THE DIMENSIONALITY OF COGNITIVE STRUCTURE: A MIRT APPROACH AND THE USE OF SUBSCORES

By

Yi-Ling Cheng

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Educational Psychology and Educational Technology—Doctor of Philosophy
Measurement and Quantitative Methods—Doctor of Philosophy

2016

ABSTRACT

THE DIMENSIONALITY OF COGNITIVE STRUCTURE:
A MIRT APPROACH AND THE USE OF SUBSCORES

By

Yi-Ling Cheng

The present study explored the dimensionality of cognitive structure from two approaches. The first approach used a famous relation between Visual Spatial Working Memory (VSWM) and calculation to demonstrate the multidimensional item response analyses when true dimensions are unknown. The second approach explored the detectability of dimensions by using person fit indices. Findings from the first study demonstrated that there were shared dimensions between VSWM and calculation. Additionally, error analysis with the MIRT approach revealed only one dimension, the Number dimension, of VSWM is related to subtraction and division. These results showed the usefulness of the MIRT analysis in analyzing cognitive structure. The result of the second study revealed that person fit indices, however, were insensitive to the simulated dimensions when estimated ability parameters were used. More sensitive detection indices need to be developed. The present study combines several dimensionality analyses with perspectives from diverse areas in the hope of providing insights into the dimensionality of cognitive structures.

ACKNOWLEDGEMENTS

The dissertation would not be possible without many people's helps.

My parent's wisdom and love helped me through every moment during the graduate life. Thanks for their patience and understanding when I decided to have the second major on doctoral study.

I could not thank enough to my EPET advisor Dr. Kelly Mix and MQM advisor Dr. Mark Reckase. Without Dr. Kelly Mix's guidance in experimental research and cognitive development, I will not be able to see the beauty of it. I was honored to work with Dr. Kelly Mix's lab, in which I learned enormously about experiments and data collection. It was my privilege to work with Dr. Mark Reckase, who leads me to the world of measurement and shapes the psychometrical perspective on my research development. I also want to thank my committee members. Dr. Cary Roseth has provided unique perspectives from children development on the dissertation. The statistical discussion with Dr. Spyros Konstantopoulos was extremely helpful. Dr. David Z. Hambrick suggestion from the perspective of cognitive psychology is very beneficial on the psychological aspects of the dissertation. Dr. Tenko Raykov is not my committee member, but the discussions with him have inspired me in many aspects about statistics and measurements. I also want to thank for his kind help in many occasions.

I would like to thank many colleagues, for their prudent advices over the years. With their encouragements, I was able to overcome the obstacles and move on to next stage. Particularly, I want to thank Dr. Chueh-an, Hsieh, who helped me greatly when I first started Ph.D. study.

I want to thank children and parents participated in the studies that I have worked with. I really appreciate the opportunity to learn from them.

I would like to thank the free software environment R. The open and free resources have provided many helpful statistical packages. I hope one day I can also contribute to the community.

Finally, I would like to thank College of Education and Institute of Education Sciences. Without their generous financial support I would not be able to finish the dual major.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

Dimensionality

The idea of dimensionality has been studied extensively in psychology (Carroll, 1993; Cattell, 1941; Spearman, 1927; Thurstone, 1947), measurement areas (Ackerman, Gierl, Walker, 2003; Reckase, 1997; 2009) and neural science (O'Toole, Jiang, Abdi, Pénard, Dunlop, & Parent, 2007). These studies may have used different terminology because they were in different disciplines, but the interest in exploring the concept of dimensionality has never died down. The influence of these studies has gone deeper and wider. For example, studies in measurement have generated research about one form of estimation after another and one model after another to understand it better. Other than studies in measurement, researchers in other areas have studied problems that are implicitly related to dimensionality without knowing that they were. A basic example that appears in many areas is that researchers have often studied how some constructs can predict another construct by identifying the shared variance or explained variance. The shared variance among studied variables is that these variables cross paths on one dimension in this universe to a varying degree. Another example can also be found in the education or psychological field is training studies. Researchers frequently explore an intervention and wish to see how it results in improving an outcome. Although it is often not outlined in the article itself, we cannot expect such an outcome to materialize unless the intervention and the outcome have something in common, meaning the overlaps on their dimensions. Finally, studies in neural imaging are often used to show how different brain areas are activated to separate the cognitive processes of particular constructs. These cognitive processes can also be seen as some aspect of dimensions.

Overall, many studies are related to the understanding of dimensionality, although they may not explicitly give it this name. On these studies that actually specified the interests on the dimensionality of constructs, however, often focused on the interpretation of the validity of the constructs. Therefore, in previous studies the idea of exploring dimensionality mostly lays in the understanding of the validity of the construct, which was sometimes limited by the way in which it was defined at the first place. Though these studies contributed to refining the boundaries of the constructs, not many studies went beyond this to consider what the elements underlying these constructs are.

Among all the studies on dimensionality, the most popular topic is studies about the structure of cognitive abilities. The growing interest in understanding cognitive abilities is developed from its predictive power on students' STEM achievement (i.e., science, technology, engineering, and mathematics; Best, Miller, & Naglieri, 2011; Deary, Strand, Smith, & Fernandes, 2007; Gustafsson and Balke, 1993; Passolunghi & Lanfranchi, 2012; Rohde & Thompson, 2007; Spinath, Spinath, Harlaar & Plomin, 2006). This relationship impacts on national wealth and economic growth, and as Rindermann and Thompson's study (2011) points out, these economic outcomes are largely the result of the development of STEM-based employment sectors. Therefore, improving students' cognitive abilities is not only critical for children's STEM academic achievement, but also has important implications for societal outcomes.

Though the influence of this relationship on social outcomes appears in adulthood, the relationship itself actually develops in early childhood. Studies demonstrate that early cognitive gain (i.e., before school enrollment) is critical for later mathematics skills (Campbell, Pungello, Miller-Johnson, Burchinal, & Ramey, 2001; Clark, Pritchard, & Woodward, 2010). This relation

progresses further and deeper when children enter school. For example, Floyd, Evans, & McGrew (2003) find that multiple cognitive abilities (including fluid intelligence, working memory, processing speed) are consistently associated with mathematics calculation skills and mathematics reasoning skills in persons from six to 19 years old. These intertwined, predictive relations between cognitive abilities and academic achievements from early childhood to later life suggest that improving children's early cognitive abilities has a long-lasting effect on their academic achievement, in addition to raising overall national growth.

General and Specific Approaches

General and specific abilities

Although the relationship between children's cognitive abilities and academic achievement seems to be strong in general, not every type of cognitive ability predicts academic achievement (Floyd, Evans, & McGrew, 2003). As a result, the aim of many studies is finding the most predictive cognitive ability. One popular approach is to identify whether general or specific abilities are more critical to children's achievement. This distinction between general and specific abilities comes from three strata theory (Carroll, 1993; Cattell, 1941; Spearman, 1927; Thurstone, 1947). General ability refers to the general intelligence underlying all cognitive abilities, whereas specific abilities refers to more narrowly defined cognitive abilities, such as processing speed, visual spatial thinking, or working memory (Carroll, 1997). Studies have found that general ability is the one and only important factor in predicting academic achievement (Deary, Strand, Smith, & Fernandes, 2007; Rohde & Thompson, 2007). The correlation is so high (r = .82; Frey & Detterman, 2004) that some even consider academic achievement tests such as the Scholastic Achievement Test (SAT) to be isomorphic to general ability tests.

Surprisingly, even though the relationship between general ability and academic achievement is strong, some studies can still identify the unique predictive power of specific abilities when testing their influences on academic achievements. For example, Gustafsson and Balke (1993) measured the cognitive abilities and academic achievement of 866 ninth graders with 16 aptitude tests and 17 course grades. They found that specific factors, in addition to the general factor, accounted for significant variance in achievement test scores. The same result is found with younger children. Passolunghi and Lanfranchi (2012) tested 70 kindergarten and first grade students on their mathematics achievement and found that specific cognitive abilities predicted mathematics achievement, in addition to general ability. These studies, however, leave open the question whether this finding is consistent for the same person over time. To explore this question, Geary (2011) conducted a longitudinal study of students from first grade to fifth grade and found that after controlling for general intelligence, some specific abilities (e.g., processing speed), but not all specific abilities contributed to different academic achievements (e.g., reading or mathematics). Overall, the literature not only supports the conclusion that general ability is a significant predictor of academic achievement, but also showed the contributions of specific abilities to academic achievements. Nonetheless, it seems as though the predictive power of a specific ability may not be consistent across different samples and different developmental stages.

Diagnostic testing and training

Similar conclusions about the predictive power of general and specific abilities are made in the area of diagnostic testing. Researchers often find that subtests of the intelligence test (e.g., the Wechsler Intelligence Scale for Children) fail to provide useful diagnostic information over and above the composite score (McDermott, Fantuzzo, & Glutting, 1990). One might argue that

the unidimensionality of intelligence tests made it less likely that any significant predictive power of specific abilities would be found. However, when researchers further examined data consisting of multiple subject matters (e.g., reading, mathematics…etc.), the same conclusion was found (e.g., Sinharay, Haberman, & Puhan, 2007).

Researchers further considered that this finding might have resulted from using certain types of subscore in previous studies. Most of them used the subscores that were based on Classical Test Theory (CCT)（e.g., raw scores from Classical Test Theory (CCT) in McDermott et al., 1990; Sinharay et al., 2007). However, recent developments in modern testing theory have helped researchers to find that subscores offered additional information when multidimensional item response models (MIRT) were applied to generate subscores (Thissen & Edwards, 2005; Yao & Boughton, 2007). Although the usefulness of subscores is still under debate, this shows that it is possible to extract useful information from subscores in addition to the total score.

Despite uncertainty about the predictability of specific abilities, general ability (g) remains the most powerful predictor of academic achievement in both applied measurement and psychological testing. Therefore, researchers in the area of cognitive training hypothesized that training *g* should generate improvement in most cognitive tasks, including academic achievement. The results are, however, mixed (Jaeggi Buschkuehl, Jonides, & Shah, 2011; Redick, Shipstead, Harrison, Hicks, Fried, Hambrick, Kane, & Engle, 2013). Despite its strong predictive power, training *g* has limited effects on other tasks (for a review, see Shipstead, Redick, & Engle, 2012) or even on improving itself (te Nijenhuis, van Vianen, & van der Flier, 2007).

There are several possible explanations for this. Some researchers argue that training programs are not designed properly. For example, te Nijenhuis, van Vianen, and van der Flier

(2007) analyzed tasks using factor analysis and found that most training programs are not *g*-loaded. The authors argue that, as a result, *g* cannot be improved. However, this seems an unclear reason to disqualify these training programs because the training was adapted from the tests that had been used to measure general ability (e.g., IQ tests).

Another possible reason for the discrepancy is that training a higher order function such as *g* to improve itself requires a combination of many cognitive processes. This combination itself may be burdening people with a large cognitive load. Training such a complex process would probably lead to learning less about each specific aspect of the task. As a result, individuals may demonstrate less mastery on the trained tasks, let alone improved overall ability. This process is similar to mastering a sport such as basketball. Athletes learn how to move the ball, how to pass a ball, and how to score a basket. The mastery of each specific process is needed to ultimately master basketball as a whole. This step-by-step learning process is likely similar in the brain. Learning one specific cognitive process at a time may help to encourage the thought process needed for a task or may even transfer to other tasks that have applied these specific processes. This type of training might then provide a better outcome than the learning of many different processes at one time.

Some studies have provided supportive evidence that training specific abilities indeed improves general ability. For example, some researchers investigated training that was more focused on a smaller part of *g*; for example, more specific cognitive abilities (e.g., working memory) that are significantly correlated with *g*. It was found that training sometimes improves *g*. However, when further examining the transfer effect to academic achievement such as mathematics or reading skills, the results are inconclusive (Jaeggi, Buschkuehl, Jonides, & Shah, 2011).

These inconclusive results suggest that the rationale of using significant correlations between two constructs in the expectation that training one will improve the other may be unfounded. Specifically, when two constructs are related, it suggests that they share a certain amount of variance. However, no testing or training is certain to be effective unless the specific shared variance is identified.

The Structure of Cognitive Abilities

The mixed results described above suggest the need to further examine the overall structure of cognitive abilities. More specifically, it is important to examine the relationships between general and specific abilities, the relationships between cognitive ability and academic achievement, and the way in which the specific sub-dimensions of these abilities are divided.

Hierarchical structure of cognitive ability

In the past, the most common way to study the structure of cognitive abilities was through factor analysis (e.g., Carroll, 1993). Over the years, researchers have come to a similar conclusion from analyzing the interrelations of various cognitive tasks and different populations—that the structure of cognitive abilities is hierarchical. The hierarchical structure has also represented the relations between general and specific abilities. However, different theories may label the same structures differently. An example of these theories is the three strata theory (Carroll, 1993; Cattell, 1941; Spearman, 1927; Thurstone, 1941). The hierarchical structure starts from the highest level: general ability (*g*). Under *g* is a second level consisting of eight or ten broad abilities, such as *Gf* (fluid intelligence) and *Gc* (crystallized intelligence), *Gv, Gs, Glr, Ga, Gsm*, and *Gq*. For instance, *Gf* is the ability to reason or engage in high order thinking, and *Gc* reflects the knowledge and experience that were obtained in the past, such as

mathematics skills (Cattell, 1941). Last, the lowest level consists of sixteen different abilities (e.g., processing speed or auditory processing; see Figure 1).

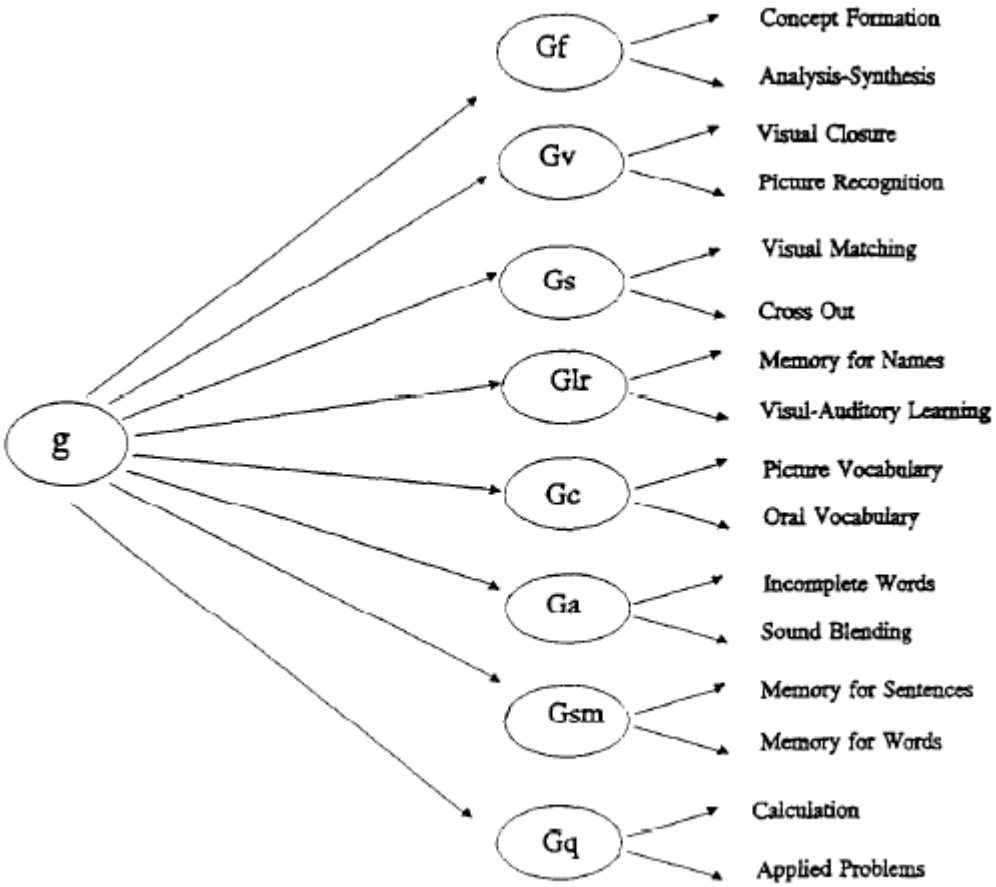

Figure 1. The three-stratum theory.

Figure 1. The structure of three-stratum theory. *Adapted from "The Three-Stratum Theory of Cognitive Abilities: Test of the Structure of Intelligence Across the Life Span," by Bickley, Keith, and Wolfle, 1995, intelligence, 20(3), p. 314.*

Nevertheless, because such a structure is generated from factor analysis, people have argued that it could be an artificial product of the statistical factor model and not directly related to the cognitive processes (e.g., Wechsler, 1950). Therefore, a somewhat different approach proposed to use neuropsychological methods to study brain connectivity on cognitive processes. However, this approach also suggests a hierarchical structure of cognitive abilities (Das, Kirby & Jarman, 1975; Kirby & Das, 1990; Luria, 1971; Naglieri & Otero, 2011). The difference is that this model theorizes (as PASS theory) the hierarchical structure of the model from lower level functions such as attention, simultaneous and successive, to higher-level functions, such as planning.

Although there are both brain imaging studies and behavioral evidence for a hierarchical structure of cognitive abilities, researchers do not always agree on defining the divisions of cognitive abilities. Most arguments lie in the division of the unit or in the levels within the hierarchy, neither of which is crystal clear. For obvious reasons, the three strata and PASS theories vary in their definitions and categorization on each level. Even within the same theory, disagreements can arise. For example, the supporters of the three strata theory still dispute which specific abilities should be included in each level (e.g., Carroll, 2003; Johnson & Bouchard, 2005). Among the sub-divisions, for example, the boundary between academic achievement and cognitive ability is not well defined. Specifically, in PASS theory, it was suggested that different achievements required different combinations of cognitive processes (e.g., planning and attention; Naglieri, & Rojahn, 2004). This implies that PASS theory posits each academic achievement as several sub-divisions of cognitive processes. Similar perplexity can be found in the three strata theory, where the broad factors *Gsm* and *Gq* are classified as second level cognitive abilities, but at the same time are also often classified as latent factors of academic achievement (e.g.,

Kaufman, Reynolds, Liu, Kaufman, McGrew, 2012). This unclear division is prevalent in these theories and questions the role of academic achievement in the structure of cognitive abilities.

Is academic achievement part of cognitive abilities?

Many studies have found significant relationships between academic achievement and cognitive abilities (Best, Miller, & Naglieri, 2011; Deary, Strand, Smith, & Fernandes, 2007; Gustafsson and Balke, 1993; Passolunghi and Lanfranchi, 2012; Rohde & Thompson, 2007; Spinath, Spinath, Harlaar & Plomin, 2006). The range of correlations is wide, from .40 to .90. Despite their spread, these fairly high correlations made researchers wonder whether cognitive ability and academic achievement are isomorphic. When the correlation is high (r = .83; Kaufman et al., 2012), 31% of the variance is still left unaccounted for, which demonstrates that these two constructs are not the same. However, even the lowest correlation, r = .40, accounts for at least 16% of variance, suggesting these two constructs share certain number of cognitive processes. A logical reason for this similarity is that they are both influenced by the same general ability. As evidence for this reasoning, Lynn and Meisenberg (2010), using factor analysis, found after correcting the unreliability (using attenuation procedure from Ferguson, 1971) that the correlation between the general factor extracted from cognitive tests and the general factor extracted from scores on achievement tests over the course of one school year is 1.0. This perfect relationship suggests that these two general factors might actually represent the same latent construct. Therefore, each score for an academic achievement (e.g., reading, mathematics, or science) can possibly be viewed as a sub-dimension under the broader construct of general ability.

The level of processing

The indistinct boundary between academic achievement and cognitive abilities is not the only issue related to the categorization of dimensions. Overall, the categorizations of these dimensions do not stay constant in different studies. To resolve this issue, therefore, the methodology used to divide these dimensions is critical and therefore needs further clarification. It is important to evaluate how dimensionality is defined, how it is studied psychometrically and biologically, and how these methods can be integrated.

The definition of dimensionality. The identification of construct dimensionality depends on one's perspective. Specifically, constructs can be viewed in some circumstances as unidimensional but in others as multidimensional. A construct is considered unidimensional when psychologically it is theorized as the same cognitive function. Statistically, it also needs to be high inter-item correlations within the construct. If both conditions are met, an individual's score on the construct can be placed on a continuum. However, sometimes several separated clusters of items/subsets sometimes appear within a given construct. When they do, the items/subsets are closely related within the clusters but not between the clusters. In this case, this construct is therefore viewed as multidimensional. This is one way to define the dimensionality of cognitive ability. Yet, in a hierarchical structure, a general ability may be unidimensional at its higher level but can also be viewed as multidimensional when the components of general ability are considered. This also can be a case to each component of general ability. Each can be considered unidimensional, but may also be multidimensional if the lower level dimensions are the ones being identified in this component.

One example of this is spatial ability. The dimensionality of spatial ability can be described in both unidimensional and multidimensional terms. When researchers conducted

factor analysis with several representative visual spatial tasks, a one-factor model was the best fit, suggesting unidimensionality (e.g., Lohman, 1979; Mix et al. in press). However, the components of spatial ability, mental rotation for example, when examined, might yield evidence to suggest that they are multidimensional. To illustrate, a dimension of mental rotation can be that the additional understanding of the figure structure adds another dimension to an individual's thought processes (Shepard & Metzler, 1988). This suggests that when examining a series of spatial tasks together, the shared variance among all spatial tasks reflects a shared common dimension. However, when examining items within each task, some items may demand different cognitive understanding from that demanded by other items. Each spatial task can be considered to be unidimensional at one level and multidimensional when examined at another. Therefore, the issue of dimensionality concerns the level at which the researchers examine a given construct (Reckase, 2009).

It may seem as though researchers studying the same level of dimensionality should arrive at similar conclusions, but they do not always appear so. Researchers sometimes disagree when identifying the dimensionality of cognitive abilities even at the same level, for several reasons. The cognitive abilities were mostly classified with factor analysis. Statistically, the categorization of the constructs depends on whether the common variance is strong enough to form an independent factor. Therefore, the technique of factor analysis is critical to the differentiation of the cognitive abilities and therefore informs the theoretical construction of cognitive abilities. For example, a review by Willis, Dumont, & Kaufman (2011) indicates that Thurstone and Spearman disagreed over whether a general factor existed because they used different factor analysis techniques. Another example can be found in Carroll's review (2003). When confirmatory factor analysis was invented, Gustafsson and Undheim (1996) were able to

suggest the possible existence of different types of broad abilities, *Gf* (fluid intelligence) and *Gc* (crystallized intelligence), while previous exploratory factor analysis sometimes found only a single general factor (Thurstone, 1947). Hence, the development of methods also had influence on whether the levels of cognitive abilities were detected.

The methodology for studying dimensionality. The dimensionality of cognitive abilities has been studied extensively with factor analysis. One representative study that was conducted by Carroll (1993) reanalyzed 400 studies to confirm the hypothesis of a hierarchical model of cognitive ability. A different way of identifying dimensionality can also be informed with the findings from psychological dual-task studies or cognitive error analysis. In dual-task studies, researchers tried to discover what two tasks people can do or cannot do at the same time, researchers gain clues about whether these two tasks share the same functional processes in the human brain (Pashler, 1994). Recent developments in brain imaging research have extended this approach by determining whether two tasks activate similar brain areas and are therefore using the same processes (e.g., Hubbard, Piazza, Pinel & Dehaene, 2005). In cognitive error analysis studies, researchers attempted to discover participants' thinking processes through analyzing their error patterns (Radatz, 1979; Tatsuoka, 1984). This pattern recognition analysis also has been made to connect with statistical models such as item response theory (e.g.,Tatsuoka, 1990).

To see if the convergent evidence of cognitive structure can be obtained from different perspectives, some researchers have tried to map the cognitive structure from statistical analysis on findings from brain studies. For example, Shaw et al. (2006) successfully associate higher general intelligence with the thicker cerebral cortex in later childhood. Comparisons from different studies have sometimes brought supportive evidence to the structure of cognitive abilities. For example, Guilford (1975, 1988) proposes that memory can be divided into memory

recording and memory retention. Recent findings of brain imaging studies suggested that memory is indeed processed in two different systems: one for appearance and one for location (Darling, Della Sala, & Logie, 2009). Although this is somewhat different from Guilford's categorization of memory, Darling et al.'s finding supports Guilford's claim that memory has subdivisions.

There are also differences emerging from the psychometric and neuroimaging approaches and these differences contribute to understand the dimensionality of cognitive abilities. For example, the review by Harris, Hirsh-Pasek and Newcombe (2013) compares and contrasts the studies of tasks in mental rotation and mental folding from both perspectives. These researchers find that, although factor analysis had suggested these two tasks were distinct, the brain activations while performing both tasks are in similar areas. Das, Kirby, and Jarman (1975) also examined the findings from brain connectivity studies that suggest that simultaneous and successive cognitive processes are separated (e.g., Luria, 1971), and they used factor analysis to confirm that these two were indeed separate latent factors.

The connections between methods. Although the findings of brain imaging studies and dimensionality analysis (e.g., factor analysis) are sometimes different, the statistical analyses to produce the result in each approach are not. Brain imaging studies identify the shared variance by examining the correlation between a participant's item response and brain activity (e.g., Shaw et al., 2006). The tasks shared enough significant variances in their activation to suggest that they may share cognitive processes (for a review, see Hubbard, Piazza, Pinel & Dehaene, 2005). Similarly, dimensionality analysis uses the covariance among people's performances of a task to identify those tasks that have shared dimensions. The difference made by statistically analyzing the items is that it can provide a more complete picture of the structure of these items/tasks by

analyzing the performance responses simultaneously, whereas, in brain imaging studies, researchers use the paired association tasks on a group of item responses to provide physical, as opposed to statistical, evidence of shared cognitive processes. Each approach, however, provides unique information about the cognitive processes, as well as converging or divergent evidence for understanding the components of cognitive structure. An example that may help to explain what is meant here is the distinction between a visual spatial working memory task, a task asking people the location of targets on briefly shown pictures, and calculation, a task asking people to carry out basic mathematical operations. While visual spatial working memory (VSWM) and calculation are separate constructs, studies have demonstrated their significant correlations (Alloway & Passolunghi, 2011; Ashkenazi , Rosenberg-Lee, Metcalfe Swigart & Menon, 2013; Bull, Espy, &Wiebe, 2008; Kyttala, Aunio, Lehto, Van Luit,, & Hautamaki, 2003; ; Mclean & Hitch, 1999). However, a study conducted by Zago and Tzourio-Mazoyer (2002), based on previous significant relations between these two, showed that, although VSWM and calculation shared common activation in the right-hemisphere frontal-parietal system in the brain, it was also indicated that calculation activated a left inferior parietal area which might not shared much with VSWM. This suggests that VSWM and calculation do not share all of their respective cognitive processing. However VSWM and calculation are not complete separate processes because previous studies have found them to be significantly correlated (e.g., Alloway & Passolunghi, 2011). It is not clear, though, which components are shared between VSWM and calculation, and further examination is therefore required to understand this relationship. An item level analysis would help break down the components, revealing the pattern of relationships between the items of VSWM and items of calculation.

Item response models

A common way to do item analysis is to apply item response theory (IRT) models. Specifically, the interactions between the person's characteristics and the item parameters produce the probability of correct responses on each item (Reckase, 1997). Therefore, the difficulty and discrimination parameter and the person's ability parameters are considered together (Reckase, 2009), and IRT provides information on the interactions between persons and items simultaneously. Item response models are applied mainly across a whole field. For example, recent studies have used item response models to understand learning (Embretson, 1991) and growth change (McArdle, Grimm, Hamagami, Bowles, & Meredith, 2009), and to explore the structure of a cognitive ability such as phonological awareness (Schatschneider, Francis, Foorman, Fletcher, & Mehta, 1999). An extension of IRT models, multidimensional item response models, was also used to identify the dimensionality of a ninth grade mathematics test (Ackerman, Gierl, Walker, 2003), and a test for English language learners (Reckase & Xu, 2014) or the dimensionality of mathematics knowledge (Reckase et al., 2012). Item response models have been applied extensively in the construction of achievement tests, but they are not often applied to exploring the structure of cognitive abilities, which is a major goal of the present study.

The Present Study

To understand further the structure of cognitive abilities with item response processes, I propose two approaches in the current study. The first approach targets a specific and well-known connection between cognitive abilities and academic achievement (VSWM and Mathematics Calculation) to study the condition when the true number of dimensions is

unknown. The purpose is to explore how the constructs might be interrelated at a multidimensional level. In this first part, I examine how strong the two constructs (VSWM and Calculation) can interact. When two tasks are closely related but can still be defined as different constructs, several methods were proposed to identify what are generally shared dimensions between them. The present study hopes to provide understanding of the connections among dimensionality of cognitive abilities.

In the second approach, I examine the possibility of detecting dimensions in subgroups when the true numbers of dimensions are known. The purpose of doing so is to test whether there are sensitive indices that can provide information to support the use of subscores. Specifically, the second part of the study asks whether it would be possible to detect the specific characteristics of an examinee's irregular item responses when a general ability dominates in the examinees' performance. Although diagnostic testing has been shown to identify students' weakness and strengths in some cases (e.g., Betts, Hahn & Zau, 2011), the decision whether to report subscores, such as mathematics or reading, has been a subject of debate in previous studies (Sinharay, Punah, & Haberman, 2011). As a solution, the second part of the study proposes to use sensitivity indices previously developed for detecting unusual patterns in examinees' responses. The application of these indices aims to test whether the values of subscores can be found, even when the data are well represented by a unidimensional model.

The remainder of this paper is ordered in the following way. Chapters Two and Three provide a literature review for the two approaches of the study. Chapters Four and Five explain the methods and results from the two enquiries. Finally, Chapter Six discusses the results, the limitations of each study, and the possibilities for future research.

CHAPTER TWO

THE DIMENSIONALITY BETWEEN VSWM AND CALCULATION

Introduction

Spatial abilities are suggested to be the critical cognitive abilities that influence a child's achievement in mathematics (Newcombe, 2010; Rohde & Thompson, 2007; Wai, Lubinski, & Benbow, 2009). Visual Spatial Working Memory (VSWM), in turn, is one of the spatial abilities especially important to mathematics. Although many studies have shown a significant relationship between the two (e.g., Alloway & Passolunghi, 2011; Ashkenazi , Rosenberg-Lee, Metcalfe Swigart & Menon, 2013; Bull, Espy, &Wiebe, 2008; Kytta¨la¨, Aunio, Lehto, Van Luit,, & Hautamaki, 2003; Mclean & Hitch, 1999; Rasmussen & Bisanz, 2005; St Clair-Thompson & Gathercole, 2006; Simmons, Wills, & Adams, 2012), training VSWM to improve mathematical ability has not yielded overall positive conclusions so far.

The multidimensionality of VSWM (Mammarella, Pazzaglia , Cornoldi, 2008; Pickering, 2001; Vecchi, Monticellai, & Cornoldi, 1995) and mathematics (Ashcraft, 1982; Dehaene & Cohen, 1995, 1997; Dehaene et al., 2003) may be the reason that training has not so far been successful. In order to determine which aspect of VSWM should be trained to improve mathematics ability, the shared dimensions between the two should be fully understood. The proposed study aims to examine this relationship on the basis of a simple form of VSWM task, which is a static VSWM, and one mathematics achievement task, a calculation test. The relationship between the static VSWM and calculation, which appears to be mixed in previous studies (Ashkenazi, Rosenberg-Lee, Metcalfe, Swiigart, Menon, 2013; Berg, 2008; Simmons, et al., 2012; Lee & Kang, 2002; Rasmussen et al., 2005), motivates the present study. The mixed results described above suggest the need to further examine the relationship between VSWM and

mathematics before training studies are conducted. Using a multidimensional item response theory approach, the current study aim to find the shared dimensions (if any) between VSWM and mathematics calculation. To do so, it is important to examine the generalized effect between cognitive tasks, the specific relationships between VSWM and mathematics, and the possible dimensionality of the two.

The Generalized Effect

Previous studies have asked whether training in one task can have a generalized effect on other tasks (Wright, Thompson, Ganis, Newcombe, & Kosslyn, 2008). An example would be using VSWM training to improve mathematics achievement. However, such effects have been elusive (Chase & Ericsson, 1981; Holmes, Gathercole, & Dunning, 2009; Jaeggi et al., 2011; Kyttala, Kanerva & Koresbergen, 2015; St Clair-Thompson, Stevens, Hunt, & Bolder, 2010; Van der Molen et al., 2010). VSWM is found to be trainable even when other cognitive abilities are not (e.g., Owen, Hampshire & Grahn, 2011). Even so, the generalizability of VSWM training is inconsistent across measures. For example, in an experiment reported by St Clair-Thompson et al., (2010), children aged five to eight were trained in working memory techniques, including image rehearsal strategies, for six to eight weeks. The training did not improve their performance in the active VSWM task (the block recall task), but it did improve mental calculation performance. This suggested that training VSWM to improve mathematics performance is possible, even when the training did not improve other type of VSWM tasks. This specific relationship between VSWM and mathematics might be bidirectional, such that mathematics can improve VSWM as well. For example, Lee, Lu, and Ko (2007) discovered that a year's mathematics training (given once a week) in the use of the mental mathematics abacus (a Chinese mathematics calculation tool) improved the performance of 12-year-old children in

simple spatial span tasks, but not in complex spatial span tasks (a task that involves both an equation and a presentation of dots in squares). Overall, training one type of VSWM task does not necessarily improve other types of VSWM tasks, but it might improve some types of mathematics tasks; additionally, training specific mathematics tasks can improve specific types of VSWM. These studies showed that VSWM and mathematics have some cross-generalizability, but it is task-specific.

Perhaps some processes shared between the trained tasks and improved measures underlie the above generalizable examples. Previously, studies had posited some potential mechanisms behind a successful transfer. An example is Wallace and Hofelich's (1992) study. In the study they had some participants complete training in mental rotation, who were then tested in a geometric analogies task. They also had the other groups of participants complete training in a geometric analogies task, and then tested them in a mental rotation task. Specifically, in the geometric analogies task, the participants were asked to view several shapes, all of which could be combined as an object which still involved one missing piece; they were then asked to also identify which of five options was the missing piece. In the mental rotation task, the participants were asked to identify whether the orientation of objects was standard or mirrored. They found that training in either task improved the performance in both, even though these two tasks did not share similar contexts. The possible mechanism, as explained by Wallace and Hofelic (1992), is Process Theory (Kosslyn, 1983), which stresses the importance of similar cognitive processes on task transfer. The argument for this effect is that when two tasks emphasize the same processes, training in either of them should be able to produce a generalized effect on the other. Additionally, this study demonstrated that the improvement between cognitive tasks can be bidirectional.

This phenomenon is not limited to mature adults. Kloo and Perner (2003), for example, found that training children ages three to four in an executive control task and a theory of mind task led to improvement on both. Executive control was trained and assessed with a Dimensional Change Card Sorting task (DCCS; Frye, Zelazo, & Palfai, 1995), and the theory of mind test was trained for and assessed using false-belief tasks. The improvement of both suggests that the generalization of training effects is possible with children. Kloo and Perner (2003) made a similar argument to Wallace and Hofelic's (1992), suggesting that there may be a specific underlying common factor in these two improved tasks.

Drawing on both studies, it seems that one should first identify the connection that links two tasks together if one expects to find a generalized effect between them. Once the connection is targeted correctly, the training effect can be such a powerful tool that it overcomes the surface dissimilarity, and carries over the training effect so that it can be applied to multiple tasks. Therefore, a generalized effect between VSWM and mathematics would have to be preceded by finding links between VSWM and mathematics performance.

The Relationship between VSWM and Mathematics Achievement

In general, VSWM and mathematics performance seem to be associated at many levels and across different populations. For example, in a sample of children with a mathematics disability, neural imaging studies have found co-existing deficits of VSWM and arithmetic (Ashkenazi et al., 2013). Even in typical children, studies have also found a close relationship between VSWM and mathematics (e.g., Bull et al., 2008). Indeed, VSWM may be more closely related to mathematics achievement compared to other cognitive abilities, such as verbal ability or central executive ability. For example, Bull et al. (2008) conducted a longitudinal study to examine the relationships between verbal, spatial, and mathematics abilities. They found that

different VSWM tasks (e.g., cosi span forward and cosi span backward) predicted mathematics achievement at different time points when reading ability was controlled. As further evidence for this relationship, St. Clair-Thompson and Gathercole (2006) examined the relations among specific working memory tasks, including shifting, updating, inhibition, verbal working memory, VSWM, and school achievement among children aged 11 to 12. They discovered that inhibition (r = .36, after controlling for VSWM) and VSWM (r = .50, after controlling for inhibition) had significant relationships with children's mathematics achievement.

It is not surprising that VSWM and mathematics are connected. VSWM and mathematics achievement are part of different broad factors of intelligence (*Gv* and *Gq*, respectively), and both factors are influenced by general ability. The shared influence might explain why some studies have shown that the relationship between VSWM and mathematics is mediated by general intelligence (e.g., Kyttala & Lehto, 2008). However, in some cases, the relationship between VSWM and mathematics is stronger than can be explained by their connection to general ability. An example of this is that VSWM was a significant predictor of mathematics ability in nine-year-old children, whereas non-verbal general ability was a non-significant predictor (Szucs, Devine, Soltesz, Nobes, & Gabriel, 2014). These studies showed that VSWM and mathematics might share a stronger link than their connection to general ability. Such a relation also showed developmental significance that gains in VSWM predicted later mathematics learning (Li & Geary, 2013).

One reason for this closer association may be that VSWM is required at every step in mathematics problem solving. For example, Männamaa, Kikas, Peets, & Palu (2012) discovered that VSWM was significantly correlated with all mathematics cognitive domains: recalling, computing, and applying. Alternatively, however, the same study also discovered that four out of

five latent factors of Working Memory (including latent VSWM factor) are not directly associated with latent mathematics factors. Although these two findings seem to be contrary, both of them could be correct—if the dimensionality of VSWM is taken into consideration. Specifically, it could be possible that different dimensions of VSWM tasks were correlated with different mathematics domains in the researchers' initial analysis, but when the researchers tried to extract a unidimensional latent factor from the VSWM task they used, this common factor might not associate with the mathematics latent factor domains much. Some studies, when they applied a multidimensional view of VSWM, have discovered similar findings that not all types of VSWM tasks are related to all types of mathematics problems (e.g., Kyttala & Lehto, 2008).

However, it also appears that the identification of dimensionality of VSWM is not consistent across different studies. One study splits VSWM into static and dynamic processing (Vecchi et al., 1995), whereas another study splits it into passive simultaneous VSWM, passive sequential VSWM, and active VSWM (Holmes, Adam, Hamilton, 2008; Kyttala & Lehto, 2008; Mckenzie et al., 2003). Different divisions of VSWM correlate with their own types of mathematical performance. For example, Kyttala and Lehto (2008) found that passive simultaneous VSWM predicts success at mental arithmetic problems, active VSWM predicts problems of geometrical understanding, and passive sequential VSWM predicts mathematics word problems in children ages 15 to 16.

What is more surprising is that very specific sub-areas of VSWM and mathematics interact differentially. For example, although mathematics calculation is significantly related to VSWM overall (Ashkenazi, Rosenberg-Lee, Metcalfe, Swiigart, & Menon, 2013; Berg, 2008; Simmons et al., 2012; Lee & Kang, 2002; Rasmussen et al., 2005), it also appears that not all types of calculation operations are related to VSWM. Specifically, Lee and Kang (2002)

observed that when they asked participants to do a dual cognitive processing task procedure (e.g., to hold images in their minds while doing calculation), asking people to hold visual spatial images impaired their ability to do subtraction, but not their ability to do multiplication. Similarly, Simmons, Wills, and Adams (2012) found that VSWM was significantly related to multiplication (r = .33) in children aged seven to eight, but that the same VSWM task was not related to addition (r = .11) in children aged five to six, even though both multiplication and addition are considered as part of mathematics calculation.

Reflecting on the fact that the connections are different when examining this relation at different levels, it is unlikely that VSWM training will have an effect on all types of mathematics problems. VSWM and mathematics are related overall but these correlations are specific to tasks, such that only the training of the related VSWM can improve the related mathematics skill. To give an example, it would be impossible to expect that training in a passive type of VSWM can guarantee significant improvement in geometry achievement in children ages15 to 16, in that they are not significantly related in the literature. Consequently, training will be effective to the extent that the links between VSWM and the types of mathematics problems are captured. However, without a systematic examination of the dimensions of VSWM and mathematics, it may be difficult to see what these connections are.

The Dimensionality of VSWM

As is the case for many cognitive tasks, the dimensionality of VSWM can be identified as both unidimensional and multidimensional. Some studies consider VSWM as a unitary construct, particularly when VSWM is compared with other working memory tasks. For example, a single, independent factor of VSWM can be isolated through confirmatory factor analysis when other working memory tasks are in the same model (Kane, Hambrick, Tuholski, Wilhelm, Payne,

& Engle, 2004; Miyake, Friedman, Rettinger, Shah & Hegarty, 2001). Similarly, Kane et al. (2004) were able to separate VSWM from general working memory, and they confirmed that one VSWM latent factor can be extracted from three VSWM tasks (rotation span, symmetry span, and navigation span). A unidimensional view of VSWM defines it as a psychological construct that temporarily holds visual and spatial information (Baddeley, 1986; 2000; Quinn et al, 2008; Reuhkala, 2001). However, even with this unidimensional view of VSWM, different subsets of VSWM (e.g., static VSWM and dynamic VSWM) were used in different studies, causing the variations on the use of VSWM tasks across those different studies. The composition of VSWM therefore can be viewed as multidimensional. An example that can be illustrated here is that the cognitive processes of static VSWM and dynamic VSWM are independent (Vecchi et al., 1995), meaning these processes can be considered two separate dimensions within VSWM.

Furthermore, when examining VSWM more closely across many different VSWM tasks, there are more than two sub-dimensions. For example, Mammarella, Pazzaglia, and Cornoldi (2008) examined the components of VSWM using confirmatory factor analysis on third and fourth graders. They found that four dimensions co-exist within VSWM: sequential-spatial, simultaneous-spatial, visual, and visuo-spatial active factors. In their study, the division of dimensionality of VSWM seemed to be task specific, such that each factor could be extracted from two or three similar VSWM tasks, suggesting that these similar tasks shared a common dimension. However, in each VSWM task, the cognitive processes can be hypothesized to reveal more than one dimension. The hypothesis, theorized from Kosslyn's (1983) classical work on mental images, speaks of two types of spatial processing within image generation. One is with a categorical spatial relationship, and the other is with a coordinating spatial relationship. A categorical spatial relationship is the categorizing processing by which people remember the

appearance of stimuli, and a coordinating spatial relationship is the location recoding process by which people remember where stimuli are located. This suggests that people understand spatial relationships through two different systems, according to their appearances and the location of the objects under scrutiny.

Similarly, Vecchi et al. (1995) found that the amount of information and the structure (the location of the objects) of the VSWM stimuli are two important factors that influence the storage capacity of VSWM. Concluding from these behavioral studies, how people remember these changes in structure and how people remember the number of stimuli may be allocated to different dimensions. This multidimensional view of VSWM is not only supported by researchers in behavioral studies (e.g., Logie and Van Der Meulen, 2009), but also from neural imaging studies. A recent neural study has demonstrated that there might be two separate cognitive subsystems operating within VSWM (Darling, Della Sala, Logie & Cantagallo, 2006). Overall, these findings point to the possibility that multiple dimensions may exist within one type of VSWM task.

Few studies, however, have investigated how or whether these dimensions in one type of VSWM (e.g., static VSWM, the basic form of VSWM) can be separated. Based on the literature, it can be speculated that there may be at least two dimensions in static VSWM, depending on the location and appearance of the stimuli in the task. Furthermore, studies also imply the possibility of global VSWM processing configuration (e.g.,Borst, Ganis,, Thompson & Kosslyn, 2012) when people combine different memory systems when performing VSWM tasks. Consequently, possible interactions between these two memory systems might lead to an increase in the dimensionality of the VSWM task.

The Dimensionality of Calculation

Although mathematics achievement has been considered as multidimensional (for a review, see Mix & Cheng, 2012) with hierarchical levels (Männamaa, Kikas, Peets, & Palu, 2011) overall, few studies have looked into the cognitive requirements for performing a single mathematics task, which may also require a multidimensional perspective. A few studies have clearly demonstrated how separate cognitive processes operate within a single mathematics calculation task (Singley & Anderson, 1989; Rickard, Healy, & Bourne, 1994). Among these, it is found that learning one type of mathematics calculation does not transfer successfully to another type of mathematics calculation without some variations. This suggests additional understanding is required, and therefore posits the possible existence of multidimensionality even in a single mathematics calculation task.

The difficulty of transfers on calculation skills has shown some specific directions (Singley & Anderson, 1989; Rickard, Healy, & Bourne, 1994). For example, Rickard et al. (1994) found no transfer of skills between division problems, even when both questions required the same rules but with slightly different formats (the participants practiced 56 / 7 = 8 and were tested on 56 / 8 = 7). However, the transfer did take place between similar multiplication problems (e.g., 6 x 8 = 48 vs. 8 x 6 = 48). A similar pattern can be observed when comparing addition and subtraction in other studies. For example, Campbell et al. (2006) found that skill transfer took place between addition problems (6 + 8 = 14 vs. 8 + 6 = 14), but not for subtraction (14 - 6 = 8 vs. 14 - 8 = 6). Such a transfer effect seems to be found only in particular situations in mathematics, which may suggest that the mental representations of different types of mathematics calculation can be quite specific even though the problems that do not transfer are quite similar on the surface (e.g., 57 / 7 = 8 vs. 56 / 8 = 7). The above studies used adults as

participants; it is unknown whether a similar transfer difficulty is held to be true in children. However, it seems that for children, transfer between calculation problems required even more additional understanding. For example, Gilmore (2006) tested eight- and nine-year-old children's understanding of inverse problems (4+ ? = 6). They found that the children's performance was affected by several dimensions within the tested tasks (e.g., conceptual mathematics understanding and the locations of the missing parts of the equations). The difficulty of transferring these processes across children and adults in similar calculations may result from the complexity of the cognitive demands in different types of mathematics problems.

Specifically, a review by Ashcraft (1982) claims that fact retrieval, carrying, and encoding in mathematics requires different kinds of brain function. If this is so, addition and multiplication may be related to fact retrieval, whereas subtraction and division are not, and therefore they belong to different cognitive processes in calculation. Neurodevelopmental evidence further supported that mathematics representations are multi-faceted and that different types of mathematics representation are associated with specific areas of the brain. A triple-code model proposed by Dehaene and his colleagues (Dehaene & Cohen, 1995, 1997; Dehaene et al., 2003) suggests that there are three brain areas responsible for the codes of representation in mathematics tasks: the bilateral posterior superior parietal lobe (PSPL), the bilateral horizontal segment of the intraparietal sulcus (HIPS), and the left angular gyrus (AG). The PSPL involves tasks such as manipulating numbers or their orientation; for example, moving numbers when executing subtraction equations or when carrying out more complex operations such as borrowing. The HIPS involves the non-verbal quantity system, used for comparison or estimation tasks. The AG involves learning or retrieving rote verbal facts in mathematics, such as multiplication tables. Overall, the diversity of brain pathway associations in mathematics

problems suggests that different cognitive process might be called upon in solving different mathematics problems.

Laboratory and physiological evidence both suggest multidimensionality in VSWM and mathematics, respectively. There are also studies that demonstrate that there might be multiple connections between VSWM and mathematics problem solving. However, less is known about how the multidimensionality of VSWM and calculation interact. From the understanding of previous studies, the relations between VSWM and calculation have mixed results. It is less likely that 1) an unidimensional model fits across both tests, or 2) each test belongs to separate dimensions. Alternatively, the relation is likely to be intertwined. For example, hypothetically, the items of VSWM that require the memory of location may be likely in the same dimension as missing terms problems because they both require identification of the locations. It is further possible that items of VSWM that require more storage might be on the same dimension as multi-digit division problems because they both require more memory spaces. To examine these hypotheses, a detailed item analysis on the division of dimensionality between VSWM and Mathematics is required.

A Measurement Approach: Multidimensional Item Response Theory

Previously, several approaches have been developed to examine the shared connections between VSWM and mathematics at an item response level. One approach is from cognitive psychology. Researchers often examine the level of connection between two tasks by using the dual task paradigm or error analysis. For example, in terms of dual tasks, if the two tasks interfere with each other when they are presented at the same time, they are considered to share the same cognitive processes. Lee and Kang (2002) observed that in dual processing tasks, holding visual spatial images impairs the ability to do subtraction. A second approach is to

identify locations of brain activity when doing tasks. When both tasks activate the same brain areas, it is inferred that those two tasks may share the same neural connections in the brain. For example, Ashkenazi et al. (2013) found such a connection between VSWM and symbolic number processing: the right intraparietal sulcus was activated during both tasks in typical children aged seven to nine.

Although both approaches provide evidence for the possible shared cognitive processes of the paired VSWM task and mathematics task, it might be difficult to use brain-imaging studies to understand how the items and person interacted on an overall scale simultaneously. Second, item response analysis in experimental design is often limited to a small sample size. Therefore, to be able to generalize a representative understanding of this relation, a larger-scale study is needed.

A more sufficient technique to understand what is commonly distributed among items is using a psychometric approach. A prevalent model for this application is an extension from IRT models: multidimensional item response theory (MIRT). Specifically, by applying a MIRT model, it is possible to examine item responses in a test simultaneously within a large sample. In this approach, any close proximity among the items is fully revealed. That is, the test items that have the closest proximity can be categorized when they are assessing in in the same dimensions (Reckase, 2009).

Some have considered item factor analysis and MIRT to be the same concept. However, item factor analysis is more concerned with the common variance among variables, and it removes the differences between items, whereas MIRT is more concerned with the item characteristics of a test. In psychological studies, where the unidimensional assumption of IRT is often violated, MIRT was developed to understand the cognitive structure that was formed from

the complex interaction between persons and item responses over two dimensions. In contrast to the function of the data reduction method in factor analysis, MIRT seeks to provide information that will help with the understanding of the structure of the data matrix using as many dimensions as are needed to represent relationships (Reckase, 2009).

MIRT has been applied across different studies of dimensionality. A primary purpose of the use of MIRT is the analysis of dimensionality of a test (e.g., Ackerman et al., 2003; Walker, Zhang & Surber, 2008; Reckase, McCrory, Floden, Ferrini-Mundy & Senk, 2015).). In such cases, the multidimensionality of the test often comes from the influence of another ability. For example, Walker et al. (2008) have demonstrated that students' reading ability influences their probability of success in certain items of the mathematics test. Ackerman et al. (2003), when analyzing a ninth grade algebra test, also discovered that although the test content was considered to be unidimensional, the items were multidimensional in that some items were only measuring spatial ability. In some cases, MIRT also can be used to identify competing models. For example, Leighton, Gokiert, and Cui (2007) applied nonparametric methods to identify the dimensionality of a science assessment; the purpose was to distinguish whether a content-based model or a psychological process-based model fit better with the dimensionality of the tests. Overall, these studies demonstrate the practicality of MIRT when the multidimensionality among test items is hypothesized. Therefore, to understand the dimensionality between VSWM and calculation, the current study applies MIRT in its approach.

Research Questions

1. What is the dimensional structure of the VSWM task?
2. What is the dimensional structure of the mathematics calculation test?
3. Do VSWM and mathematics calculation share dimensions?

CHAPTER THREE

ARE SUBSCORES WORTH REPORTING?

The Use of Subscores

The proposal to use subscores comes from the practical needs of examinees and practitioners, who often notice the useful information revealed by subscores, even when the total scores do not show any significant differences between different examinees. For example, Lawrence and Curley (1989) discovered that although the total scores between males and females are quite similar, these two groups have different subscores for their reading comprehension in the SAT verbal test. Haladyna and Kramer (2004) also found that there are consistent subscore differences worth reporting to the participants in a National Dental Examination (2004) test program. In both cases, the test results carried high stakes, and detailed information on subscores seemed to be in need to help examinees improve their future performance.

Although it seems straightforward to report the differences between subscores and total scores to satisfy a market need, the subscores need to be reported following pre-determined procedures, and the psychometric value of using subscores must first be demonstrated. It is suggested that inaccurate subscores may have a worsening effect on students' remedial actions if they are fed false information (Sinharay et al., 2011). In addition, the standards for educational and psychological testing suggest that reporting subscores calls for adequate reliability and validity as criteria (Sinharay et al., 2011;AERA, APA & NCME, 1999-2008). Overall, the use of subscores have supported for reliability for intended used and for the validity of inference from the scores.

There has been a continuing debate about the value of using subscores. In Classical Test Theory, researchers mostly failed to find their value in practice (e.g., Sinharay, 2010). In particular, researchers have shown that in order for subscores to have added value, the necessary condition is that the observed subscores should be more predictive than the predictor that was generated from the total score on the true score of the subscore. Otherwise, this would suggest that subscores generate more error by reporting its observed value (Sinharay, 2010). To evaluate this, Sinharay (2010) developed a criterion for judging this condition. The criterion works by using the mean square error (MSE) on the predictors to generate the error variance for the subscores and the total score. By using these two values, researchers are able to compare the proportional reduction in mean square error (PRMSE). A higher PRMSE value of subscores than its value in the total score suggests that the subscores have added value beyond just reporting the total score. Using this approach, it shows that in most conditions, subscores do not offer any added value. This was also confirmed by Babenko (2013), who tested three different Classical Test Theory approaches: attenuation, r' (which is the corrected correlation); the proportional reduction of the mean squared error, PRMSE (Haberman, 2008; Sinharay et al., 2007); and the augmented method (Kelley, 1923). Babenko (2013) found that none of the approaches demonstrated that subscores confer additional added value over and above that of the total score.

This finding leads researchers to question when or whether it is worthy to report subscores. To test this, Sinharay, Puhan and Haberman (2010) used the two-parameter logistic multidimensional item response (MIRT) model to generate simulated multidimensional data, and they experimented with the conditions when the subscores may have further added to value. By using estimated parameters from operational data in simulation, they found that in order to have

subscores reported, several conditions must be satisfied in advance. First, a sufficient number of items (say, 20) should be included in each subset. The subsets should be distinct from each other (say, the correlation should be less than $r = .85$), so that the psychometric properties of the subscores meet acceptable standards. However, because these two conditions are necessary requirements, if for example subscores have a correlation of .70, then they might not have any added values even if the subscore has only 10 items. Therefore, to generate meaningful subscores, a long test that has low correlations among subsets would be required.

Another issue is that many current tests designed to screen participants over a single continuum would probably not be able to provide essential, reliable, or valid subscores information because the items that did not fit unidimensional models would be deleted from the test (Yao & Boughton, 2007). The attempt to fit unidimensional model is prevalent in measurement area, but as Tate (2004) posits, harmful results may be generated when researchers apply a unidimensional model to data that are multidimensional. Finding methods to generate meaningful subscore information has become another potential solution to this problem.

Multidimensional Item Response Theory (MIRT) Approach

Following this notion, another group of researchers tried to show the merit for using a multidimensional item response approach to generate subscores. Specifically, researchers compare different methods on either simulated or real data. For example, Haberman and Sinharay (2010) found that by using subscores generated with the two parameter logistic MIRT-based models approach, the added values of the subscores were slightly improved over the value of the raw total score. However, in their five datasets, it appeared that the data that were unidimensional (Test B and Test D) did not always shown improved precision with MIRT model based subscores. Other researchers have found similar results. Stone, Ye, Zhu, and Lane (2010),

using an eighth grade unidimensional mathematic test, showed that the augumented score reports subscores more precisely than do the subscores generated by the MIRT models.

Overall, although it seems as though the result of the model fit is based on the psychometric nature of the data, MIRT models show some promising effect. This has motivated researchers to investigate different estimation methods to stabilize the outcome of reliable subscore estimations from the MIRT models. For example, de la Torre and Patz (2005) propose that using a hierarchical Bayesian framework can improve the estimations in MIRT testing. Regardless of how many estimations methods are proposed, the estimation methods still require a good fit with the MIRT models, which would require the data to be multidimensional in nature.

The Psychometric Property of the Data

The decision of reporting subscores seems then to be defined by the nature of the constructs that are being measured. Haberman and Sinharay (2010) conclude that when subscores are valuable to report, the subscores and total score are usually the result of measuring different constructs. However, this is not always the case.

As introduced earlier in Chapter one, Researchers often find that subtests of the intelligence test (e.g., the Wechsler Intelligence Scale for Children) or achievement test (e.g., a composite test of reading, language, and science) fail to provide useful diagnostic information over and above the composite score (McDermott, Fantuzzo, & Glutting, 1990; Sinharay et al., 2007). Considering the hierarchical structure of cognitive abilities, this might not be a surprise. For example, Templin and Bradshaw (2013) compared several unidimensional and multidimensional diagnostic models to fit data with a hierarchical structure. It appears that when the data were hierarchically multidimensional, such that students needed to master the pre-requisite traits before they could master other higher order traits, the results still showed that the

unidimensional models fit better. Therefore, the linearity of the traits made it impossible to show any added value from tested subscores with those traits, because they would be masked by the linear relationship of the attributes. Consequently, students who had never mastered the low level skills would suffer from the fact that they were notified only of their overall scores.

Furthermore, detecting the difference between subscores and the total score can be beneficial for determining whether there are differing psychological processes for each subgroup or individual child. These different processes could be divided into two categories: children who overachieved compared to their estimated ability, and children who underachieved compared to their estimated ability. Both situations could happen with the curriculum modifications. Some research has demonstrated that differences in curriculum led to differences in student performance (e.g., Geier et al., 2008). The first category could create a subdimension. For example, some children may show overachievements on test scores if they were the only group being taught the concepts that were targeted in the exam. Another example of overachievement could be that some groups of children are more familiar or comfortable with the context of some test items than other children. For example, if a group of items were based on basketball rules, children who are familiar with basketball could grasp the concepts better and therefore potentially perform better. Conversely, the second category could also create a subdimension if a small subgroup of children did not receive proper education while all other children did. An example of this might be that a group of children did not attend class some days if they were all sick because of an infectious disease that was passing around the class. Reporting subscores might be able to reflect theses variations if there is any information that can be taken into consideration. Therefore, methods that could detect whether differences in subscores might indicate certain subgroup differences that would be masked if only looking at the total score.

Last, researchers have suggested a different approach that subscores reporting might not be useful for everyone (Raymond & Feinberg, 2015), but only for those students whose total scores cannot predict their true subscores. Therefore, deciding subscores' values by comparing these with total scores might not be the best method for those individuals, and finding alternative procedures to determine those subgroups' differences to report could still be meaningful for subgroups' future remediation (e.g, Raymond & Feinberg, 2015).

Person Fit Indices

To sum up, the current methods for detecting the use of subscores may not offer much value when unidimensional model is a good fit. In order to explore this problem and add understanding to the gaps in the previous literature, an alternative measure is needed to determine whether it is possible to extract information from individual differences the data that fits a unidimensional model well.

To detect individual differences in testing, one possible approach is through the methods for detecting outliners. Specifically, a line of research in measurement has been trying to generate optimal appropriateness indices to identify irregular response patterns in the data (Drasgow, Levine, & McLaughlim, 1987; Levine & Drasgow, 1984). Irregular response patterns could result either from cheating behaviors (which increase the correct response rates) or from participants' omitting behaviors (which lower the correct response rates). Both patterns are threats to test reliability and validity. Among all the appropriateness indices proposed in previous studies, the *Lz* and *T2* index was shown to be most efficient in detecting whether examinees are showing unusual response patterns (Drasgow et al., 1987). It is worth noting that although the detecting rate of *Lz* can very depending on the ability range of participants, the detecting rates generally are high. For example, *Lz* showed a 94-99% detection rate regarding 30% suspiciously

high scores on low ability group (0- 30% ability range), and a 92-98% detection rate with 30% suspiciously low scores with a high ability group (65%-100%) and 70-82% detecting rate with a middle ability group (31-64 ability range). *T2* has also shown a 96%-98% detection rate regarding 30% suspiciously a high scores on a low ability group (0- 30% ability range), and a 90-99% detection rate with 30% suspiciously low scores with a high ability group (93%-100%) and 60-81% detecting rate with a middle ability group (31-64 ability range). Given its relatively high detection rate compared to other indices, as well as the ease of using it in future testing applications, these two indexes was chosen to be our primary detection method in the current study.

The goal of using *Lz* and *T2* is to determine whether the sensitivity of the index could be utilized for detecting subscore differences. The construction of *Lz* is as follow (Drasgow et al, 1987):

u_i= dichotomously scored item response (1 or 0) for item *i* (*i*=1.2……..n)

$\hat{a}, \hat{b}, \hat{c}$ are item parameters estimates. *D* is the constant 1.702

$$Lz = \frac{\ell o - M(\hat{\theta})}{[S(\hat{\theta})]^{1/2}}$$

The logarithm of the three-parameter logistic likelihood function evaluated at the maximum likelihood estimate $\hat{\theta}$ of θ.

$$\ell o = \sum_{i=1}^{n} [u_i \log P_I(\hat{\theta}) + (1 - u_i) \log Q_i(\hat{\theta})]$$

$$Q_i(\hat{\theta}) = 1 - P_i(\hat{\theta})$$

$$P_i(\hat{\theta}) = \hat{c_i} + \frac{1 - \hat{c_i}}{1 + \exp[-D\hat{a_i}(\theta - \hat{b_i})]}$$

The conditional expectation of ℓ_o:

$$M(\hat{\theta}) = \sum_{i=1}^{n} \left[P_i(\hat{\theta}) \log P_i(\hat{\theta}) + Q_i(\hat{\theta}) \log Q_i(\hat{\theta}) \right]$$

The conditional variance:

$$S(\hat{\theta}) = \sum_{i=1}^{n} P_i(\hat{\theta}) Q_i(\hat{\theta}) \{ \log [P_i(\hat{\theta}) / Q_i(\hat{\theta})] \}^2$$

Overall, *Lz* is comparing the estimated logistic likelihood function with its conditional expectation value and then divided by its conditional variance to generate the standardized value. The rationale focusing *Lz* is that it has demonstrated its sensitivity in identifying unusual patterns within data. The application of the *Lz* index may be able to detect these even from a seemingly unidimensional data structure and may be to help researchers to, for example, pinpoint the children who have unusually low or high responses for some items according to their performances on other items.

The Formula of *T2* is as follow (Tatsuoka, 1984).

$$G_i = \frac{1}{N} \sum_{j=1}^{N} P_{ij}(\hat{\theta}_j), \text{ and } \bar{G} = \frac{1}{n} \sum_{i=1}^{n} G_i,$$

$$T2 = \frac{\sum [P_i(\hat{\theta}) - u_i][G_i - \overline{G_i}]}{[\sum P_i(\hat{\theta}) Q_i(\hat{\theta}) (G_i - \bar{G})^2]^{1/2}}$$

Although previous studies provided good detecting rates on both *Lz* and *T2*, It is possible that we might not obtain the same promising result. A review conducted by Meijer and Sijtsma (2001) has suggested that type of misfitting patterns, test length, and trait level (theta) could vary the result of person fit indices. The misfitting patterns of the current study are less powerful comparing to previous studies. Although the *b* parameters were shifted between ±2 range, it is still less strong than altering the responses to all zero. Secondly, it is worth noting that in previous studies, person fit indices often use true θ, and previous studies (Snijders, 2001; Magis, Raiche, & Beland, 2012, Nering, 1995) have pointed out that *Lz* is not an optimal index when

trying to apply it to real data. In our studies, we however use estimate θ to be closer to a real life condition on large scale testing.

Research Questions

1. Can person fit indices detect the subscores created by the differences between subgroups?
2. If so, under what conditions can person fit indices detect the differences?

CHAPTER FOUR

STUDY ONE RESULTS

Method

Participants

537 third graders participated in this study. Data was from the IES funded project: Spatial Ability as a Malleable Factor for Mathematics Learning. The data were collected in spring 2013 and spring 2014. Thirty-three schools, across rural, suburban, and urban areas in Michigan and Chicago, have participated this study. The rate of average free/reduced lunch was 42.87% (range from 0%-99%). The mean age was 9.05 year (SD = .36). There were 238 boys in the sample.

Measures

VSWM. Visual Spatial Working Memory was adapted from the Kaufman ABC test (Kaufman & Kaufman, 1983). The Kaufman VSWM task measures the static types of VSWM. It has been used widely in previous studies (e.g., Mariani & Barkley 1997; Wang, Woodin, Kreps-Falk, & Moss, 2000). In each trial a 14 cm x 21.5 cm grid is divided into two different numbers of squares (e.g., 3 x 3 and 4 x 3). On each grid, different objects are shown inside the squares. The locations of these objects on the grids are what the children need to remember and mark on their sheets. To vary the difficulty of the items, the numbers of objects were between two and seven and were randomly displayed on the squares in each trial. There were 17 items in this test.

For group administration, the original paper version was transformed into a computerized version that was displayed on the computer. Children viewed these images from 100 centimeters. At the beginning of each task, children were told that the stimulus would be displayed for only 5 seconds, and they needed to remember where these pictures were. Each slide was set to disappear after 5 seconds and the screen arrangement was replaced with an identical blank grid

(with no objects inside the squares). When the pictures were made to disappear from the screen, the children marked where these pictures were on their booklet. The test was first introduced with three practice items. The children received feedback on the correctness of their answers on these three items. They were also allowed to compare their results to the correct answers. Before the last practice item they were also told that when the practice session was over there would be no more feedback; each picture would be given only once so they needed to look at them carefully. It took approximately 20 to 25 minutes for each child to complete the test. The reliability of the test on the sample that was used here was $\alpha = .67$.

Calculation. The Calculation tests were adapted from the Michigan focal point on matched grade level mathematics questions. There were 24 items. The questions included both single digit and multi-digit problems. These problems included addition, subtraction, multiplication, division, fraction and missing terms problems (A list of these items is in Table 2). The test was administered in groups. Children were given paper and pencil copies of these questions. It took approximately 20 minutes for the children to complete the test. Test reliability on the sample that was used here was $\alpha= .75$.

Analytic procedure

As the primary purpose of the study is to examine the dimensionality of cognitive structure within cognitive abilities, the current study used two types of cognitive abilities: visual spatial working memory and calculation to explore their possible shared dimensions and hope to reveal the cognitive processes. The item responses from both measurements were used for the analysis. The data analysis plan closely followed the MIRT analysis steps from Reckase, McCrory, Floden, Ferrini-Mundy and Senk (2015). The dimensionality analysis was also carried out with more confirmatory approaches, based on the results of expert reviews, convergent

evidence from the literature review of neural imaging studies, and error analysis on the patterns of item responses. These analyses are described in more detail below.

Data analysis: exploratory approaches

Identifying the dimensions. The first analysis was to examine the item responses from these two tests separately for tests of unidimensionality. This is because it only makes sense for further multidimensional item analysis if both of the tests presented multidimensionality. To do so, DIMTEST (Stout, Nandakumar, Junker, Chang, & Steidinger, 1992) was carried out to test whether the matrix of item responses is well fitting with a single person parameter (one θ). So, item responses from VSWM were tested for whether a single VSWM construct can represent this dataset and item response from calculation also were tested for whether a single calculation construct can represent this calculation dataset. After that, both datasets were combined to test whether a single general ability construct can well represent this combined dataset (denoted as CVC for the following analysis).

Dimensionality. If the unidimensionality is rejected, the item responses from CVC then were used to proceed with further MIRT analysis to determine the number of dimensions in the data. First, a parallel analysis was conducted to compare the eigenvalues between the CVC data and the parallel-simulated dataset. This parallel dataset was simulated to have the same sample size and the same item proportional correctness but the relationships among the items were random. The eigenvalues from both datasets were compared. The numbers of eigenvalues in CVC that were bigger than the eigenvalues from simulated dataset were used for initial indication of the number of dimensions to support this dataset. After that, an exploratory MIRT analysis was done using a MIRT package in R (Chalmers, 2012). The *a* parameters generated from the MIRT program were also compared with the *a* parameters generated from the same

estimated procedure from FlexMIRT (Cai, 2012) to confirm similar parameter values were generated from different programs. These *a* parameters were then used to calculate angles between variables. The generated angular distances were then used to cluster the items. Ward's method of cluster analysis in R was used for the clustering procedure.

Data analysis: confirmatory approaches

The distribution of these items on the dimensions could also vary in several different scenarios. One possibility is that even though there are three dimensions, the data still present a simple structure. In this case, items don't cross load on dimensions. For example, there might be two dimensions that exist in these VSWM items and one dimension exists in the calculation items. Another possibility is that these three dimensions could come from a more complex structure but items don't cross load on different dimensions, such as one dimension is spatial, one dimension is calculation, and another dimension is the interaction between spatial ability and mathematics ability (such as this interaction creates a different dimension). The third possibility is that each item is distributed in multiple dimensions. For example, these items all require memory of location, in which this "memory of location" could be one dimension. Some of these items might also require "counting", which will be another dimension. In this case, the dimensions in this dataset are quite a complex structure.

To give a clearer picture on these dimensions, content analyses were also conducted to generate several possible explanations for these dimensions. Ackerman et al. (2003) suggested that a content analysis might provide directions for further assessments of dimensionality. Therefore, the current study conducted content analyses to construct confirmatory models. These contents analyses included an expert review, a review of neural imaging studies, and an error analysis.

Content analyses. In the expert review, three content experts reviewed the contents of the VSWM and calculation tests and identified the possible dimensions based on the reviewers' expertise. They coded items according to the cognitive process of the items into three different categories: number (mathematics), location (spatial), or both. In the neural review, a through literature review was conducted to decide each item's category in a neural model. If the results within the literature conflicted, the majority was followed. In the error analysis, the 537 item responses are recoded to two different directions. One was recoded with whether the children correctly answered the number of items in each VSWM stimulus, and the other was recoded with whether the children correctly identified the location of these items in each VSWM stimulus.

Confirmatory MIRT Analysis. The exploratory approaches are meant only to provide the possible dimensionality of the dataset. A two-parameter logistic MIRT model was applied to provide further understanding of the cognitive structure of these two tasks. The results tell us whether the calculation performance requires certain high VSWM dimensions. To conduct this step of analysis, confirmatory MIRT analyses (e.g., Mplus) were used.

Results

Exploratory approaches

Multidimensionality of the data. First, to identify whether there are shared dimensions between VSWM and calculation, the multidimensionality of VSWM and calculation need to be confirmed separately. Separate tests for unidimensionality were carried out with DIMTEST 2.0 for each test. The null hypothesis of unidimensionality was rejected for both VSWM, $t = 3.5471$, $p = .0002$, and calculation, $t = 3.1479$, $p = .0002$. These results indicate that multidimensionality might exist in VSWM and calculation separately. A third test for unidimensionality was also

carried out for the data that combined both VSWM and calculation (denoted as CVC below). This test also rejected the null hypothesis of unidimensionality, $t = 5.9791$, $p <. 0001$.

Dimensionality. Because the multidimensionality of the CVC was confirmed, the next step was to identify how many dimensions are in the data. To do so, parallel analysis (Horn, 1965) and hierarchical cluster analysis were carried out. Parallel analysis provides information about the possible numbers of dimensions of the data by comparing the real data with simulated data, which is a parallel comparison with another dataset with the same sample size, the same number of items, and the same item difficulty distribution among items, but no relations among these items. To simulate such a dataset, I randomly replaced the item responses in each item in a loop to create a random distribution of the responses. For example, each item has 537 responses (0 or 1). For example, for item four there are 537 responses. 100 responses are zero and 437 responses are one. I randomly replaced these responses 537 times, so that the item still had the same difficulty but there will be no relations with other items. In this way, the proportion of correct scores was intact, which ensures the simulated data has the same item difficulty distribution as the original data. The simulation went through 100 replications. Both real data and simulated data each generated their eigenvalues from their correlation matrices. All the procedures were conducted in R 3.2.4. The results indicated that the biggest first four eigenvalues in the real data were 5.60, 2.28, 2.00, and 1.56. For simulated data, the mean and standard deviation of the first three biggest eigenvalues were 1.57(0.03), 1.50(0.03), and 1.46(0.03). It appears that the first three eigenvalues of the real data are much bigger when compared to simulated datasets. Therefore, the result suggests that the interactions between persons and items can be represented by three dimensions.

To confirm this number of dimensions, a hierarchical cluster analysis was also carried out on the CVC. The analysis was carried out with two different approaches. One approach is using the item score (binary responses) to generate a distance matrix, and another approach is using angular distance generated from a MIRT model estimating *a* parameters with six dimensions. In the approach utilizing MIRT *a* parameters, the six dimensions were decided by doubling the numbers of dimensions from the parallel analysis. Both approaches were conducted with the hierarchical cluster analysis package in R 3.2.4 with Ward's algorithm. Both approaches showed three clusters, but the combination of items in the clusters were slightly different between the two methods. Based on the results of the parallel analysis and hierarchical cluster analysis, the number of dimensions seems to be established as three dimensions.

Confirmatory approaches

For the next step, confirmatory MIRT analysis was carried out to investigate the items located on each dimension. Note that although the number of dimensions was identified, these methodologies were exploratory approaches. Therefore those results only suggest there are three major dimensions in this data. To identify the meaning of the dimensions from the perspective of theoretical understanding, content analyses were conducted to further assist the constructs on hypothesized cognitive process in the combined data.

Expert review. Three experts reviewed the items from both VSWM and calculation. They categorize these items into spatial (the location of stimulus), mathematics (quantify numbers), or an interaction between these two processes. Two of the three experts were graduate students who have done research related to spatial ability and mathematics performance, and the third expert was a professor who has expertise in this area. Prior to the review, the three experts had a meeting to discuss what possible dimensions might exist in CVC. Then each expert

individually categorized items into three categories: spatial process, mathematical process, and if the item has both processes then it was categorized as both. The percentage of agreement was calculated among three experts. Because the agreement is quite low (39%), the models suggested from each expert were carried out separately to identify the best-fit model.

For model comparisons, because these three models were not nested, the approach to use chi-square difference tests to check the relative model fit is not an option here. The best model fit therefore was identified by finding the lowest Akaike information criterion (AIC; Akaike, 1974), Bayesian information criterion (BIC; Schwarz, 1976), and Sample Size Adjusted Bayesian Information Criteria (SABIC) from these models. These criteria were suggested by previous studies (Bozdogan,1987; Burnham & Anderson, 2004; Lubke & Neale, 2006; Raftery, 1995; Vrieze, 2012). Using these indices, it appears that the second model is the best fit (see Table 1).

Table 1.

Model comparisons among three expert models

Models	Number of parameters	AIC	BIC	SSABIC
Model 1	85	19097.599	19461.909	19192.091
Model 2	84	18952.050	19312.073	19045.429
Model 3	85	19039.304	19403.614	19133.796

Table 2.

The item distribution of Expert 2 Model

Dimensions	VSWM	Calculation
Spatial (location)	vswm3x3_85 vswm3x3_95 vswm3x3_105 vswm4x3_55 vswm4x3_66 vswm4x3_76 vswm4x3_87 vswm4x3_97 vswm4x3_107 vswm4x3_117	
Mathematics (Number)		C1_1+6 C11_347-129 C12_8753-3497 C3_378+146 C4_4098+1567 C9_7-3 M1_7+?=10 M2_5-?=2 M3_6+3=5+? M4_8-4=6-?
Both	vswm3x3_54 vswm3x3_64 vswm3x3_73 vswm4x3_12 vswm4x3_24 vswm4x3_34 vswm4x3_43	C10_26-17 C13_2/5=1/5+?/5 C14_4/7-1/7=?/7 C15_42/7 C16_800/80 C2_24+18 C5_1/2=?/6 C6_3/4=6/? C7_40x3 C8_50x12 M5_9x?=63 M6_56/? =8 M7_48x2=24x? M8_105/5=210/?

*the 3x3 and 4x3 in VSWM represented the grids, the first digit after that is the item order in that grid, the second digit represents the number of objects in the grid. For example, vswm4x3_43

means that it is a stimulus has 4x3 grid and it was the fourth item given in 4x3 grid and there were three objects in this grid.

Neural imaging review. Although expert review provided psychological justification for the categorization of these items, reviewing findings from neuroimaging studies could potentially provide different insights over the cognitive processes associated with each of these items. Therefore, a different categorization was developed derived from findings of neural imaging studies was adapted to identify the cognitive processes in these items. One common method of categorization in neuroimaging studies is using hemispheric specialization such as identifying the lateralization of different brain functions. To do so, previous neuroimaging studies were reviewed in order to classify items into three dimensions: left hemisphere, right hemisphere, or bilateral. Hemispheric specialization was suggested by previous studies to understand the mechanism (Semmes, 1968) and to identify the cognitive encoding of tasks from neural perspectives (e.g., Slotnick & Moo, 2006). The literature review began by searching for neuroimaging studies that used VSWM and calculation in a typical population. For the categorization of VSWM items, although there are some discussions about the multidimensionality of VSWM (e.g., categorization vs. coordination), the nature of the VSWM task that was used here did not generate the categorization responses similar to previous studies because we asked children to indicate the location of the pictures rather than using a spatial frame of reference (e.g., above or below a reference point). Therefore, VSWM task was divided based on previous studies on its relation of subitizing (Piazza, Fumarola, Chinello, & Melcher, 2011) and counting (Kyttälä, Aunio, Lehto, Van Luit, & Hautamäki, 2003). Subitizing is the ability to correctly identify the quantity of a small set of items in a brief moment (the quantity is usually smaller than 4). (Kaufman, Lord, Reese, & Volkmann, 1949). It was recalled that in our

VSWM task, the number of objects within items was between two and seven. Therefore the items were categorized with four or fewer objects as subitizing and items with more than four objects as non-subitizing.

In previous studies, subitizing has showed bilateral advantages (Dehaene and Cohen, 1994; Delvenne, Castronovo, Demeyere, & Humphreys, 2011; Vuokko, Niemivirta, & Helenius, 2013), while a few studies suggested a right hemisphere advantage (e.g., Pasini & Tessari, 2000). In here the present study followed what was said in most studies. The rest of VSWM items were categorized as non-subitizing type of items, which might require more on counting and more spatial memory. Previous studies have demonstrated right hemisphere advantage on spatial memory (Thomason et al., 2009; Weintraub & Mesulam, 1987). Therefore, we categorized the rest of items are on right hemisphere while "subitizing items" are bilateral.

Regarding hemispheric specialization in the calculation test, a meta-analysis by Arsalidou and Taylor (2011) concluded that addition has more left hemisphere activity, subtraction is bilateral, and multiplication has more right hemisphere activity. However, as Arsalidou and Taylor (2011) indicated, because not many studies have analyzed calculation with division as a separate operation, they were not able to conclude the hemispheric specialization for division. There is, however, evidence suggesting that division and multiplication are highly correlated (Lefevre & Morris, 1999) and multidigit multiplication and division have some overlapping activations in similar brain areas (Fehr, Code, & Herrmann, 2007). Therefore in this study, multiplication and division are categorized together. For fractions, to the author's best knowledge, no literature has addressed how fractions activate which brain area(s), so it is difficult to categorize its hemispheric specialization. For this reason, the four fractions were removed from the comparison between expert models and the hemispheric specialization model.

Specifically, Model 2 (which was the best fit from the previous expert review session) was used as a comparison with this hemispheric specialization model (Model 4). Therefore fraction items were taken out from Model 2 (which is Model 2.1 here) for further comparisons. One might wonder whether Model 2 will still be the best model compared to Models 1 and 3 after taking out the four fraction items. To test this, this analysis was re-run on Models 1, 2, and 3, and Model 2 was still found to be the best. Therefore, Model 2.1 (which is Model 2 with four fraction items removed) and Model 4 (the hemispheric specialization model) were compared with AIC and BIC indices. The results suggested Model 2.1 is the better model (see Table 3).

Table 3.

Model comparison between Expert 2.1 Model and neural imaging model

Models	Number of parameters	AIC	BIC	SSABIC
Model 2.1	76	16565.990	16891.725	16650.476
Model 4	81	19206.949	19554.114	19296.993

Error analysis. Another type of confirmatory approach is to identify the dimensions through error analysis in advance. Previous studies have used this type of analysis to understand cognitive processes (Radatz, 1979;. Tatsuoka, 1984, 1990). To do so, the children's original VSWM responses were recoded and separated into two different datasets. In one set of data, these responses were recoded by whether children were correct about the number of pictures, and another was recoded by whether they correctly marked the location. Both responses were coded as a dichotomous response (0 and 1). To begin with, only the items that have under 85% correct rates in the original dataset were recoded, because the items that have high correct rates would

have correct responses on both datasets. This would have caused these two dimensions to be very similar if items over 85% correct were included. The recoding procedure proceeded as follows: for the dataset that was recoded by the numbers (denoted Number VSWM), the score was counted as 1 when the child answered correctly on the numbers of pictures within each stimulus, even though the location of the pictures were marked wrong in original responses. For the dataset that was recoded by the location, the original responses were used, because when a child correctly marked the location, these responses matched what were originally coded in the VSWM dataset. This location dataset is denoted as Location VSWM.

To identify whether these two datasets are too similar to begin with (if they are very similar then the attempt to separate the dimensions with error analysis might have failed), the correlation of the sum scores between Number VSWM and Location VSWM was calculated. The correlation was $r\ (537) = .62$. Compared to the criteria for strong correlation $r = .8$, it is not high. After this step, each dataset was combined with the calculation data and was processed with the MIRT analysis (denoted as Number VSWMCT and Location VSWMCT).

Figure 2 and Figure 3 showed two different sets of clusters by using the *a* parameters that were generated from MIRT analyses with Number VSWMCT and Location VSWMCT, separately. Surprisingly, while Location VSWM grouped only two calculation items, the Number VSWM grouped with subtraction, multiplication, and division items. This might suggest that the Number dimension of VSWM is what really correlated with arithmetic in previous studies (e.g. De Smedt, Janssen, Bouwens, Verschaffel, Boets, & Ghesquière, 2009; Rasmussen & Bisanz, 2005), and the location dimension of the VSWM might not have a high correlation with most of the calculations.

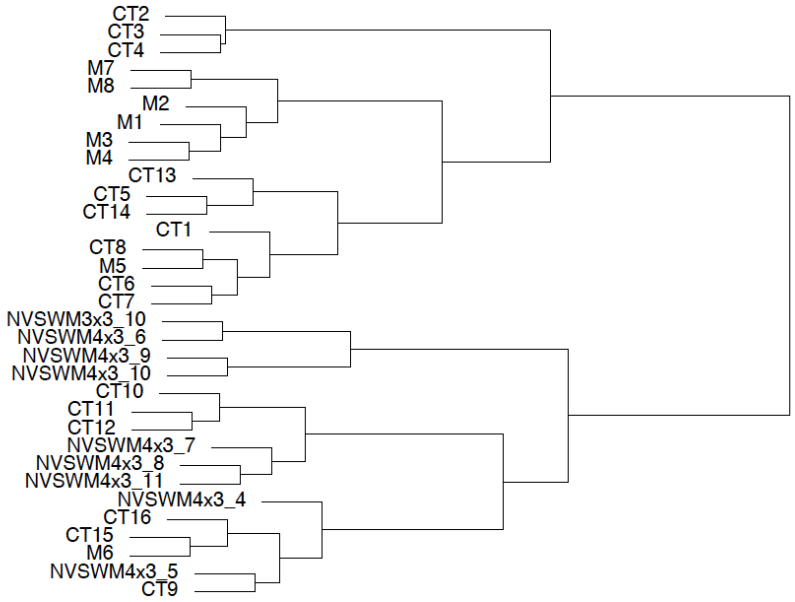

Figure 2. Number VSWM and Calculation. *NVSWM is Number VSWM. All "CT "and "M" items are calculations.*

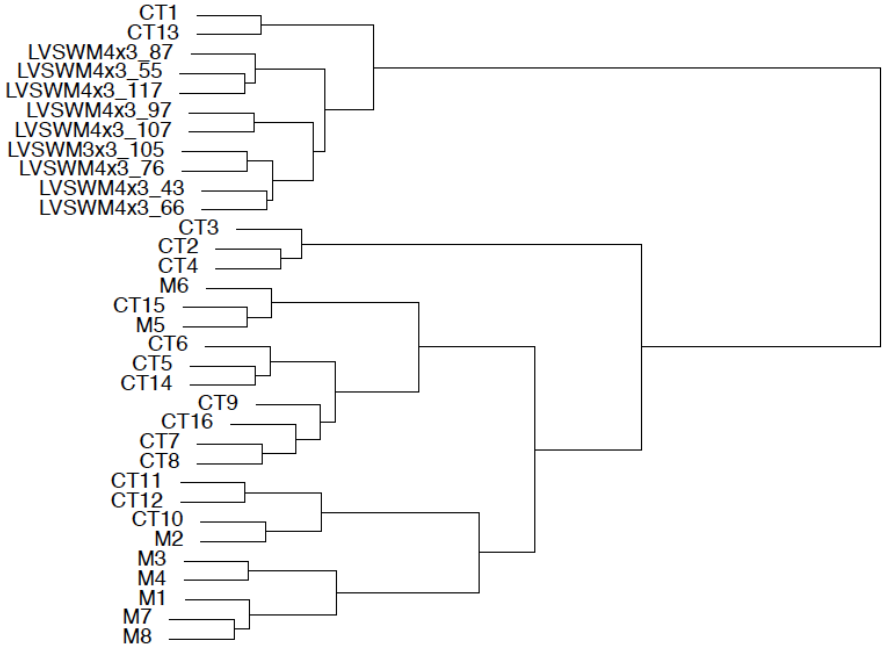

Figure 3. Location VSWM and calculation. *LVSWM is Location VSWM. All "CT "and "M" items are calculations.*

Discussion

The number of dimensions

The result suggested that there is converging evidence that multiple dimensions exist in this CVC dataset from both parallel analysis and cluster analysis. If VSWM and the calculations are both unidimensional, the CVC dataset should generate one dimension for spatial ability and another dimension for mathematics calculation, which would show as two dimensions here for CVC. The result therefore is somewhat different from what was found in some previous studies, that VSWM was unidimensional (e.g., Miyake, Friedman, Rettinger, Shah & Hegarty, 2001), but was supported by several other claims that VSWM and the calculations are multidimensional (e.g., Darling, Della Sala, Logie & Cantagallo, 2006; Männamaa, Kikas, Peets, & Palu, 2011).

Another piece of supportive evidence for the existence of three dimensions is that, aside from the results of parallel analysis and cluster analysis, it appears the relative model fit of three dimensions is better than two dimensions (e.g., for Expert 2 Model with two dimensions, BIC was 19515.830, and with three dimensions, BIC was 19312.073). Specifically, this occurs when comparing the model fit between two dimensions (which lump these interaction items with both spatial and mathematics dimensions because they might weigh equally on both processes) and three dimensions (which put these interaction items in an independent dimension).

What are these three dimensions?

Among all of these models, the Expert 2 Model is the best one. When reexamining the categorization of the Expert 2 Model, it appears that the Expert 2 Model has quite a similar categorization with the neural imaging model (see Table 2). Recall in the initial expert review that these items were categorized as more spatial or mathematics, and as more number or location. Such a categorization is a heuristic process. To help interpret the result, the finding

from neural imaging studies was used. Specifically, by identifying which functions are not grouped together in the same areas, their underlying cognitive processes can possibly be identified. In this study, this would mean identifying the functions that are not in the same dimensions. The categorization of the Expert 2 Model is presented in Figure 1, which shows several clear patterns. First, it appears that there might be a distinction between VSWM stimuli that have 2 to 4 pictures and the VSWM stimulus that has 5 to 7 pictures, suggesting these two might be separate cognitive processes. Previous studies have shown that there is a relation between VSWM and subitizing (Piazza, Fumarola, Chinello, & Melcher, 2011). Perhaps when children were processing the VSWM items with 2 to 4 pictures, they also used their subitizing ability, and when they were processing VSWM items with 5 to 7 pictures, they switched to use their counting ability. Evidence from previous studies also suggests that subitizing and counting are two different cognitive processes (Pasini & Tessari, 2001).

It also appears that in terms of calculation, addition and subtraction make up one dimension, while multiplication and division make up another dimension. This is also supported by the fact that neural studies have found addition and multiplication are on separate dimensions (Arsalidou & Taylor, 2011). However, the difference between the neural imaging studies and the Expert 2 Model is the categorization of subtraction. In the neural imaging model, subtraction and addition are in separate dimensions, which is different from what was defined in the Expert 2 Model that had subtraction and addition in the same dimension. However, because the Expert 2 Model is statistically better when compared to neural models, this might suggest that subtraction could be more likely to function on the left side rather than bilaterally. One possible reason could be that in previous neural imaging studies the subjects are mostly adults, but the subjects in the present study is children ages 9 and 10. However, overall, the neural imaging model and the Expert 2

Model are mostly similar despite the small difference in subtraction. This possible difference between children and adults also corresponds with general evidence that has shown that despite developmental changes in the brain, the overall core system is similar, regardless of the operations of the calculations. However, there is a shift in brain regions when specific cognitive processes occur (Kawashima et al., 2004).

Finally, in terms of shared dimensions between VSWM and the calculations, it appears that the ability to subitize items might be important for more difficult calculations. For example, all of the fraction, multiplication, and division items are in the same dimension as the subitizing in VSWM. One possible explanation could be that these "difficult" calculation items and the subitizing of VSWM items both require more processing in the right hemisphere. Therefore, they are more likely to be located in the same dimension. The finding from the error analysis also supported this finding. Recall that while Location VSWM was not grouped with most of the calculation items, Number VSWM was grouped with subtraction and division items. Apparently, the ability to correctly grasp or subitize the numbers of items on a target, in a brief given time, is related to these calculation items. When performing these calculation items, it is possible that children maintain the necessary elements in the process of calculation on their memory board. For example, one strategy of calculating 347-129 is to split it into six numerical elements, as in (300+40+7) – (100+20+9), and two operated symbols (+ and -). If the child did not grasp these elements of the equation well, they might not produce correct the answer on the calculation.

Overall, these findings might suggest that when training VSWM on children to improve their calculation performance, perhaps training does not need to focus on the location aspect of the VSWM. Instead, perhaps the training could focus on the number aspect of VSWM. For example, children can be trained by using exactly the same VSWM training but altering the

procedure by asking children to remember how many pictures are in the grid instead of marking their locations.

CHAPTER FIVE

STUDY TWO RESULTS

Simulation Study Design

The second study explored the possibility of reporting meaningful subscores by utilizing appropriate standardized measurement indices. Since the purpose of the current study is using the sensitivity of person fit indices, the most sensitive person fit indices were chosen. Although many person fit indices have been developed, Drasgow et al. (1987) showed *Lz* and *T2* are the most sensitive indices. However, studies also have demonstrated that *Lz* does not have standard normal distribution when true abilities are replaced with estimated ability parameters and has suggested a formula for correction for *Lz*, denoting *Lz** (Snijders, 2001; Magis, Raiche, & Beland, 2012). Therefore, *Lz** is used instead of *Lz* here along with *T2*. Also, as the true ability is unknown, the current study uses estimated ability parameters.

The research questions of the current study are two-folded. First, can person fit indices detect the differences in the atypical dataset of subscores that was created by the subgroups' differences? If so, in what conditions can the indices detect the differences? To do so, the indices were first applied to a series of the simulated data to find out in what conditions the indices are sensitive to detect the difference. We also applied the indices to test their performances with a real data application. Specifically, the indices were applied to the dataset from Study 1 to examine the detection rate.

Data simulation

The simulated data was generated by using the Wingen program 3.0 (Han, 2007; Han & Hambleton, 2007). The purpose is to create samples that will be a good fit with unidimensional data overall, while embodying a set of items that represented an atypical subgroup. To do so, two

samples were simulated for each condition. First, an *N*=1000 normal sample was simulated, and another atypical *N*=100 sample was simulated later. The atypical sample was used to replace a certain percentage of the items in a subgroup in the normal sample. The data for the atypical sample were simulated in a way that they have an unusually low or high probability of correct response patterns, regardless of the students' ability level in this particular subgroup. Note that no matter what purpose the sample was intended to generate (normal or atypical), the simulated data generation here used the same set of ability θ parameters and a parameters. The only difference between the original sample and the replaced sample is the +2 and -2 shift on the b parameters.

All the parameters were generated from a normal distribution, but the means and standard deviations were different in each type of parameters. . I applied a two parameter logistic IRT model to generate two sets of data in each condition, based on the same ability parameter θ. The ability parameter θ for simulated examinees was generated from the normal distribution *N* (0, 1) first. The normal dataset *N*=1000 was generated first. In this *N*=1000 normal sample, the *a* parameters were generated from ~*N* (1, .2). Then the *b* parameters were generated from ~*N* (0, .7). In the atypical *N*=100 sample, the same set of *a* parameters were used, but the mean of *b* parameters was generated with a shift on normal distribution from *N* (0, .7) to *N* (-2, .7) and N (2, .7). I randomly replaced the 100 simulated subjects in the normal sample with the atypical sample. The procedure was the same with +2 and -2 *b* parameters.

The conditions were varied by the percentage of atypicality (low 5%, middle 15%, high 30%) in items, the percentage of participants in the total sample (5%, 10%, 20% of 1,000), the total number of items (30 items, 45 items, 60 items, 75 items, 90 items), and the shift of *b* parameters. Therefore, 180 conditions were generated.

The percentage of replacement in the samples was aligned with what Drasgow et al. (1987) used in their study. The number of manipulations of items were also aligned with what had been used in previous studies (e.g., Haberman & Sinharay, 2010). The subtests in previous test programs had 13 to 35 items while the total number of items was around 49 to 119 questions. Therefore we divided our conditions into five categories (30 items, 45 items, 60 items, 75 items and 90 items). The sample size of 1000 was used here because Sinharay (2010) found in his simulation study that there was no difference between a sample size of 1000 and 4000 in the results.

Analytic procedure

In each condition three different analyses were conducted. First, the data in each condition were tested to see the fit for the unidimensional model. Note that although the original and replaced data were both generated by using the IRT model, it is possible that replacing part of the sample essentially changed its dimensionality dramatically enough to cause multidimensionality. Therefore, each modified dataset is checked for its dimensionality to see if the unidimensionality is still held in each condition. The reason for doing this was to make sure that the simulated dataset was properly generated to fit the unidimensional model well and that the simulation had successfully generated the proper psychometric property needed for the study.

Secondly, to estimate the new ability parameters and item parameters on the sample that we had already replaced with some atypical subgroups, we used Ltm package in R. To check the validity, the parameters from Ltm were compared with IRTpro (Cai, Thissen, & du Toit, 2011) and their correlation was quite high ($r > .90$). A new set of θ parameters were generated for each condition. Thirdly, using these new θ parameters with true a and b parameters, the Lz^* index was used to calculate the person fit value for each person in the simulated data. Because the Lz^*

index is a standardized index, the criteria for identifying whether the *Lz** index had detected an atypical sample that have values over ± 1.65. Thus any value over the absolute value of 1.65 would indicate atypical responses. The *Lz** index was written by using R program 3.2.4. However, in 2015, there is also a person fit package in R developed. The *Lz* and *Lz** functions in that program were also used to compare the detecting rates that were generated from the present study. The results were the similar.

Real data application

The person fit index was also applied to the dataset produced from Study 1 to check its function when applied to authentic data. 537 third graders participated in that study. Data were obtained from IES funded project: Spatial Ability as a Malleable Factor for Mathematics Learning. The data were collected in the Michigan and Chicago areas in spring 2013 and spring 2014.

Simulation Study Results

The results are presented in Table 1 to Table 4. Table 1 and Table 2 showed -2 and +2 b parameters shift when using indexes *Lz**. Table 3 and Table 4 showed -2 and +2 b parameters shift when using indexes *T2.* Each number indicated the average from 5 replications with its standard deviation denoted as (standard deviation).

Table 4.
Lz with -2b parameters (rate of detection)*

Replaced 5% Items					
Percentage people replaced	30Items	45Items	60Items	75Items	90Items
5%	18% (3)	14% (1)	11% (2)	15% (1)	16% (2)
10%	19% (1)	17% (2)	14% (2)	16% (1)	16% (2)
20%	18% (1)	16% (1)	14% (1)	14% (1)	16% (1)
Replaced 15% Items					
Percentage people replaced	30Items	45Items	60Items	75Items	90Items
5%	17% (1)	15% (1)	14% (2)	14% (2)	16% (3)
10%	21% (2)	15% (2)	16% (3)	15% (1)	16% (2)
20%	22% (1)	16% (2)	18% (2)	15% (1)	16% (1)
Replaced 30% Items					
Percentage people replaced	30Items	45Items	60Items	75Items	90Items
5%	17% (2)	19% (2)	19% (4)	17% (1)	16% (4)
10%	21% (2)	17% (3)	21% (3)	17% (1)	18% (3)
20%	22% (2)	18% (2)	23% (2)	19% (1)	18% (2)

Table 5.
Lz with +2 b parameters (rate of detection)*

Replaced 5% Items					
Percentage people replaced	30Items	45Items	60Items	75Items	90Items
5%	16% (2)	14% (2)	11% (2)	14% (2)	14% (2)
10%	17% (1)	17% (1)	12% (2)	14% (1)	14% (1)
20%	17% (1)	16% (1)	13% (1)	14% (0)	14% (1)
Replaced 15% Items					
Percentage people replaced	30Items	45Items	60Items	75Items	90Items
5%	16% (2)	16% (1)	10% (3)	11% (1)	20% (2)
10%	17% (2)	18% (1)	12% (2)	14% (1)	18% (2)
20%	18% (2)	16% (1)	13% (2)	16% (2)	21% (2)
Replaced 30% Items					
Percentage people replaced	30Items	45Items	60Items	75Items	90Items
5%	16% (3)	18% (2)	13% (4)	13% (3)	26% (5)
10%	19% (2)	19% (1)	15% (2)	17% (2)	24% (4)
20%	18% (1)	20% (2)	17% (2)	22% (3)	26% (3)

Table 6.
T2 with -2 b parameters (rate of detection)

Replaced 5% Items					
Percentage people replaced	30Items	45Items	60Items	75Items	90Items
5%	7% (2)	8% (2)	13% (2)	9% (3)	11% (1)
10%	6% (1)	9% (2)	11% (1)	10% (1)	10% (1)
20%	8%(1)	9% (1)	10% (1)	10% (0)	10% (1)
Replaced 15% Items					
Percentage people replaced	30Items	45Items	60Items	75Items	90Items
5%	8% (1)	7% (1)	11% (3)	8% (2)	12% (2)
10%	8%(2)	8% (1)	11% (1)	10% (1)	11%(2)
20%	9% (1)	10% (1)	10% (1)	10% (1)	11% (2)
Replaced 30% Items					
Percentage people replaced	30Items	45Items	60Items	75Items	90Items
5%	8% (1)	10% (2)	14% (3)	8% (2)	12% (2)
10%	10%(1)	11% (1)	12% (1)	8% (2)	11% (2)
20%	10% (1)	11% (1)	10% (0)	8% (1)	11% (1)

Table 7.
T2 with +2 b parameters (rate of detection)

	Replaced 5% Items				
Percentage people replaced	30Items	45Items	60Items	75Items	90Items
5%	6% (2)	8% (1)	9% (3)	9% (2)	10% (1)
10%	7% (2)	10% (1)	9% (1)	10% (1)	9% (1)
20%	10%(1)	11% (0)	8% (1)	10% (1)	10% (1)
	Replaced 15% Items				
Percentage people replaced	30Items	45Items	60Items	75Items	90Items
5%	6% (2)	8% (0)	11% (1)	8% (2)	10% (2)
10%	7%(2)	9% (1)	10% (1)	10% (2)	8%(1)
20%	9% (2)	11% (1)	10% (1)	10% (1)	9% (1)
	Replaced 30% Items				
Percentage people replaced	30Items	45Items	60Items	75Items	90Items
5%	8% (3)	11% (2)	10% (3)	8% (2)	9% (1)
10%	8%(2)	12% (1)	11% 2)	8% (1)	9% (1)
20%	8% (1)	13% (2)	10% (1)	10% (1)	7% (1)

Also, because the current study has two (*Lz** and *T2*) by two (-2 *b* parameter and + 2 *b* parameter) by three (5%, 15%, 30% item replacement) by three (50, 100, and 200 people replacement) by five (30, 45, 60, 75, 90 item length) factors, these factors generated 180 conditions. When a specific condition is referred in the context, it is denoted as (index, shift in *b* parameters, percentage of people replaced, the number of replaced items, total item numbers) in a basket. For example, for the condition using *Lz**, -2 *b* parameter, 100 people replaced, 15%items replaced, 60 item in total will be denoted as (*Lz**, -2, 100, 9, 60).

Type 1 error rates

The type 1 error rates here refer to the misdetection rate on these non-replaced item responses in the original data. On average the type 1 error rates were around 12% -18% across different conditions with *Lz** index. On average the type 1 error rates were around 8% -11% across different conditions with *T2* index. The type 1error rates tend to go up when higher number of total items were used with *Lz** but not with *T2* (e.g., in two of the replications of 90 items conditions, the type 1 error rate was 17% and 20% with *Lz**). The higher type 1 error rate in the present study might relate to the estimated parameters that were used here.

Rate of detection between *Lz and *T2***

In terms of rate of detections between these two person fit indices, it appears that *Lz** has a better rate of detections than *T2* has across all conditions. *Lz** generally has an 11%~26% rate of detection while *T2* generally has a 6%~12% rate of detection. Although this rate of detecting is far smaller than what was found in Dragow et al. (1987), the rate of detection of *T2* is usually smaller than *Lz* (which is the uncorrected version of *Lz**) in Dragow et al. (1987). For example, in Dragow's 15% replacement conditions, *Lz* has a 72% rate of detection and *T2* has a 54% rate

of detection. Therefore it seems to be reasonable that *Lz** also had a better rate of detection than *T2* had.

Rate of detection between -2 *b* and +2 *b* parameter shift.

In terms of the rate of detection between the shift of -2 and +2 on the *b* parameters among conditions, the result showed that the rates of detection of the corresponding conditions are mostly symmetrical. What symmetrical means here is that when the rate of detection in the correspond conditions between -2 and +2 are in close range. For example, (*Lz**, +2, 50, 23, 75) and (*Lz**, -2, 50, 23, 75) both have 14% rate of detection. However, in some conditions +2 *b* parameters conditions tend to show a better rate of detection than -2 *b* parameters conditions. For example, in the +2 *b* parameter (*Lz**, +2, 200, 30, 90) condition, higher rate of detection is shown than in the corresponding condition of -2 *b* parameter (*Lz**, -2, 200, 30, 90). The difference is (26% vs. 18%).

Rate of detection among different percentages of replacements on people

When replaced with a different percentage (5%, 10%, 20% of 1000 simulated subjects) of atypical people, it appears that the rate of detection does go up. However, when the number of replaced items is low (e.g., 5% of the total), even though 20% of the people were replaced, the rate of detection seems to be similar. For instance, the difference of rate of detection ranges between 0% and 1 % across conditions. However, when 30% of the total is replaced, the rates of detection tend to rise, from 5% of people to 20% of people across conditions. For example, (*Lz**, +2, **200**, 23, 75) has a 9% greater rate of detection than (*Lz**, +2, **50**, 23, 75). However, in general, despite the proviso that replacing 20% of people tended to produce a higher percentage of people being detected, the percentage of people being detected does not change dramatically.

Rate of detection among different numbers of replaced items and total items.

It appears that when a higher number of items was replaced, the rate of detection did go up using *Lz**, but this was not the case for *T2*. In addition, even in the condition of having a total of 30 items, it seems that nothing changed much, even though the number of replaced items went up regardless of whether *Lz** or *T2* was used. However, in other conditions, with 45, 60, 75 and 90 items in total, the rate of detection went up in a linear way as more items were included. For example, in the condition using *Lz**, a shift on +2 *b* parameter appeared to increase when the length of the total item increased as well (45 items: 4%, 60 items: 4 %, 75 items: 8 %, and 90 items: 12%). In general, the rate of detection is not high compared to what was found in previous studies when people used these person fit indices to detect cheating or slipping behaviors.

Real data application

The *Lz** index also was performed on the CVC data from the first study. The detecting rate of CVC dataset is 17%, which is not great. This is surprising considering that CVC is a multidimensional dataset. However, it further demonstrated that the insensitivity of *Lz** index when estimated parameters were used.

Discussion

Although the person fit index has proven in previous studies to be quite sensitive (e.g., Dragow et al., 1987), the current study did not find similar results. One possible reason may be that previous studies mostly used true ability parameters, while the current study used estimated ability parameters. Since the true parameters are usually unknown, the use of true parameters in previous studies is often criticized (e.g., Magis et al. 2012). However, the low rate of detection on both simulated data and real data suggested that either *Lz** and *T2* were not sensitive enough to detect the changes in the simulations when estimated parameters were used.

The second possible result is that the shift of difficulty in the present study may be less severe than the shift in previous studies. For example, Dragow et al. (1987) shifted the percentage of altered responses to be all correct responses (for example, if a sequence of response was originally 0,1,1,0,0, the response was then shifted to 1,1,1,1,1). This is the maximum amount of "treatment" when altering the original responses. Therefore the treatment was much more severe than the treatment in the present study. In addition, this type of strong treatment was used to test whether a higher rate of detection is possible even with estimated parameters. The results suggest that the improvement is better (e.g., 18%~53% across (*Lz**, +2, 200, 27, 90 condition with replications) but still not as high as previous studies. The true parameters were also used in one condition to test whether the detection rate would become higher, and it did (went up to 70%). However, it will not make sense for the current study used the true θ parameters or a stronger alteration, since true ability parameters are unknown and stronger alteration is not realistic in the real life. These findings also imply that perhaps the main reason for the low rate of detection is the sensitivity of the person fit indices for the simulated conditions in the present study.

Therefore, a different way of thinking about this result is to ask whether a more sensitive index can be developed. After all, an effective index should be able to identify the person responsible for bizarre response patterns if they are detected more than a certain number of times (e.g., 3 times or 5 times), notwithstanding the presence of random errors. However, one might need to ask whether a more severe condition could be representative of the conditions in actual tests. Since the goal of the present study was to see whether person fit indices can detect a small number of unusual sub-scores, it does not seem realistic to increase the degree of irregularity in response patterns. When comparing the response patterns with the people who are

detected and not detected, it seems that the person fit indices sometimes identify people who have severe responses. For example, the *Lz** detected a simulated subject that had five bizarre responses but did not pick out a simulated subject who had eight mildly bizarre responses. This may have something to do with the cumulative index values gathered from each item. However, eight mildly bizarre responses from a child's test might also indicate the need for some cautious remediation as well. For example, if a child missed some instructions in part of the mathematics curriculum, but had a good understanding of other mathematics contexts, the misfiting response patterns might not severe. Therefore, a sensitive index can be beneficial only if it can identify this child correctly. Perhaps it would reap more profits to also identify such a child than to just identify a child who makes one severely bizarre response, as a result perhaps of a clerical error.

CHAPTER 6

GENERAL DISCUSSION

Implication

The present study explores the dimensionality of cognitive structure. As the previous chapter illustrated, identifying dimensionality can be critical for training and testing. dimensionality can be useful when identifying the necessary training components for an intervention. For example, a cognitive training program might fail if the program did not offer training on the components that are in the same dimension as the targeted improvement of ability. Also, dimensionality can also be helpful when diagnosing certain score drops on sub-scores that may reflect the fact that students missed certain instructional episodes. However, despite its usefulness, the true number of dimensions is hard to identify. This is because a meaningful dimension can be identified only when it has a sound theoretical construct and a proper psychometrical property. Therefore, to recover the true number of dimensions, the construct needs to be well defined on the basis of all sources of evidence and the methodology to detect dimensions also needs to be reliable, valid and sensitive.

Multidimensionality analysis

The present study addresses these issues from several perspectives. First, when the true number of dimensions is called into question, it asks what the reasonable methods are to construct the best possible models to help researchers get close to the truth. The first phase of the present study has demonstrated several possible methods by using an example. Recalling what was discussed at the beginning of the introduction, the dimensions between cognitive ability and academic achievement were not clear. Such an unclear relationship is reflected in the vague outcomes of cognitive training (e.g., Jaeggi et al., 2011; Redick et al., 2013). Specifically,

training specific abilities seems to have some effect, but it was unclear what the training program should target because not all specific cognitive abilities are predictive. However, in this situation, when the true number of dimensions is unknown, a possible way forward seems to be to examine the overlapping dimensionality between cognitive ability and academic achievement.

To solve this problem, in the first study, the tests from two traditional psychological constructs, VSWM and calculations, were combined and re-analyzed by using their item responses from a perspective of cognitive processes. Several types of MIRT analyses were proposed to address this situation. The results indicated that with psychometric models, three dimensions were detected in this combined dataset MIRT analysis, and further content analysis was proposed to identify their theoretical constructs. Several possible theoretical models were then proposed and a better model was eventually projected after a process of model competition. Using this process, the possible candidates for dimensions were revealed. For example, subitizing and counting processes represent the two dimensions of VSWM. Note that the items belonging to these two dimensions of VSWM are the same in their general testing procedure and the surface similarity of the stimuli being used. Another finding from the present study was the partition of calculation items, such that addition and subtraction occupied one dimension, while multiplication and division occupied another. More importantly, these partitions generated from MIRT analysis are also partially parallel to the findings from neural imaging studies (Arsalidou & Taylor, 2011), specifically on the categorization between addition and multiplication. Findings from these neural imaging studies provided a theoretical standpoint for these dimensions while the MIRT analysis provided evidence of the psychometrical properties on them. The results showed that the collaboration between neural science and measurement brought insights into the understanding of dimensions. The third proposed method comes from the analysis of error in

item responses. By isolating two different types of error patterns in the item responses, two dimensions in VSWM were separated. Different patterns of groupings were observed when Number VSWMCT and Location VSWMCT were submitted separately for analysis.

Design of cognitive training

While the validity of the results still needs to be corroborated by another independent study, these MIRT analyses brought insights into the dimensionality of the tasks and point to possible directions for future research. Overall, there is a repeated pattern in the analysis that not all dimensions of VSWM are related to calculation. Perhaps not all dimensions of the VSWM task have to be trained in order to gain improvement in calculation. One direction is that these findings can in turn inform the design of cognitive training. For example, considering the fact that calculation and VSWM only shared dimensions on subitizing and multiplication/division, it might be possible that when conducting VSWM training, asking children to subitize "how many objects there are" is more practical and effective than asking children to "mark where these objects are." Another direction that could be considered is how to train subitizing within VSWM contexts. Previous studies have suggested that subitizing is related to the capacity of VSWM (Piazza, Fumarola, Chinello, & Melcher, 2011; also see the discussion in Cutini & Bonato, 2012). However, the time interval for subitizing probably should be shorter than the usual VSWM tasks for each stimulus. For example, in this study, the VSWM stimulus was shown for five seconds, but for the purpose of training subitizing, perhaps restricting the time interval to one or two seconds would be sufficient.

Lastly, the results also suggest that within calculation, only subtraction and division are related to the Number VSWM, therefore it makes sense to train in Number VSWM to improve subtraction and division. This relation seems to suggest that the ability to simultaneously

perceive multiple objects is particularly important for these two operations within calculation (compared to addition and multiplication). Perhaps in order to compute addition or multiplication, the psychological process requires fewer steps for attention to multiple objects compared to subtraction and division. For example, to compute 6x7=42, a child might recruit the answer from memory, but to compute a similar problem such as 42/7=?. It requires at least one more step. That is, a child uses his/her memory from multiplication to reverse the equation while simultaneously maintaining all elements of the equation in his/her memory board. Previous studies showed that learning from addition and multiplication is difficult to transfer to compute subtraction and division (Campbell et al., 2006). Therefore, although these inverse operations seem to be similar in their psychological processes, the actual learning outcome said otherwise. Perhaps the attention to multiple simultaneous steps is the key difference here between addition and subtraction, and multiplication and division as well.

Detectability of dimensions

While the first study explored multiple approaches in conditions when the true dimensions were unknown, the second study addressed the detection of dimensions. Specifically, it asked whether there is a possible method of detecting dimensions when they exist only on a subgourp of sample. Recalling that previous studies had difficulty on finding the proper psychometric properties of the (dimensions) sub-scores, the second study started from a slightly different approach and asked whether, if the dimensions existed, they could be detected. In this perspective, the true dimensions need to be known beforehand so that what is being detected can not be questioned. To do so, the second study created simulations to generate the condition that the dimensions are embedded in unidimensional models. These were then tested with person fit indices. However, it turns out that even with the highest replacement rate of atypical items, the

detecting on these dimensions (sub-scores) is not high (the highest rate of detection rate was 26%) compared to previous studies. The use of estimated parameters and less severe treatment may account for this result. Recalling that in our simulations, the shift of *b* parameter was not comparable to the most severe condition in previous studies (which shifts every response from 0 to1), however, it is still a significant change because the shift was two standard deviations below and above the means of the original parameters. However, it still reflects the difficulty of detecting sub-scores when they are small and only exist in subgroups. Only in the condition of 30% misfitting responses and 200 people replaced was there an increase in the detection rate. This result corresponded with what Meijer and Sijtsma (2001) have suggested that type of misfitting patterns, test length, and trait level (θ) could influence the result of person fit indices. This also implies that the dimensions from the first group might not be smaller dimensions and therefore can well represent the dataset when conducting dimensionality tests. However, in terms of detecting the dimensions that were generated from the misfiting responses of subgroups, a more sensitive index should be developed to help identify the dimensions.

General ability vs. specific abilities

The third question that the present study addressed was whether the dimensions (sub-scores) still can be observed when general ability scores (total scores) are presented. The result seems to depend on the analyses and conditions. In the first study, while VSWM and calculation are both under the influence of "*g*", the dimensions were still presented with MIRT analysis. One might argue that this is because the cluster analysis started from high dimensions (where every item is a dimension) and thus multiple dimensions could be presented, possibly the opposite direction to that in the factorial analysis method which first extracted what was common to the two constructs. Although this may have been the case, recalling that three clusters eventually become

two clusters, but the cluster indices C-index (one of the best cluster indices, see Milligan & Cooper, 1985 simulation study) indicated that 3 clusters are better than 2 clusters (3 clusters vs. 2 clusters: 0.03 vs. 0.27, smaller number is better). So the conclusion about three dimensions holds here. In other words, the dimensions in VSWM and Calculation may be significant; hence it appears their existence in these analyses, even under the power of general ability. However, the dimensions from the second study were much smaller ones and therefore the rate of detection was not high. This suggested that, when the dimensions existed only in some subgroups, person fit indices might not be an ideal procedure to detect. Although the overall rate of detection was not high, the comparisons among tests of different lengths still suggested that longer tests helped to uncover the existence of dimensions. In addition to this finding, the comparison between different percentages of replaced items also suggested that when a higher number of subgroups possess this dimension, the dimensions may be easier to detect.

Limitations

There are several limitations in the present study. First, because the first study used a particular dataset of VSWM and calculation, the results cannot be generalized to all the relations between cognitive ability and academic achievement. However, for tasks that apply the same psychological concepts, the results might be informative. For example, some previous studies have demonstrated significant relations among several spatial tasks compared to other spatial tasks (Mammarella et al., 2008). In particular, Mammarella et al. (2008) found that the visual patterns task (Della Sala et al., 1997), which is a task very similar to the VSWM task used in present study, along with another two spatial tasks, dots reproduction and static mazes (Pickering et al., 1989), contributed significantly to a passive, simultaneous spatial factor in their model. Therefore, it is possible to extend the finding of VSWM from the current study to the use of

these three spatial tasks. The findings of the current study on the relations between VSWM and calculation might also be able to extend to other grades depending on whether the significant relations were found in previous studies. For example, Berg (2008) found a significant relation between VSWM and calculation in sixth graders, while Rasmussen and Bisanz (2005) only found a significant relation in preschoolers but not first graders.

Furthermore, the technique of MIRT analysis that was demonstrated in the present study can be applied to understanding the dimensions of other constructs. In addition, new approaches were adapted instead of previous forms of MIRT analysis. The validity of these new methods still needs to be examined. Second, the sample size in the first study is on the low end for IRT analysis (Reckase, personal communication), which may generate some error in the results of the first study. However, because convergent evidence was found from a different analysis of the dimensionality, it appears that even though some error is present, it is likely to be marginal. Third, the neural imaging models were constructed from previous findings in the literature, including a meta-analysis from previous studies. While the tasks used in the neural studies may have similar concepts to those underlying the tasks we used in the first study, they are still different tasks. Our study may show signs of parameter drift. It is possible that a different conclusion might be derived if an actual MRI scan was made using the tasks in the present study. Furthermore, when constructing the neural imaging models, even though we mainly followed the methods observed in previous studies, most studies in this area were conducted with adult subjects, whereas our subjects were children. Individual differences, then, may be expected from the present study.

The second study also has several limitations. First of all, because not all person fit indices were tested, it is possible that there are other person fit indices that would show better rates of

detection, though we used the two most sensitive indices shown in previous studies (Dragow, 1987). The second limitation is that the present study used a specific range of IRT parameters. For example, we chose '*a*' parameter on the basis that the mean is 1 and the *SD* is .2, and the '*b*' parameters on the basis that the mean is .7 and the *SD* is 1 and θ had a mean of 0 and an *SD* of 1. Although the range of *a* and *b* parameters was not a critical factor in previous review (Meijer & Sijtsma, 2001) and these parameters were also normally distributed in the current study, previous studies have demonstrated that high ability participants could be easier to detect than subjects with other ability levels (Dragow et al., 1987). It is possible that a different range of parameterization might generate different results. However, because the present study was more interested in a normal distortion of θ parameters, participants with particularly high or low ability levels were not tested. Perhaps further research can be done with a different range of parameters than those in the present study.

Future Directions for Research

From a theoretical standpoint, it has already been established that studies in many areas are all closely related to the dimensions of the cognitive structure. However, different disciplines discussed this topic in slightly different ways (e.g., total scores vs. sub-scores in measurement/general ability vs. specific ability in cognitive psychology/common areas vs. specific areas in neural imaging studies) and cross communications among these different disciplines has not been well established yet. A few examples have showed, however, that when the connection is made, better insight can be brought to the understanding of dimensionality of constructs (e.g., Harris, Hirsh-Pasek and Newcombe, 2013; O'Toole, Jiang, Abdi, Pénard, Dunlop, & Parent, 2007). Therefore, one direction for future studies of dimensionality should be to devote themselves to making connections between related studies from different disciples.

While the present study established some findings by merging studies from education, cognitive science, and measurement, still further improvement is possible. For example, when constructing neural imaging models we used finding from past neural imaging studies. One way to improve on this might be to connect MIRT analysis with actual neural imaging procedures. For example, perhaps the activation response outputs from MRI scanning could be used instead of using the paper-and-pencil responses elicited from traditional tests.

From a practical standpoint, the present study on dimensionality has useful applications in several areas. For example, one area that has been popular in previous dimensionality studies is diagnostic testing. Perhaps one possibility is to extend the findings from the current dissertation by developing more sensitive measurement indices to identify children's weaknesses and strengths in academic learning. Specifically, a more sensitive index could be developed to identify the mismatched results between children's underlying abilities and their performance response patterns. A second plan is to develop a research project that aims to use advanced quantitative methods and sensitive measurements to detect the efficiency of educational training. For example, one possibility would be to apply quantitative methods in order to identify whether training or intervention is effective and to evaluate the reliability for the expected sample of examinees and validity of the influences made from the outcome measures. In combining these two plans, one might be able to develop thorough diagnostic testing to identify students' weaknesses and then give them custom-designed cognitive training to improve their academic outcomes, such as mathematics.

A second practical direction that could be derived from dimensionality studies is their application to automated item generation (AIG). AIG is a new and growing method of development items by using specific algorithms to generate large numbers of high quality test

items (Alves, Gierl, & Lai, 2010). The development of AIG requires both a deep understanding of the cognitive process of item dimensionality and the ability to select effective algorithms from measurements. Studies have used the understanding of dimensionality to develop AIG methods on figural matrices (e.g., Arendasy & Sommer, 2005). While this line of research seems to be promising, further studies will be needed to better understand the validity of inferences from the test scores and reliability scores obtained from tests produced using AIG methods.

Overall, research that has worked on this aspect has made a great impact on the design of methods for testing, cognitive training, and understanding of the function of human brains. However, even though many attempts have been made, the separation of dimensionality still seems a controversial issue. Different studies with different datasets have often found different results with different analyses. Secondly, because the developments between theoretical understanding and the advance of statistical techniques are often not synchronized, studies have claims that they have theoretical reason to pursue certain separation on dimensions but unable to find statistical support for it, or statistical models were developed without considering the real world situation, further research is needed on collaborating both. The present study combines several dimensionality analyses with perspectives from different areas in the hope of providing some insights into the dimensionality of cognitive structures.

REFERENCES

REFERENCES

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational Measurement: Issues and Practice*, *22*(3), 37-51.

Adelman, N.E. et al. (2002) A Developmental fMRI study of the Stroop color–word task. Neuroimage 16, 61–75.

American Educational Research Association, American Psychological Association, and National
Council on Measurement in Education. (1999-2008). Standards for educational and Psychological testing. Washington, DC: American Psychological Association.

Alloway, T. P., & Passolunghi, M. C. (2011). The relationship between working memory, IQ, and mathematical skills in children. *Learning and Individual Differences, 21*, 133–137.

Alves, C. B., Gierl, M. J., & Lai, H. (2010). Using automated item generation to promote principled test design and development. *American Educational Research Association, Denver, CO, USA*.

Ashcraft, M. H. (1982). The development of mental arithmetic: A chronometric approach. *Developmental Review, 2*, 213–236.

Ashkenazi, S., Rosenberg-Lee, M., Metcalfe, A.W.S., Swigart, A.G., Menon, V., 2013. Visual-spatial working memory is an important source of domain-general vulnerability in the development of arithmetic cognition. *Neuropsychologia* 51, 2305–2317.

Arendasy, M., & Sommer, M. (2005). The effect of different types of perceptual manipulations on the dimensionality of automatically generated figural matrices. *Intelligence*, *33*(3), 307-324.

Arsalidou, M., & Taylor, M. J. (2011). Is 2+ 2= 4? Meta-analyses of brain areas needed for numbers and calculations. *Neuroimage*, *54*(3), 2382-2393.

Babenko, O. I. (2013). Methods for determining whether subscores reporting is warranted in large-scale achievement assessments. (Order No. AAINR89443, Dissertation Abstracts International Section A: Humanities and Social Sciences,

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

Baddeley,A.D.(2000) *Short-term and working memory*. In The Oxford Handbook of

Memory (Tulving, E. and Craik, F.I.M., eds), pp. 77–92, Oxford University Press, Oxford.

Best, J. R., Miller, P. H., & Naglieri, J. A. (2011). Relations between executive function and academic achievement from ages 5 to 17 in a large, representative national sample. *Learning and individual differences*, *21*(4), 327-336.

Berg, D. H. (2008). Working memory and arithmetic calculation in children: The contributory roles of processing speed, short term memory, and reading. *Journal of Experimental Child Psychology, 99*, 288–308.

Betts, J. R., Hahn, Y., & Zau, A. C. (2011). *Does Diagnostic Mathematics Testing Improve Student Learning?*. Public Policy Instit. of CA.

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.

Blair C, Razza RP. (2007) Relating effortful control, executive function, and false-belief understanding to emerging mathematics and literacy ability in kindergarten. Child Development, 78, 647–663.

Blakemore, S. J., & Choudhury, S. (2006). Development of the adolescent brain: implications for executive function and social cognition. *Journal of child psychology and psychiatry*, 47(3‐4), 296-312.

Bosco, A., Longoni, A. M., Vecchi, T. (2004). Gender effects in spatial orientation: Cognitive profiles and mental strategies. *Applied Cognitive Psychology, 18,* 519-532.

Borst, G., Ganis, G., Thompson, W. L., & Kosslyn, S. M. (2012). Representations in mental imagery and working memory: Evidence from different types of visual masks. *Memory & cognition*, *40*(2), 204-217.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345-370.

Bull, R., Espy, K. A., & Wiebe, S. A. (2008). Short-term memory, working memory, and executive functioning in preschoolers: Longitudinal predictors of mathematicsematic achievement at 7 years. *Developmental Neuropsychology*, 33, 205−228.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference understanding AIC and BIC in model selection. *Sociological methods & research*, *33*(2), 261-304.

Cai, L., Thissen, D., & du Toit, S. H. C. (2011). IRTPRO: Flexible, multidimensional, multiple categorical IRT modeling [Computer software]. Lincolnwood, IL: Scientific Software International.

Cai, L. (2012). flexMIRT: Flexible multilevel item factor analysis and test scoring[Computer software]. *Seattle, WA: Vector Psychometric Group, LLC*.

Campbell, F. A., Pungello, E. P., Miller-Johnson, S., Burchinal, M., & Ramey, C. T. (2001). The development of cognitive and academic abilities: Growth curves from an early childhood educational experiment. *Developmental Psychology, 37*(2), 231-242. doi:http://dx.doi.org/10.1037/0012-1649.37.2.231

Campbell, J. I. D., Fuchs-Lacelle, S., & Phenix, T. L. (2006). Identical elements model of arithmetic memory: Extension to addition and subtraction. *Memory & Cognition, 34*, 633-647.

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

Carroll, J. B. (1997). The three-stratum theory of cognitive abilities. *Contemporary intellectual assessment: Theories, tests, and issues.*(pp. 122-130) Guilford Press, New York, NY.

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

Cattell, R. B. (1941). Some theoretical issues in adult intelligence testing. *Psychological Bulletin, 38*(592), 10.

Caviola, S., Mammarella, I. C., Cornoldi, C., & Lucangeli, D. (2009). A metacognitive visuospatial working memory training for children. *International Electronic Journal of Elementary Education*, *2*(1), 122-136.

Cerella, J., & Hale, S. (1994). The rise and fall in information-processing rates over the life span. *Acta psychologica*, *86*(2), 109-197.

Cestari, V., Lucidi, A., Pieroni, L. and Rossi-Arnaud, C. (2007) Memory for Object Location: A Span Study in Children. *Canadian Journal of Experimental Psychology, 61*(1), 13-20.

Chase, W. G., & Ericsson, K. A. (1982). Skill and working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 16, pp. 1–58). New York, NY: Academic Press. doi:10.1016/S0079- 7421(08)60546-0

Cheng, Y. L., & Mix, K. S. (2014). Spatial training improves children's mathematics ability. *Journal of Cognition and Development*, *15*(1), 2-11.

Christal, R. E. (1958). Factor analytic study of visual memory. *Psychological Monographs: General and Applied*, *72*(13), 1.

Clark, C. A., Pritchard, V. E., & Woodward, L. J. (2010). Preschool executive functioning abilities predict early mathematics achievement. *Developmental psychology*, *46*(5), 1176.

Cutini, S., & Bonato, M. (2012). Subitizing and visual short-term memory in human and non-human species: a common shared system?. *Number without language: comparative psychology and the evolution of numerical cognition*, *129*.

Darling, S., Della Sala, S., Logie, R. H., & Cantagallo, A. (2006). Neuropsychological evidence for separating components of visuo–spatial working memory. *Journal of neurology*, *253*(2), 176-180.

Das, J. P., Kirby, J., & Jarman, R. F. (1975). Simultaneous and successive synthesis: An alternative model for cognitive abilities. *Psychological Bulletin, 82*(1), 87-103. doi:10.1037/h0076163

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35*(1), 13-21.

Dehaene, S., and Cohen, L. (1995). Towards an Anatomical and Functional Model of Number Processing Mathematical *Cognition 1*, 83-120. *Psychology, 62*, 417–425.

Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology, 20*, 487–506.

de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics*, *30*(3), 295-311.

Delvenne, J. F., Castronovo, J., Demeyere, N., & Humphreys, G. W. (2011). Bilateral field advantage in visual enumeration. *PLoS One*, *6*(3), e17743

Demetriou, A., & Efldides, A. (1994). Hierarchical Models of Intelligence and Educational Achievement '. *Intelligence, mind, and reasoning: Structure and development*, *106*, 45.

De Smedt, B., Janssen, R., Bouwens, K., Verschaffel, L., Boets, B., & Ghesquière, P. (2009). Working memory and individual differences in Mathematics achievement: A longitudinal study from first grade to second grade. *Journal of experimental child psychology*, *103*(2), 186-201.

Drasgow F, Levine MV, McLaughlin ME. (1987) Detecting inappropriate test scores with optimal and practical appropriateness indices. *Apply Psycholology Measure*, 59-79.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, *56*(3), 495-515.

Fehr, T., Code, C., & Herrmann, M. (2007). Common brain regions underlying different arithmetic operations as revealed by conjunct fMRI–BOLD activation. *Brain research*, *1172*, 93-102

Ferguson, G. A. (1971). *Statistical analysis in psychology and education.* London: McGraw-Hill.

Floyd, R. G., Evans, J. J., & Mcgrew, K. S. (2003). Relations between measures of Cattell Horn-Carroll (CHC) cognitive abilities and mathematics achievement across the school age years. *Psychology in the Schools, 40*(2), 155-171. doi:10.1002/pits.10083

Frey, M. C., & Detterman, D. K. (2004). Scholastic assessment or *g*? The relationship between the Scholastic Assessment Test and general cognitive ability. *Psychological Science*, *15*(6), 373-378.

Fry, A. F., & Hale, S. (2000). Relationships among processing speed, working memory, and fluid intelligence in children. *Biological Psychology*, *54*(1), 1-34.

Frye, D., Zelazo, P., & Palfai, T. (1995). Theory of mind and rule-based reasoning. *Cognitive Development, 10*, 483–527.

Fuchs, L.S., Fuchs, D., Hosp, M.K., & Hamlett, C.L. (2003). The potential for diagnostic analysis within curriculum-based measurement. *Assessment for Effective Intervention, 28*, 13–22.

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology, 40*(2), 177–190.

Gallistel, C.R., Gelman, R., 1992. Preverbal and verbal counting and computation. *Cognition 44*, 43–74.

Gallistel, C.R., Gelman, R., 2004. *Mathematical cognition.* In: Campbell, J.I.D. (Ed.), *Handbook of Mathematical Cognition*. Psychology Press, New York.

Geary, D. C. (2011). Cognitive predictors of achievement growth in mathematics: a 5-year longitudinal study. *Developmental Psychology*, *47*(6), 1539.

Geier, R., Blumenfeld, P. C., Marx, R. W., Krajcik, J. S., Fishman, B., Soloway, E., & Clay-Chambers, J. (2008). Standardized test outcomes for students engaged in inquiry-based science curricula in the context of urban reform. J*ournal of Research in Science Teaching, 45*(8), 922–939. http://dx.doi.org/10.1002/tea.20248

Giedd, J. N., Blumenthal, J., Jeffries, N. O., Castellanos, F. X., Liu, H., Zijdenbos, A., . . . Rapoport, J. L. (1999). Brain development during childhood and adolescence: longitudinal MRI study. *Nature Neuroscience, 2*(10), 861-863.

Gilmore, C. K. (2006). Investigating children's understanding of inversion using the missing number paradigm. *Cognitive Development, 21*, 301-316.

Guilford, J. P. (1975). Varieties of creative giftedness, their measurement and development. *Gifted Child Quarterly, 19*, 107–121.

Guilford, J. P. (1988). Some changes in the structure-of-intellect model. *Educational and Psychological Measurement, 48*, 1–4

Gustafsson, J. E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research*, *28*(4), 407-434.

Gustafsson, J., & Undheim, J. O. (1996). *Individual differences in cognitive functions* Macmillan Library Reference Usa Prentice Hall International, New York London, NY.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33*, 204–229.

Haberman, S. J., Sinharay, S., & Puhan, G. (2005). Subscores for institutions. (ETS Research Rep. No. RR-06-13). Princeton, NJ: ETS.

Haberman, S. J., Sinharay, S., & Chon, K. H. (2013). Assessing item fit for unidimensional item response theory models using residuals from estimated item response functions. *Psychometrika, 78(*3), 417-440. doi:http://dx.doi.org/10.1007/s11336-012-9305-1

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika*, *75*(2), 209-227.

Haberman, S. J., & Sinharay, S. (2013). Does subgroup membership information lead to better estimation of true subscores?. British Journal of Mathematical and Statistical Psychology, 66(3), 452-469.

Hamilton C. J., Coates R. O., & Heffernan T. (2003). What develops in visuo-spatial working memory development? *European Journal of Cognitive Psychology.15*, 43–69.

Han, K. T. (2007). WinGen: Windows software that generates IRT parameters and item responses. *Applied Psychological Measurement, 31*(5), 457-459.

Han, K. T., & Hambleton, R. K. (2007). User's Manual: WinGen (*Center for Educational Assessment Report No. 642)*. Amherst, MA: University of Massachusetts, School

of Education.

Haladyna, T. M., & Kramer, G. A. (2004). The validity of subscores for a credentialing test. *Evaluation and the Health Professions*, 24(7), 349–368.

Harris, J., Hirsh-Pasek, K., & Newcombe, N. S. (2013). Understanding spatial transformations: similarities and differences between mental rotation and mental folding. *Cognitive processing*, *14*(2), 105-115.

Holmes, J., Adams, J. W., & Hamilton, C. J. (2008). The relationship between visuospatial sketchpad capacity and children's mathematical skills. *European Journal of Cognitive Psychology, 20*, 272–289.

Holmes, J., Gathercole, S. E., & Dunning, D. L. (2009). Adaptive training leads to sustained enhancement of poor working memory in children. *Developmental Science, 12*, F9–F15.

Holmes, J., & Gathercole, S. E. (2013). Taking working memory training from the laboratory into schools. *Educational Psychology, 34*(4), 440–450.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*(2), 179-185.

Hubbard, E. M., Piazza, M., Pinel, P., & Dehaene, S. (2005). Interactions between number and space in parietal cortex. *Nature Reviews Neuroscience, 6,* 435–448.

Hyde, Janet S., Elizabeth Fennema, and Susan J. Lamon. "Gender Differences in Mathematics Performance: A Meta-Analysis." *Psychological bulletin* 107.2 (1990): 139-55. ProQuest. Web. 8 Nov. 2014.

Iseman, J. S., & Naglieri, J. A. (2011). A cognitive strategy instruction to improve mathematics calculation for children with ADHD and LD: A randomized controlled study. *Journal of Learning Disabilities*, *44*(2), 184-195.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences of the United States of America,* 105, 6829–6833. doi:10.1073/pnas.0801268105

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Shah, P. (2011). Short- and long-term benefits of cognitive training. *Proceedings of the National Academy of Sciences of the United States of America,* 108, 10081– 10086. doi:10.1073/pnas.1103228108

Jaeggi, S. M., Studer-Luethi, B., Buschkuehl, M., Su, Y.-F., Jonides, J., & Perrig, W. J. (2010). The relationship between n-back performance and matrix reasoning—Implications for training and transfer. *Intelligence,38*, 625–635.

doi:10.1016/j.intell.2010.09.00

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

Kail, R. (1991). Developmental change in speed of processing during childhood and adolescence. *Psychological bulletin*, *109*(3), 490.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuo-spatial memory span and reasoning. *Journal of Experimental Psychology: General,* 133, 189-217.

Kaufman, A. S., & Kaufman, N. L. (1983). *Kaufman Assessment Battery for Children*. Circle Pines, MN: American Guidance Service.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American journal of psychology*, *62*(4), 498-525.

Kaufman, S. B., Reynolds, M. R., Liu, X., Kaufman, A. S., & McGrew, K. S. (2012). Are cognitive g and academic achievement g one and the same g? An exploration on the Woodcock–Johnson and Kaufman tests. *Intelligence*, *40*(2), 123-138.

Kelley, T. L. (1923). A new method for determining the significance of differences in intelligence and achievement scores. Journal of Educational Psychology, 14, 300–303.

Kirby, J. R., & Das, J. P. (1990). A cognitive approach to intelligence: Attention, coding and planning. *Canadian Psychology/Psychologie canadienne*, *31*(4), 320.

Kloo, D., & Perner, J. (2003). Training transfer between card sorting and false belief understanding: Helping children apply conflicting descriptions. *Child Development, 74,* 1823–1839.

Kosslyn, S. M. (1983). Ghosts in the mind's machine: Creating and using images in the brain. New York: Norton.

Kyttälä, M., Aunio, P., Lehto, J. E., Van Luit, J., & Hautamäki, J. (2003). Visuospatial working memory and early numeracy. *Educational and Child Psychology*, *20*(3), 65-76.

Kyttälä, M., Kanerva, K., & Kroesbergen, E. (2015). Training counting skills and working memory in preschool. *Scandinavian journal of psychology*, *56*(4), 363-370.

Kyttala, M., & Lehto, J. E. (2008). Some factors underlying mathematical performance:

The role of visuospatial working memory and non-verbal intelligence. *European Journal of Psychology of Education, 23*, 77−94.

Lawrence, I. M., & Curley, E. W. (1989). Differential item functioning for males and females on SAT-verbal reading subscores items: Follow-up study. (No. ETS-RR-89-22).

Lee, K., & Kang, S. (2002). Arithmetic operation and working memory: Differential suppression in dual tasks. *Cognition, 83*, B63−B68.

Lee, Y., Lu, M., & Ko, H. (2007). Effects of skill training on working memory capacity. *Learning and Instruction, 17*, 336–344. doi:10.1016/ j.learninstruc.2007.02.010

LeFevre, J. A., & Morris, J. (1999). More on the relation between division and multiplication in simple arithmetic: Evidence for mediation of division solutions via multiplication. *Memory & cognition*, *27*(5), 803-812.

Leighton, J. P., Gokiert, R. J., & Cui, Y. (2007). Using exploratory and confirmatory methods to identify the cognitive dimensions in a large-scale science assessment. *International Journal of Testing, 7*(2), 141-189.

Levine, M., & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika*, 53, 161-176.

Li, Y., & Geary, D. C. (2013). Developmental gains in visuospatial memory predict gains in mathematics achievement. *PloS one*, *8*(7), e70160.

Lohman, D. F., Stanford University. School of Education. Aptitude Research Project, & United States. Office of Naval Research. (1979). *Spatial ability: A review and reanalysis of the correlation literature*. Stanford, CA: School of Education, Stanford University.

Logie, R. H., & Van Der Meulen, M. (2009). Fragmenting and integrating visuospatial working memory. *The visual world in memory*, 1-32.

Lubke, G., & Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood?. *Multivariate Behavioral Research*, *41*(4), 499-532.

Luria, A. R. (1971). The origin and cerebral organization of man's conscious action. In *Proceedings of the Nineteenth International Congress of Psychology* (Vol. 19, pp. 37-52).

Lynn, R., & Meisenberg, G. (2010). National IQs calculated and validated for 108 nations. *Intelligence*, *38*(4), 353-360..

Mammarella IC, Pazzaglia F, Cornoldi C. (2008) Evidence for different components in

children's visuospatial working memory. *British Journal of Developmental Psychology. 26*, 337–355.

Mannamaa, M., Kikas, E., Peets, K., & Palu, A., (2012). Cognitive correlates of mathematics skills in third-grade students. Educational Psychology: *An International Journal of Experimental Educational Psychology, 32*(1), 21-44.

Mariani, M.A., & Barkley, R.A. (1997). Neuropsychological and academic functioning in preschool boys with attention deficit hyperactivity disorder. *Developmental Psychology, 13*, 111-129

McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P., & Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological methods,* 14(2), 126.

McDermott, P. A., Fantuzzo, J. W., & Glutting, J. J. (1990). Just say no to subtest analysis: A critique on Wechsler theory and practice. *Journal of Psychoeducational Assessment*, *8*(3), 290-302.

McLean, J. F., & Hitch, G. J. (1999). Working memory impairments in children with specific arithmetic learning difficulties. *Journal of Experimental Child Psychology*, *74*, 240–260.

Mehrens, W. A., & Phillips, S. E. (1986). Detecting impacts of curricular differences in achievement test data. *Journal of Educational Measurement, 23*(3), 185-196.

Mehrens, W. A., & Phillips, S. E. (1986). Detecting impacts of curricular differences in achievement test data. *Journal of Educational Measurement,* 23(3), 185-196.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*(2), 107-135.

Meyer, M. L., Salimpoor, V. N., Wu, S. S., Geary, D., and Menon, V. (2009).Differential contribution of specific working memory components to mathematical skills in 2nd and 3rd graders. *Learning& Individual Difference. 20*(2), 101-109.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*(2), 159-179.

Mix, K. S., & Cheng, Y. L. (2011). The relation between space and mathematics: developmental and educational implications. *Advances in child development and behavior*, *42*, 197-243.

Mix, K. S., Levine, S. C., Cheng, Y-L., Young, C. Konstantopolous, K., Hambrick, Z., & Ping, R. (in press). Developmental relations among specific spatial and Mathematics abilities: A factor analytic approach. *Journal of Experimental Psychology:*

General.

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General*, *130*(4), 621.

Morrison, A., & Chein, J. (2011). Does working memory training work? The promise and challenges of enhancing cognition by training working memory. *Psychonomic Bulletin & Review,* 18, 46–60. doi:10.3758/ s13423-010-0034-0.

Naglieri, J. A., & Rojahn, J. (2004). Construct validity of the PASS theory and CAS: Correlations with achievement. *Journal of Educational Psychology, 96*(1), 174-181.

Naglieri, J. A., & Otero, T. (2011). Cognitive Assessment System: Redefining intelligence from a neuropsychological perspective. *Handbook of pediatric neuropsychology*, 320-333.

Newcombe, N. S. (2010). Picture this: Increasing mathematics and science learning by improving spatial thinking. *American Educator*, 8, 29–43.

Nering, M. L. (1995). The distribution of person fit using true and estimated person parameters. *Applied Psychological Measurement, 19*(2), 121-129.

O'Toole, A. J., Jiang, F., Abdi, H., Pénard, N., Dunlop, J. P., & Parent, M. A. (2007). Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. *Journal of cognitive neuroscience*, *19*(11), 1735-1752.

Owen A. M., Hampshire A., Grahn J. A., Stenton R., Dajani S., Burns A. S., Howard R. J., Ballard C. G. (2010). Putting brain training to the test. *Nature* 465, 775-778.

Pashler, H. (1994). Dual-task interference in simple tasks: data and theory. *Psychological bulletin*, *116*(2), 220.

Pasini, M., & Tessari, A. (2001). Hemispheric specialization in quantification processes. *Psychological research*, *65*(1), 57-63.

Passolunghi, M. C., & Lanfranchi, S. (2012). Domain‐specific and domain‐general precursors of mathematical achievement: A longitudinal study from kindergarten to first grade. *British Journal of Educational Psychology*, *82*(1), 42-63.

Piazza, M., Fumarola, A., Chinello, A., & Melcher, D. (2011). Subitizing reflects visuo-spatial object individuation capacity. *Cognition*, *121*(1), 147-153.

Pickering, S. J. (2001). The development of visuo-spatial working memory. *Memory, 9,* 423–432.

Quinn, J.G., (2008). Movement and visual coding: the structure of visuo-spatial working memory. *Cognitive. Process. 9* (1), 35–43.

Radatz, H. (1979). Error analysis in mathematics education. *Journal for Research in mathematics Education*, 163-172.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, *25*, 111-164.

Rasmussen, C., & Bisanz, J. (2005). Representation and working memory in early arithmetic. *Journal of Experimental Child Psychology, 91*, 137–157.

Raymond & Feinberg (2015) "Subscores aren't for everyone: Alternative Strategies for Evaluating Subscore Utility." Annual meeting of the National Council on Measurement in Education in Chicago, IL, April 15-16, 2015.

Reckase M.D. (2009). *Multidimensional Item Response Theory.* Springer-Verlag, New York.

Reckase, Mark D. 1997. The Past and Future of Multidimensional Item Response Theory. *Applied Psychological Measurement,* 21, 25--36.

Reckase. M, D, McCrory, R., Floden R. E., Ferrini-Mundy, J., & Senk, S. L (2015). A Multidimensional Assessment of Teachers' Knowledge of Algebra for Teaching: Developing an Instrument and Supporting Valid Inferences, Educational Assessment, 20:4, 249-267, DOI: 10.1080/10627197.2015.1093927

Reckase, M. D., & Xu, J. R. (2015). The Evidence for a Subscore Structure in a Test of English Language Competency for English Language Learners. *Educational and Psychological Measurement*, 75, 805-825.

Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., ... & Engle, R. W. (2013). No evidence of intelligence improvement after working memory training: a randomized, placebo-controlled study. *Journal of Experimental Psychology*: General, 142(2), 359.

Ree, M. J., & Earles, J. A. (1991). Predicting training success: Not much more than *g*. *Personnel Psychology*, *44*(2), 321-332.

Reuhkala M. (2001) Mathematical skills in ninth-graders: Relationship with visuo-spatial abilities and working memory. *Educational Psychology.* 21, 387–399.

Rickard, T. C., Healy, A. F., & Bourne, L. E., Jr. (1994). On the cognitive structure of basic arithmetic skills. Operation, order, and symbol transfer effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1139-1153.

Rivera, S. M., Reiss, A. L., Eckert, M. A., & Menon, V. (2005). Developmental changes in mental arithmetic: evidence for increased functional specialization in the left inferior parietal cortex. *Cerebral Cortex,* 15(11), 1779-1790.

Rindermann, H., & Thompson, J. (2011). Cognitive Capitalism The Effect of Cognitive\ Ability on Wealth, as Mediated Through Scientific Achievement and Economic Freedom. *Psychological Science, 22*(6), 754-763.

Rohde T.E., Thompson L.A. (2007) Predicting academic achievement with cognitive ability. *Intelligence, 35*, 83–92.

Schatschneider, C., Francis, D. J., Foorman, B. R., Fletcher, J. M., & Mehta, P. (1999). The dimensionality of phonological awareness: An application of item response theory. *Journal of Educational Psychology*, *91*(3), 439.

Semmes, J. (1968). Hemispheric specialization: A possible clue to mechanism. *Neuropsychologia*, *6*(1), 11-26.

Shaw, P., Greenstein, D., Lerch, J., Clasen, L., Lenroot, R., Gogtay, N., . . . Giedd, J. (2006). Intellectual ability and cortical development in children and adolescents. *Nature, 440,* (7084), 676-679.

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

Shipstead, Z., Redick, T. S., & Engle, R. W. (2012). Is working memory training effective?. *Psychological bulletin, 138*(4), 628.

Siegler, R. S. (1994). Cognitive variability: a key to understanding cognitive development. *Current Directions in Psychological Science*, 3, 1-5.

Siegler, R. S. (2007). Cognitive variability. *Developmental Science,* 10, 104-109.

Simmons F. R., Willis C., & Adams A.M. (2012) Different components of working memory have different relationships with different mathematical skills. *Journal of Experimental Child Psychology,* 111, 139–155. doi: 10.1016/j.jecp.2011.08.0

Singley, K. & Anderson, J.R., (1989). *The transfer of cognitive skill*. Cambridge, MA: Harvard University Press.

Sinharay, S. (2010). When can subscores be expected to have added value? Results from operational and simulated data. *ETS Research Report Series*, *2010*(2), i-28.

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice, 30*(3), 29–40.

Sinharay, S. (in press). *When can subscores be expected to have added value*? Results from operational and simulated data. Princeton, NJ: ETS.

Sinharay, S., Haberman, S. J., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice*, *26*(4), 21–28.

Slotnick, S. D., & Moo, L. R. (2006). Prefrontal cortex hemispheric specialization for categorical and coordinate visual spatial memory. *Neuropsychologia*, *44*(9), 1560-1568.

Spearman, C., & Jones, L. W. (1950). *Human ability: A continuation of "the abilities of man"*. London: Macmillan.

Spearman, C. E. (1927). The nature of "intelligence" and the principles of cognition. London: Macmillan.

Spinath, B., Spinath, F. M., Harlaar, N., & Plomin, R. (2006). Predicting school achievement from general cognitive ability, self-perceived ability, and intrinsic value. *Intelligence*, *34*(4), 363-374.

St. Clair-Thompson, H. L., & Gathercole, S. E. (2006). Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. *Quarterly Journal of Experimental Psychology, 59,* 745–759.

St Clair‐Thompson, H.L, Stevens, R., Hunt, A., & Bolder, E. (2010). Improving children's working memory and classroom performance. *Educational Psychology*, *30*(2), 203-219.

Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2010). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. Applied Measurement in Education, 23, 63-86.

Stout, W., Nandakumar, R., Junker, B., Chang, H. H., & Steidinger, D. (1992). DIMTEST: A Fortran program for assessing dimensionality of binary item responses. *Applied Psychological Measurement*, *16*(3), 236-236.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction,* 4(4), 295-312.

Szűcs, D., Devine, A., Soltesz, F., Nobes, A., & Gabriel, F. (2014). Cognitive components of a mathematical processing network in 9‐year‐old children. *Developmental Science, 17*(4), 506-524.

Tate, R. L. (2004). Implications of multidimensionality for total score and subscore performance. *Applied Measurement in Education*, 17(2), 89–112.

Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika*, *49*(1), 95-110.

Tatsuoka, K. K. (1984). Changes in error types over learning stages. *Journal of educational psychology*, *76*(1), 120.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. *Diagnostic monitoring of skill and knowledge, acquisition*, 453-488.

Templin, J., & Bradshaw, L. P. (2014). Hierarchical diagnostic classification models: A family of models for estimating and testing attribute hierarchies. *Psychometrika*, 79, 317-339.

te Nijenhuis, J., van Vianen, A. E. M, & van der Flier, H. (2007). Score gains on g-loaded tests: No g. *Intelligence*, 35(3), 283-300.

Thissen D. & Edwards. M. C. (2005) Enhancing the Diagnostic Value of Large- Scale Achievement Tests: Technical Developments and Applications," Annual meeting of the National Council on Measurement in Education in Montreal, PQ, Canada, April 12-14, 2005.

Thomason, M. E., Race, E., Burrows, B., Whitfield-Gabrieli, S., Glover, G. H., & Gabrieli, J. D. (2009). Development of spatial and verbal working memory capacity in the human brain. *Journal of cognitive neuroscience*, *21*(2), 316-332.

Thurstone, L. L. (1947). *Multiple-factor analysis: A development and expansion of the vectors of the mind*. Chicago, Ill: The University of Chicago Press.

Van der Molen, M. J., Van Luit, J. E. H., Van der Molen, M. W., Klugkist,I., & Jongmans, M. J. (2010). Effectiveness of a computerised working memory training in adolescents with mild to borderline intellectual disabilities. *Journal of Intellectual Disability Research, 54*, 433–447.

Vecchi, T., Monticellai, M.L., Cornoldi, C. (1995) Visuo-spatial working memory: structures and variables affecting a capacity measure. *Neuropsychologia,* 33, 1549–1564.

Völkl-Kernstock S, Willinger U, Feucht M. (2006) Spatial perception and spatial memory in children with benign childhood epilepsy with centro-temporal spikes (BCECTS). Epilepsy Res, 72, 39–48.

Wai, J., Lubinski, D., & Benbow, C. P. (2009). Spatial ability for STEM domains: Aligning over 50 years of cumulative psychological knowledge solidifies its importance. *Journal of Educational Psychology, 101*, 817– 835.

Walker, C. M., Zhang, B., & Surber, J. (2008). Using a multidimensional differential item functioning framework to determine if reading ability affects student

performance in mathematics. Applied Measurement in Education, 21, 162-181.

Wallace B., Hofelich B. (1992) Process generalisation and the prediction of performance on mental imagery tasks. *Memory Cognition, 20*, 695–704.

Wang, P. P., Woodin, M. F., Kreps‐Falk, R., & Moss, E. M. (2000). Research on behavioral phenotypes: velocardiofacial syndrome (deletion 22q11. 2). *Developmental Medicine & Child Neurology*, *42*(6), 422-427.

Wechsler, D. (1950). Cognitive, conative, and non-intellective intelligence. *American Psychologist*, 5(3), 78.

Weintraub, S., & Mesulam, M. M. (1987). Right cerebral dominance in spatial attention: Further evidence based on ipsilateral neglect. *Archives of neurology*, *44*(6), 621-625.

Willis, J. O., Dumont, R., & Kaufman, A. S. (2011). Factor-analytic models of intelligence. In R. J. Sternberg & S. B. Kaufman (Eds.), *The Cambridge handbook of intelligence* (pp. 39–57). Cambridge, England: Cambridge University Press.

Wright, R., Thompson, W. L., Ganis, G., Newcombe, N. S., & Kosslyn, S. M. (2008). Training generalized spatial skills. *Psychonomic Bulletin & Review, 15*, 763–771.

Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, *358*(6389), 749-750.

Vrieze, S. I. (2012). Model selection and psychological theory: a discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological methods*, *17*(2), 228.

Vuokko, E., Niemivirta, M., & Helenius, P. (2013). Cortical activation patterns during subitizing and counting. *Brain research*, *1497*, 40-52.

Yao, L., & Boughton, K. A. (2007). A multidimensional item response modeling approach for improving subscale proficiency estimation and classification. *Applied Psychological Measurement, 31*, 83−105.

Zago, L., & Tzourio-Mazoyer, N. (2002). Distinguishing visuospatial working memory and complex mental calculation areas within the parietal lobes. *Neuroscience Letters,* 331,45–49.