount of misfit measured in the mean-square values (Bond & Fox, 2015).

Examination of infit mean-squares has often been used to determine which items show best fit to the model and remove less fitting items (Wright & Linacre, 1994). This method is often the method of choice for item reduction with scales developed using IRT (Wright & Linacre, 1994). Adams and Khoo (1996) noted that outfit statistics were more susceptible to distortion with outlying examinees and suggested that infit statistics would be more suitable in analysing items as these estimations were more robust. When deciding optimal cut-off points for mean-square values, Wright and Linacre (1994) find the range between 0.5 to 1.5 as being useful in providing information about the accuracy of measurement. Mean-square values of below 1.0 are considered an indicator of item overfit whereupon the data obtained is more predictable than should be expected, and values over 1.0 are indicative of item underfit which suggests the data is more random than the model expects (Wright & Linacre, 1994). A reasonable range of mean-square values for high-stakes multiple-choice assessments as suggested by Wright and Linacre (1994) lies between 0.8 to 1.2 to depict optimal fit of data to model.

3.7.3 Distractor analysis

In promoting accurate measurement, Siroky and Di Leonardi (2015) offer six tips for refining test items, one of which relates to analysing test items based on test results. One aspect for the test developer to consider would be distractor analysis (Siroky & Di Leonardi, 2015). This is echoed in Haladyna (2004) where it was posited that a relationship exists between a candidate's choice of test score and their total score on the test. Haladyna (2004) proposed several reasons for conducting a distractor analysis. Firstly, running a distractor analysis would help to optimise the number of distractors for any given test item (Haladyna, 2004). Only the best distractors would be retained or faulty distractors can be identified for further refinement. In addition, distractor analysis provides information regarding the performance of a test item thus guiding the researcher in making decisions about retaining, rejecting, or revising items (Haladyna, 2004). Finally, distractor analysis can serve as a diagnostic tool to identify why an item is not performing as expected (Haladyna, 2004).

Generally, distractor analysis allows the researcher to observe a rough estimate of the test taker's ability based on their choice of distractor (Irvin, Alonzo, Lai, Park, & Tindal, 2012). It is expected that test takers with high levels of ability will be more likely to select the correct answer and distractors will be selected by those of lower ability (Irvin et al., 2012). This is related to the effectiveness of a distractor where DiBattista and Kurzawa (2011) stated that in order for an item to effectively differentiate test takers of high versus low ability, those who obtain higher test scores are expected to select the distractor less often than the correct answer as compared to those who obtain lower test scores. In this manner, the effectiveness of a distractor is measured in terms of its correlation with the test taker's total test score, or the point-biserial correlation (Haladyna, 2004) where the value of the correlation should be negative (DiBattista & Kurzawa, 2011). The higher the test taker's score on the test, the less likely they would be to choose

the distractor as the correct answer thus resulting in a negative correlation. Distractors of poor quality show positive or no correlations with the total test score and thus, are described as not functioning well as they are distracting to test takers of higher ability (DiBattista & Kurzawa, 2011).

In addition, for a distractor to be effective, it should be plausible and therefore attractive to test takers of lower ability who are assumed to not know the correct answer (Siroky & Di Leonardi, 2015). It is essential that the distractor is selected by some test takers (Irvin et al., 2012) because this contributes to the discriminatory power of the distractor (Haladyna, 2004). An obviously wrong distractor will not be effective in luring low ability test takers away from the correct answer (Haladyna, 2004) and increases the chances of low ability test takers obtaining the correct answer by guessing. This is because the probability of selecting the correct answer by chance is greatly increased when one option is easily eliminated in the test takers' decision-making process.

3.8 Pilot Study

A pilot study was conducted with 149 first-year undergraduate students from the Department of Psychology in the said public university. Participants had a mean age of 20.14 (SD = 1.03) and their ages ranged between 18-24 years old. The output tables may be referred to in appendix B.

3.8.1 Sample of pilot study

Convenience sampling was used due to time and accessibility constraints. Participants were only recruited from the Department of Psychology due to the ease of administering the scale through their designated portal for experimental recruitment. Although limited in generalizability (Bordens & Abbott, 2011), this method of sampling allowed for sampling of a subset of the overall sample in preparation for the actual data collection process. In addition, the statistical measures being used to analyse the psychometric properties of the sample are known to be sample-independent (Yu, 2017);

hence, the lack of great variation in the sample could be accommodated without greatly affecting the overall results. Bordens and Abbott (2011) noted that although much research has been carried out on social science undergraduates, little differences are found between students from various courses at the undergraduate level in applicability of results. Therefore, this method of sampling is suitable based on the intended population of study.

3.8.2 Design of pilot study

The pilot study utilized a cross-sectional survey design similar to the actual study where the data was collected in a one-shot administration (Sedgwick, 2014) across various races, ages, and genders. This allowed the researcher to collect data from a 'cross-section' of the population such that the sample would be more representative of the overall population. Participants signed up for the study via the online portal for experiments known as iPsy. In exchange for their participation, they were awarded 0.5% extra credit that could be allocated to their subjects to increase their grades. Links were distributed to participants via email in batches of 30 participants per day over 5 days. Out of 150 participants who signed up for the study, only one participant failed to respond despite participants being reminded a week ahead of the cut-off time for participation.

3.8.3 Results of pilot study

Construct validity of the items was determined using confirmatory factor analysis in Mplus (Version 7.0; Muthén & Muthén, 2012). Mplus was chosen to run the factor analysis due to the ability of the software to provide model estimation for binary items based on tetrachoric correlations (Muthén & Muthén, 2012). Goodness-of-fit for the six-factor model of visual-spatial aptitude was determined using three fit indices, namely the chi-square goodness-of-fit statistic, the Steiger-Lind Root Mean Square Error of Approximation (RMSEA), and the Comparative Fit Index (CFI).

The chi-square test presents one of the oldest methods of determining goodness of fit. Good model fit is determined at $p>0.05$ (Hooper et al., 2008). The RMSEA is a popular index of model fit due its parsimonious nature (Hooper et al., 2008). Good model fit is determined at RMSEA<0.06 (Hu & Bentler, 1999, as cited in Hooper et al., 2008). A probability estimate is also given in Mplus similar to PCLOSE which provides a test of the null hypothesis for the RMSEA where $p>0.05$ allows the conclusion that the model fit is 'close' (Kenny, 2015). The CFI is an adjusted form of the Normed Fit Index (NFI) which compares the chi-square statistic of the model of that with the null model (Bentler & Bonnett, 1980, as cited in Hooper et al., 2008). The CFI is adjusted for sample size constraints and good fit is indicated by CFI>0.95 (Hu & Bentler, 1999, as cited in Hooper et al., 2008).

The chi-square test showed good model fit ($\chi^2(390)=401.416$, $p=.334$). This was supported by RMSEA that showed good model fit (RMSEA=.014, 90% CI [0.000, 0.033]) and was supported by a probability of RMSEA<.05 of 1.00 indicating that model fit was 'close' (Kenny, 2015). CFI indicated good model fit as well (CFI=.982). Based on output from the fit indices, the six-factor model of visual-spatial aptitude is considered to be a good representation of the latent constructs. Please refer to appendix C for full Mplus output.

Regression coefficients were examined to determine the extent to which each item loads significantly within the model. Standardised regression coefficients were analysed to ensure that all items were accounted for. Table 3.2 shows the estimates for all items.

Table 3.2

Standardised Regression Coefficients for Items in MAT-D(VS)

Item No.	Item Name	Estimate	Standard Error	Two-tailed p-value
1	FGP1	0.178	0.139	0.200
2	FGP2	0.489	0.127	0.000**
3	FGP3	-0.210	0.145	0.147
4	FGP4	0.644	0.130	0.000**
5	FGP5	0.272	0.137	0.048*

6	OA1	0.475	0.150	0.002*
7	OA2	0.086	0.141	0.539
8	OA3	-0.061	0.130	0.642
9	OA4	0.248	0.143	0.082
10	OA5	1.181	0.300	0.000**
11	PS1	0.270	0.112	0.016*
12	PS2	0.545	0.109	0.000**
13	PS3	0.594	0.121	0.000**
14	PS4	0.080	0.135	0.555
15	PS5	-0.194	0.132	0.142
16	SO1	0.464	0.100	0.000**
17	SO2	0.652	0.091	0.000**
18	SO3	0.650	0.101	0.000**
19	SO4	0.581	0.107	0.000**
20	SO5	0.703	0.097	0.000**
21	SV1	0.394	0.105	0.000**
22	SV2	-0.594	0.098	0.000**
23	SV3	0.677	0.089	0.000**
24	SV4	0.799	0.085	0.000**
25	SV5	0.062	0.115	0.592
26	VD1	0.659	0.079	0.000**
27	VD2	0.921	0.049	0.000**
28	VD3	0.925	0.054	0.000**
29	VD4	0.331	0.112	0.003*
30	VD5	0.572	0.101	0.000**

Note. *. $p < .05$. **. $p <. 001$.

A total of 22 items loaded significantly within the model with $p<.05$. One item showed significant but negative factor loading (item SV2). The items were retained for subsequent analysis before determining if they should be removed. The R-squared value for items that loaded significantly was examined to determine the variance explained by each item and thus, determine its contribution to the measurement model. Table 3.3 displays the squared multiple correlations for items that fit significantly to the model.

Table 3.3

Squared Multiple Correlations for Items in MAT-D(VS)

Item No.	Item Name	Estimate	Standard Error	Two-tailed p-value
2	FGP2	0.239	0.124	0.055
4	FGP4	0.415	0.168	0.013*
5	FGP5	0.074	0.075	0.323
6	OA1	0.226	0.143	0.114
10	OA5	Undefined	Undefined	-
11	PS1	0.073	0.060	0.227
12	PS2	0.297	0.119	0.012*
13	PS3	0.353	0.143	0.014*

16	SO1	0.215	0.093	0.021*
17	SO2	0.426	0.119	0.000**
18	SO3	0.423	0.132	0.001*
19	SO4	0.337	0.124	0.006*
20	SO5	0.494	0.136	0.000**
21	SV1	0.156	0.083	0.059
22	SV2	0.353	0.117	0.002*
23	SV3	0.458	0.120	0.000**
24	SV4	0.638	0.135	0.000**
26	VD1	0.434	0.104	0.000**
27	VD2	0.848	0.091	0.000**
28	VD3	0.856	0.100	0.000**
29	VD4	0.109	0.074	0.138
30	VD5	0.327	0.116	0.005*

Note. *. $p < .05$. **. $p <. 001$.

A total of 15 items contributed significantly to the variance explained of the constructs at $p<.05$. SV2 still contributed significantly to the underlying construct despite the earlier negative factor loading. It was retained for further distractor analysis to determine qualities of the item that could be amended. Items that did not fit to the model structure and those that did not contribute significantly to the constructs were removed from the scale, and new items were generated to replace them. Therefore, a total of 15 items as listed in table 3.4 were retained for distractor analysis.

Table 3.4

Items Retained After Confirmatory Factor Analysis

FGP	PS	SO	SV	VD
FGP4	PS2	SO1	SV2	VD1
	PS3	SO2	SV3	VD2
		SO3	SV4	VD3
		SO4		VD5
		SO5		

Correlations between the constructs were examined to ensure that all were significantly related to the underlying construct of visual-spatial aptitude. The correlations between the constructs are shown in table 3.5 below. Standardized correlation coefficients were examined, similar to the above. All constructs correlate significantly with strength of correlations ranging from 0.700 to 0.909 except for between

all constructs and OA. Thus, items in the construct OA will be completely removed and new ones generated. The correlation between PS and SV at 1.124 shows that the two constructs may be extremely correlated and multicollinearity may exist (Jöreskog, 1999). However, as many PS items are due to be changed, the subsequent outcome may be altered.

Table 3.5

Correlations between Constructs within the MAT-D(VS)

	FGP	OA	PS	SV	SO	VD
FGP	-					
OA	0.227	-				
PS	0.846**	0.045	-			
SV	0.829**	0.069	1.124**	-		
SO	0.700**	-0.015	0.828**	0.909**	-	
VD	0.753**	-0.144	0.852**	0.724**	0.730**	-

Note. **. $p < .001$.

To determine the reliability of the scale, a Rasch analysis was conducted using Quest (Version 2.1; Adams & Khoo, 1996). Quest was chosen to run the Rasch analysis due to the features of the program which allows outlying persons to be manually removed to ensure proper model fit for the distractor analysis to be conducted. Three reliability measures were produced; item reliability, person reliability, and internal consistency. In addition, infit mean square values were analysed to determine the goodness-of-fit for each item and person within the probabilistic model.

The first run of Quest showed that all items fit to the model with infit mean square values ranging between 0.77 to 1.22 except VD2 (Infit MNSQ=0.76). 37 cases with infit values outside the accepted range of 0.77 to 1.30 were removed before the second run of analysis was conducted. With the removal of 37 cases, all items fit to model with infit between 0.78 to 1.15. A third run of analysis was required after removing six more cases that did not fit to model leaving all items fitting to model with infit between 0.79 to 1.15.

Item reliability was high (item reliability=0.94) and person reliability was moderate (person reliability=0.69). Internal consistency was also moderate (α=0.68). Though the coefficient of internal consistency was lower than the acceptable range of 0.70 to 0.90 suggested by Tavakol and Dennick (2011), it may have been due to the items identified in the CFA that did not load within the model. As such, the internal consistency coefficient may be different with appropriate changes to the items.

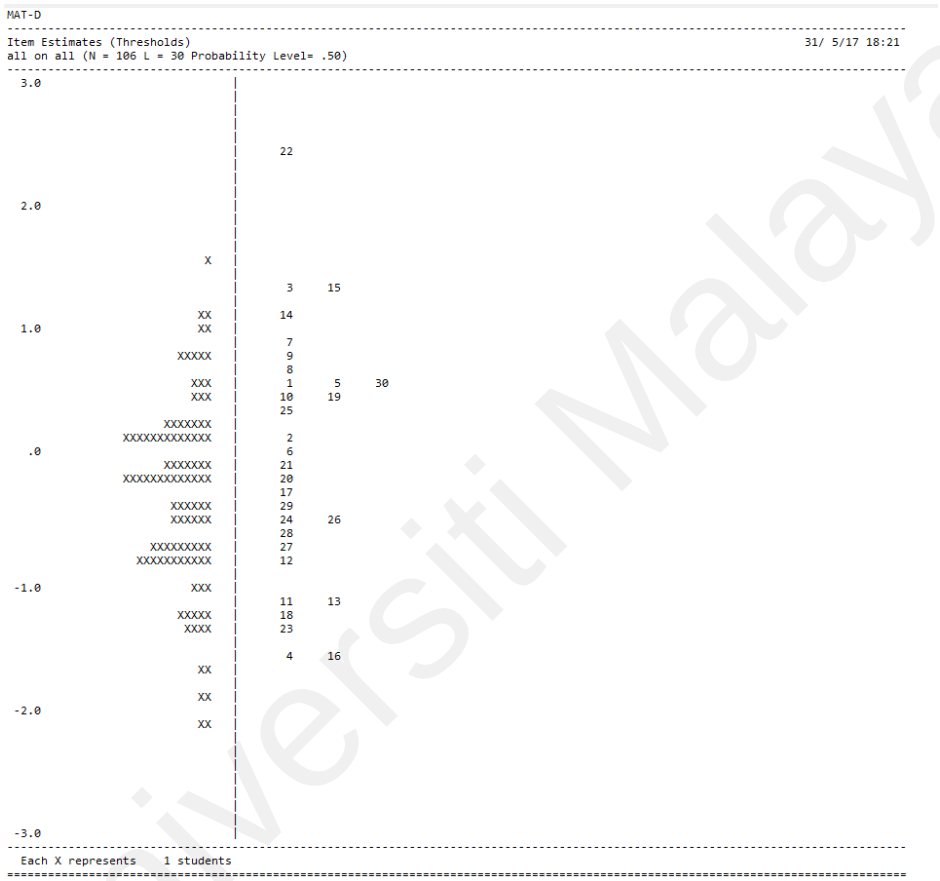

Figure 3.2. Wright map for MAT-D(VS)

Students scored a mean of 13.39 out of 30 with a standard deviation of 4.28. The Wright map as shown in Figure 3.1 shows that most items are fairly matched to the respondents' ability with the exception of SV2 which seems to be measuring extremely high ability. Table 3.6 summarizes fit values and thresholds for all items. 'Thresholds' represent the level of item difficulty with corresponding infit and outfit mean-squares. Infit and outfit t-statistics represent the standardized fit statistics after application of the

Wilson-Hilferty transformation (Adams & Khoo, 1996). Full Rasch model output from Quest can be referred to in appendix D.

Table 3.6

Threshold Values, Infit, and Outfit Statistics for MAT-D(VS)

Item No.	Item Name	Thresholds	Infit Mean Squared (INFIT MNSQ)	Outfit Mean Squared (OUTFIT MNSQ)	Infit t-statistic	Outfit t-statistic
1	FGP1	0.58	1.08	1.12	0.9	0.7
2	FGP2	0.22	0.93	0.97	-1.0	-0.2
3	FGP3	1.41	1.11	1.32	0.7	1.2
4	FGP4	-1.56	0.92	0.87	-0.6	-0.6
5	FGP5	0.58	0.96	1.11	-0.4	0.7
6	OA1	0.00	1.12	1.15	1.8	1.2
7	OA2	0.93	0.99	0.94	-0.1	-0.2
8	OA3	0.67	1.11	1.12	1.1	0.7
9	OA4	0.77	1.11	1.16	1.0	0.9
10	OA5	0.53	1.05	1.13	0.6	0.8
11	PS1	-1.10	1.15	1.25	1.6	1.5
12	PS2	-0.79	1.01	1.02	0.2	0.2
13	PS3	-1.15	0.88	0.80	-1.3	-1.3
14	PS4	1.10	1.13	1.39	1.0	1.7
15	PS5	1.41	1.03	1.22	0.3	0.9
16	SO1	-1.56	1.08	1.06	0.7	0.4
17	SO2	-0.28	0.98	0.96	-0.3	-0.3
18	SO3	-1.24	0.91	0.89	-0.9	-0.6
19	SO4	0.44	0.97	0.93	-0.4	-0.4
20	SO5	-0.20	0.93	0.92	-1.1	-0.6
21	SV1	-0.08	1.02	1.03	0.3	0.3
22	SV2	2.50	1.10	2.07	0.4	1.9
23	SV3	-1.40	0.89	0.84	-1.0	-0.9
24	SV4	-0.53	0.85	0.83	-2.3	-1.4
25	SV5	0.35	1.06	1.10	0.8	0.7
26	VD1	-0.45	0.91	0.90	-1.4	-0.8
27	VD2	-0.70	0.79	0.74	-3.1	-2.1
28	VD3	-0.62	0.79	0.75	-3.2	-2.1
29	VD4	-0.41	1.12	1.14	1.8	1.1
30	VD5	0.58	0.97	0.92	-0.3	-0.4

The Rasch analysis also provided information regarding the quality of items in terms of their fit to the Rasch model through examination of the infit mean-squares and the perceived level of difficulty based on the threshold values. Item difficulty is represented as 'Thresholds' in Quest output and the corresponding standard error is

reported in 'Error'. Adams & Khoo (1996) have noted that the 'thresholds' in the output of Quest are similar to that of Thurstonian thresholds. In addition, standard error is generally considered to be a measure of item precision where items with larger standard error values are less precise in their measurements (Bond & Fox, 2015). The following section will discuss the quality of the 15 items retained from the factor analysis from these two aspects.

All items showed good model fit as evidenced by infit mean-square values between 0.79 to 1.15 (please refer to highlighted rows in table 3.6 for specific item values). Perceived difficulty of the items ranged from 'very easy' to 'difficult'. These classifications were based on the value of the threshold (logit). Table 3.7 provides the difficulty classification of the 15 retained items with corresponding threshold values and standard error values.

Table 3.7

Thresholds, standard error values, and perceived difficulty for 15 retained items

Item No.	Item Name	Thresholds	Standard Error	Perceived Difficulty
4	FGP4	-1.56	0.24	Very easy
12	PS2	-0.79	0.21	Easy
13	PS3	-1.15	0.22	Very easy
16	SO1	-1.56	0.24	Very easy
17	SO2	-0.28	0.21	Easy
18	SO3	-1.24	0.23	Very easy
19	SO4	0.44	0.22	Moderate
20	SO5	-0.20	0.21	Easy
22	SV2	2.50	0.40	High
23	SV3	-1.40	0.23	Very easy
24	SV4	-0.53	0.21	Easy
26	VD1	-0.45	0.21	Easy
27	VD2	-0.70	0.21	Easy
28	VD3	-0.62	0.21	Easy
30	VD5	0.58	0.22	Moderate

Distractor analysis was conducted on the 15 items retained from the factor analysis to determine the qualities of distractors and to decide if amendments to item distractors was necessary. The measure of distractor effectiveness is seen in the point-

biserial value indicating the product-moment correlation of the chosen response and the raw total test score (Adams & Khoo, 1996). In interpreting the distractor analysis, point-biserial correlations for the options of each item were examined. Results of the distractor analysis are shown in table 3.8. Key options (the correct answer to the item) are shown in the highlighted cells with the bold values.

Table 3.8

Point-biserial values for each distractor by item in MAT-D(VS)

Item	A	B	C	D
Item 4: FGP4	**.40**	-.28	-.16	-.19
Item 12: PS2	-.13	-.16	-.18	**.31**
Item 13: PS3	-.15	**.48**	-.43	-.17
Item 16: SO1	-.19	**.20**	-.10	.00
Item 17: SO2	-.34	-.15	**.37**	-.04
Item 18: SO3	**.41**	-.23	-.08	-.30
Item 19: SO4	-.17	-.14	-.12	**.36**
Item 20: SO5	-.19	-.19	**.43**	-.20
Item 22: SV2	-.26	**-.09**	.37	-.22
Item 23: SV3	**.44**	-.34	-.20	-.15
Item 24: SV4	-.07	**.52**	-.39	-.29
Item 26: VD1	-.21	**.45**	-.18	-.28
Item 27: VD2	-.42	-.19	**.60**	-.24
Item 28: VD3	**.61**	-.32	-.35	-.18
Item 30: VD5	-.06	-.11	-.23	**.36**

From the distractor analysis, it was seen that all distractors for all the selected items were functioning well with point-biserial values that were negative for distractors and positive for the key. With the exception of SO1 (item 16) and SV2 (item 22), all the items were retained without change. For SO1 (item 16), distractor D(r_{pbD}=0.00) was problematic although distractors A(r_{pbA}=-0.19) and C(r_{pbC}=-0.10) were functioning well. The distractor confused students of higher ability. Hence, the item was retained with modification to distractor D. For SV2 (item 22), distractor C was problematic (r_{pb}=0.37) as well as the key B (r_{pb}=-0.09). This could have been a factor that contributed to the perceived 'high' level of difficulty and could also be the reason for the negative factor loading obtained in the confirmatory factor analysis. Re-examination of the item in the

scale revealed that a mistake was made in digitization; hence, the stem showed the incorrect rotation. The item was rectified and maintained for re-testing.

In summary, from the results of the factor analysis, Rasch analysis, and distractor analysis, a total of 15 items were removed and replaced with new items. Of the remaining 15 items, one item (SO1) required amendments to a distractor, one item was digitized wrongly and subsequently rectified, and 13 items were retained with no changes. New items in FGP were amended to add curved lines within the shapes as it was found that curves may aid in improving viability of test items due to the grouping according to continuity (Wagemans et al., 2012). The revised MAT-D(VS) maintained the format of five items per subscale for the six constructs.

3.9 Summary

This chapter presented a brief overview of the methodological framework of the current study. A cross-sectional survey design was utilised in administering the 30-item, six-construct MAT-D(VS) to a sample of 203 first-year undergraduate students of a local private university. The test development process was outlined, and results of the pilot study were presented. Based on the output of the confirmatory factor analysis using Mplus, and the Rasch and distractor analysis using Quest, 15 items were removed to be replaced with new items from the initial 30-item scale. Of the 15 remaining, one was retained with changes to the faulty distractor, one was re-digitised following the identification of a digitization mistake, and 13 items were retained with no changes. Amendments will be made to the scale before administration to a new sample. The statistical analyses that will be conducted are confirmatory factor analysis using Mplus, and Rasch analysis and distractor analysis using Quest.

CHAPTER 4

FINDINGS

4.1 Introduction

The results of the confirmatory factor analysis, Rasch analysis, and distractor analysis are presented in this chapter. Model fit of the MAT-D(VS) is discussed in relation to fit indices and factor loadings of each individual item are examined. The second part of the chapter presents the findings of the Rasch analysis. Degree of fit to the model based on infit and outfit statistics is discussed prior to presentation of the findings from the distractor analysis. Recommendations for item amendments or item removal will be made to determine next steps in the development of the MAT-D(VS).

4.2 Confirmatory Factor Analysis

In determining the construct validity of the MAT-D(VS), data was input into Mplus (Version 7.0; Muthén & Muthén, 2012) for confirmatory factor analysis to be conducted. A six-factor model of visual-spatial aptitude was tested as per the previous versions of the MAT (Mamauag et al., 2016). As in the pilot study, the fit indices that were used to determine model fit were the chi-square goodness-of-fit statistic, the Steiger-Lind Root Mean Square Error of Approximation (RMSEA), and the Comparative Fit Index (CFI). Good model fit is determined at the value of $p>0.05$ for the chi-square test (Hooper et al., 2008), RMSEA<0.06 (Hu & Bentler, 1999, as cited in Hooper et al., 2008) and PCLOSE $p>0.05$ (Kenny, 2015), and CFI>0.95 (Hu & Bentler, 1999, as cited in Hooper et al., 2008). Confidence intervals for the RMSEA were provided at 90% CI by Muthén & Muthén (2012) based on the recommendations of Kenny (2015) (Muthén, 2012).

Good model fit was obtained according to these three indices. The results of the chi-square test showed good model fit ($\chi^2(390)=384.100$, $p=.575$). This was further supported by the RMSEA (RMSEA<.001, 90% CI [0.000, 0.023], PCLOSE=1.00) and

the CFI indices (CFI=1.000). Based on these indices, the six-factor structure of visual-spatial aptitude represents a good structural representation of the latent construct. Full Mplus output is appended in appendix G. To determine the factors loadings of the individual items within the model, standardized regression coefficients were examined so that loadings of each item were accounted for. Estimates, standard error, and significance values for all items are shown in Table 4.1.

Table 4.1

Standardised Regression Coefficients for Items in MAT-D(VS) [Revised]

Item No.	Item Name	Estimate	Standard Error	Two-tailed p-value
1	FGP1	0.502	0.094	0.000**
2	FGP2	0.479	0.106	0.000**
3	FGP3	0.513	0.099	0.000**
4	FGP4	0.829	0.079	0.000**
5	FGP5	0.525	0.093	0.000**
6	OA1	0.037	0.136	0.784
7	OA2	0.855	0.230	0.000**
8	OA3	0.126	0.169	0.456
9	OA4	-0.055	0.145	0.707
10	OA5	0.373	0.131	0.004*
11	PS1	0.100	0.106	0.343
12	PS2	0.432	0.089	0.000**
13	PS3	0.532	0.097	0.000**
14	PS4	0.644	0.086	0.000**
15	PS5	0.450	0.104	0.000**
16	SO1	0.472	0.114	0.000**
17	SO2	0.475	0.093	0.000**
18	SO3	0.439	0.110	0.000**
19	SO4	-0.217	0.104	0.037*
20	SO5	0.412	0.098	0.000**
21	SV1	0.375	0.098	0.000**
22	SV2	0.109	0.108	0.316
23	SV3	0.629	0.096	0.000**
24	SV4	0.610	0.088	0.000**
25	SV5	0.487	0.089	0.000**
26	VD1	0.747	0.061	0.000**
27	VD2	0.628	0.079	0.000**
28	VD3	0.837	0.049	0.000**
29	VD4	0.809	0.058	0.000**
30	VD5	0.845	0.057	0.000**

Note. *. $p < .05$. **. $p <. 001$.

25 items loaded significantly within the model at $p<.05$. Of these 25 items, 23 loaded significantly at $p<.001$. One item significantly loaded within the model but showed negative factor loadings (item SO4); the item was retained for subsequent analysis of its properties. To determine the contribution of individual items to the measurement model, the R-squared value was examined for items that had significant factor loadings in order to determine the variance explained by each item. The squared multiple correlation values for each significant item is displayed in table 4.2.

Table 4.2

Squared Multiple Correlations for Items in MAT-D(VS) [Revised]

Item No.	Item Name	Estimate	Standard Error	Two-tailed p-value
1	FGP1	0.252	0.094	0.007*
2	FGP2	0.230	0.101	0.023*
3	FGP3	0.263	0.102	0.010*
4	FGP4	0.687	0.132	0.000**
5	FGP5	0.276	0.097	0.005*
7	OA2	0.730	0.392	0.063
10	OA5	0.139	0.097	0.153
12	PS2	0.186	0.076	0.015*
13	PS3	0.284	0.104	0.006*
14	PS4	0.415	0.110	0.000**
15	PS5	0.202	0.094	0.031*
16	SO1	0.222	0.108	0.039*
17	SO2	0.226	0.088	0.011*
18	SO3	0.193	0.096	0.045*
19	SO4	0.047	0.045	0.297
20	SO5	0.169	0.081	0.036*
21	SV1	0.140	0.073	0.056
23	SV3	0.395	0.120	0.001*
24	SV4	0.373	0.107	0.001*
25	SV5	0.238	0.087	0.006*
26	VD1	0.557	0.091	0.000**
27	VD2	0.394	0.099	0.000**
28	VD3	0.700	0.082	0.000**
29	VD4	0.654	0.094	0.000**
30	VD5	0.714	0.097	0.000**

Note. *. $p < .05$. **. $p <. 001$.

21 items provided significant contributions to the variance explained of the constructs at $p<.05$. Item SO4 that loaded negatively within the model was not a significant contributor to the variance explained of the underlying construct. Hence, it

was disregarded for further analysis. Items that did not load within the model nor contribute significantly to the variance explained were disregarded for further analysis and were to be removed in the next revision of the scale. Therefore, a total of 21 items as shown in table 4.3 were retained for Rasch and distractor analysis.

Table 4.3

Items Retained after Confirmatory Factor Analysis on MAT-D(VS) [Revised]

FGP	PS	SO	SV	VD
FGP1	PS2	SO1	SV3	VD1
FGP2	PS3	SO2	SV4	VD2
FGP3	PS4	SO3	SV5	VD3
FGP4	PS5	SO5		VD4
FGP5				VD5

Correlations between the constructs of visual-spatial aptitude were examined to determine the relationship to the main underlying construct. Table 4.4 presents the values of the correlations between the constructs. As per the above, the standardized correlation coefficients were analysed. Most constructs show significant moderate to high correlations ranging between 0.674 to 0.993 with the exception of OA which shows a low of 0.388 to a high of 0.881. In addition, the correlations between the constructs and OA show weaker statistical significance than among the others. The correlation between FGP and SO demonstrates a high degree of correlation at r=1.001 which may be indicative of multicollinearity (Jöreskog, 1999). The implications of these findings will be further discussed in chapter five.

Table 4.4

Correlations between Constructs within the MAT-D(VS) [Revised]

	FGP	OA	PS	SV	SO	VD
FGP	-					
OA	0.388*	-				
PS	0.854**	0.673*	-			
SV	0.674**	0.881*	0.917**	-		
SO	1.001**	0.578*	0.993**	0.929**	-	
VD	0.783**	0.501*	0.818**	0.790**	0.742**	-

Note. *. $p < .05$. **. $p <. 001$.

4.3 Rasch Analysis

Rasch analysis using Quest (Version 2.1; Adams & Khoo, 1996) was run to determine reliability of the scale. The first run of Quest showed good fit of items to the model with infit mean-square values ranging from 0.81 to 1.26 with the exception of SO4 (Infit MNSQ=1.32). A total of 47 cases were removed before the second run of analysis to minimise the influence of misfitting persons on the psychometric properties of the items. In the second run of analysis, all items fit to the model with infit mean-square values ranging from 0.80 to 1.27. A further three cases were removed before the final analysis was conducted on the remaining 153 cases. In the third run of analysis, all cases fit to model (Infit MNSQ between 0.77 and 1.28) and all items fit to model (Infit MNSQ between 0.79 and 1.26). Item reliability was high with a reliability estimate of 0.97 and person reliability was fair at 0.78. Internal consistency was also fair (α=0.77). The mean score of 153 valid cases in the model was 16.72 out of 30 (SD=4.87).

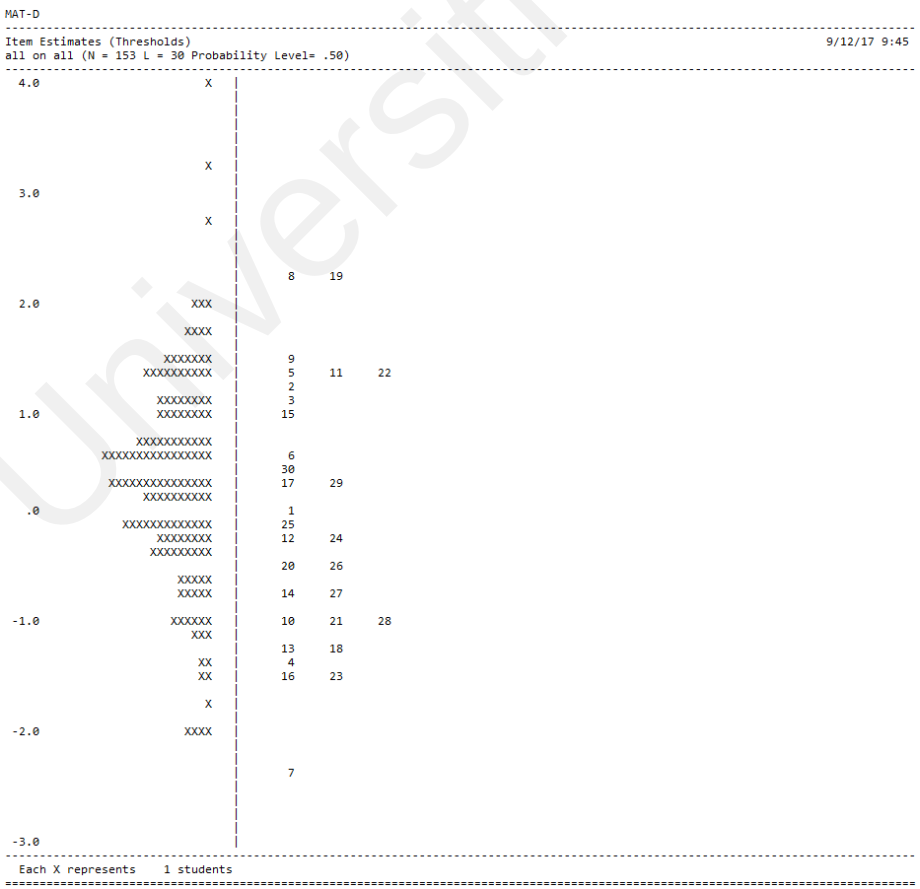

Figure 4.1. Wright map for MAT-D(VS) [Revised]

Figure 4.1 shows the Wright map from the analysis. From the Wright map, it can be seen that the test items are rather fairly matched to the ability of majority of examinees with the exception of item OA2 which is too far below the general ability measured. Table 4.5 presents the fit values, t-statistics and threshold values for all items. Full output of the Rasch model can be seen in appendix H.

Table 4.5

Threshold Values, Infit, and Outfit Statistics for MAT-D(VS) [Revised]

Item No.	Item Name	Thresholds	Infit Mean Squared (INFIT MNSQ)	Outfit Mean Squared (OUTFIT MNSQ)	Infit t-statistic	Outfit t-statistic
1	FGP1	0.03	0.99	1.00	-0.1	0.1
2	FGP2	1.14	0.97	1.03	-0.3	0.3
3	FGP3	1.10	0.98	1.10	-0.3	0.7
4	FGP4	-1.34	0.87	0.67	-1.0	-1.5
5	FGP5	1.31	1.02	0.99	0.3	0.0
6	OA1	0.60	1.25	1.47	3.9	3.3
7	OA2	-2.42	0.95	0.93	-0.1	0.0
8	OA3	2.21	0.98	1.10	-0.1	0.5
9	OA4	1.46	1.13	1.38	1.3	2.0
10	OA5	-0.91	1.11	1.01	1.1	0.1
11	PS1	1.35	1.09	1.12	1.1	0.7
12	PS2	-0.25	1.04	1.05	0.6	0.4
13	PS3	-1.24	0.95	0.88	-0.4	-0.5
14	PS4	-0.69	0.89	0.85	-1.3	-0.9
15	PS5	0.97	0.99	1.06	-0.1	-0.4
16	SO1	-1.48	1.00	1.09	0.0	0.4
17	SO2	0.36	1.00	0.95	0.0	-0.3
18	SO3	-1.20	1.05	1.05	0.4	0.3
19	SO4	2.21	1.26	1.44	1.7	1.6
20	SO5	-0.51	1.10	1.18	1.2	1.2
21	SV1	-0.99	1.10	1.10	0.9	0.5
22	SV2	1.31	1.10	1.29	1.1	1.7
23	SV3	-1.43	0.94	0.76	-0.4	-1.0
24	SV4	-0.19	0.91	0.87	-1.4	-0.9
25	SV5	-0.04	1.00	0.96	0.1	-0.3
26	VD1	-0.51	0.90	0.79	-1.3	-1.4
27	VD2	-0.65	0.91	0.85	-1.0	-0.9
28	VD3	-0.91	0.79	0.66	-2.2	-1.9
29	VD4	0.27	0.87	0.82	-2.3	-1.5
30	VD5	0.48	0.84	0.77	-2.8	-1.9

Quality of items was determined based on fit to the Rasch model (infit mean-square values) and perceived difficulty of items. As per Adams and Khoo (1996), the measure of difficulty given as 'Thresholds' with corresponding standard error. The measure of standard error provides information regarding the precision of the item where it is seen that higher standard error indicates an item that is less precise in measuring latent ability (Bond & Fox, 2015). It was seen that the 21 items retained from the confirmatory factor analysis showed good fit within the Rasch model as evidence by infit mean-square values between 0.79 to 1.10 for all items (specific item values can be referred to in table 4.5). Perceived difficulty of the items ranged from 'very easy' to 'difficult', for which classifications were based on thresholds values (logit). Table 4.6 presents the classification of items for perceived difficulty based on the corresponding thresholds and standard error values for the 21 items retained after confirmatory factor analysis. A total of five items fell into the category of 'very easy', eight items in the category of 'easy', five items of 'average' difficulty, and the remaining three items could be considered 'difficult'.

Table 4.6

Thresholds, standard error values, and perceived difficulty for 21 retained items

Item No.	Item Name	Thresholds	Standard Error	Perceived Difficulty
1	FGP1	0.03	0.18	Moderate
2	FGP2	1.14	0.19	Difficult
3	FGP3	1.10	0.19	Difficult
4	FGP4	-1.34	0.22	Very easy
5	FGP5	1.31	0.19	Difficult
12	PS2	-0.25	0.18	Easy
13	PS3	-1.24	0.22	Very easy
14	PS4	-0.69	0.19	Easy
15	PS5	0.97	0.18	Moderate
16	SO1	-1.48	0.23	Very easy
17	SO2	0.36	0.18	Moderate
18	SO3	-1.20	0.21	Very easy
20	SO5	-0.51	0.19	Easy
23	SV3	-1.43	0.23	Very easy
24	SV4	-0.19	0.18	Easy
25	SV5	-0.04	0.18	Easy
26	VD1	-0.51	0.19	Easy

27	VD2	-0.65	0.19	Easy
28	VD3	-0.91	0.20	Easy
29	VD4	0.27	0.18	Moderate
30	VD5	0.48	0.19	Moderate

4.4 Distractor Analysis

The 21 items retained from the results of the confirmatory factor analysis were subject to a distractor analysis to examine the qualities of distractors and to determine necessary amendments for further refinement of the scale. Table 4.7 outlines the findings of the distractor analysis. In determining efficacy of distractors, the point-biserial correlation is examined. This indicates the product-moment correlation between the selected option and the raw total test score (Adams & Khoo, 1996). It is expected that distractors are negatively correlated but the key will show a positive correlation (DiBattista & Kurzawa, 2011). Poorly functioning distractors will show either a positive correlation or no correlation (DiBattista & Kurzawa, 2011).

Table 4.7

Point-biserial values for each distractor by item in MAT-D(VS)[Revised]

Item	A	B	C	D
Item 1: FGP1	-.21	-.25	**.40**	-.10
Item 2: FGP2	**.39**	-.04	-.11	-.33
Item 3: FGP3	.00	-.24	-.20	**.38**
Item 4: FGP4	**.50**	-.30	-.35	-.14
Item 5: FGP5	-.06	-.21	**.34**	-.10
Item 12: PS2	-.14	-.28	-.12	**.34**
Item 13: PS3	-.17	**.39**	-.32	-.20
Item 14: PS4	-.11	-.40	-.23	**.49**
Item 15: PS5	**.38**	-.10	-.35	-.01
Item 16: SO1	-.23	**.31**	-.18	-.07
Item 17: SO2	-.16	-.18	**.40**	-.25
Item 18: SO3	**.29**	-.22	.03	-.21
Item 20: SO5	-.13	-.21	**.28**	-.06
Item 22: SV3	**.42**	-.27	-.21	-.20
Item 23: SV4	-.17	**.49**	-.40	-.14
Item 24: SV5	-.13	-.22	-.31	**.39**
Item 26: VD1	-.35	**.50**	-.23	-.16
Item 27: VD2	-.35	-.16	**.47**	-.21
Item 28: VD3	**.59**	-.29	-.38	-.28
Item 29: VD4	-.24	-.33	**.53**	-.12
Item 30: VD5	-.13	-.23	-.37	**.56**

From the distractor analysis, it could be seen that most items contained well-functioning distractors as evidenced by the positive point-biserial correlation values for the answer key (highlighted cell with values in bold) and negative point-biserial values for the distractor options. 19 of the 21 items were retained without amendments excepting FGP3 (item 3) and SO3 (item 18). For FGP3 (item 3), although distractors B and C showed good functioning, distractor A could be seen as problematic (r_{pbA}=0.00). As such, the item required amendments to distractor A. Similarly with SO3 (item 18), although distractors B (r_{pbB}=-0.22) and D (r_{pbD}=-0.21) were functioning well, distractor C was problematic (r_{pbC}=0.03). SO3 was also retained with amendments to distractor C. Therefore, from the total 21 items retained based on the results of the confirmatory factor analysis, 19 were retained with no amendments necessary and two items (FGP3 and SO3) needed amendments to one distractor each.

4.5 Summary

This chapter presented the results of three analysis that were done to determine the psychometric properties of the MAT-D(VS) subsequent to revision after the pilot study. From the confirmatory factor analysis, better model of the six-factor structure of visual-spatial aptitude was found. A total of 21 items loaded significantly within the model and all constructs showed significant correlations among one another. The Rasch model showed good fit for all items and acceptable values for all reliability indices. Classification of items according to their threshold estimated difficulty levels showed that the difficulty of items ranged from 'very easy' to 'difficult' with most items being considered 'easy'. Distractor analysis on the 21 items retained post-confirmatory factor analysis showed that 19 items could be retained without any necessary amendments, but two items required amendments to one distractor option each. The following chapter will

discuss the results obtained in relation to the earlier research questions and objectives and make suggestions for the next stage of scale development.

CHAPTER 5

DISCUSSION AND CONCLUSION

5.1 Introduction

This chapter will present an in-depth discussion of the results obtained from the study. First, a brief summary of the results from chapter four will be presented. These results will be discussed in relation to the research objectives and questions outlined in chapter one. Parallels with extant literature will be drawn. Theoretical and practical implications of results and future directions of the scale will also be considered. The chapter will end with an overall summary of the study and final conclusions.

5.2 Summary of Findings

The confirmatory factor analysis showed that the model was a good fit with the obtained data according to the chi-square goodness-of-fit test ($\chi^2(390)=384.100$, $p=.575$), Steiger-Lind RMSEA (RMSEA<.001, 90% CI [0.000, 0.023], PCLOSE=1.00) and the Comparative Fit Index (CFI=1.000). An examination of the individual factor loadings showed that 25 items loaded significantly within the model at $p<.05$ with 23 of these items showing significant loadings at $p<.001$. Of the 25 items that loaded significantly, 21 items provided significant contributions to the variance explained within the model. These 21 items were retained to be examined within the Rasch model.

A total of five items were retained from the construct figure-ground perception (FGP), four items from progressive series (PS), four items from spatial orientation (SO), three items from spatial visualisation (SV), and five items from visual discrimination (VD). No items from object assembly (OA) loaded significantly within the model and contributed significantly to the variance explained. All constructs showed significant inter-construct correlations although the correlation of FGP-OA was seen as low at $r=0.388$ and the correlation of FGP-SO was indicative of multicollinearity ($r=1.001$).

From the Rasch analysis, 50 cases were removed before all items showed fit to the Rasch model with infit mean-square values between 0.79 and 1.26. Item reliability was high at 0.97 with acceptable person reliability (0.78) and internal consistency (α=0.77) values. The subsequent distractor analysis showed that two items still required amendments to distractors (items FGP3 and SO3) while 19 items were retained with no required amendments. Overall, based on the threshold (logit) values, it was seen that five items could be considered 'very easy', eight items were considered 'easy', five items were of 'average' difficulty, and three items were 'difficult'.

5.3 Comparison with Pilot Study Results

The findings from the actual study show an improvement in the psychometric properties of the scale when compared to the results of the pilot study. Table 5.1 summarises the changes in the scale from pilot to actual study. From the confirmatory factor analysis, it was seen that model fit improved according to all three indices. More items showed significant factor loadings in the actual study (25 items at $p<.05$) compared to the pilot study (22 items at $p<.05$). In addition, more items showed significant contributions to the variance explained within the model (21 items at $p<.05$ [actual study] compared to 15 items at $p<.05$ [pilot study]). However, based on the factor analysis results, two noteworthy cases were not retained from the pilot study (despite originally showing model fit). Item SO4 loaded negatively within the model of the actual study and then showed no significant contribution to variance explained whereas item SV2 showed no significant factor loading within the model after being amended based on the pilot study results.

The inter-construct correlations showed improvements from the pilot to the actual study where, in the pilot study, OA was not significantly correlated with any of the other constructs and PS-SV showed strong evidence of multicollinearity (r=1.124), the actual study showed significant correlations between all constructs providing evidence of

construct validity (to be further discussed in the following section). However, it could be seen that the FGP-OA correlation was somewhat weak (r=0.388) and there was the suggestion of multicollinearity between FGP and SO (r=1.001).

Within the Rasch model, it could be seen that items from the pilot study showed slightly better fit (Infit MNSQ between 0.79 to 1.15) compared to the actual study (Infit MNSQ between 0.79 to 1.26) but both rounds of analysis showed that items fit within the model as acceptable values are typically between 0.77 to 1.30 (Adams & Khoo, 1996). Reliability values showed improvement from the pilot (Item reliability=0.94, person reliability=0.69, α=0.68) to the actual study (Item reliability=0.97, person reliability=0.78, α=0.77). From the distractor analysis, it was seen that 19 items were retained with no amendments necessary in the actual study compared to 13 items from pilot study.

Table 5.1

Summary of Changes in Psychometric Properties of MAT-D(VS) from Pilot to Actual Study

Analysis	Pilot study	Actual study
Confirmatory factor analysis	χ^2(390)=401.416, p=.334 RMSEA=0.14 90% CI [0.000, 0.033] PCLOSE=1.00 CFI=.982	χ^2(390)=384.100, p=.575 RMSEA<.001 90% CI [0.000, 0.023] PCLOSE=1.00 CFI=1.000
Factor loadings	22 items at p<.05 18 items at p<.001	25 items at p<.05 23 items at p<.001
Contribution to variance explained	15 items at p<.05	21 items at p<.05
Items retained after CFA	FGP4 PS2 SO1 SV2 VD1 PS3 SO2 SV3 VD2 SO3 SV4 VD3 SO4 VD5 SO5	FGP1 PS2 SO1 SV3 VD1 FGP2 PS3 SO2 SV4 VD2 FGP3 PS4 SO3 SV5 VD3 FGP4 PS5 SO5 VD4 FGP5 VD5
Inter-construct correlations	No sig. correlations of OA Multicollinearity of PS-SV (r=1.124)	All sig. correlations Weak correlation FGP-OA (r=0.388)

		Multicollinearity of FGP-SO (r=1.001)
Rasch analysis	43 cases removed Infit MNSQ of items between 0.79 to 1.15 Item reliability 0.94 Person reliability 0.69 Internal consistency 0.68	50 cases removed Infit MNSQ of items between 0.79 to 1.26 Item reliability 0.97 Person reliability 0.78 Internal consistency 0.77
Distractor analysis	13 items retained no change SO1 requires distractor amendments SV2 mistake in digitization	19 items retained no change FGP3 requires distractor amendments SO3 required distractor amendments

5.4 Discussion of Research Objectives and Research Questions

5.4.1 Research objective 1

Revisiting research objective 1, the study was conducted to determine the psychometric properties of the Multiple Aptitude Test – Form D (Visual-Spatial) [MAT-D(VS)]. This put forth the research question of what are the psychometric properties of the MAT-D(VS)? In answering this question, three sub-objectives and sub-questions were formulated. The results of the actual study will be discussed in relation to each of these sub-questions.

5.4.1.1 RQ1.1 To what extent does the MAT-D(VS) exhibit construct validity, such that the items load sufficiently on the underlying factors?

From the study, it was seen that the six-factor structure of visual-spatial aptitude was a good representation of the underlying construct. This was evidenced by the good model fit obtained in the actual study according to the chi-square goodness-of-fit test, Steiger-Lind Root Mean Square Error of Approximation (RMSEA) and Comparative Fit Index (CFI).

The results of the chi-square goodness-of-fit test showed that the model represented the underlying construct well because the value of $p>.05$ indicated that there was no significant difference between the sample covariance matrix as estimated from the available dataset and the model-fitted matrices which are theoretical in nature (Hu & Bentler, 1999). Thus, it could be seen that the data-generated matrix was very similar to the theoretical matrix. Given that the distribution of test scores from the scale were normal (please refer to chapter three, page 60 for normality results and appendix F for full normality output), the chi-square test was an appropriate measure of model fit given its heavy reliance on assumptions of normality (McIntosh, 2006, as cited in Hooper et al., 2008).

The good model fit obtained in the study was further supported by the RMSEA value of less than 0.001 and bolstered by the PCLOSE test which tested the null hypothesis of the RMSEA index. The RMSEA is a popular test of model fit given its favour for parsimony whereupon it will show better fit for the simplest possible model and confidence intervals can be estimated to show the highest probability of the range of possible RMSEA values at 90% CI (Hooper et al., 2008). Hence, the statistical evidence from the RMSEA shows support for the six-factor structure of visual-spatial aptitude as being sufficient to capture the essence of the underlying construct.

Finally, looking at the results from the CFI, the value of 1.00 indicates very good model fit as the range of values that the CFI can achieve is between 0.00 to 1.00 with larger numbers considered to be better (Iacobucci, 2010). The CFI presents an index of incremental fit whereupon the CFI compares the fit between two models; the hypothesized model is fit to the data and then compared against the fit of a baseline model which usually specifies zero correlations between the constructs (Iacobucci, 2010). In this manner, it provides an estimation of relative model fit where it is relative to the fit of the baseline model (Iacobucci, 2010). It also adjusts for parsimony (Iacobucci, 2010) and is,

thus, seen as complementary to the RMSEA (Rigdon, 1996). Generally, researchers favour a strict cut-off point of CFI≥0.95 to prevent acceptance of misspecified models as the CFI is very robust and powerful (Hu & Bentler, 1999; Rigdon, 1996). Hence, CFI=1.00 in tandem with the RMSEA and chi-square results provide strong evidence that the six-factor structure well represents the underlying construct of visual-spatial aptitude.

In determining the usefulness of items within the model, the factor loadings and squared multiple correlation value were examined. It was seen that more than half the initial items showed sufficient factor loadings to support the construct validity of the model. Examination of standardised loading values was in line with recommendations of Chua (2014) to ensure that each and every item could be accounted for within the model. The squared multiple correlation values were used as an indicator of the variance explained by the items within the model (Chua, 2014). Examination of these values was necessary to ensure that only items that contributed significantly to the variance explained by the model would be retained (Abdi, 2007). From these values, a total of 21 items were considered useful in measuring the underlying constructs. However, it was seen that none of the items from OA were retained. Although two OA items (OA2 and OA5) showed significant factor loadings, they were not significantly contributing to the variance of the main construct of OA. As such, no items from OA were seen as providing useful information for measurement in this scale.

However, from examining the inter-construct correlations, it could be seen that all constructs including OA were significantly correlated with one another. Based on Mukaka's (2012) guide for interpreting coefficient size, the coefficients showed a range from moderate to high correlations with the exception of FGP-OA which was classified as low. Table 5.2 indicates the classification of correlation size for contructs.

Table 5.2

Correlation Coefficient Size for Constructs in the MAT-D(VS) according to Mukaka (2012)

	FGP	OA	PS	SV	SO	VD
FGP	-					
OA	Low	-				
PS	High	Moderate	-			
SV	Moderate	High	Very high	-		
SO	***	Moderate	Very high	Very high	-	
VD	High	Moderate	High	High	High	-

Although the OA items in the pilot study revealed that OA was not significantly correlated to the other constructs, the revised items showed that the construct itself is correlated despite the items not being good representations of the constructs. As such, it will be necessary to revise the items in the following cycle of scale development.

The correlation between FGP-SO is marked *** due to the correlation value exceeding 1.00 ($r_{FGP\text{-}SO}$=1.001). The common interpretation of correlation coefficients is that correlations represent a standardized value of covariance and are, hence, bound by the limits of ±1.00 (Field, 2013). However, Jöreskog (1999) argues that the value may exceed 1.00 should the constructs approach singularity; in other words, multicollinearity is indicated as the two concepts are so similar as to be indistinguishable. At the item level, it can be seen that all FGP items loaded significantly within the model and four out of five SO items were significant within the model. Given that the indication of multicollinearity is small (exceeds 1.00 by a value of 0.001), one new item is necessary to be generated for SO, and one item each from FGP (FGP3) and SO (SO3) require amendments to distractors, the probability of multicollinearity in the revised version is reduced. This is similar to the issue in the pilot version of the scale where multicollinearity was also indicated for PS-SV ($r_{PS\text{-}SV}$=1.124) but was no longer indicated once revisions to the scale had been done.

5.4.1.2 RQ1.2 What is the reliability of the MAT-D(VS) in terms of person and item characteristics?

From the study, three indices of reliability were generated to determine the reliability of the scale. Internal consistency was reported as a measure of the degree to which items within a scale purportedly measure the same underlying construct (Tavakol & Dennick, 2011). It is typically seen as a measure of inter-relatedness of items and reported in the form of Cronbach's alpha (α; Tavakol & Dennick, 2011). The authors recommend that scales have alpha values between 0.70 to 0.90 as values higher than 0.90 have occasionally been seen as indicators of item redundancy (Tavakol & Dennick, 2011). As such, the current scale shows a good degree of consistency among items (α=0.77).

The second index of reliability generated is person reliability which can be considered as similar to Cronbach's alpha (Linacre, n.d.) but is better described by Bond and Fox (2015) where it is defined as the consistency of responses by the same set of individuals if they are administered a set of parallel test items. Thus, person reliability serves as a measure of confidence that the researcher can have regarding the reliability of the ability estimates provided by the sample (Bond & Fox, 2015). In the same manner of internal consistency, person reliability can be subject to influences of test length where too few items would result in lower reliability estimates (Bond & Fox, 2015). Though no cut-off score is given as to what constitutes 'high' reliability, given the parallel to Cronbach's alpha, it can be assumed that the same distinctions apply (Linacre, n.d.). Hence, reliability of the scale estimates is further supported by the person reliability value of 0.78 for the MAT-D(VS).

Finally, reliability is also given as item reliability which is seen as the degree to which estimates of item difficulty will remain similar when administered to a different sample (Bond & Fox, 2015). In other words, will an item reported to measure high ability

really show the same difficulty level when administered to another sample of individuals (Linacre, n.d.). Item difficulty tends to be affected by the spread of ability levels within the sample where a sample with a narrow range of ability (very high achieving students/very low achieving students) would result in lower item reliability (Bond & Fox, 2015). Given the current sample used for scale development, it can be seen that the spread of ability in students is sufficient to yield a high item reliability value of 0.97. Thus, the three indicators taken together show statistical support for the reliability of the items on the scale and the appropriateness of the sample used for development.

5.4.1.3 RQ1.3 What is the item fit of the MAT-D(VS) in terms of person and item characteristics?

The first run of Quest analysis showed that most items fit to the model with infit mean-square values between 0.81 to 1.26 with the exception of SO4 (Infit MNSQ=1.32). However, this was run with outlying cases still within the model. As per the recommendations of Sumintono and Widhiarso (2013), outlying cases with infit mean-square values outside the accepted range of 0.77 to 1.30 were removed before the final model results could be determined. This resulted in the removal of a total of 50 cases before all items fit to the model with infit mean-square values between 0.79 to 1.26 and all cases fit to model with values between 0.77 to 1.28.

Wright and Linacre (1994) put forth recommendations of reasonable mean-square values for various types of assessments. Generally, it is seen that multiple-choice questionnaire items should show item fit within the range of 0.70 to 1.30 and a narrower range of 0.80 to 1.20 for high-stakes multiple-choice tests (Wright & Linacre, 1994). Given that the purpose of the current scale is to guide career exploration in a career guidance setting (Mamauag et al., 2016), the wider range of infit mean-square values can be applied as it is not being used for high-stakes decision-making. As such, the range of item fit obtained by items in the MAT-D(VS) show good and accepted fit values as per

Wright and Linacre's (1994) recommendation (Recommended value: 0.70-1.30; scale values: 0.79-1.26). This suggests that the behaviour of the items in terms of accuracy of measurement falls within an acceptable range of randomness; the items are neither too predictable (an indicator of item redundancy in measurement) nor too unpredictable (an indicator of too much randomness which doesn't provide accurate measurement) (Wright & Linacre, 1994).

Furthermore, the Rasch analysis allowed the estimation of difficulty levels for each item, thus allowing for an estimation of how well the test is matched to the ability level of the examinees (Bond & Fox, 2015). The Wright map produced from the Rasch model showed that the distribution of item difficulty approximated the distribution of the examinee ability, and these distributions were visually approximately normal. This was further supported by the normality test of the score distribution which indicated that the distribution of scores on the scale were approximately normal. In this manner, it can be said that the test is relatively fair as it is accurately matched to the abilities of the sample of examinees.

5.4.2 Research objective 2

RO2 focused on an examination of the quality of item distractors in the MAT-D(VS) using distractor analysis. Thus, the research question proposed was what are the qualities of the distractors found in the items of the MAT-D(VS)? The results of the distractor analysis showed that 19 of the items from the initial 21 items from the confirmatory factor analysis were retained with no changes. This was an increase from the earlier pilot study that retained 13 items with no changes. Item SO1 was still retained after amendments were made to the distractor after the pilot was concluded. These improvements supported the recommendations of Haladyna (2004) which suggested a distractor analysis as necessary in efforts to produce good items. In addition, the new FGP items were all retained (albeit with distractor amendments necessary for FGP3) thus

supporting the findings of Wagemans et al. (2012) that suggested using curved lines in FGP items would improve the item properties as it allows grouping according to continuity.

Taking a closer look at the items with less effective distractors, distractor A of FGP3 (item 3) was seen as confusing to students of higher ability where students who had greater estimated visual-spatial aptitude tended to select distractor A over the key option D. This may be due to the similarity between the stimulus and the object shown in the distractor as highlighted in figure 5.1. Students of higher ability may have seen distractor A first and assumed that the stimulus may have been shown in a distorted form thus quickly selecting the distractor without considering other options. Amendments to the distractor may include removing the similar-looking object. For SO3 (item 18), distractor C (as shown in figure 5.2) was seen as ineffective as it confused students of higher ability. It is possible that students assumed it to be a bottom view of the stimulus (even though there is no bottom layer illustrated) and thus chose that as the answer instead of the key option A. Future amendments may include removing option C with a different image/perspective that is less open to students' interpretation.

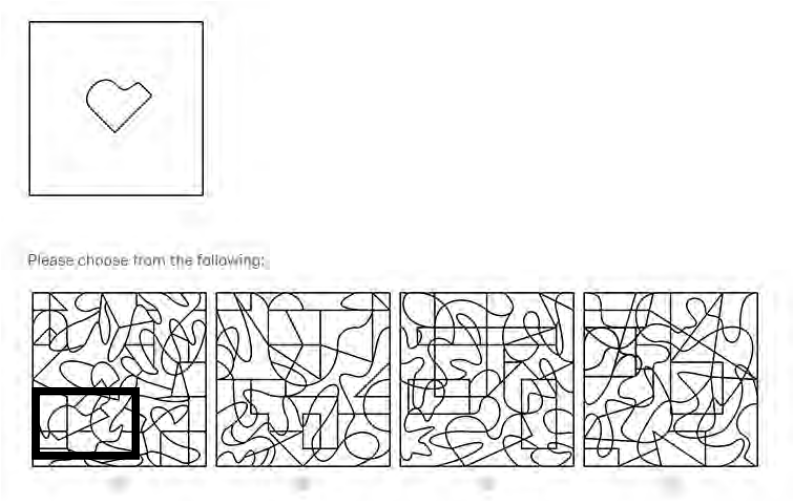

Figure 5.1. FGP3 (item 3) with object in distractor highly similar to stimulus (outlined)

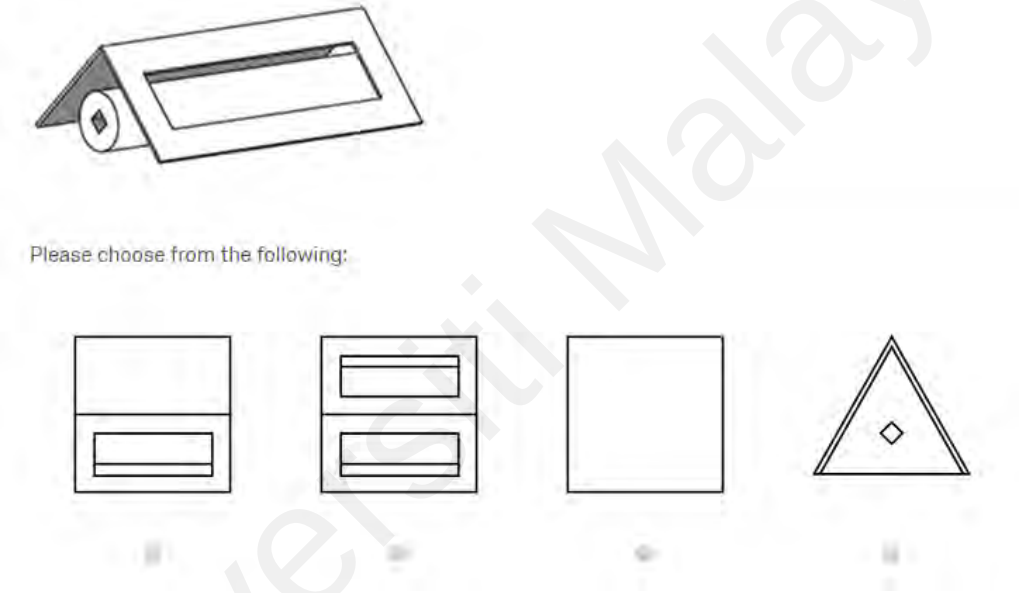

Figure 5.2. SO3 (item 18). Option C is an ineffective distractor

The use of the distractor analysis is also characteristic of the test development process by DeVellis (2003) that is being applied in the current study. DeVellis (2003) recommends amendments that are informed by evaluation of results when developing a new scale. In this manner, the distractor analysis allows for refinements to be made at the finest level, the level of distractors. This allows the researcher to develop a precise and accurate measure such that it is able to differentiate low- versus high-performing examinees from the choice of distractors themselves (DiBattista & Kurzawa, 2011).

5.5 Parallels with Extant Literature

The results of the study provide further support for Carroll's three-stratum theory of intelligence. The six-factor structure of visual-spatial aptitude established in this study show support for the existence of Stratum I and Stratum II abilities as posited by Carroll. The six constructs of visual-spatial aptitude showed sufficient correlation to one another as to be indicative of the underlying construct of visual-spatial aptitude, thus supporting the idea that measuring specific abilities reflects the underlying more general ability (Carroll, 2003). In addition, the correlations between constructs taken together with the individual factor loadings support Carroll's idea that the Stratum I abilities should be linearly independent despite being correlated (Carroll, 2003).

Furthermore, it is supported by Hegarty and Waller (2005) and Linn and Peterson (1985, as cited in Yilmaz, 2009) that suggests examining spatial ability should be done at the level of separate abilities. The six-factor structure established in this study also provides evidence for a set of abilities that can be used as an appropriate measure of visual-spatial ability. As mentioned by Johnson & Bouchard (2005), there is no specific set of abilities that can be taken as the only abilities that are used to assess visual-spatial ability. The evidence for these six constructs as being good representations of the underlying construct can provide a good starting point for other assessments to developed in measuring visual-spatial aptitude. These six chosen constructs have also shown much literature support as sub-constructs of visual-spatial aptitude (Magno, 2007) and, taken together with the model fit evidence of the confirmatory factor analysis, suggest that these constructs are suitable to be used as a basis for design of scales in this area of research. It is also beneficial as the scores can be identified as a sum of clusters or provided for each individual construct which helps with identification of latent strengths and shortcomings, as suggested by Anastasi and Urbina (1997).

The use of factor analysis in determining the structure of the underlying construct supports the initial work of Carroll (1993) who used this technique in exploring his initial theory development. It has been noted by D'Oliveira (2004) that few researchers have attempted an effort on the scale of Carroll's but this study forms a good starting point in attempting to determine the structure of human abilities in the area of visual-spatial aptitude particularly given that the use of factor analysis is a well-documented technique in establishing support for this theory (Bickley et al., 1995). In addition, the identification of the six-factor structure in this study also provides some clarity to the field based on problems highlighted by D'Oliveira (2004). Not only does the support for the structure show that the six constructs are suitable in encompassing a measurement of visual-spatial aptitude, the generated operational definitions and terms used provide an indication of how these constructs can be addressed in the future thus providing support for the standardized use of the terms identified in this study.

As used in this study, the DeVellis (2003) guidelines for scale development are seen as beneficial in directing the development process of the MAT-D(VS). The current development process is shown to be effective in generating items of good quality, with improvements seen in every subsequent iteration of scale development. From the pilot to the main study, it can be seen that the psychometric properties of the scale are gradually improving, although further amendments are necessary as evidenced in the results particularly pertaining to object assembly. This provides support for the use of the DeVellis (2003) guidelines as being suitable in various forms of scale development.

5.6 Theoretical Implications

In this study, it was seen that the results were able to show support for the three-stratum theory in multiple ways. Firstly, the current study showed empirical support for the six-factor structure of visual-spatial aptitude. From a theoretical standpoint, the findings provided support for the existence of Stratum I and Stratum II abilities under

Carroll's three-factor theory as shown in the significant factor loadings for the items that was found in the confirmatory factor analysis. In particular, it supports the idea that Carroll's theory is appropriate to be applied in determining the structure of abilities and can be extended to other types of Stratum II abilities aside from visual-spatial aptitude as the same statistical technique was originally used by Carroll (1993). In addition, it provides support for the applicability of Carroll's theory within a context different from that in which it was originally developed; namely, it can be evidenced that Carroll's theory is applicable within an Asian context as much of Carroll's (1993) work when developing his theory was centred in an American context. Therefore, the findings of the current study provide a useful extension of Carroll's work in a novel, Asian context.

Furthermore, the generation of new operational definitions which are grounded in extant literature build on the findings of previous studies to provide updated definitions of the identified constructs such that all important aspects of relevant literature are accounted for. The new definitions are further validated by the findings of the study thus providing an extension of the field whereupon further studies can utilise the definitions afforded in this study to generate more items that are able to measure these identified constructs. In addition, it provides greater standardization within the field where these constructs and definitions can be used by other researchers to generate new scales that will gather further evidentiary support for Carroll's three-factor theory that underlies the current study which aligns with the recommendations of D'Oliveira (2004).

5.7 Practical Implications

From a practical standpoint, the current study sets the foundation for the development of the MAT-D(VS) where the current version already demonstrates desirable psychometric properties of high validity and reliability. Though further amendments to the individual items will be necessary in subsequent iterations of the test

development process, it provides a clear framework from which further items may be generated.

By developing this scale with clear validity and high reliability as established through the use of empirically sound methods, it will become a useful assessment tool in enhancing the career guidance process by providing counsellors with accurate information regarding the students' abilities in the area of visual-spatial competence. It will also benefit the career guidance process as counsellors are better able to advise students on ways to leverage and grow their natural abilities such that they can achieve success in finding and retaining work in the future.

In addition, the tool can also help identify specific areas of weakness. Given that certain constructs have been shown to benefit certain job classes more than others (Magno, 2009), it can assist the counsellor in identifying jobs that may not be suitable for the student or identify skills that need to be improved should the student still be interested in pursuing that particular career (Krumboltz & Vidalakis, 2000). Furthermore, as accurate self-reflection is a necessary part of the career development process (Savickas et al., 2009), this tool can help students better understand their abilities and choose careers that will allow them to showcase their best possible talents.

5.8 Future Directions

The next step in the development of the MAT-D(VS) will be to carry out amendments for distractors of items FGP3 and SO3, generate one new item each for constructs PS and SO, generate two new SV items, and generate new OA items. The second revision of the scale will then have to be tested and the same tests of validity and reliability should be conducted, namely, confirmatory factor analysis, Rasch analysis, and distractor analysis. This is to ensure that all items will load well within the model and show desirable psychometric properties. The construct of topology may also be re-

examined to develop a cohesive operational definition and good items to capture the essence of the construct.

Following that, norming procedures should be conducted to determine cut-off scores for accurate interpretation of the scores. This is to ensure that the information given to examinees is an accurate and understandable reflection of their abilities. The VS scale should also be tested together with other items from the MAT such as the more difficult verbal analogy and number-letter series items taken from the item bank to ensure that the difficulty level is well matched. In addition, the scale can also be further tested with differential item functioning analysis to ensure that the items show no bias to any group with regard to age, ethnicity or gender.

Aside from that, more items can be generated according to the given operational definitions in order to create parallel forms. The development of parallel forms will benefit test administration procedures to reduce the probability of cheating or guessing if examinees are distributed different but equivalent forms of the scale. Finally, with a large enough sample size, two- or three-parameter IRT model analysis can be applied. This is to determine the other parameters that influence examinee performance such as item discrimination and the guessing factor. In this manner, only the best performing items from all three aspects will be maintained such that more accurate information regarding examinees latent abilities can be obtained.

5.9 Conclusion

This study was conducted with the aim of developing a scale of visual-spatial aptitude known as the MAT-D(VS) which would be a fair match to the ability levels of foundation and first-year undergraduate students. The resulting scale showed desirable psychometric properties of high validity as established using confirmatory factor analysis and high reliability as established using Rasch analysis. Distractor analysis was conducted to ensure that all items had well-functioning distractors. The findings of the study

supported a six-factor structure of visual-spatial aptitude consisting of the constructs figure-ground perception (FGP), object assembly (OA), progressive series (PS), spatial orientation (SO), spatial visualisation (SV), and visual discrimination (VD). After one round of the revisions, the 30-item scale showed good item loadings for 25 items. 21 items were retained for further analysis based on their contribution to the variance explained within the model. Based on the Rasch analysis, all items showed desirable fit to the Rasch model with high reliability indices. The distractor analysis revealed the need for amendments to distractors of two items. Based on the cyclical process outlined by DeVellis (2003) for scale development, future revisions of the scale are necessary to ensure further improvement of the psychometric properties of the scale before standard setting and norming procedures can be conducted. Overall, the findings of this study support the theory of the structure of abilities posited by Carroll (1993) and extend its applicability within an Asian context. In terms of practicality, the scale developed presents an assessment tool for career guidance that is well-contextualised to an Asian population.

REFERENCES

Abdi, H. (2007). Multiple correlation coefficient. In N. J. Salkind (Ed.), *Encyclopedia of measurement and statistics*. Thousand Oaks, CA: Sage Publishing. Retrieved from https://www.utd.edu/~herve/Abdi-MCC2007-pretty.pdf.

Ackerman, P. L. & Heggestad, E. D. (1997). Intelligence, personality and interests: Evidence for overlapping traits. *Psychological Bulletin*, 121(2), 219-245.

Adams, R. J. & Khoo, S-T. (1996). *Quest: The interactive test analysis system*. Victoria, Australia: The Australian Council for Educational Research Ltd.

Amla H. M. Salleh. (1984). *An investigation of the reliability, validity, and translation of Holland's Self Directed Search for utilization by a Malaysian population* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses Global. (ProQuest document ID: 303311846).

Amla Mohd. Salleh. (2010). *Pendidikan kerjaya dan pembangunan modal insan.* Selangor: Penerbit UKM.

An, X & Yung, Y-F. (2014). Item response theory: What it is and how you can use the IRT procedure to apply it. *Proceedings of the SAS Global Forum 2014*, Paper SAS364-2014. Retrieved from https://support.sas.com/resources/papers/proceedings14/SAS364-2014.pdf.

Anastasi, A. & Urbina, S. (1997). *Psychological Testing* (7th Ed.). New Jersey: Prentice-Hall, Inc.

Arulmani, G., Bakshi, A. J., Flederman, P., & Watts, A. G. (2011). Editorial. *International Journal for Educational and Vocational Guidance*, *11*(2), 61-64. doi: https://doi.org/10.1007/s10775-011-9204-5.

Azura Abas & Hashini Kavishtri Kannan. (2017, November 21). Those picking up their UPSR results on Thursday, brace yourselves for something different. *New Straits Times.* Retrieved from

https://www.nst.com.my/news/nation/2017/11/305854/those-picking-their-upsr-results-thursday-brace-yourselves-something.

Baker, F. B. (2001). *The basics of Item Response Theory* (2nd Ed.). United States: ERIC Clearinghouse on Assessment and Evaluation.

Beauducel, A. & Herzberg, P. Y. (2006). On the performance of Maximum Likelihood versus Means and Variance Adjusted Weighted Least Squares estimation in CFA. *Structural Equation Modelling*, *13*(2), 186-203. doi: 10.1207/s15328007sem1302.

Bellows, R. M. (1941). Matching youth and jobs. *The Phi Delta Kappan*, 23(5), 201-203. Retrieved from www.jstor.org/stable/20330899.

Bickley, P. G., Keith, T. Z., & Wolfle, L. M. (1995). The three-stratum theory of cognitive abilities: Test of the structure of intelligence across the life span. *Intelligence*, *20*(3), 309-328. doi: 10.1016/0160-2896(95)90013-6.

Blum, D., Holling, H., Galibert, M. S., & Forthmann, B. (2016). Task difficulty prediction of figural analogies. *Intelligence*, *56*, 72-81. doi: 10.1016/j.intell.2016.03.001.

Bond, T. G. & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd Ed.). New York: Routledge.

Bordens, K. S. & Abbott, B. B. (2011). *Research designs and methods: A process approach* (8th Ed.). New York: McGraw-Hill.

Busciglio, H. H., Palmer, D. R., King, I. H., & Walker, C. B. (1994). *Creation of new items and forms for Project A Assembling Objects test* (ARI Technical Report 1004). Alexandria, VA: U.S. Army Research Institute for the Behavioural and Social Sciences. (AD A282 727).

Butner, J. E., Gagnon, K. T., Geuss, M. N., Lessard, D. A., & Story, T. N. (2015). Utilizing topology to generate and test theories of change. *Psychological Methods*, *20*(1), 1-25. doi: 10.1037/a0037802.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge: Cambridge University Press.

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports g and about 10 broad factors. In H. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 5-21). Amsterdam: Pergamon.

Cattell, R. B. (1978). *The scientific use of factor analysis in behavioural and life sciences*. New York: Plenum Press.

Catts, H. W., Fey, M., Zhang, X., & Tomblin, J. B. (2001). Estimating the risk of future reading difficulties in kindergarten children: A research-based model and its clinical implementation. *Language, Speech, and Hearing Services in Schools*, *32*(1), 38-50.

Chong, N., Arosh, S. G., Mamauag, M. F. M., Teo, S. W., Tai, Y. S., & Ooi, H. J. (2016, July). *Defining constructs of the Five Factor model using a Malaysian context: Towards developing an assessment tool for personality*. Paper presented at the 31st International Congress of Psychology, Yokohama, Japan. Abstract retrieved from http://onlinelibrary.wiley.com/doi/10.1002/ijop.12345/epdf.

Chua, Y. P. (2014). *Kaedah dan statistik penyelidikan buku 5: Ujian regresi, analisis faktor dan analisis SEM* (2nd Ed.). Selangor: McGraw-Hill Education (Malaysia) Sdn. Bhd.

Cohen, R. J. & Swerdlik, M. E. (2009). *Psychological testing and assessment: An introduction to tests & measurement* (7th Ed.). United States of America: The McGraw-Hill Companies, Inc.

Darwishah, N., Chong, N., Mamauag, M. F. M., Tai, Y. S., Arosh, S. G., Teo, S. W., Ooi, H. J., Wong, I., & Pang, W. T. (2016, July). *Defining 21st Century Skills in the Malaysian context: Towards developing an assessment tool for employability*

and career readiness. Paper presented at the 31st International Congress of Psychology, Yokohama, Japan. Abstract retrieved from http://onlinelibrary.wiley.com/doi/10.1002/ijop.12345/epdf.

de Freitas, E. & McCarthy, M. J. (2014). (Dis)orientation and spatial sense: Topological thinking in the middle grades. *PNA*, *9*(1), 41-51. doi: 10.1080/026432900380463.

DeVellis, R. F. (2003). *Scale development: Theory and applications* (2nd Ed.). California: Sage Publications, Inc.

DiBattista, D. & Kurzawa, L. (2011). Examination of the quality of multiple-choice items on classroom tests. *The Canadian Journal for the Scholarship of Teaching and Learning*, *2*(2), Article 4. doi: 10.5206/cjsotl-rcacea.2011.2.4.

D'Oliveira, T. C. (2004). Dynamic spatial ability: An exploratory analysis and a confirmatory study. *The International Journal of Aviation Psychology*, *14*(1), 19-38. doi: 10.1207/s15327108ijap1401_2.

Englehard, G., Jr. (2013). *Invariant measurement: Using Rasch models in the social, behavioural and health sciences*. New York: Routledge.

Field, A. (2013). *Discovering statistics using IBM SPSS Statistics* (4th Ed.). London: Sage Publications Ltd.

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2012). *How to design and evaluate research in education* (8th Ed.). New York: McGraw-Hill.

FreeCADweb.org. (2017). FreeCAD [Computer software]. Retrieved from https://www.freecadweb.org/.

Fung, D. & Wong, P. S. L. (2012). Using career education and career services to enhance employability: A case of the Hong Kong Polytechnic University. *Asian Journal of Counselling*, *19*(1), 75-96.

Fuss, T. & Schluessel, V. (2015). Something worth remembering: Visual discrimination in sharks. *Animal Cognition*, *18*(2), 463-471. doi: 10.1007/s10071-014-0815-3.

Furr, R. M. (2011). *Scale construction and psychometrics for social and personality psychology*. London: SAGE Publications Ltd.

Ghose, T. & Palmer, S. E. (2016). Gradient cuts and extremal edges in relative depth and figure-ground perception. *Attention, Perception, & Psychophysics*, *78*(2), 636-646. doi: 10.3758/s13414-015-1030-2.

Gillie, S. & Isenhour, M. G. (2003). *The educational, social, and economic value of informed and considered career decisions*. America's Career Resource Network Association: United States of America. Retrieved from http://www.bridges.com/us/training/resnwhitepapers/ACRN_ICCD.pdf.

Godoy, F. & Rodríguez, A. (2004). Defining and comparing content measures of topological relations. *GeoInformatica*, *8*(4), 347-371. doi: 10.1023/B:GEIN.0000040831.81391.1d.

Gottfredson, L. S. (2003). The challenge and promise of cognitive career assessment. *Journal of Career Assessment*, *11*(2), 115-135. doi: 10.1177/1069072702250415.

Hair, J. F. Jr., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th Ed.). New Jersey: Pearson Education, Inc.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd Ed.). New Jersey: Lawrence Erlbaum Associates, Inc.

Harris, J., Newcombe, N. S., & Hirsh-Pasek, K. (2013). A new twist on studying the development of dynamic spatial transformations: Mental paper folding in young children. *Mind, Brain, and Education*, *7*(1), 49-55. doi: 10.1111/mbe.12007.

Hegarty, M. & Waller, D. A. (2005). Individual differences in spatial abilities. In P. Shah & A. Miyake (Eds.), *The Cambridge handbook of visuospatial thinking* (pp. 121-169). New York: Cambridge University Press.

Ho, Y.-F. (2008). Reflections on school career education in Hong Kong: Responses to Norman C. Gysbers, Darryl Takizo Yagi, and Sang Min Lee & Eunjoo Yang. *Asian Journal of Counselling*, *15*(2), 183-205.

Hooper, D., Coughlan, J. & Mullen, M. R. (2008). Structural Equation Modelling: Guidelines for determining model fit. *The Electronic Journal of Business Research Methods*, *6*(1), 53-60.

Hu, L.-t. & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modelling*, *6*(1), 1-55. doi: 10.1080/10705519909540118.

Iacobucci, D. (2010). Structural equations modeling: Fit indices, sample size, and advanced topics. *Journal of Consumer Psychology*, *20*, 90-98. doi: 10.1016/j.jcps.2009.09.003.

Irvin, P. S., Alonzo, J., Lai, C.-F., Park, B. J., & Tindal, G. (2012). *Analyzing the reliability of the easyCBM Reading Comprehension measures: Grade 7* (Technical Report #1206). Oregon: Behavioural Research and Teaching.

Ivie, J. L. & Embretson, S. E. (2010). Cognitive process modeling of spatial ability: The assembling objects task. *Intelligence*, *38*(3), 324-335. doi: 10.1016/j.intell.2010.02.002.

Jin, L. (2017). The current status of career services and professionals in Mainland China's educational settings. In J. Y. Hyung, B. Hutchison, M. Maze, C. Pritchard, & A. Reiss (Eds.), *International practices of career services, credentials, and training* (Chapter 4). Broken Arrow, Oklahoma: National Career Development Association. Retrieved from https://www.ncda.org/aws/NCDA/page_template/show_detail/139017?layout_name=layout_details&model_name=news_article.

Johnson, W. & Bouchard, T. J., Jr. (2005). The structure of human intelligence: It is verbal, perceptual, and image rotation (VPR), not fluid and crystallized. *Intelligence*, *33*(4), 393-416. doi: 10.1016/j.intell.2004.12.002.

Jöreskog, K. G. (1999). *How large can a standardized coefficient be?* Retrieved from http://www.ssicentral.com/lisrel/techdocs/HowLargeCanaStandardizedCoefficientbe.pdf.

Katsuoloudis, P., Jovanović, V., & Jones, M. (2014) A comparative analysis of spatial visualization ability and drafting models for industrial and technology education students. *Journal of Technology Education*, *26*(1), 88-101. doi: 10.21061/jte.v26i1.a.6.

Kell, H. J., Lubinski, D., & Benbow, C. P. (2013). Who rises to the top? Early indicators. *Psychological Science*, *24*(5), 648-659. doi: 10.1177/0956797612457784.

Kenny, D. A. (2015). *Measuring model fit*. Retrieved from http://davidakenny.net/cm/fit.htm.

Kimchi, R. & Peterson, M. A. (2008). Figure-ground segmentation can occur without attention. *Psychological Science*, *19*(7), 660-668. doi: 10.1111/j.1467-9280.2008.02140.x.

Krumboltz, J. D. & Vidalakis, N. K. (2000). Expanding learning opportunities using career assessments. *Journal of Career Assessment*, *8*(4), 315-327.

Kunda, M., McGreggor, K., & Goel, A. K. (2012). Reasoning on the Raven's Advanced Progressive Matrices test with iconic visual representations. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society*, Sapporo, Japan. pp. 1828-1833. Retrieved from https://pdfs.semanticscholar.org/6aaf/af2eecccac91735be67937b79e52faa71475.pdf.

Lin, C.-H. & Chen, C.-M. (2016). Developing spatial visualization and mental rotation with a digital puzzle game at primary school level. *Computers in Human Behaviour*, *57*, 23-30. doi: 10.1016/j.chb.2015.12.026.

Lin, C.-H., Chen, C.-M., & Lou, Y.-C. (2014). Developing spatial orientation and spatial memory with a treasure hunting game. *Educational Technology & Society*, *17*(3), 79-92.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878. Retrieved from http://www.rasch.org/rmt/rmt162f.htm.

Linacre, J. M. (n.d.). *Reliability and separation of measures*. Retrieved from http://www.winsteps.com/winman/reliability.htm.

Linacre, J. M. & Wright, B. D. (1989). The "length" of a logit. *Rasch Measurement Transactions*, *3*(2), 54-55. Retrieved from https://www.rasch.org/rmt/rmt32b.htm.

Magno, C. (2009). Taxonomy of aptitude test items: A guide for item writers. *The International Journal of Educational and Psychological Assessment*, *2*, 39-53.

Mamauag, M. F. M., Tai, Y. S., Arosh, S. G., Teo, S. W., & Ooi, H. J. (2016, July). *Establishing the psychometric qualities of the Multiple Aptitude Test using the Rasch model*. Paper presented at the 31st International Congress of Psychology, Yokohama, Japan. Abstract retrieved from http://onlinelibrary.wiley.com/doi/10.1002/ijop.12345/epdf.

Martinelli, C. & Shergill, S. S. (2015). Clarifying the role of pattern separation in schizophrenia: The role of recognition and visual discrimination deficits. *Schizophrenia Research*, *166*, 328-333. doi: 10.1016/j.schres.2015.06.004.

Mathewson, J. H. (1999). Visual-spatial thinking: An aspect of science overlooked by educators. *Science Education*, *83*(1), 33-54. doi: 10.1002/(SICI)1098-237X(199901)83:1<33::AID-SCE2>3.3.CO;2-Q.

Ministry of Education. (2013). *Malaysia Education Blueprint 2013-2025 (Preschool to Post-Secondary Education)*. Putrajaya: Ministry of Education Malaysia.

Mukaka, M. M. (2012). A guide to the appropriate use of correlation coefficient in medical research. *Malawi Medical Journal*, *24*(3), 69-71. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3576830/.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin*, *105*(3), 430-445.

Muthén, L.K. (2012, April 2). Confidence intervals for RMSEA (Other than 90%) [Online forum comment]. Message posted to https://www.statmodel.com/discussion/messages/9/2418.html?1333635935.

Muthén, L.K. & Muthén, B.O. (2012). *Mplus User's Guide* (7th Ed.). Los Angeles, CA: Muthén & Muthén.

Newcombe, N. & Shipley, T. F. (2015). Thinking about spatial thinking: New typology, new assessments. In J. S. Gero (Ed.), *Studying visual and spatial reasoning for design creativity* (179-192). Heidelberg: Springer Dordrecht.

O'Hare, L. & McGuinness, C. (2015). The validity of critical thinking tests for predicting degree performance: A longitudinal study. *International Journal of Educational Research*, *72*, 162-172. doi: 10.1016/j.ijer.2015.06.004.

Office of Career, Technical, and Adult Education. (2014). *Career guidance and counseling programs*. Retrieved from http://www2.ed.gov/about/offices/list/ovae/pi/cte/cgcp.html.

Olkun, S. (2003). Making connections: Improving spatial abilities with engineering drawing activities. *International Journal of Mathematics Teaching and Learning*, *3*(1), 1-10. doi: 10.1501/0003624.

Ong, S. L. (2010). Assessment profile of Malaysia: High-stakes external examinations dominate. *Assessment in Education: Principles, Policy, and Practice*, *17*(1), 91-103. doi: 10.1080/09695940903319752.

Personnel Testing Division, Defense Manpower Data Center. (2008). *ASVAB Technical Bulletin No. 3: CAT-ASVAB Forms 5-9*. Retrieved from http://official-asvab.com/docs/asvab_techbulletin_3.pdf/

Peterson, M. A. (2015). Low-level and high-level contributions to figure-ground perception. In J. Wagemans (Ed.), *The Oxford Handbook of Perceptual Organization* (pp. 259-280). Oxford, United Kingdom: Oxford University Press.

Pope, M., Musa, M., Singaravelu, H., Bringaze, T., & Russell, M. (2002). From colonialism to ultranationalism: History and development of career counseling in Malaysia. *The Career Development Quarterly*, *50*(3), 264-276. doi: 10.1002/j.2161-0045.2002.tb00902.x.

Quaiser-Pohl, C. Neuburger, S., Heil, M., Jansen, P., & Schmelter, A. (2014). Is the male advantage in mental-rotation performance task independent? On the usability of chronometric tests and paper-and-pencil tests in children. *International Journal of Testing*, *14*(2), 122-142. doi: 10.1080/15305058.2013.860148.

Qualtrics [Computer software]. (2017). Retrieved from https://www.qualtrics.com/homepage/.

Quek, A.-H. (2001, March). *Career guidance and counselling in Malaysia: Development and trends*. Paper presented at the 9th Asian Regional Association for Career Development (ARACD) Conference, Singapore.

Raven, J. (2000). The Raven's progressive matrices: Change and stability over culture and time. *Cognitive Psychology*, *41*(1), 1-48. doi: 10.1006/cogp.1999.0735.

Rigdon, E. E. (1996). CFI versus RMSEA: A comparison of two fit indexes for structural equation modelling. *Structural Equation Modeling*, *3*(4), 369-379. doi: 10.1080/10705519609540052.

Roitman J. D. & Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *The Journal of Neuroscience*, *22*(21), 9475-9489.

Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice*, *17*, 321–335. http://dx.doi.org/10.12738/estp.2017.1.0270.

Savickas, M. L., Nota, L., Rossier, J., Dauwalder, J.-P., Duarte, M. E., Guichard, J., Soresi, S., Van Esbroeck, R., & van Vianen, A. E. M. (2009). Life designing: A paradigm for career construction in the 21st century. *Journal of Vocational Behavior*, *75*(3), 239-250. doi: 10.1016/j.jvb.2009.04.004.

Sedgwick, P. (2014). Cross sectional studies: Advantages and disadvantages. *The BMJ*, *348*. doi: http://dx.doi.org/10.1136/bmj.g2276.

See, C. M. & Ng, K.-M. (2010) Counseling in Malaysia: History, current status, and future trends. *Journal of Counseling & Development*, *88*(1), 18-22. doi: 10.1002/j.1556-6678.2010.tb00144.x.

Shamama-tus-Sabah, S., Gilani, N., & Iftikhar, R. (2012). Ravens Progressive Matrices: Psychometric evidence, gender, and social class differences in middle childhood. *Journal of Behavioural Sciences*, *22*(3), 120-131.

Siroky, K., & Di Leonardi, B. C. (2015). Refine test items for accurate measurement: Six valuable tips. *Journal for Nurses in Professional Development*, *31*(1), 2-8. doi: 10.1097/NND.0000000000000123.

Stumpf, H., Mills, C. J., Brody, L. E., & Baxley, P. G. (2013). Expanding talent search by including measures of spatial ability: CTY's Special Test Battery. *Roeper Review*, *35*, 254-264. doi: 10.1080/02783193.2013.829548.

Sumintono, B. & Widhiarso, W. (2013). *Aplikasi model Rasch untuk penelitian ilmu-ilmu sosial*. Yogyakarta: TrimKom Publishing House.

Super, D. E. (1983). Assessment in career guidance: Toward truly developmental counselling. *Personnel & Guidance Journal*, *61*(9), 555-562.

Tavakol, M. & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53-55. doi: 10.5116/ijme.4dfb.8dfd.

The GIMP Team. (2017). GNU Image Manipulator Program (GIMP) [Computer software]. Retrieved from https://www.gimp.org/.

Tien, H.-L. (1997, August). *The vocational interest structure of Taiwanese high school students*. Paper presented at the 105th Annual Convention of the American Psychological Association, Chicago, United States of America. Retrieved from https://files.eric.ed.gov/fulltext/ED413421.pdf.

Tien, H.-L. S. (2009). Cultural encountering: The applicability of Holland's typology in Taiwan. *Asian Journal of Counselling*, *16*(2), 193-226.

Tien, H.-L. S. & Wang, Y.-C. (2016). Career counseling research and practice in Taiwan. *The Career Development Quarterly*, *64*(3), 231-243. doi: 10.1002/cdq.12057.

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A Century of Gestalt Psychology in Visual Perception I. Perceptual Grouping and Figure-Ground Organization. *Psychological Bulletin*, *138*(6), 1172-1217. doi:10.1037/a0029333.

Wong, P. W. L. (2017). Career and life planning education in Hong Kong: Challenges and opportunities on the theoretical and empirical fronts. *Hong Kong Teachers' Centre Journal*, *16*, 125-149.

Wright, B. D. & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370. Retrieved from http://www.rasch.org/rmt/rmt83b.htm.

Wright, B. D. & Stone, M. H. (1979). *Best test design*. Chicago: MESA PRESS.

Yang, J. C. & Chen, S. Y. (2010). Effects of gender differences and spatial abilities within a digital pentominoes game. *Computers & Education*, *55*(3), 1220-1233. doi: 10.1016/j.compedu.2010.05.019.

Yeo, L. S. & Lee, B.-O. (2014). School-based counseling in Singapore. *Journal of Asia Pacific Counseling*, *4*(2), 69-79.

Yeo, L. S., Tan, S. Y., & Neihart, M. F. (2012). Counseling in Singapore. *Journal of Counseling & Development*, *90*(2), 243-248. doi: 10.1111/j.1556-6676.2012.00031.x.

Yilmaz, H. B. (2009). On the development and measurement of spatial ability. *International Electronic Journal of Elementary Education*, *1*(2), 83-96.

Yu, C-H. (2017). *A simple guide to Item Response Theory (IRT) and Rasch modeling*. Retrieved from http://www.creative-wisdom.com/computer/sas/IRT.pdf.

Yuen, M., Leung, S. A., & Chan, R. T. H. (2014). Professional counseling in Hong Kong. *Journal of Counseling and Development*, *92*(1), 99-103. doi: 10.1002/j.1556-6676.2014.00135.x.

Zacharapoulos, G., Binetti, N., Walsh, V., & Kanai, R. (2014). The effect of self-efficacy on visual discrimination sensitivity. *PLoS ONE*, *9*(10), 1-10. doi: 10.1371/journal.pone.0109392.

Zhang, W., Hu, X., & Pope, M. (2002). The evolution of career guidance and counselling in the People's Republic of China. *The Career Development Quarterly*, *50*(3), 226-236. doi: 10.1002/j.2161-0045.2002.tb00898.x.

Zhang, X. & Lin, D. (2015). Pathways to arithmetic: The role of visual-spatial and language skills in written arithmetic, arithmetic word problems, and nonsymbolic arithmetic. *Contemporary Educational Psychology*, *41*, 188-197. doi: 10.1016/j.cedpsych.2015.01.005.

Zhou, X., Li, X., & Gao, Y. (2016). Career guidance and counselling in Shanghai, China: 1977 to 2015. *The Career Development Quarterly*, *64*(3), 203-215. doi: 10.1002/cdq.12055.