Exploring the Aurora Battery, a gifted identification tool in a small sample of 4^{th}, 5^{th} and 6^{th} graders.

Samuel D. Mandelman

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2013

ABSTRACT

Exploring the Aurora Battery, a gifted identification tool

in a small sample of 4th, 5th and 6th graders.

Samuel D. Mandelman

The objective of this dissertation is to offer a series of analyses that contribute to the validation of the Aurora Battery, a cognitive assessment based on Robert J. Sternberg's theory of Successful Intelligence that is currently under development. Convergent validity will be examined by exploring how objective and subjective measures of the battery converge, through the novel application of the Correlated Trait Correlated Method-1, a specialized confirmatory factor analysis model that allows subjective measures to be compared against an objective measure. The predictive validity of Aurora will be shown by highlighting Aurora's ability to help predict students' school grade point average through latent growth curve models that are extended into path models. Divergent validity will be demonstrated by establishing sensitivity and specificity between the Aurora Battery and the TerraNova tests. Finally, the current state of the field of giftedness and possible future directions will be discussed.

Chapters

Chapter 1: Introduction

Chapter 2: CTCM-1 analysis comparing objective and subjective measures of Aurora.
Based on: Mandelman, S. D., Geiser, C., Mei, T., & Grigorenko, E. L. (2013). Who knows best? Applying the CTC(M-1) model to the Aurora Battery. Manuscript in preparation.

Chapter 3: Predicting GPA using Aurora.
Based on: Mandelman, S. D., Barbot, B., Mei, T., & Grigorenko, E. L. (2013). Predicting Grades and Their Change with Aurora. Manuscript in preparation.

Chapter 4: Sensitivity and Specificity analysis comparing the TerraNova and Aurora.
Based on: Mandelman, S. D., Barbot, B., Tan, M., & Grigorenko, E. L. (2013). Addressing the 'Quiet Crisis': Gifted Identification with Aurora. *Educational and Child Psychology 30*, 101-109.

Chapter 5: The current state of the field of gifted education.
Based on: Mandelman, S. D., & Grigorenko, E. L. (2013). Questioning the Unquestionable: Reviewing the evidence for the efficacy of gifted education. *Talent Development and Excellence, 5*, 125–137.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

ברוך ד׳ שהחיינו וקיימנו והגיענו לזמן הזה, אשר רק בעזרתו ובחסדו השגתי תואר זה,

ויהי רצון שעל ידי השגתי זה אזכה להגדיל תורה ולהאדירה.

At the culmination of a process such as this, one must look back and reflect as to how they got to this point. I would like to take this opportunity to thank those who have made this dissertation and degree possible. Firstly, I would like to thank my mentor Dr. Elena L. Grigorenko, who challenged me constantly to develop to the greatest extent possible, and who gave me professional opportunities one could only dream of. What I have attained in my graduate career is to her credit. I would like to thank Dr. John B. Black, my dissertation sponsor, who allowed me to pursue research in the area of my interest. Thank you to my dissertation committee members Dr. John B. Black, Dr. Elena L. Grigorenko, Dr. Matthew S. Johnson, Dr. Gary Natriello and Dr. Steven T. Peverly for serving on my committee and offering great insight and invaluable feedback that improved and enriched this dissertation. I would also like to thank Mr. Gary Ardan of the Office of Doctoral Studies for all his help throughout this process.

Throughout most of my time at Teachers College I was privileged to be a member of the EGLab of the Child Study Center at the Yale School of Medicine. It was there, under the direction of Dr. Elena L. Grigorenko that I was introduced to and had the opportunity to work on The Aurora Battery, which is the focus of this dissertation, as well as being afforded the chance to pursue other research. This dissertation would not have been possible without my colleagues and collaborators at the Child Study Center and beyond. I would like to thank my friend and colleague Ms. Mei Tan, the Aurora Project Director, who allowed me to work on many facets of this project, and helped facilitate the data collection and scoring of the data presented within this dissertation, as well as working with me on many other papers. Her endless hours of help, support and editing enhanced so much of what I have to show for my graduate career. The scholarship presented in this dissertation would not have been possible without the close collaboration of Dr. Baptiste Barbot and Dr. Christian Geiser. Their expertise formed this dissertation into what it is. I would like to thank Mrs. Ethel Gottlieb for her expert copy editing that made this dissertation readable.

I would like to thank Mr. Stanley Frankel, who was kind enough to connect me with the wonderful administration of Hillel Day School where the data for this dissertation was collected. I would like to thank the administration of the Hillel Day School, Mr. Steve Freedman, Head of School and Mrs. Joanne Sokol, Assistant Head of School, who could not have possibly been more helpful with every aspect of this project. Thank you to the parents, students and teachers of Hillel Day School, who participated in this project.

I would like to thank my friends at TC with whom I worked and commiserated with, and who helped me through this very long process. I would also like to thank friends from outside of TC who help make the years that I was pursuing this degree more tolerable.

יש לי הרבה הכרת הטוב למורי ורבי שליט״א על הכל במשך כל השנים.

One cannot possibly get to the stage of doing a dissertation, let alone finishing one, without a great deal of support and encouragement from ones' family. I would like to thank Mrs. (Bubby) Alice Berlin for making this degree possible. I would like to thank my in laws Dr. Jeffrey and Mrs. Caren Goldenberg for their support and encouragement throughout the entire process. Thank you to my mother Mrs. Ann G. Mandelman, who despite many very challenging years of schooling and otherwise, always believed in me and had fullest confidence that I could achieve anything I wanted to. I would also like to thank the rest of my family members for all that they have done for me over the years.

No words can possibly express my appreciation and gratitude to my wife, Chavie, who has been my strongest supporter through every step of this process, and who has sacrificed so much so I could complete this dissertation and degree. She has been there to listen, advise and has done everything in her power to make getting this degree any bit easier. Most importantly she has been the most incredible mother to our son Moshe. To you my Moshe, you have brought so much into my life. You have provided much needed balance and fun while working on my dissertation, and have clarified for me why it was that I was pursuing this degree.

DEDICATION

This dissertation is dedicated to my son **Moshe**, who has brought unbridled joy into my life, and who has taught me more about learning and development than any class or book possibly could.

Chapter 1

Gifted children are our nation's most precious natural resource as they represent our nation's future (Sternberg, 1996b). This sentiment is held by countries around the world as they recognize the great importance and value of gifted children (and even explicitly refer to gifted students as natural resources see McCann, 2005; Milgram, 2000). The increasing interest in gifted education, and education in general, around the world is important for the advancement of humanity. It is even more important for those countries that wish to remain leaders of the free world, to ensure that their countries' intellectual capital is developed to the fullest extent possible.

Many countries have made, and are continuing to make, great advances in their educational policies and programming for the gifted, making this an extremely pressing issue for the U.S. if they wish to maintain their position globally. In addition, the political, societal, and intellectual elite have created a global knowledge economy (Thompson, 2004; Tomlinson, 2008). In the past, the U.S. has responded to global intellectual competition with increased focus on educating the nation's gifted. It was the launch of Sputnik that spurred the U.S. government to pass The National Defense Education Act (NDEA), passed by Congress in 1958 to support large scale gifted education initiatives. With other countries making advances in gifted programming, the U.S. must evaluate its practices regarding gifted education, if they wish to maintain an edge in the development of human capital.

One of the greatest challenges facing the field of gifted education is the identification of gifted students. This dissertation will focus on exploratory research validating a cognitive assessment battery that is currently under development and that can be used in gifted identification. The purpose of this investigation is not only to further the validation of the assessment, but also to serve as a basis for future research on this battery. The goal of the battery, that is being developed, which will be described in detail in the coming pages, is to be able to capture individual differences in students' abilities. This assessment has particular usefulness in gifted identification, since the abilities that it taps are not the same abilities that are assessed by conventional measures, such as the Wechsler Intelligence Scale for Children (WISC) or standardized achievement tests such as the TerraNova. Yet, they are abilities that can lead to success in school and beyond (Sternberg & Grigorenko, 2004; Sternberg & The Rainbow Project Collaborators, 2006). Those found to have gifts in the areas that the battery covers, whether or not they would be identified as gifted using conventional tests, deserve the opportunity to develop these abilities further. It is the objective of this dissertation to contribute to the validation of the Aurora Battery by establishing its convergent, predictive and divergent and validity. Convergent validity will be examined in Chapter 2 by exploring how objective and subjective measures of the battery converge. The predictive validity of Aurora will be shown by highlighting Aurora's ability to help predict students' school grade point average in Chapter 3. Finally, divergent validity will be demonstrated in Chapter 4 by establishing sensitivity and specificity indexes that demonstrate the disparity between those identified as gifted using different measures.

The challenge of formally identifying unique intellectual profiles can be traced back nearly a century ago and led to the development of standardized tests of intelligence. While this may sound like a mere historical anecdote, it is sobering that a survey of the literature reveals that most of the tools being used today to identify gifted children, are revisions of, or largely based upon theoretical frameworks and instruments constructed a century ago (Mandelman, Tan, Aljughaiman, & Grigorenko, 2010). While these theoretical frameworks may still be psychometrically sound, instruments based on them continue to exclude individuals who possess gifts worthy of attention and nurturing. These exclusions have been and continue to be one of the greatest objections to gifted education, an issue that has been referred to by the U.S. government as a "quiet crisis" (U.S. Department of Education, 1993). The U.S. Government in the Jacob K. Javits Gifted and Talented Education Act (U.S. Department of Education, 1993) explicitly states, when defining gifted and talented, that "Outstanding talents are present in children and youth from all cultural groups, across all economic strata, and in all areas of human endeavor" (p.26). The identification methods that are being used most widely today (*g* based assessments) are not in line with this definition and continue to exclude many with diverse abilities and backgrounds.

Triarchic Theory of Successful Intelligence

Since the time that psychometric theories of intelligence were introduced, there have been many attempts to better define and classify human intelligence (for a review of theories of intelligence see Sternberg, 2005a; Sternberg, Jarvin, & Grigorenko, 2011). Particularly since the 1980's there has been great interest in alternative, more encompassing models of human intelligence and ability. One such model is Robert J. Sternberg's Triarchic Theory of Successful Intelligence (1985, 1988, 1996a, 1999b, 2005a). According to his theory, intelligence relates to three realms of the individual experience: the internal world of a person, one's experience, and the external world. The internal world of a person is the information processing capacity that the person possesses. Sternberg further divides this information processing into three specific components: 1) higher order cognition, including meta-cognitive and executive processes; 2) the lower level abilities referred to as performance components, which are controlled by the higher order abilities; and 3) the knowledge acquisition components that allow the person to learn how to approach and solve problems.

The next aspect that Sternberg describes is how one's experience interacts with their intelligence. The simple logic presented suggests that the tasks which we are familiar with and have encountered in the past require less cognitive resources than those that are novel. How one's intelligence interacts with the external world, the last aspect put forth in this theory, considers how one interacts with one's environment. Intelligence can determine how one adapts to one's environment, how one changes the current environment to make it new, and, finally, how one chooses one environment over another.

The next part of Sternberg's theory, which is most widely known, concerns the interplay between three abilities which together with the aforementioned components combine to make up Successful Intelligence. The three abilities highlighted in this theory are analytical, creative and practical. Analytical abilities are involved in analyzing, evaluating, judging, and comparing and contrasting. Analytical abilities are exhibited in reasoning and logical thinking as they are exercised in activities such as debating, research, and mathematical problem-solving. Creative abilities are reflected in the capacity to generate new ideas, create and design new things. Creative abilities are particularly well measured by problems assessing how well an individual copes with relative novelty. Practical abilities are involved when individuals apply their abilities to the kinds of problems that confront them in daily life, such as on the job or in the home. Practical abilities are exercised in leadership and other social interactions, as well as in the adaptation and application of knowledge in real world problem-solving. A successfully intelligent person does not necessarily have to possess high levels of each of these abilities to be considered intelligent; rather one must recognize one's own strengths and weaknesses and create compensatory strategies that rely on those strengths.

In sum, successful intelligence is defined as the integrated set of abilities needed to attain success in life, however an individual defines it within his or her sociocultural context. Successfully intelligent people adapt to, shape, and select environments through a balance in their use of analytical, creative and practical abilities. According to this view, intelligence and success are defined beyond performance in school to the broader context of what happens in life. Therefore, early recognition of, and teaching to, these component abilities of intelligence can set children on a road to success that will last well beyond their time in school. The assessment battery, which will be the focus of this dissertation, is a new assessment for intelligence that is under development and is based on a broader, more complete theory of intelligence—the theory of successful intelligence (Sternberg, 1999a, 2005b)

The importance of a better understanding of student abilities—by teachers, parents, administrators and students themselves—cannot be overstated. As it stands now, student evaluation is largely based on academic achievement (as indicated either by class grades or by standardized test scores) and teacher assessment. Both of these, though useful and valuable for certain purposes, are actually quite limited in perspective. These limitations lead to unrecognized and un-nurtured ability, and for at least some students, may have devastating long-term effects. The assessment that will be explored is a multi-tool, multi-informant battery—*The Aurora Battery*— based on the theory of successful intelligence(Sternberg, 1999a, 2005b). Its purpose is not only to provide a much-needed broader perspective of student ability, but also to re-define what is important to address and nurture in school for individuals' long-term success toward achieving life goals.

Aurora's aim is to effectively broaden the scope of identification in the school setting by assessing important cognitive abilities (namely creative and practical) that have been overlooked in the past. It can also be used as a tool to identify areas of cognitive strength and weakness in individuals. Based on the widely accepted premise that cognitive strengths and weaknesses can be identified, Aurora can be utilized as a guide for adjusting the delivery and assessment of curriculum material, and for addressing areas of thinking that need practice and development. A secondary but no less important purpose of Aurora is to challenge the status quo and bring to the forefront of the field of education the importance of abilities that are not traditionally recognized in the school setting. These abilities can and do affect, academic performance and successful outcomes both in school and in life.

Over the past half a century the world has changed significantly. One of the greatest driving forces behind this change is the advancement of technology. This technological advancement has contributed to rapid globalization. With growing globalization, the economies around the world, in turn, have had to change as well. Whereas formerly geographic borders were often seen as trade barriers, in the current information age this is no longer the case. With the advancement of technology and globalization there has been the development of a global knowledge economy. In the past, countries relied heavily on natural resources and industry to drive their economies; now many of these economies have become knowledge economies that are driven by human capital and knowledge (Powell & Snellman, 2004).

With the development of global knowledge economies has come the recognition that the most important commodity is human ability. Many countries have responded to this change with educational reforms and programming designed to keep the country competitive within the global knowledge economy (Tomlinson, 2008). The extent of the competition for the brightest and most capable has been called a “global war for talent” (P. Brown & Tannock, 2009). Just as the economies themselves have changed, so have the abilities needed to be successful. Some of the abilities that are necessary to compete in a global knowledge economy are analytical, problem solving, creativity, communication skills and tacit knowledge (The World Bank, 2003). The skills included by The World Bank (2003) as being vital skills for the new economy are some of the very skills that comprise Successful Intelligence and highlights the importance of identifying and developing these skills.

Description of the Battery

The Aurora Battery is currently under development and the studies presented here will help to further the validation of this battery and to serve as a basis for further research. Aurora's goal is to provide a multifaceted view of a child's intellectual profile, as expressed through a range of abilities. It is designed to meet the needs of parents who are concerned with their children's intellectual performance, and for teachers, counselors, schools or school systems in their attempt to identify individual differences among their student body.

The Aurora Battery (Chart, Grigorenko, & Sternberg, 2008) is composed of multiple modules that involve multiple informants, including a group administered maximal performance assessment, a parent rating scale (Aurora-*i*, formerly a parent interview), a teacher rating scale (Aurora-*r*), and a self-report rating scale (Aurora-*s*, Mandelman, Tan, Kornilov, Sternberg, & Grigorenko, 2010). It measures abilities in the areas of analytical, creative and practical thinking. Each module of the battery is structured on the same grid of abilities and domains, and each is composed of items or subtests that examine an ability—analytical, creative, and practical—in one of three domains; verbal, numerical, and visual-spatial.

Aurora-*a* is composed of 17 sub-tests that are designed to assess the child's analytical, creative, and practical abilities. There are 6 subtests for the assessment of analytical ability; 5 for the assessment of creative ability; and 6 for the assessment of practical ability (see Appendix A

Table 1. The Aurora Battery Structure

	Analytical	Creative	Practical
Images - Figural	Floating Boats: Identify matching patterns among connected boats. (5 items)(MC)	Book Covers: Interpret an abstract picture and invent a story to accompany it. (5 items) (OE) Multiple Uses: Devise three new uses for each of several household items. (5 items)(OE)	Paper Cutting: Identify the proper unfolded version of a cut piece of paper. (10 items)(MC) Toy Shadows: Identify the shadow that will be cast by a toy in a specific orientation. (8 items)(MC)
Words - Verbal	Words That Sound the Same (Homophone Blanks): Complete a sentence with two missing words using homonyms. (20 items) (RW) (Limited) Metaphors: Explain how two somewhat unrelated things are alike. (10 items)(OE)	(Inanimate) Conversations: Create dialogues between objects that cannot typically talk. (10 items) (OE) Interesting (Figurative) Language: Interpret what sentence logically comes next after one containing figurative language. (12 items)(MC)	(Silly) Headlines: Identify and explain an alternative “silly” meaning of actual headlines. (11 items)(RW) Decisions: List elements given in a scenario on either “good” or “bad” side of a list in order to make a decision. (3 items) (RW)
Numbers - Numerical	Number Cards (Letter Math): Find the single-digit number that letters represent in equations. (5 items) (RW) Story Problems (Algebra): (before any Algebra training) devise ways to solve logical math problems with two or more missing variables. (5 items) (RW)	Number Talk: Imagine reasons for various described social interactions between numbers. (7 items)(OE)	Maps (Logistics Mapping): Trace the best carpooling routes to take between friends’ houses and destinations. (10 items) (RW) Money (Exchange): Divide complicated “bills” appropriately between friends. (5 items) (RW)

Note. MC: Multiple Choice; OE: Open-ended items that need to be scored by an individual using a rating scale; RW: Answers are either Right or Wrong for example questions). The subtests represent multiple formats—multiple choice (5 subtests), short answer (scored right or wrong; 7 subtests), and open-ended response questions, which are rated by trained expert judges (5 subtests). Kornilov, Tan, Elliott, Sternberg, and Grigorenko (2012) reported a median internal consistency coefficient for Aurora of .67 in a British sample, while in the current sample the median Cronbach's was alpha .79.

Aurora-*g* is composed of 9 sub-tests to assess the *g*-factor of intelligence, as in conventional tests. All of its subtests are multiple- choice. There are 3 verbal, 3 numerical, and 3 figural subtests (see Appendix A for an example of a question). The three areas that are explored in each domain are analogies, classification tasks, and series tasks (i.e. in the verbal domain, there is a verbal series subtest, a verbal analogy subtest, and a verbal classification subtest).

Aurora-*i* (parent), Aurora-*r* (teacher) (Chart, et al., 2008) and Aurora-*s* ([self] Mandelman, Tan, Kornilov, et al., 2010) are parallel rating scales that are filled out by parents, teachers and students. The scales consist of a total of 40 questions with the questions evenly divided between the abilities of memory, analytical, practical and creative (see Appendix B for example questions). The scales were carefully designed to allow for exactly parallel scales to be used. The use of multiple informants in the identification process has been suggested to help overcome the under-identification of diverse groups of gifted students (Ford & Trotman, 2000; VanTassel-Baska, Feng, & De Brux, 2007).

The Aurora Battery is unique in the theory it employs, its design, format and administration. It has a strong potential to fill a substantial gap in school practice by providing information on a child's cognitive ability that is accessible to, and usable by, educators in the classroom. The battery does not simply attempt to measure by proxy a proposed underlying psychological construct, like many of the instruments that we regularly use. Another characteristic that differentiates Aurora is that it is based upon the best practice in using multiple informants. The coming chapters are meant to contribute to the validation of the Aurora Battery by establishing convergent, predictive and divergent validity. In Chapter 2 convergent validity will be demonstrated by exploring how objective and subjective measures of the battery converge. The predictive validity of Aurora will be shown by highlighting Aurora's ability to help predict students' school grade point average in Chapter 3. Finally in Chapter 4, divergent validity will be established by establishing sensitivity and specificity indexes that demonstrate the disparity between those identified as gifted using different measures.

Chapter 2

Introduction

The fact that identification of the gifted is one of the most widely written about and critical topics in the field of gifted education (S. W. Brown et al., 2005; Dai, Swanson, & Cheng, 2011) does not mean that the discussion has brought about a great deal of change in gifted identification. Gifted children have been identified using IQ tests for one hundred years (Pfeiffer, 2012; Sternberg, et al., 2011) and those IQ test have changed little over that time period (Mandelman, Tan, Aljughaiman, et al., 2010; Sternberg, et al., 2011). IQ tests have been, and are still, the most common means used by schools and states to identify gifted students (S. W. Brown, et al., 2005; McClain & Pfeiffer, 2012). The problem is not that IQ test haven't changed or their age per se; the problem is whom they identify—or to be more precise—whom they do not identify. There continues to be great potential that is never identified and lost because of the tests being used. Students, whose abilities extend outside of what the relatively narrow *g* based IQ test measure, are often not identified. Students from minority groups who are, and have historically been, under-represented in gifted programs, as well as those students with diverse cognitive profiles (Brody & Mills, 1997; Ford & Trotman, 2000; Maker, 1996; McCoach, Kehle, Bray, & Siegle, 2001) are not identified as a result of the continued reliance on these measures. The problem is not limited to the use of IQ tests themselves but also to the way the scores that they produce are used (McClain & Pfeiffer, 2012), and how high IQ scores have become synonymous with giftedness (Pfeiffer, 2012). Moreover, many rely on IQ tests as the single measure to identify giftedness. Borland (2009) contends that equating high IQ to giftedness is one of the reasons that minorities are consistently under-represented in gifted programs and refers to it as one of the field's greatest failings.

Despite the many issues with IQ tests, their continued widespread use can be attributed to a number of factors. In fact, Sternberg, et al. (2011) before offering alternative methods for identification, present 15 possible reasons as to why IQ tests and high IQ has maintained its dominance as part of giftedness and gifted identification. Among those reasons are their existence for more than 100 years that have allowed people to become familiar with IQ tests, to the extent of defining intelligence as what IQ tests measure (Boring, 1923). IQ test are also viewed as objective measures that attempt to neatly quantify a complex construct. Additionally, IQ tests generally have high reliabilities and good criterion validity. The relative ease of administration and scoring as well as the time and money saved over the use of more time-consuming and expensive identification methods have also contributed to their continued dominance. However, these benefits must be weighed against the profound disadvantages of IQ tests.

The greatest limitation of IQ tests is that the models that they are based on and the abilities that they measure are simply too narrow. Sternberg (2010) explains that to succeed in life in the real world you need a more varied set of abilities than those represented on IQ tests, which are the one that are most valued by schools. Renzulli's (2005) recognition of this led him to distinguish between two types of giftedness: "schoolhouse giftedness" and "creative-productive giftedness." In this distinction schoolhouse giftedness is most aligned with what IQ test measure, but even those abilities that are needed to succeed in school are not represented fully on IQ tests, and certainly not those abilities that comprise non-school related giftedness. Renzulli (2005) further points out that even the fathers of psychometric conceptions of intelligence and giftedness, such as Terman and Thorndike, themselves acknowledged the great limitations of the IQ tests.

Two prominent examples of theories giftedness that include abilities beyond the traditionally valued analytical abilities are Renzulli's Three Ring Conception of Giftedness (1978, 1986, 2005) and Sternberg's Triarchic Theory of Successful Intelligence (1985, 1988, 1996a, 1999b, 2005a). In Renzulli's model, which is the most widely known and probably the most influential model of giftedness, the three rings that converge to make up this theory are above-average ability, high levels of task commitment, and high levels of creativity. In Sternberg's model the abilities that are highlighted are analytical, creative and practical among other higher order abilities. The ability, which both of these modern encompassing models of giftedness have in common is creativity.

A thorough review of the literature on gifted identification, reveals a topic of great concern to researchers (Pfeiffer, 2003) and practitioners (S. W. Brown, et al., 2005) in the field of giftedness alike: the use of IQ tests as the sole measure in identifying giftedness and the failure to use multiple criteria that includes input from multiple informants. The use of multiple criteria and information from multiple sources is one of the most universally agreed upon practices in the field of gifted education (see Borland, 2009; Briggs, Reis, & Sullivan, 2008; Brody & Mills, 1997; S. W. Brown, et al., 2005; Coleman, 2003; Ford & Trotman, 2000; Friedman-Nimz, 2009; Maker, 1996; McClain & Pfeiffer, 2012; McCoach, et al., 2001; Nielsen, 2002; Pfeiffer, 2001, 2003; Pfeiffer, 2012; Pfeiffer & Blei, 2008; VanTassel-Baska, et al., 2007). The premise behind the use of multiple informants is that since human cognitive abilities are complex and multidimensional, so too, should be the tools that are used to identify these abilities (Friedman-Nimz, 2009; McClain & Pfeiffer, 2012; Pfeiffer, 2012). Besides the need to use multiple objective assessments to try to capture some of the complexity of giftedness, ratings of students from others are extremely valuable. Gifted behaviors manifest themselves differently across contexts. Eliciting information about students from people who interact with the student in different contexts is helpful in getting a better sense of the student and a more comprehensive understanding of their abilities.

While the use of multiple criteria in the identification process is considered to be the best practice for all students (Coleman, 2003; National Association for Gifted Children, 2012), it is of particular importance in identifying students from groups that have been historically under-represented in gifted programs (VanTassel-Baska, et al., 2007). This practice has been specifically argued for when identifying twice exceptional students (Brody & Mills, 1997; McCoach, et al., 2001; Nielsen, 2002) and students from diverse backgrounds (Briggs, et al., 2008; Ford & Trotman, 2000; Maker, 1996; VanTassel-Baska, et al., 2007).

Pfeiffer (2003), when suggesting how to improve the gifted identification process, argues the need for, in his words “identification packages”. These are based on sound theories of giftedness that offer multiple tools for multiple informants, and are culturally and linguistically sensitive. The Aurora Battery (Chart, et al., 2008), the assessment examined in this dissertation, fulfills these needs.

While the use of multiple informants in the gifted identification process is widely agreed upon and promulgated and is considered to be the best practice, there is no literature that includes an empirical examination of the use of multiple informants in gifted identification. The use of multiple informants, which allows for many pieces of information from different sources to be available in the identification process, does not come without challenges. When multiple informants are used, their ratings often differ and differ significantly. While the literature on psychopathology is replete with studies looking at cross-informant agreement (e.g. Achenbach, Krukowski, Dumenci, & Ivanova, 2005) there is little to no literature in the field of gifted identification. This scarcity may be due to the somewhat limited use of multiple informants to allow for studies of this nature. Questions that commonly arise when examining multi-rater data include the accuracy of the ratings, the possibility of shared rater bias, and whether differences exist in the accuracy of the ratings depending on the subject being rated. When multiple sources of information other than standardized tests are used, the question arises as to whether (and to which extent) these other sources converge with standardized tests and with each other as opposed to providing a unique perspective that is not shared with other sources of input.

Research Questions:

To what extent do the subjective Aurora teachers, parents & students rating scales (Aurora-*r*,*s*,*i*) converge with Aurora's objective assessment of abilities (Aurora-*a, g*) ?

What is the degree of agreement between raters when rating student abilities? How much of the agreement between raters reflects true convergent validity as opposed to reflecting a shared rater bias?

Materials and Method

Participants

The participants who took Aurora (N = 145, 69 female, 76 male, Mean *(M)* = 10.29, Standard Deviation (*SD)* = 1.0) in this study were 4th, 5th and 6th graders attending a suburban Midwestern United States private parochial school (4th graders, n=50, 29M/21F, *M* 9.20, *SD* .40; 5th graders, n = 48, 23M/25F, *M* = 10.35, *SD* = 0.48, 6th graders, n = 47, 24M/25F, *M* 11.38, *SD* .49).

Instruments

The Aurora Battery described above.

Procedure

Parental consents along with students' assents from all of the participants were collected in accordance with Human Subjects Committee protocol of Yale University. The Aurora testing took place over three days in a large group setting. Aurora was administered over three consecutive days; each day consisted of two 45 minute testing periods separated by a 15-minute break between them. On the first day of testing, before taking Aurora-*a*, students were given 20 minutes to complete Aurora-*s*. Aurora-*r* was completed by teachers during their weekly professional development session. Aurora-*i* was sent home to parents to complete at their convenience.

Statistical Analyses

Confirmatory factor analysis (CFA) is currently the most widely used approach to analyze multitrait-multimethod (MTMM) data due to its ability to separate measurement error from true construct validity and method specificity, as well as other advantages that have been detailed elsewhere (e.g., Eid, Lischetzke, & Nussbeck, 2006). Being that the Aurora Battery is an MTMM instrument, a CFA-MTMM analysis was employed in the present study to examine to what degree the objective and subjective assessments of giftedness converge.

Step 1: Simple CFA model. In the CFA-MTMM modeling approach, the two-step procedure described in Geiser, Eid, West, Lischetzke, & Nussbeck (2012) was used. In Step 1, simple CFA models for the self, parent, and teacher ratings were estimated. An example of such a simple CFA model is shown in Figure 1 for numeric analytical abilities. Figure 1 shows that there is a separate latent factor for each type of rater in this model. Hence, the factors are rater-specific. All latent factors can be correlated. The higher the correlations between the rater-specific factors in the Step 1 CFA model, the higher the agreement between the three rater types.

Figure 1

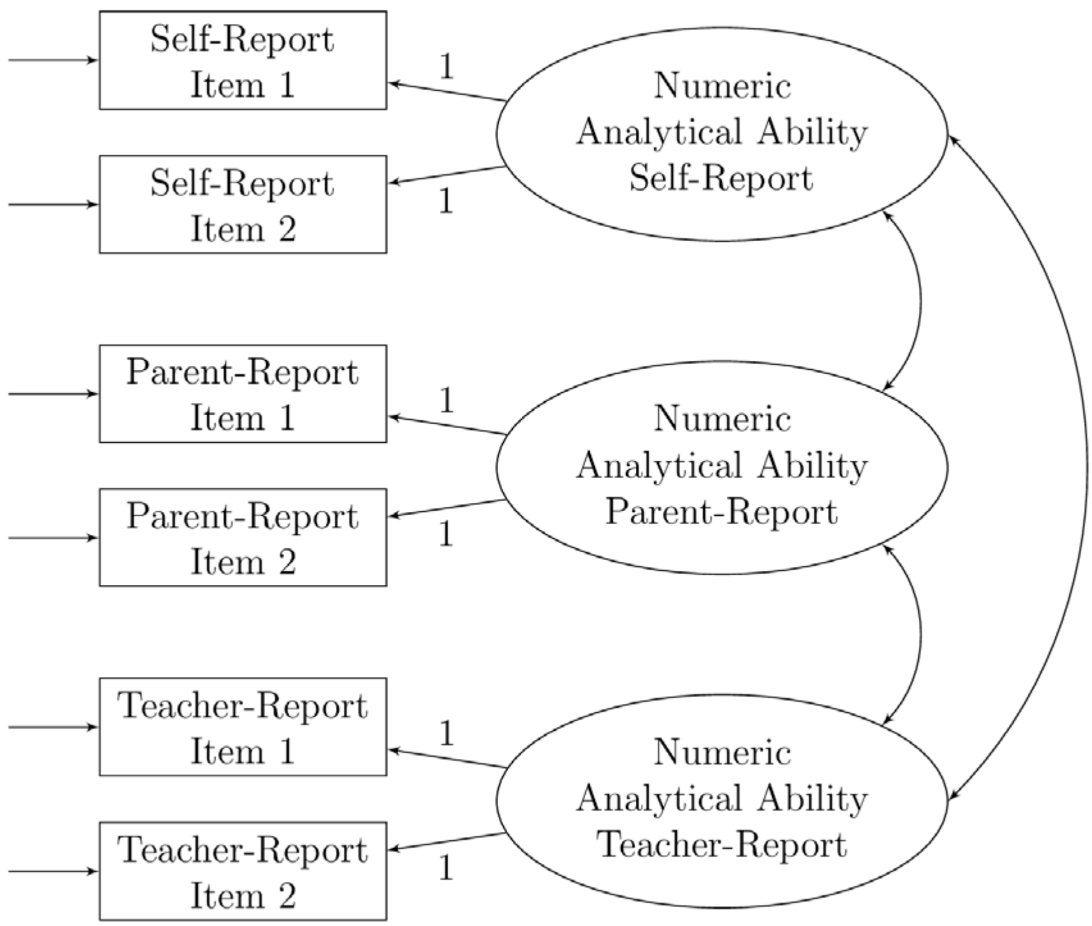


Figure 1. Simple CFA Model for Measuring Inter-Rater Agreement for the Construct Numeric Analytical Ability. Correlations between rater factors indicate the degree of agreement between raters.

The simple CFA model is useful to quantify the degree of agreement between raters (in terms of the zero-order correlations between rater-specific factors). However, this model does not tell us whether this agreement reflects true convergent validity (shared true ability variance) or just a shared bias across raters. That is, even when ability ratings are highly correlated, one cannot be sure that these correlations actually represent true ability variance. Therefore, the simple CFA model in Step 2 was extended by including the objective Aurora ability measures and contrasting these measures against the latent rater factors. This allowed for testing whether the zero-order correlations between raters reflected true convergent validity or just a shared rater bias.

Step 2: CTC(M-1) model. Eid (2000; Eid, Lischetzke, Nussbeck, & Trierweiler, 2003) proposed a CFA-MTMM approach that allows one to contrast a reference (or "gold standard") method, such as objective ability tests against other (non-reference) methods used in the study, in this case, ability ratings. This approach is known as the *Correlated Traits-Correlated (Methods – 1)* [*CTC(M–1)*] *model* and can be seen as an extension of the simple CFA model. More specifically, the CTC(M–1) model allows one to find out to what extent the scores provided by different raters (e.g., Aurora self, parent, and teacher reports) converge with an objective assessment of abilities (e.g., the Aurora-*a*, -*g* test batteries). Furthermore, the CTC(M–1) model allows quantifying the degree of convergent validity (consistency) and method specificity in terms of proportions of explained variance, and it can be used to examine whether rater biases differ for different abilities. Given that the objective Aurora test scores provided a clear "gold standard" measure in the present study, the CTC(M-1) approach was perfectly suited to examine the research questions. In the following section, a more detailed description of the CTC(M-1) approach is provided and its use in Step 2 of the present study.

Figure 2 shows a path diagram of a CTC(M-1) model for the Aurora construct *numeric analytical ability* measured by the four assessment methods used in Aurora. In the CTC(M-1) model, the reference method defines the meaning of the latent ability factor (Geiser, Eid, & Nussbeck, 2008). In the present study, the objective Aurora test scores served as the reference method. Therefore, in the example shown in Figure 2, there is no method factor for the objective tests (because the objective tests are seen as “unbiased” with regard to children’s true abilities), and the latent ability factor can thus be interpreted as *numeric analytical ability measured by objective tests*. This latent ability factor can also be seen as the common true score variable underlying the observed test scores pertaining to the reference method (here: the objective tests). It reflects individual differences on the objective tests, corrected for random measurement error.

The remaining methods (self, teacher, and parent ratings) are contrasted against the reference factor in a latent regression analysis (Eid, et al., 2003). The method factors for each of the three types of raters represent this contrast and are interpreted as residual factors. The method factors are by definition uncorrelated with the reference factor and reflect the measurement-error-free rater variance that is unique to each type of rater—in the sense that it is not shared with the reference method. Hence, the CTC(M-1) model partitions the error-free rater variance into two parts: One part that can be predicted from (is shared with) the true scores underlying the objective test scores and one part that cannot be predicted from (is not shared with) the test scores (but may potentially be shared with other raters, indicating a common rater bias, see below).

Using the CTC(M-1) model, a straightforward way to quantify the degree of convergent validity (*consistency*) between raters and objective tests is to examine the squared standardized factor loadings (SSLs) of the non-reference method indicators on the reference factor. The SSLs

Figure 2

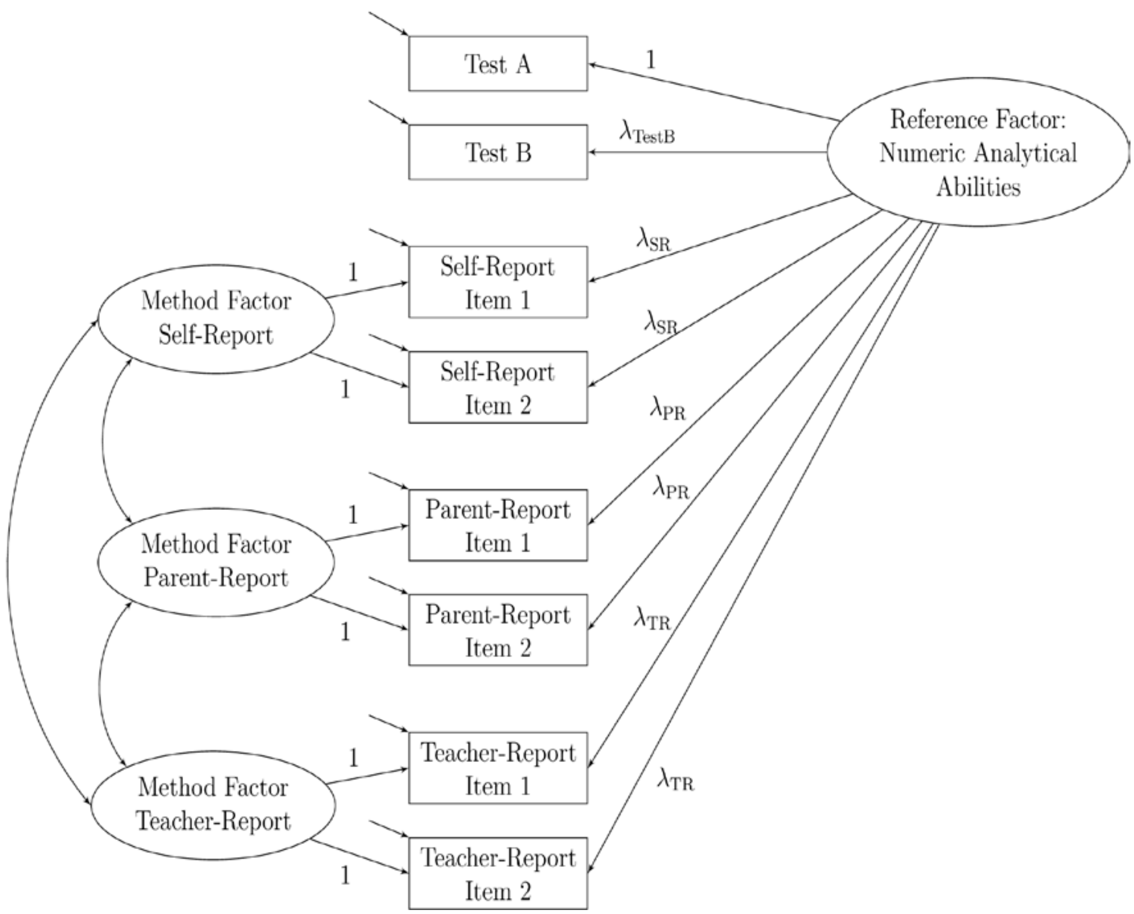


Figure 2. Example of a Domain-Specific CTC(M-1) Analysis (here: for numeric analytical ability). The reference factor is defined by objective tests of numeric analytical ability. The self, parent, and teacher ratings are contrasted against the reference factor. Correlations between method factors indicate whether raters share a common bias.

λ = reference factor loading. δ = method factor loading.

on the reference factor indicate the proportion of (error-free) individual differences that the raters share with the reference method. *Method-specificity* on the other hand can be quantified by examining the SSLs on the method factors. The SSLs on the method factors indicate the proportion of (error-free) individual differences that is *not* shared with the objective method.

The correlations between method factors in the CTC(M-1) model have an interesting meaning. They represent partial correlations between raters, corrected for variance that two types of raters share with the objective test scores. Significant positive correlations between the method factors in the CTC(M-1) model indicate that two types of raters share a common perspective that is not shared with the objective tests. For example, parent and teachers might share the same positivity bias (Halo effect), causing them to systematically overestimate students' performance relative to the students' actual test performance.

The question of whether zero-order correlations between raters (as estimated in the Step 1 CFA model) reflect true convergent validity or just a common bias can therefore be examined by comparing the zero-order correlations obtained in the Step 1 CFA model to the partial correlations between method factors in the corresponding CTC(M-1) analysis. If the method factor correlations in the CTC(M-1) model are substantially lower than the correlations between the rater-specific factors in the Step 1 CFA model, then the zero-order correlations between raters are due to true convergent validity [the inter-rater correlations are then accounted for by the common ability factor in the CTC(M-1) model]. In contrast, if the correlations between the rater method factors in the CTC(M-1) model remain similarly high compared to the zero-order correlations in the Step 1 CFA model, then the agreement between raters mainly reflects a shared rater bias, rather than true convergent validity. The present study is the first to apply the CTC(M-

1) approach to a multimethod design for identifying gifted children that uses an objective ("gold standard") comparison method.

In the present study, separate CFA models for each Aurora construct were examined. This was done because the available sample size of N = 145 limited the size of the structural equation models that could be estimated. Two different series of models were analyzed. In the first set of analyses, the convergent validity and method-specificity of the ratings with regard to domain-specific abilities was looked at. That is, separate models for each domain-specific ability (e.g., verbal analytical ability; numeric analytical ability; figural analytical ability) were evaluated.

In the second set of analyses, general (i.e., non-domain specific) abilities were looked at. In order to do so, the domain-specificity from the four Aurora abilities by using indicators pertaining to different domains in a single model were purged. To illustrate, Figure 2 represents an example of a domain-specific analysis, whereas Figure 3 is an example of a non-domain specific analysis. In Figure 2, analytical abilities are assessed based on numerical material only. This can be seen from the fact that all observed variables in this model deal with numeric content. In Figure 3, analytical abilities are measured not only based on numeric, but also verbal and figural material. Therefore, the latent ability and method factors in Figure 3 represent more broadly defined ability factors that do not depend on the specific content of the indicators.

There is evidence in the literature that rater-specific biases might vary strongly by domain such that, for example, numeric abilities appear to be easier to judge than figural or verbal abilities (Freund & Kasten, 2012). By conducting both domain-specific and non-domain specific analyses, more detailed information with regard to the convergent validity and method specificity of self, teacher, and parent ratings were obtained.

Figure 3

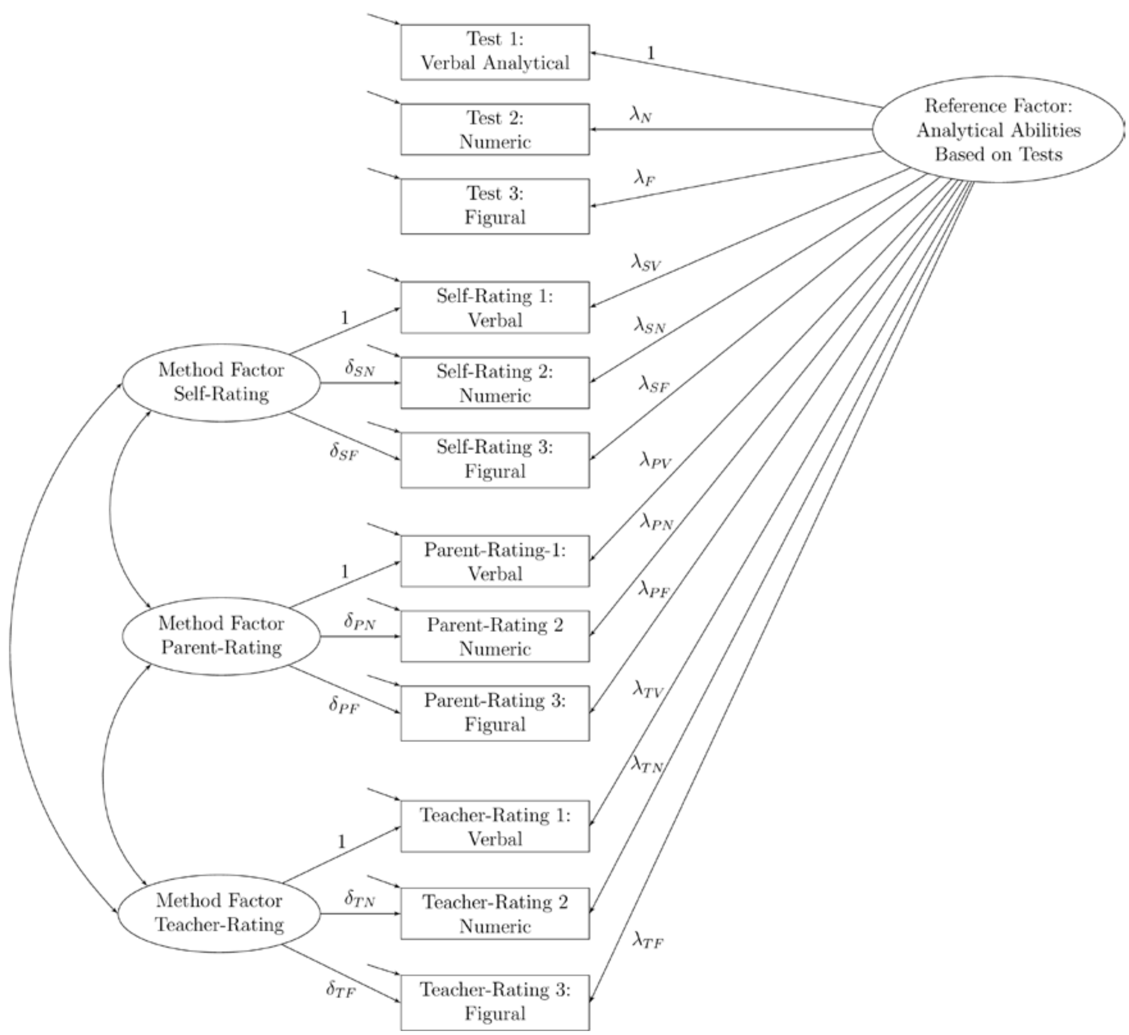


Figure 3. Example of a Domain-Unspecific CTC(M-1) Analysis (here: for analytical ability). The reference factor is defined by objective tests of analytical ability. The self-, parent, and teacher ratings are contrasted against the reference factor. Correlations between method factors indicate whether raters share a common bias.

λ = reference factor loading. δ = method factor loading.

For the domain-specific analyses, two domain-specifically worded items of the rater questionnaires directly as indicators for each rater type, respectively were used. For the domain-unspecific analyses, the scores on the two parallel items for each of the three domains (verbal, numeric and figural) were averaged and used the three resulting item parcels as indicators in the models, respectively.

For the memory subscale of Aurora, no objective tests were available and hence CTC(M-1) models with objective test scores as reference method could not be estimated. Therefore, the focus was on the Step 1 CFA models for this construct. Even though the simple CFA model did not allow for the study of convergent validity with regard to objective memory tests, it was still informative, because it allowed for the examination of convergence between raters (in terms of the latent correlations between rater-specific factors). For two domain-specific constructs (figural analytical ability and numeric creativity), only one test was available as an objective indicator of the reference factor. The CTC(M-1) model is underidentified in this case unless the loading and error variance of the test can be fixed a priori. Therefore, the procedure recommended by Hayduk (1996) was followed and identified the models in these two cases by setting the unstandardized loadings of the tests to 1 and their error variances to values that were calculated based on the known test variances and reliabilities using the formula Variance(error) = Variance(test) + (1 – Reliability). In all other cases, the loadings and error variances were freely estimated, except for arbitrary scaling constraints. All analyses were carried out in Mplus 6 (Muthén & Muthén, 1998-2010) using robust weighted least squares (WLSMV) estimation for analyses that used item-level (i.e., categorical) variables and maximum likelihood (ML) estimation for analyses that included only scale-level (i.e., continuous variables). Missing data were included using full information estimation (Enders, 2010).

Results

CFA Analyses

Domain-Specific Analyses. Table 2 summarizes the results for the Step 1 CFA models for the domain-specific analyses. The goodness-of-fit statistics indicated that the model generally fit the data well, with eight out of 12 analyses showing a non-significant chi-square value. The standardized factor loadings were generally substantial for all raters. However, there was a tendency for self-report measures to consistently show lower standardized loadings than parent and teacher reports, indicating lower reliabilities of self-ratings compared to parent and teacher reports across all abilities. Teacher reports in general showed the highest standardized loadings, with most loadings being greater than .9.

With regard to the factor correlations, it can be seen that there was low agreement between self and other reports, except for numeric memory, numeric analytical ability, numeric creativity, and numeric practical ability, for which substantial correlations between self and parent reports were found. Self and teacher correlations were generally very small and non-significant, except for numeric memory, for which there was a small but significant correlation ($r = .26$, $p < .05$). On the other hand, there was substantial agreement between parent and teacher ratings for the majority of constructs. The highest level of convergence for these raters was found for verbal memory, numeric memory, numeric analytic ability, and verbal practical ability, all of which showed significant latent correlations > .5. Consistently non-significant or weak correlations between parent and teacher ratings for figural abilities were observed.

Table 3 contains the results of the CTC(M-1) analyses for the domain-specific ability constructs. The fit statistics indicated that the CTC(M-1) models fit the data well in general, with nine out of 12 analyses showing a non-significant chi-square value. In the models estimated for

Table 2. Summary of Simple CFA(Step 1) Analyses for Domain-Specific Abilities

Domain	Goodness of fit			Standardized factor loadings			Factor correlations		
	Chi-square (*df* = 9)	RMSEA	CFI	Self report	Parent report	Teacher report	Self-parent	Self-teacher	Parent-teacher
Memory									
Verbal	7.64, *p* = .57	0	1	0.71	0.92	0.96	0.1	0.04	.62***
Numeric	21.04, *p* = .01	0.1	1	0.82	0.88	0.97	.66***	.26*	.60***
Figural	7.18, *p* = .62	0	1	0.7	0.95	0.95	0.13	-0.21	.29*
Analytical ability									
Verbal	9.56, *p* = .39	0.02	1	0.7	0.92	0.96	-0.06	-0.11	.47***
Numeric	10.00, *p* = .35	0.03	1	0.72	0.93	0.99	.60***	0.15	.56***
Figural	15.08, *p* = .09	0.07	1	0.75	0.84	0.97	-0.06	-0.1	0.05
Creativity									
Verbal	20.70, *p* = .01	0.1	1	0.7	0.88	0.96	-0.02	-0.13	.44***
Numeric	5.32, *p* = .81	0	1	0.74	0.91	0.99	.62***	0.14	.41***
Figural	32.32, *p* < .001	0.14	1	0.79	0.79	0.92	0.01	-0.15	-0.18
Practical ability									
Verbal	14.58, *p* = .10	0.07	1	0.57	0.79	0.86	-0.22	-0.27	.62***
Numeric	15.57, *p* = .08	0.07	1	0.69	0.8	0.95	.48***	0.05	.42***
Figural	26.42, *p* = .002	0.12	1	0.74	0.73	0.9	0.29	-0.05	0.16

Note. * p < .05; ** p < .01; *** p < .001. All factor loadings were significantly different from zero (p < .05) verbal analytical ability assessed via Aurora-*a* as well as numeric practical ability, some of the rater variables loaded more strongly onto the reference factors than did the objective test scores. Consequently, the reference factor in these two cases was not defined by the objective tests as it should be, but rather by rater variables. This made the interpretation of the results ambiguous. The problem could be explained by the fact that the Aurora tests used as indicators for these constructs were rather heterogeneous and therefore did not result in a strong common factor. Therefore the CTC(M-1) results for these two cases were not interpreted. (However, for verbal analytical abilities, the Aurora-*g* tests yielded reliable results that are reported in Table 3 and discussed below.)

Analyses of the standardized reference factor loadings (SRLs) consistently revealed zero or very low convergent validity of self-reports relative to the objective tests. For some constructs, the self-report items even showed negative loadings on the reference factor, indicating that children who showed a lower performance on the tests tended to estimate their performance as higher, while children with a stronger performance tended to estimate their performance to be lower. However, this effect was relatively weak, and most self-report SRLs were simply not significantly different from zero. The only construct for which self-reports SRLs were positive and significantly different from zero (indicating true convergent validity) was numeric analytical ability. Even for numeric analytical ability, however, the self-report loadings were rather small (SRLs = .22 and .23), indicating weak convergent validity (consistency = .05).

Teacher ratings overall showed the highest level of convergent validity relative to the Aurora tests. The highest level was found for numeric analytical ability assessed via Aurora-a (SRLs = .75, consistency = .56). Teachers also showed a noticeable level of convergence with

Table 3. Summary of CTC(M-1) (Step 2) Analyses for Domain-Specific Abilities

Domain	Goodness of fit			Standardized reference factor loadings			Standardized method factor loadings			Method factor correlations		
	Chi-square	RMSEA	CFI	Self report	Parent report	Teacher report	Self report	Parent report	Teacher report	Self-parent	Self-teacher	Parent-teacher
Analytical ability												
Verbal [a]	15.19, $df = 17$, $p = .58$[c]	0.00[c]	1.00[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]
Verbal [b]	19.89, $df = 24$, $p = .70$	0	1	.04 (n.s.)	0.45	0.4	0.7	0.8	0.87	-0.1	-0.14	.34*
Numeric [a]	17.04, $df = 17$, $p = .45$	0	1	0.22	0.57	0.75	0.69	0.73	0.64	.55*	-0.14	0.17
Numeric [b]	34.97, $df = 24$, $p = .07$	0.06	1	0.23	0.5	0.44	0.68	0.78	0.88	.54**	0.01	.42***
Figural [a]	22.78, $df = 12$, $p = .03$	0.08	1	-.20 (n.s.)	-.11 (n.s.)	-.00 (n.s.)	.72	0.83	0.97	-0.1	-0.11	0.05
Figural [b]	32.06, $df = 24$, $p = .13$	0.05	1	-0.2	.12 (n.s.)	0.19	0.72	0.83	0.95	-0.02	-0.06	0.02
Creative Ability												
Verbal	24.31, $df = 17$, $p = .11$	0.06	1	-0.01	0.36	0.49	0.7	0.8	0.82	-0.01	-0.14	0.29
Numeric	7.96, $df = 12$, $p = .79$	0	1	-.05 (n.s.)	.17 (n.s.)	0 (n.s.)	0.73	0.89	0.99	.65***	0.14	.42***
Figural	40.48, $df = 17$, $p = .001$	0.1	0.97	.30 (n.s.)	-.14 (n.s.)	.10 (n.s.)	0.73	0.77	0.91	0.09	-0.21	-0.16
Practical ability												
Verbal	22.06, $df = 17$, $p = .18$	0.05	0.98	.05 (n.s.)	.18 (n.s.)	0.44	0.56	0.76	0.74	-0.25	-.37*	.60***
Numeric	24.33, $df = 17$, $p = .11$[c]	0.06[c]	1.00[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]
Figural	40.45, $df = 17$, $p = .001$	0.1	0.94	.16 (n.s.)	-.10 (n.s.)	.37	0.72	0.72	0.82	0.33	-0.16	0.24

Note. [a] Model used Aurora-*a* tests as indicators of the reference factor; [b] Model used Aurora-*g* tests as indicators of the reference factor. [c] Both tests showed lower loadings on the reference factor than some of the rater variables in these analyses. The results for these analyses are therefore not interpretable. * $p < .05$; ** $p < .01$; *** $p < .001$; n.s. = not significantly different from zero ($p > .05$). All factor loadings were significantly different from zero ($p < .05$) unless otherwise indicated.

the tests of verbal creativity (SRLs = .49, consistency = .24), verbal practical ability (SRLs = .44, consistency = .19), and verbal analytical ability assessed via Aurora-*g* (SRLs = .40, consistency = .16). The lowest level of convergent validity of teacher ratings was found for figural analytical ability, numeric creativity, and figural creativity, for which the SRLs were not significantly different from zero.

Similar to self- and teacher ratings, parent ratings showed the highest level of convergent validity for numeric analytical ability (SRLs = .57, consistency = .32). Furthermore, parent ratings showed slightly higher consistencies than teacher ratings for numeric analytic abilities assessed via Aurora-*g* (SRLs = .50, consistency = .25). Parent ratings also showed relatively high consistencies for verbal creativity (SRLs = .36, consistency = .13). For all other constructs, parents showed consistencies that were not significantly different from zero.

The question of whether agreement among raters reflects true convergent validity or just a shared rater bias can be answered by comparing the zero-order correlations between raters reported in Table 2 to the corresponding method factor correlations in the CTC(M-1) analyses presented in Table 3. It can be seen that relative to the simple CFA model, the correlations between parent and teacher ratings were substantially reduced in the CTC(M-1) model for a number of constructs. This was especially the case for verbal analytical ability (r = .47 vs. .34), numeric analytical ability assessed via Aurora-*a* (r = .56 vs. .17), numeric analytical ability assessed via Aurora-*g* (r = .56 vs. .42), and verbal creative ability (r = .44 vs. .29). This shows that for these constructs, the correlations between ratings partly reflected true convergent validity and not just a shared rater bias.

In contrast, for numeric creativity (r = .41 vs. .42) and verbal practical abilities (r = .62 vs. .60), the correlations between parent and teacher ratings remained high and virtually unchanged in the CTC(M-1) model, indicating that for these constructs, the zero-order rater correlations reflected just a shared rater bias. Similarly, the substantial correlations between self- and parent ratings for numeric analytic ability (r = .60 vs. .55 for Aurora-*a* and .54 for Aurora-*g*) and numeric creativity (r = .62 vs. .65) did not change much in the CTC(M-1) model, showing that parents' and children's shared view reflected a shared bias rather than shared convergent validity for these constructs.

In summary, the results for the domain-specific analyses provided a clear picture. There was a substantial amount of agreement between parent and teacher ratings for the majority of constructs, whereas self-report measures showed very low convergence with teacher ratings, and only some convergence with parent ratings. Self-ratings also showed the lowest level of convergence with the objective tests. Consistently across raters, the highest level of convergent validity was for numeric analytical abilities. This construct was also the only one for which self-report measures showed non-zero convergent validity. Overall, teacher ratings showed non-zero convergent validities for a larger number of constructs than self- and parent ratings. Correlations among raters reflected mainly shared rater biases for parents and self-ratings, whereas the same types of correlations partly reflected true convergent validity for parent and teacher ratings.

Domain Non-Specific Analyses. Table 4 summarizes the simple CFA results for the domain-unspecific analyses for memory, analytical, creative and practical abilities. It can be seen that three out of four models showed an excellent fit to the data. The self-report loadings fluctuated very strongly within each construct and were in part very small and non-significant. For memory, all self-report loadings were non-significant, indicating that no common memory ability factor could be established for self-reports. In summary, the results for self-reports indicated that general abilities may not be reliably measured based on self-reports in this age group.

In contrast, parent and especially teacher ratings showed fairly substantial factor loadings, indicating a sufficient level of reliability of these ratings to measure general abilities. In terms of inter-rater agreement, a similar picture was found for the domain-specific ability analyses. Self-reports were not significantly correlated with other reports. The only exception was practical ability, for which a significant *negative* correlation between self- and teacher reports (r = -.29, p < .05) was found. In contrast, parent and teacher reports showed significant positive correlations for three of the four domain-unspecific abilities. The highest correlation between parents and teachers was found for memory (r = .53, p < .001). The only ability for which the parent and teacher factors were not significantly correlated was creativity (r = .23, p > .05).

The CTC(M-1) models for the three general abilities showed a more modest goodness-of-fit compared to the domain-specific analyses (see Table 5). The weaker fit can be explained by the fact that (1) more variables were included in these models, including the degrees of freedom of the models and (2) more heterogeneity with regard to the indicators was introduced in the domain-unspecific as compared to the domain-specific analyses. The CTC(M-1) model for creativity fit the data well. However, an analysis of the parameter estimates of this model revealed that the rater variables loaded more highly on the reference factor than did the objective test scores, rendering the interpretation of the reference factor ambiguous. This could be explained by the fact that the two tests measuring creativity were only weakly and partly insignificantly correlated (-.02 < r < .29). Therefore, the assumption of a common creativity test factor was questionable for these data, and thus did not consider the CTC(M-1) model for creativity further.

Table 4. Summary of Simple CFA (Step 1) Analyses for Domain-Unspecific Abilities

	Goodness of fit			Standardized factor loadings			Factor correlations		
Ability	Chi-square (*df* = 15)	RMSEA	CFI	Self report	Parent report	Teacher report	Self-parent	Self-teacher	Parent-teacher
Memory	14.35, *p* = .50	0	1	.07-1.19 (all n.s.)	.77-.79	.72-.89	0.23	0.01	.53***
Analytical	11.80, *p* = .69	0	1	.34-.62	.63-.94	.77-.86	-0.1	-0.17	.39**
Creative	31.21, *p* = .008	0.09	0.9	.36-.78	.53-.80	.78-.80	-0.11	-0.17	0.23
Practical	20.97, *p* = .14	0.05	1	.18 n.s. -.66	.59-.67	.79-.84	-0.27	-.29*	.37*

Note. * $p < .05$; ** $p < .01$; *** $p < .001$; n.s. = not significantly different from zero ($p > .05$). All factor loadings were significantly different from zero ($p < .05$) unless otherwise indicated.

An analysis of the SRLs for the remaining constructs (see Table 5) revealed that self-report measures showed non-significant convergent validities for both analytical and practical abilities. Parent ratings showed some convergent validity for analytical and practical abilities. However, the results were mixed, with some of the parent report loadings being significant, whereas others were not significantly different from zero. The highest convergent validities were found for teacher reports of practical ability (SRL range .35-.55, consistencies .12-.30). Teacher ratings also showed consistent significant convergent validities for analytical ability measured by Aurora-*g* (SRL range .24-.41, consistencies .06-.17). Zero-order correlations between parent and teacher reports again partly reflected shared convergent validity, especially for analytical ability(r=.39 vs. .29 and .24).

In summary, the domain-unspecific analyses yielded lower levels of inter-rater agreement and convergent validity than the domain-specific analyses. Teacher reports again showed the highest level of convergent validity among raters, whereas no evidence of convergent validity for self-reports was found.

Discussion

In this study, the extent to which teacher, parent, and student ratings of analytical, creative, and practical abilities converged with objective assessments of these abilities and the extent to which this convergence reflected true convergence (convergent validity) rather than a shared rater biases or other error was examined. A number of consistent findings emerged from the current study. Self-ratings were found to be the least reliable and convergent, which is in contrast to teacher ratings which were found to be the most reliable and convergent.

Table 5. Summary of CTC(M-1) (Step 2) Analyses for Domain-Unspecific Abilities

Ability	Goodness of fit			Standardized reference factor loadings			Standardized method factor loadings			Method factor correlations		
	Chi-square	RMSEA	CFI	Self-report	Parent report	Teacher report	Self report	Parent report	Teacher report	Self-parent	Self-teacher	Parent-teacher
Analytical [a]	95.99, $df = 44$, $p < .001$	0.1	1	-.01-.10 (all n.s.)	.06(n.s.)-.34	.15(n.s.)-.35	.34-.68	.61-.85	.72-.81	-0.1	-0.17	.29*
Analytical [b]	151.01, $df = 105$, $p = .002$	0.06	0.94	-.11-.10 (all n.s.)	.27(n.s.)-.51	.25-.41	.34-.64	.54-.77	.68-.78	-0.16	-0.2	0.24
Creative	65.56, $df = 56$, $p = .18$[c]	0.03[c]	0.97[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]	[c]
Practical	102.52, $df = 69$, $p = .01$	0.06	0.93	.03-.09 (n.s.)	-.08-.24(n.s.)	.35-.55	.18-.65	.56-.70	.66-.72	-0.24	-.39*	0.3

Note. [a] Model used Aurora-*a* tests as indicators of the reference factor. [b] Model used Aurora-*g* tests as indicators of the reference factor. [c] The creativity tests showed lower loadings on the reference factor than some of the rater variables in these analyses. The results for these analyses were therefore not interpretable.* $p < .05$; ** $p < .01$; *** $p < .001$. n.s. = not significantly different from zero ($p > .05$). All factor loadings were significantly different from zero ($p < .05$) unless otherwise indicated

Additionally, the highest levels of convergence for all of the raters were in the numerical domain.

The lack of convergence of the self-reported ability with the objective measures and the other raters does not come as a great shock. Firstly, meta-cognition is thought to play a central role in self-rating (see Ehrlinger & Dunning, 2003; Kruger & Dunning, 1999; Mandelman, Tan, Kornilov, et al., 2010) and it is not surprising that students in grades 4-6 do not have sufficiently well-developed meta-cognition to be able to accurately judge their ability, as meta-cognition is one of the later occurring abilities in cognitive development (Kuhn, 2000). Secondly, meta-analytical studies (Freund & Kasten, 2012; Mabe & West, 1982) have found self-estimates of ability to be unreliable and largely invalid when compared to standardized assessments and found only a moderate relationship between the results of self and standardized assessments. While the inclusion of self-reports as a single source of data among multiple informants may be viewed as offering a unique and valuable perspective within the identification processes, it is important that it be included along with an awareness of its considerable weaknesses. The possible weaknesses of individual informants in the process of gifted identification are exactly why the use of multiple informants is so very important.

Another key finding was that among the three rater types, teacher ratings showed the highest level of convergent validity relative to the objective measures. Others (Bracken & Brown, 2008; Pfeiffer & Jarosewich, 2007; Pfeiffer & Petscher, 2008) have also reported high convergence of teachers' ratings with objective measures. These findings support the continued ubiquitous use of teacher rating scales in gifted identification, which are so widely used that they are second only to IQ tests in the identification of the gifted (Pfeiffer & Blei, 2008). Bracken and Brown (2008) explain that teachers' ability to accurately rate student ability is due to the large amount of time that teachers spend with the students and their interaction with the students across subject matter and context. An interesting implication of this finding is that while the Aurora battery assesses a much broader set of abilities, some of which may not necessarily be thought of as school related abilities or well represented on a traditional cognitive ability tests, teachers were able to reliably rate (demonstrated through the convergence with the objective measures) students on these abilities based on their knowledge of the students in the classroom, which highlights the relevance and place of these abilities in classrooms. Also, these findings support the continued use of teacher rating scales in gifted identification.

Finally, it was found that convergence between raters and objective tests tended to be highest for numeric analytical abilities. This is consistent with other studies that showed that numerical abilities are more easily judged than many other abilities (Freund & Kasten, 2012). This may be due to the somewhat more concrete nature of mathematics and the fact that people are particularly aware of their numerical abilities since they are among the most assessed abilities, as they are regularly evaluated in school starting at an early age. This finding demonstrates the importance of using measures that are specifically designed to, and can best draw upon the desired abilities, since a misalignment of these can lead to invalid ratings. Although the results of the current study are very informative, it does have limitations. The primary limitation of this study is its relatively small and homogenous sample. Future research with larger and more diverse samples will allow for the careful examination of other issues, such as the role that age and gender play in these findings.

The current study makes significant and novel contributions. This research represent the first time an empirical examination has been conducted on the use of multiple informants in gifted identification. As one of the most agreed upon practices in gifted education and identification, the use of multiple informants will hopefully continue to gain traction not only in theory, but in practice as well. With the expanded use of multiple informants in the identification process, it is vital that more empirical examinations of this practice be designed and conducted. The fact that it seems logical to gather information from as many sources as possible in the identification process does not mean that all of the pieces of information will neatly line up with each other, or should they be given the same weight in the decision-making process. Each piece of information used in the identification process must be carefully weighed, taking into account the strengths and weaknesses of each, and must be viewed within the context of the student profile emerges.

This study's methodological contribution is quite significant, as it marks the first time the CTC(M-1) analyses has been used on a cognitive assessment. While this method has been used in other areas of Psychology (e.g., Grigorenko, Geiser, Slobodskaya, & Francis, 2010), its application to an assessment of cognitive abilities is novel and can help pave the path forward in gifted identification. The complex nature of cognitive abilities and in turn their assessment showcase the eloquence of this approach in being able to disentangle true agreement and variance among raters from shared rater bias. The field of giftedness has suffered greatly and has been hindered by the a lack of empirical evidence supporting its practices (Mandelman & Grigorenko, 2013; Ziegler, Stoeger, & Vialle, 2012). The importance of being able to employ this kind of sophisticated modeling approach, to be able to empirically explore one of the few seemingly universally agreed practices, cannot be overstated. The careful examination of this practice and other common ones can guide and inform the future of gifted identification and education on the basis of empirical evidence. This study contributes to the validation of the Aurora Battery by offering evidence of its convergent validity.

Chapter 3

Introduction

Standardized assessments of ability and achievement are used at many levels of education. One of the reasons for the widespread use of these assessments is due to their purported ability to predict academic achievement. As a result, these tests are used as part of the admissions process at the undergraduate and graduate level, as well as at some selective high schools. The literature is replete with studies regarding the predictive validity of such assessments, including the SAT (formally known as the Scholastic Assessment Test; Bridgeman, McCamley-Jenkins, & Ervin, 2000; Kobrin, Patterson, Shaw, Mattern, & Barbuti, 2008; Shaw, Kobrin, Patterson, & Mattern, 2012), ACT (formally known as the American College Testing; Noble & Sawyer, 2002), Graduate Record Examinations (GRE; Kuncel, Hezlett, & Ones, 2001), Graduate Management Admission Test (GMAT; Talento-Miller & Rudner, 2005) and Law School Admission Test (LSAT; Stilwell, Dalessandro, & Reese, 2011), with the literature being far scarcer on the predictive validity of assessments used in the high school admissions process (Grigorenko et al., 2009). While the use of standardized tests in admission processes primarily begins in high school, these assessments may be informative well before then. In particular, standardized assessments can be used to help identify students’ strengths and weakness in various intellectual abilities, and to make projections about students’ future performance, allowing for appropriate educational provisions tailored to each student’s needs to be made based on both current achievement and probable performance. The earlier such provisions can be made, the better we may expect the student’s outcomes to be.

Although the literature on the use of standardized assessments at the elementary and middle school level is seemingly nonexistent to date, theoretical models underlying assessments that have accurately predicted academic performance at other levels of education may be extended to the middle school level. In particular, Sternberg's (1985, 1988, 1996a, 1999b, 2005a) theory of Successful Intelligence has been used as the theoretical basis for standardized assessments at the graduate (Hedlund, Wilt, Nebel, Ashford, & Sternberg, 2006), undergraduate (Sternberg, 2006, 2009, 2010), as well as high-school level (Grigorenko, et al., 2009). Assessments based on this model have proven to be successful in accurately predicting academic performance at all of these levels of education.

In the Rainbow Project (Sternberg, 2006, 2009; Sternberg, The Rainbow Project Collaborators, & The University of Michigan Business School Project Collaborators, 2004) Sternberg and his colleagues developed measures based on his theory of intelligence to supplement the SAT. These measures were administered to almost one thousand students from 15 schools across the United States. This new measure broadened the abilities measured from traditional analytical abilities which are overrepresented on SATs, to also include measures of creative and practical abilities. These study found that the inclusion of these measure were able to almost double the amount of explained variance in GPA over the SAT alone. In the Kaleidoscope Project (Sternberg, 2009, 2010; Sternberg, Bonney, Gabora, & Merrifield, 2012) which included a measure based on the Successful Intelligence model, used as a supplement to the admissions application to Tufts University, it was found that students who were admitted using this Kaleidoscope measure performed academically as well as their peers who were admitted using traditional measures such as the SAT, but were engaged in more extracurricular and leadership activities.

At the graduate level, Sternberg and his collaborators at The University of Michigan Business School (Hedlund, et al., 2006; Sternberg, et al., 2004) developed measures of practical intelligence (an ability included in the Successful Intelligence model) to augment the GMAT used in the graduate school admissions process. The results revealed that scores on the measures of practical intelligence were able to predict success in business school, above and beyond what the GMAT and GPA alone could. Finally, at the high school level, Grigorenko, et al. (2009) were able to predict GPA and its growth using a measure based on Sternberg's model as well as the Secondary School Admission Test (SSAT), a standardized tests used at selective high schools. Additionally, when self-report measures of Successful Intelligence were included in the admissions process, they were able to approximately account for 67% of the variance that was accounted for by the SSAT. In the current study, these previous finding are extended in the context of middle school, by evaluating the predictive power of the Aurora Battery for students in grades four through six.

This studies' objective is to explore Aurora's ability to predict academic performance, one year following its administration. Additionally, the study will examine if Aurora can predict the rate of change in grades. These analyses are considered to be exploratory given the sample size and the characteristics of the sample. This study constitutes a step in the validation process of the Aurora Battery and will provide the basis for future research with larger and more diverse populations.

Research Questions:

Can Aurora predict student academic performance one year following its administration?

Can Aurora predict the rate of change in grades?

Method

Participants

See Chapter 2

Measures

Aurora-*a* (Chart, et al., 2008) described above.

Student grades in Math, Science, English and Social Studies for the year (3 grading periods, December, March & June) following Aurora's administration were provided by the school for participating students. GPA was calculated by averaging the grades received for each subject, at each grading period.

Procedure and Data Analyses

Aurora-*a* was administered over two consecutive days in a large group setting; each day consisted of two 45-minute testing periods separated by a 15-minute break between them. Data analyses were conducted to estimate Aurora's predictive validity against GPA level and growth. Specifically, after a series of preliminary analyses to examine pattern of missingness, distributional features, and correlations matrix properties of the data used for planned analyses, two main sets of analyses in a Structural Equation Modeling (SEM) framework were conducted. In the first set, a Latent growth curve model (LGCM; e.g., Preacher, Wichman, MacCallum, & Briggs, 2008) was applied to grades in each subject (as well as for the GPA) throughout three grading periods during the year following the administration of Aurora. LGCM allows capturing the initial or “baseline” grade level (intercept) and the curve and amount of change (slope) in grade received across the three grading periods. Following recommendations in the literature (e.g., Grimm, Ram, & Hamagami, 2011), linear and non-linear growth curves were estimated to identify the best functional patterns of growth over the course of the year. Accordingly, *linear* LGCMs – testing the hypothesis of a constant rate of grade change throughout the year (i.e., time scores were fixed at 0, 1 and 2 for Term 1, Term 2, and Term 3, respectively) – and *“free change”* LGCMs (i.e., latent basis growth curve), providing a free estimation of the time scores at Term 2 to inform the curvature of the growth pattern across the three grading periods were compared. After specific models parameters of interest were examined for suitability for the following set of analyses (in particular, the significance of the estimated variance of the intercept and the slope), the best fitting model for each subject was selected.

In a second step, each subject LGCMs was extended in structural models in which intercept and slope were predicted simultaneously by a latent construct capturing Aurora general performance (general factor underlying the three Aurora abilities measured, Analytical, Practical and Creative), then in a model using the three Aurora ability scores as predictors of both intercept and slope for each subject LGCMs. This set of analysis was meant to distinguish Aurora’s overall predictive power over grades’ baseline level and rate of change. To do so, the relative amount of variance (R^2) of the dependent variables (grades intercept and slope) accounted for by the explanatory variables (Aurora indicators) was also estimated.

All models’ parameters were estimated using the Full Information Maximum Likelihood (FIML) algorithm as implemented in AMOS 18 (Arbuckle, 2009). FIML makes full use of the data available (data from partially complete cases contributed to the estimation of parameters that involve the missing portion of the data as well; Jeličić, Phelps, & Lerner, 2010). The overall goodness-of-fit of the models to the data was evaluated using common fit indicators and their cut-off-values in the literature (e.g., Kline, 2011). Accordingly, a non-significant Chi-Square Test (χ^2), a proportion for $\chi^2/df \leq 2$, a Comparative Fit Index (CFI) higher than .95, a Root Mean Square Error of Approximation (RMSEA) lower than .08, suggest an adequate overall model fit.

Results

Preliminary Analyses

Missing values ranged from 13% to 18% with an average of 15% across variables under investigation. However, the pattern of missing data, holding at least the assumption of missingness at random (Little's MCAR test: $\chi^2 = 96.36$, $df = 83$, $p = .150$), allowed for the use of FIML for model parameters estimation (e.g., Graham, 2009). Distributional features of the data indicated no major distributional non-normality, while properties of the correlation matrix appeared suitable for planned analyses (Bartlett's test of sphericity = 954.5, df = 105, p < .001; KMO = .86).

Modeling Grades Initial Level and Change

Table 6 presents the fit indices for each of the 10 LGCMs (one for each subject-matter and the GPA, in both a linear and free change version). As shown, most models returned an adequate fit to the data in particular in their linear version that appears to be closer to the average functional pattern of growth in the sample (see p value of the $\Delta\chi^2$ between model), at the exception of the science model which best described a non-linear growth (estimated time score at Term 2 was 2.42, $p < .001$).

Table 6. Model fit of grades' Latent Growth Models and Prediction Models for Each Subject-Matter

Model	χ^2	*df*	χ^2/df	*p*	$\Delta\chi^2$	CFI	RMSEA [90% CI]
Univariate LGCM for each Subject							
English LGCM - Free	.06	2	.03	.97	_	1	.000 [.000 - .000]
English LGCM - Linear	.64	3	.21	.89	.45	1	.000 [.000 - .062]
Math LGCM - Free [a]	1.66	1	1.66	.20	_	.99	.064 [.000 - .231]
Math LGCM - Linear [a]	1.86	2	.93	.39	.65	1	.000 [.000 - .153]
Science LGCM - Free	.12	2	.06	.94	_	1	.000 [.000 - .026]
Science LGCM - Linear	3.6	3	1.21	.31	.06	1	.036 [.000 - .146]
Social Studies LGCM - Free	3.88	2	1.94	.14	_	.98	.076 [.000 - .191]
Social Studies LGCM - Linear	4.23	3	1.41	.24	.56	.99	.050 [.000 - .151]
GPA LGCM - Free	1.84	2	.92	.40	_	1	.000 [.000 - .152]
GPA LGCM - Linear	1.9	3	.63	.59	.80	1	.000 [.000 - .112]
Prediction Models							
AuroraG->English_LGCM[b]	8.04	11	.73	.71	_	1	.000 [.000 - .063]
AuroraG->Math LGCM	17.66	10	1.76	.06	_	.97	.077 [.017 - .130]
AuroraG->Science LGCM	5.78	9	.64	.76	_	1	.000 [.000 - .062]
AuroraG->SS_LGCM	19.1	10	1.91	.04	_	.97	.075 [.000 - .126]
AuroraG->GPALGCM	13.3	10	1.33	.21	_	.99	.045 [.000 - .103]

Note. a = fit of the initial model failed, the tested version relaxed the equality of error variance at Term 1. b = due to non-significant variance on the slope factor, the Aurora general factor was only regressed to the Intercept factor. χ^2 = chi-square; *df* = degrees of freedom; *p* = *p* value of the χ^2 test; $\Delta\chi^2$ = p value of the χ^2 difference test; CFI = Comparative fit index; ΔCFI = difference in the CFI value; RMSEA = Root Mean Square Error Of Approximation; CI = 90% Confidence Interval of RMSEA value.

After selecting the best fitting model for each subject-matter and GPA, the parameters estimates for means and variance of intercept and slope were examined (Table 7). As indicated in Table 7, the means and variances of intercept were all significant, suggesting substantial individual differences in the baseline grade level in each subject-matter. In contrast, the means of the slope were all close to 0 (reflecting average stability in grade received across the three grading periods), except for Science, associated with an average rate of grade change of - .093 ($p < .001$) by grading period. In addition, with the exception of English, the estimated variances of the slopes were all significant, reflecting substantial inter-individual differences in intra-individual change over the three grading period (i.e., differences in the direction and rate of grade change over time).

Predicted Initial Level and Rate of Grade Change with Aurora

The selected LGCMs were individually extended into a path model in which intercept and slope were predicted simultaneously by Aurora general performance (as modelled by a latent construct representing the ability underlying Analytical, Creative and Practical scores). Figure 4a provides an example of this model using the GPA intercept and slope as dependent variables. The same model was tested for each subject-matter. Table 6, presents the fit indices for all models tested in this set of analyses. Goodness-of-fit was overall acceptable for all models, ranging from excellent (e.g., Science), to limited (Social Studies). Table 8 provides for each subject-matter and the GPA, the estimated standardized regression coefficients between Aurora general performance, intercept and slope[1].

[1] Because the English LGCM yielded non-significant variance of the slope factor, only the intercept was used as dependent variable in the prediction model.

Table 7. Growth Parameters Estimates for Each Selected Model

Model	Intercept		Slope	
	Mean	Variance	Mean	Variance
English LGCM - Linear	3.33***	.20***	.03	.12
Math LGCM - Linear	3.66***	.234***	-.025	.027*
Science LGCM - Free	3.65***	.098*	-.093***	.023*
Social Studies LGCM - Linear	3.54***	.204***	.004	.019*
GPA LGCM - Linear	3.54***	.093***	-.022	.006*

Note. Var. = Variance; **p* < .05, ***p* < .01, ****p* < .001.

Table 8. Standardized Beta of the Aurora General Factor, on Intercept and Slope for each.

Model	Intercept	Slope
English	.68	_
Math	.30	-.06
Science	.05	.76
Social Studies	.54	-.20
GPA	.59	.09

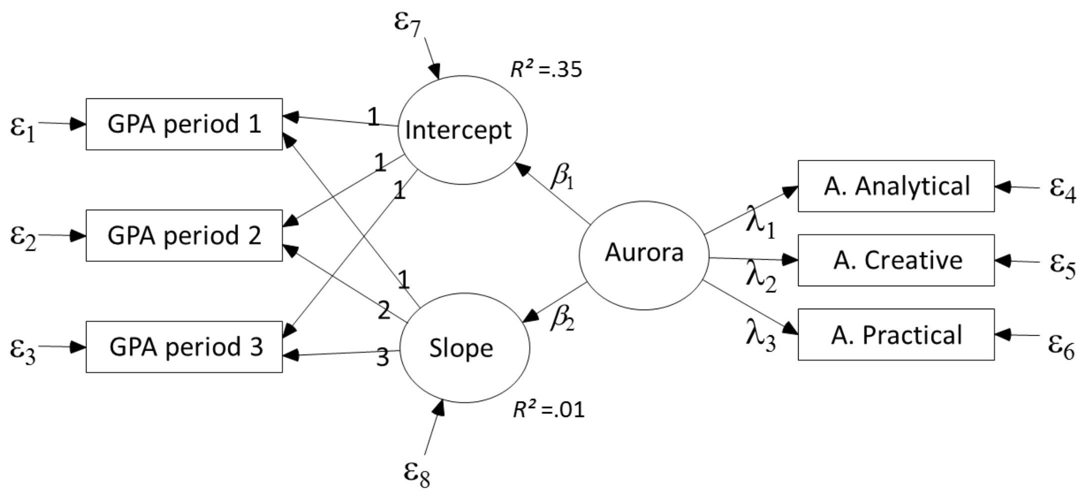


Figure 4a. Latent Growth Curve Model of GPA Predicted by Aurora Latent Construct Representing Overall Performance.

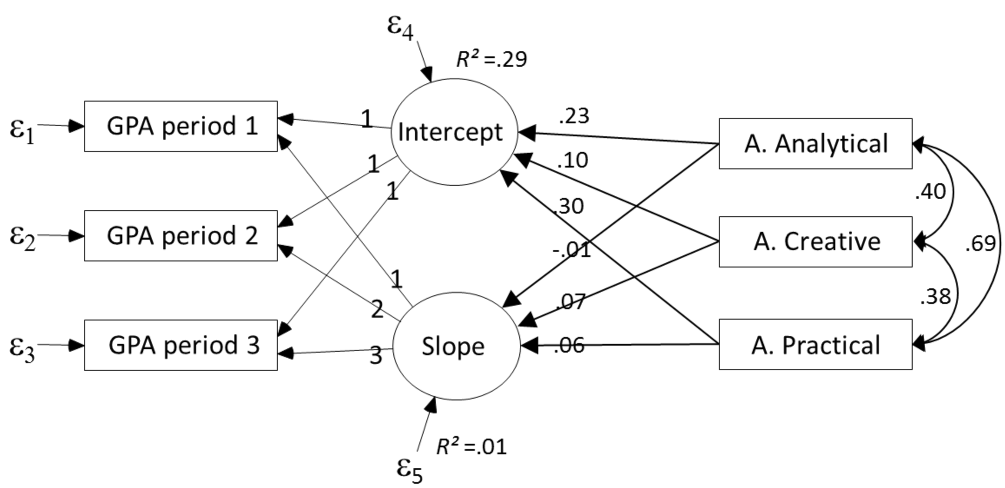


Figure 4b. Latent Growth Curve Model of GPA Predicted by Each Aurora Ability Score.

As shown in Table 8, “baseline” level in GPA was greatly predicted by Aurora, with a regression weight translating in 35% of GPA baseline level variance explained by Aurora. However, the predictive power of Aurora to baseline grade level was the function of the subject-matter under consideration, ranging from null relation (science) to high predictive power (English, with 46.2% of explained variance by Aurora). Regarding the prediction of the rate of growth across the three grading periods under investigation, results indicate only marginal predictive power of Aurora for GPA, Social Studies, and Math. This result was expected given the overall stability of grades across the three grading periods and the limited variability in rate of change in the subject outlined in the previous set of analyses. However, results suggest that Aurora greatly predicts increase in the Science grade (translating in 58% of explained variance). In other words, a higher performance on Aurora abilities measures predicts a faster increase in Science grades during the year following Aurora administration. Conversely, a lower performance on Aurora predicts a faster decrease in Science grades.

In order to examine the unique contribution of each Aurora ability over and above the remaining abilities in the prediction of grade baseline level and rate of change, an alternative model in which all Aurora indicators (Analytic, Creative, Practical) were allowed to intercorrelate, and were used to simultaneously predict GPA intercept and slope was tested. This model yielded an acceptable fit to the data (χ^2 [6] = 12.36, p = .054, χ^2/df = 2.06, CFI = .991, RMSEA [90%-CI] = .080 [.000 - .146]), given the small sample size and degrees of freedom. Figure 4b depicts the tested model and the unstandardized loadings estimated for the prediction of GPA intercept and slope by each

Aurora predictor. As suggested, the three Aurora ability indicators were mostly predictive of GPA intercept, with estimated standardized β = .30, .23 and .10 for Aurora Practical, Analytical and Creative scores, respectively, for a total of 29% of explained variance in GPA intercept.

This prediction model was fitted to each set of specific subject-matter data. However due to limited sample size, this model didn't converge with both the Math and Social Studies data. Regarding English (showing a proper fit to the data: χ^2 [10] = 9.07, p = .53, χ^2/df = .907, CFI = 1, RMSEA [90%-CI] = .000 [.000 - .080]), results indicate that all Aurora abilities contribute rather equally in English grades intercept (with standardized β ranging from .19 to .26, translating in a total of 32% of explained variance in English baseline level). Likewise, the Science model was associated with proper fit (: χ^2 [10] = 9.07, p = .53, χ^2/df = .907, CFI = 1, RMSEA [90%-CI] = .000 [.000 - .080]). Précising results obtained with the general ability model, results observed at the specific ability level suggested a major contribution of Practical ability (β = .50), Creative ability (β = .22) and Analytical ability (β = .12) in the prediction of Science grades rate of change. In other words, students with higher performance in particular on Aurora Practical abilities were increasing their grades in Science more quickly over time.

Discussion

In this chapter the Aurora Battery, based on Sternberg's (1985, 1988, 1996a, 1999b, 2005a) encompassing model of human cognitive ability, was explored as a predictor of students' academic achievement (i.e., grades in various subjects and their change over a one year period). In line with expectations, based on earlier work using assessments built on Sternberg's model at various level of education (Grigorenko, et al., 2009; Hedlund, et al., 2006; Sternberg, 2006, 2009, 2010; Sternberg, et al., 2012; Sternberg, et al., 2004), Aurora was able to predict 35% of overall academic performance (GPA) one year following its administration. Specifically, Aurora's ability to predict academic performance differed greatly depending on the subject-matter under consideration, from explaining almost no variance in baseline level grades in Science (though predicting greatly the change of Science grades over time), to being highly predictive in English with 46.2% of the variance explained.

Additionally, analyses at the level of individual abilities measured by Aurora indicated a different contribution of each ability (analytical, practical and creative) in the prediction of the GPA intercept, and suggested that Practical abilities, rather than Analytical abilities, were the best predictors. This finding is of particular importance as it highlights that abilities that are not traditionally thought of as being school related, i.e. that practical or creative abilities were able to predict the GPA intercept and deserve attention and the opportunity to develop within the classroom. This finding questions the dominance afforded to analytical abilities when it comes to school achievement.

The predictive power of these abilities was also largely dependent on the specific subject. Aurora's abilities were able to explain 32% of variance in the English baseline level, and 32% of the variance in Science grades rate of change. Once again, Practical abilities were identified as the greatest predictor of academic outcome. These finding suggests that the specific demands and nature of each subject, dictate how much of a role each individual ability plays and affects how well the abilities captured by Aurora can predict academic performance. This finding has implication for the development of targeted educational provisions to match the children's cognitive strengths and weaknesses, in the context of the specific abilities that are most involved in that particular subject. For example, if a student was struggling in English, based on these findings a program designed to improve his skills in English may include a focus on practical abilities. Creative abilities in this study did not have prominence in the prediction of GPA. A greater contribution of creative abilities can be expected in subjects that involve more creative thinking, which unfortunately tend to be less and less represented in traditional classrooms. Over the last twenty years, creativity has decreased as conformity of thought is encouraged at home and at school as opposed to creative, divergent or associative thinking (Kim, 2011). Kim (2011) suggests that some of the decrease in creativity in school may be due to the culture of standardized testing which results in the over-emphasis on analytical abilities. This often means that creative thinking is not rewarded or encouraged and opportunities for students to be creative in school are being eliminated.

Although these results are promising, the current study does have limitations. The primary limitation is the use of a rather small sample size, yielding limited statistical power to detect significant effect and intra-individual differences in the rate of change (variance of the slopes). This limitation also resulted in failure to fit relatively more complex models (i.e., individual ability level) needed to better examine the specific contribution of each ability in each subject. Future research with larger samples will allow for more careful examination of individual trajectories of grade change in relation to Aurora profiles of abilities. This study contributes to the establishment of Aurora's predictive validity.

Chapter 4

Introduction

When developing any new assessment, it is necessary to establish that the new measure does something other than what is already done by existing measures, and that there is a place and need for the new assessment. In the case of Aurora, it is necessary to demonstrate that the Aurora Battery, in fact, identifies as gifted, students that the traditional assessments do not identify. In the current study, Aurora's convergent and divergent validity was examined using the TerraNova—a conventional assessment of academic achievement (CTB/McGraw-Hill, 2010a) to explore the convergence of these tests as well as the overlap with regard to whom they identify as gifted. Data analyses were conducted to estimate Aurora's criterion validity (against the TerraNova) as well as the sensitivity and specificity of Aurora's gifted identification, in comparison to TerraNova's gifted identification. Kornilov, et al. (2012) conducted a similar analysis with two measures used in the British education system and Aurora. A weak to moderate overlap was expected in the students that were identified using the respective measures, given the theoretical differences underlying these tests and the constructs they are designed to measure.

Research Questions:

What is the convergence between Aurora's and TerraNova's scores?

To what extent do the Aurora Battery and the Terra Nova overlap in the students that they identify as gifted (& non-gifted)?

Method

Participants

See Chapter 2

Instruments

Aurora-*a* (Chart, et al., 2008) described above.

The TerraNova (CTB/McGraw-Hill, 2010a), a widely used standardized norm-referenced group achievement test, was regularly employed by the school at which the research took place to evaluate student achievement in the areas of Reading, Language, Mathematics, Science, and Social Studies. The TerraNova produces six scaled scores, one in each of the areas covered. By design, the TerraNova also yields a total composite score that is based on the reading, language and mathematics scores, to reflect a child's overall performance. In this sample, the composite score appeared to be reliable (Cronbach's alpha = .85) and was used in this study as the main external criterion for gifted identification. Extensive research with the TerraNova supports its excellent reliability, and provides strong evidence for its validity (CTB/McGraw-Hill, 2010b) and utility as a gifted identification tool (e.g. Neumeister, Adams, Pierce, Cassady, & Dixon, 2007; Pierce et al., 2006).

Procedure

Aurora-*a* was administered over two consecutive days; each day consisted of two 45-minute testing periods separated by a 15-minute break. The TerraNova was administered in a separate group session with delays ranging from 31 to 336 days (mean delay = 243.8) due to field/organizational constraints. Data analyses were conducted to estimate Aurora's criterion validity (against the TerraNova) as well as the sensitivity and specificity of Aurora's gifted identification, in comparison to TerraNova's gifted identification. These analyses were considered to be exploratory given the sample size and the characteristics of the sample. This study constitutes a step in the validation process of the Aurora Battery and will provide the basis for future research with larger and more diverse populations.

Results

Criterion Validity

The first set of analyses in this study was devoted to further estimate Aurora's criterion validity. Because of the varying delays between the test-administrations of the TerraNova and the Aurora Battery, this possible source of variation was controlled for in a set of partial correlations between both measures. Partial correlations statistically control for the possible influence of a third, outside variable (in this case the delay between the administrations) on the linear relationship between two variables (in this case the two measures) by partialing out its influence. The resulting inter-correlations are displayed in Table 9.

As expected, results indicate a medium-level, on average, convergence between Aurora's and TerraNova's scores, with the average of 35.2% and the range of 9% to 58% of shared variance between scales. Consistent with theoretical expectations, the highest correlations between TerraNova's total score and Aurora's aptitudes scores were observed with the Analytical and Practical scores (46% and 58% of shared variance, respectively) whereas the Creative abilities which are not directly measured by and are only limitedly represented in TerraNova's total score (13.6% of shared variance). In terms of Aurora's domain scores, the results support both the convergent and divergent validity of the Aurora Battery with the TerraNova, given the pattern of inter-correlation:

Aurora’s Numbers domain score is mostly related to TerraNova’s Math score, which presents numerical content (46% of shared variance), which present verbal content (24% and 30% of shared variance, respectively); while Aurora’s Words domain score is mainly associated with TerraNova’s Reading and Language scores (41% of shared variance), and is less related to TerraNova’s Math score (32% of shared variance). Aurora’s Figural domain score is only moderately related to all of the TerraNova’s domain scores (23 to 26% of shared variance), which was expected given that the figural domain is not highly represented in the items of TerraNova. As a whole, these results are highly consistent with theoretical expectations and provide preliminary evidence of Aurora’s criterion validity with a well-established widely used measure.

Sensitivity and Specificity of Aurora Scores

The second set of analyses in this study focused on the nature of gifted identification with Aurora. To identify gifted students for each of Aurora’s abilities and domains, the 90th-percentile threshold was used above which scores are considered to reflect exceptional abilities (i.e., top 10% performance), consistent with prior research (Kornilov, et al., 2012) and the guidelines of educational practice (Tan et al., 2009). The same procedure was used for the TerraNova total score. The resulting classifications (i.e., gifted vs. non-gifted) were used in a series of contingency analyses indicating the extent to which Aurora and TerraNova scores converge in identifying gifted children. Results were interpreted in terms of Aurora’s Sensitivity (percentage of individuals simultaneously classified as gifted by both measures) and Specificity (percentage of individuals simultaneously classified as non-gifted by both measures). Table 10 summarizes the Specificity and Sensitivity estimates for each Aurora ability and domain,

Table 9. Partial Correlations between TerraNova and Aurora Scores

	TerraNova			
Aurora Battery	Reading	Language	Math	Total
Analytical	.55	.62	.61	.68
Creative	.37	.30	.31	.37
Practical	.62	.66	.71	.76
Images	.48	.49	.51	.56
Words	.64	.64	.57	.69
Numbers	.49	.55	.68	.67
Ability Index	.63	.66	.67	.75

Note. N = 125. Partial correlations are controlled for administration delay between both measures. All coefficients are significant at $p \leq .001$

as well as the percentage of agreement (overlap) between both classification systems (i.e., classifications resulting from Aurora and from TerraNova).

As indicated in Table 10, both measures converge highly for the identification of non-gifted students (agreement between both classification systems ranging from 90% to 96%), while both measures yield rather divergent identification of gifted students, as evidenced by the low to moderate classification agreement rate, with the average of 42.9% and the range from 15.4% to 69% agreement. In particular, students identified as gifted through their Creative abilities as measured by Aurora and those identified as gifted by their performance in Aurora's Words domain significantly differed from those identified as gifted by The TerraNova. As a whole, the results suggest that while both Aurora and the TerraNova are able to similarly distinguish non-gifted students, the measures diverge greatly with regard to which students are identified as gifted. Despite the reasonable amount of children identified as gifted according to both tests, the divergence between the measures suggests that the Aurora Battery is able to identify gifted children in specific areas (verbal, numerical, and figural) and with specific abilities (analytical, creative and practical) that are not captured with traditional measures of academic achievement.

Discussion

Consistent with previous work presenting encouraging results with regard to Aurora's psychometric features (e.g., Kornilov, et al., 2012) the data presented here complement and extend these earlier results by suggesting Aurora's criterion validity with the TerraNova, an established measure of academic achievement, which provides evidence of both convergent and discriminant validity. Specifically, while overlap

Table 10. Summary of Sensitivity and Specificity of Aurora Scores

	TerraNova Scale Score Total			
Aurora Score	Specificity (% True Negative)	Sensitivity (% True positive)	$\chi^2(1)$	*p*
Analytical	83.7 (92.6%)	3.7 (38.5%)	12.20	.001
Creative	81.2 (90%)	1.5 (15.4%)	.36	NS
Practical	85.9 (95.1%)	5.9 (61.5%)	40.50	.001
Images	84.5 (94%)	4.7 (46.2%)	20.70	.001
Words	82.1 (90.9%)	2.2 (23.1%)	2.45	NS
Numbers	86.7 (95.9%)	6.7 (69.2%)	53.60	.001
Ability Index	85.2 (94.3%)	4.4 (46.2%)	22.05	.001

Note. Specificity Index is calculated as a % of the total sample identified as non-gifted by both measures (% True negative = % agreement in classifying the participants as non-gifted). Sensitivity Index is calculated as a % of the total sample identified as gifted by both measures (% True positive = % agreement in classifying the participants as gifted).

between some Aurora and TerraNova subscales exists, there are some abilities measured with Aurora—the creative abilities, in particular— that are only minimally represented in TerraNova's constructs. Therefore, this suggests that Aurora taps into another type of abilities, and consequently, may further contribute to a broader identification of giftedness.

In line with this hypothesis, a Sensitivity-Specificity analysis indicated a reasonable overlap between Aurora's and TerraNova's ability to distinguish non-gifted students, but only weak agreement between the measures regarding the identification of gifted children. Together, these results suggest that, although some overlap exists between Aurora and traditional measures of academic achievement, both measures tend to result in rather different identification of gifted children, which has important implications for gifted education as one must make informed choices when deciding on gifted identification tools.

Chapter 5

The field of gifted education and gifted research is at a crossroads, or, as more pointedly stated by Ziegler, et al. (2012), is in a state of crisis. Ziegler, et al. (2012) outlined a number of issues that have placed and continue to place the field in this state of crisis. Firstly, the field of gifted education and research is not taken seriously by those outside of the field, as is evident by the fact that gifted education research articles do not make it into top tier education and psychology journals, nor are citations of gifted research prevalent in those journals. This issue is not limited to fields of general and education psychology research; even neighboring disciplines often ignore the research put forth by those in the field of gifted education. Ziegler, et al. (2012) explain that the field has hurt itself by failing to demonstrate the reliability of gifted identification or the efficacy of gifted education.

Gifted education has been around for almost 100 years and the field is badly plagued by identification issues, including lack of reliability. As stated in previous chapters these identification issues can at least be explained in part by the models upon which the identification tools are built and the way those tools and the scores they produce are used. Gifted children have been identified using IQ tests for one hundred years (Pfeiffer, 2012; Sternberg, et al., 2011), and they have changed little over this time (Mandelman, Tan, Aljughaiman, et al., 2010; Sternberg, et al., 2011). They are still the most common means used by schools and states to identify gifted students (S. W. Brown, et al., 2005; McClain & Pfeiffer, 2012). Besides issues with the IQ tests themselves, the ways in which the scores are used (McClain & Pfeiffer, 2012), and the fact that many rely on IQ tests as the sole measure to identify giftedness (S. W. Brown, et al., 2005; Pfeiffer,

2003) have also contributed to problems with gifted identification. The greatest issue with IQ tests is the narrow model of intelligence upon which they are built and the fact that there are many gifted children who are never identified as a result.

It is within this very context that Aurora is being developed. Aurora's theoretical foundation goes beyond the narrow *g* based conception of intelligence and is designed to measure analytical, practical and creative abilities, which are of vital importance, particularly in today's global knowledge economy. This extension beyond *g* by definition broadens not only the base of abilities deemed important, but allows for the identification of students who have great potential and gifts that can make important contributions to society, that may otherwise never be identified or given the opportunity to develop these abilities. This is evident in the research presented in Chapter 4, which shows that Aurora and TerraNova identify different students.

One of the reasons that IQ tests have such great appeal despite their many flaws, is that they have relatively high levels of reliability and validity. One of the most important validities that IQ tests are purported to have is predictive validity. Yet measures that do not necessarily have the same weaknesses as IQ tests can also have strong predictive powers demonstrated with Aurora in Chapter 3. The widespread practice and misuse of IQ scores as the sole criteria for gifted identification is directly challenged by Aurora's multiple modules that include multiple ability tests (Aurora-*a* and Aurora-*g)* as well multiple rating scales (Aurora-*r, s, i*). Aurora's scales include input from multiple informants—teachers, parents and the student themselves—in the identification process, which is one of the most universally agreed upon practices in the field of gifted education (Borland, 2009; Briggs, et al., 2008; Brody & Mills, 1997; S. W.

Brown, et al., 2005; Coleman, 2003; Ford & Trotman, 2000; Friedman-Nimz, 2009; Maker, 1996; McClain & Pfeiffer, 2012; McCoach, et al., 2001; Nielsen, 2002; Pfeiffer, 2001, 2003, 2012; Pfeiffer & Blei, 2008; VanTassel-Baska, et al., 2007). While the use of multiple informants is commonly touted as the best practice, it has yet to be widely empirically investigated, which is in part what spurred the research presented in Chapter 2. This lack of research to support practice is in line with the larger problem of there being little empirical evidence supporting even the most common practices within gifted education (Ziegler, et al., 2012).

A question that must be contended with, that lies at the heart of the field, is what is truly the purpose of being identified as gifted. The most obvious answer to this question is to being admitted into a gifted education program and that there is some inherent benefit to being identified as gifted. This leads to the next question: Does gifted education actually work? Ziegler, et al. (2012) point out that there is a woeful lack of research supporting the efficacy of gifted education or its supposed benefit. Gifted education has had a relatively long history in the US, and has existed in one form or another for almost one hundred and fifty years (Tannenbaum, 1958). This long history may help to explain why if anyone, a layperson or a professional, is asked the fundamental question as to whether gifted education works, you will almost certainly get an immediate and emphatic “yes”! The question is on what basis is this claim made? There has been little empirical work done to support this claim and the field of giftedness has suffered greatly and been hindered by the lack of empirical evidence supporting its practices (Ziegler, et al., 2012). For a review of the existing evidence supporting gifted education see Mandelman and Grigorenko (2013).

Yet, another concern expressed by Ziegler, et al. (2012) is that there is scarce evidence that there are economic or cultural benefits to those who invest in gifted education (for further disscussion see Mandelman, Tan, Aljughaiman, et al., 2010). There is equally scarce literature that carefully evaluates long-term effects of being identified and/or educated as a gifted student (Delcourt, Cornell, & Goldberg, 2007; Subotnik, Edmiston, & Rayhack, 2007). The scarcity of empirical evidence has been used by some (Borland, 2005) to fundamentally question gifted education as a practice.

The final point that Ziegler, et al. (2012) make is the need for more empirical studies to evaluate the practices of gifted education and models of giftedness. As previously discussed there have been few rigorous studies evaluating the most common practices in gifted education. The general lack of literature on gifted education and the quality of the existing literature are disheartening, and impels one to think critically. If we believe that there is a construct of intelligence and giftedness, and that it can be identified and educational provisions should be made for those who possess it, why is there is so little empirical support for the educational provisions made? Why has more not been done to move the field forward?

One possible explanation for the lack of progress in gifted education research is the enduring disconnect between the definitions, identification methods, the educational accommodations, the desired outcomes of gifted education, and societal needs. The definition and the desired outcome must be clearly connected and inform each other, with the definition guiding the true desired outcome and vice versa. The most important question in this process is, what is the ultimate purpose of gifted education? If this purpose could be made entirely clear, then definitions would have to be updated to clearly reflect the desired outcomes. In turn, the identification tools would also have to be updated to match the definition and the desired outcome. The educational provisions made for the gifted students would be clearly in line with the desired outcome and would thus make it possible to carefully examine the efficacy of gifted education. If this happened, it would not only address the well-known issues of the definition and identification tools that draw so much criticism, it would also be able to make educational accommodations that can be empirically supported and validated, gaining the field greater credibility.

A Possible Step Forward

While not the only, nor the most definitive answer to the problem, Robert J. Sternberg's theory and the measures based on it moves in the right direction in addressing many of the aforementioned issues. Robert J. Sternberg's Triarchic Theory of Successful Intelligence (1985, 1988, 1996a, 1999b, 2005a) offers a definition that appreciates and captures the complexity of human intellectual ability, serves as the basis of a set of identification tools that clearly match the definition, suggests educational provisions that are in line with the rest of the components, and finally offers a desired outcome that is part and parcel of the definition.

Sternberg's definition of intelligence suggests a more comprehensive model of human ability that includes analytical, creative and practical abilities. These theory-based assessments (e.g., the Aurora Battery, Chart, et al., 2008; Kornilov, et al., 2012; Sternberg, 2005c, 2010; Sternberg, Grigorenko, & Jarvin, 2006; Sternberg & The Rainbow Project Collaborators, 2006) have been shown to effectively identify those who were previously unidentified by traditional methods.

The educational provisions for the gifted do not necessitate acceleration or ability grouping, as the objective of gifted education according to this model is for the individual to be exposed to rich and diverse experiences, as exposure to different kinds of information, in different ways, and in different contexts, can allow students to further develop their abilities (analytical, creative and practical) and use their strengths to compensate for their weaknesses (for a discussion of ability & environmental interaction see Barab & Plucker, 2002). The focus of this model is on the individual developing and maximizing his potential to the fullest. The desired outcome is set by the definition of successful intelligence, which defines success as what is dictated by the sociocultural context of the individual.

The research presented within this dissertation will help to further validate the Aurora Battery, which begins to address many of the current pitfalls of gifted identification. Future research with larger and more diverse samples will address the primary limitations of this dissertation. Yet Aurora clearly represents a step in the right direction for the future of gifted education and moves the field forward by highlighting these critical issues through empirical methods. This is a critical time for gifted education. Aurora is meant provide an alternative to the crisis and instead offers a new pathway of gifted identification, so that all of society will ultimately reap the benefits of gifted education.

References

Achenbach, T. M., Krukowski, R. A., Dumenci, L., & Ivanova, M. Y. (2005). Assessment of Adult Psychopathology: Meta-Analyses and Implications of Cross-Informant Correlations. *Psychological Bulletin, 131*, 361-382.

Arbuckle, J. L. (2009). *Amos 18 user's guide*: Amos Development Corporation.

Barab, S. A., & Plucker, J. A. (2002). Smart People or Smart Contexts? Cognition, Ability, and Talent Development in an Age of Situated Approaches to Knowing and Learning. *Educational Psychologist, 37*, 165-182.

Boring, E. G. (1923). Intelligence as the Tests Test It. *New Republic, 35*, 35-37.

Borland, J. H. (2005). Gifted education without gifted children. In R. J. Sternberg, Davidson, J. E. (Ed.), *Conceptions of giftedness, second edition* (pp. 1-19). New York: Cambridge University Press.

Borland, J. H. (2009). Myth 2: The Gifted Constitute 3% to 5% of the Population. Moreover, Giftedness Equals High IQ, Which Is a Stable Measure of Aptitude. *Gifted Child Quarterly, 53*, 236-238.

Bracken, B. A., & Brown, E. F. (2008). Early Identification of High-Ability Students: Clinical Assessment of Behavior. *Journal for the Education of the Gifted, 31,* 403-426.

Bridgeman, B., McCamley-Jenkins, L., & Ervin, N. (2000). *Predictions of freshman grade-point average from the revised and recentered SAT I: Reasoning test (No. 2000–1)*. New York: College Entrance Examination Board.

Briggs, C. J., Reis, S. M., & Sullivan, E. E. (2008). A National View of Promising Programs and Practices for Culturally, Linguistically, and Ethnically Diverse Gifted and Talented Students. *Gifted Child Quarterly, 52*, 131-145.

Brody, L. E., & Mills, C. J. (1997). Gifted Children with Learning Disabilities: A Review of the Issues. *Journal of Learning Disabilities, 30*, 282-296.

Brown, P., & Tannock, S. (2009). Education, meritocracy and the global war for talent. *Journal of Education Policy, 24*, 377-392.

Brown, S. W., Renzulli, J. S., Gubbins, E. J., Siegle, D., Zhang, W., & Chen, C.-H. (2005). Assumptions Underlying the Identification of Gifted and Talented Students. *Gifted Child Quarterly, 49*, 68-79.

Chart, H., Grigorenko, E. L., & Sternberg, R. J. (2008). Identification: The Aurora Battery. In J. A. Plucker & C. M. Callahan (Eds.), *Critical issues and practices in gifted education* (pp. 281-301). Waco, TX: Prufrock.

Coleman, M. R. (2003). *The Identification of Students Who Are Gifted*. Arlington, VA: Eric Clearinghouse on Disabilities and Gifted Education. (ERIC Document Reproduction Service No. ED480431).

CTB/McGraw-Hill. (2010a). TerraNova.

CTB/McGraw-Hill. (2010b). *TerraNova Technical Report*. Monterey, California.

Dai, D. Y., Swanson, J. A., & Cheng, H. (2011). State of Research on Giftedness and Gifted Education: A Survey of Empirical Studies Published During 1998—2010 (April). *Gifted Child Quarterly, 55*, 126-138.

Delcourt, M. A. B., Cornell, D. G., & Goldberg, M. D. (2007). Cognitive and affective learning outcomes of gifted elementary school students. *Gifted Child Quarterly, 51*, 359-381.

Ehrlinger, J., & Dunning, D. (2003). How chronic self-views influence (and potentially mislead) estimates of performance. *Journal of Personality and Social Psychology, 84*, 5-17.

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika*, *65*, 241–261.

Eid, M., Lischetzke, T., & Nussbeck, F. W. (2006). Structural equation models for multitrait-multimethod data. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 283–299). Washington, DC: American Psychological Association.

Eid, M., Lischetzke, T., Nussbeck, F. W., & Trierweiler, L. I. (2003). Separating trait effects from trait-specific method effects in multitrait-multimethod models: A multiple indicator CT-C(*M*-1) model. *8*, 38-60.

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford.

Ford, D. Y., & Trotman, M. F. (2000). The office for civil rights and non discriminatory testing, policies, and procedures: Implications for gifted education. *Roeper Review, 23*, 109-112.

Freund, P. A., & Kasten, N. (2012). How smart do you think you are? A meta-analysis on the validity of self-estimates of cognitive ability. *Psychological Bulletin, 138*, 296-321.

Friedman-Nimz, R. (2009). Myth 6: Cosmetic Use of Multiple Selection Criteria. *Gifted Child Quarterly, 53,* 248-250.

Geiser, C., Eid, M., & Nussbeck, F. W. (2008). On the meaning of the latent variables in the CT-C(M-1) model: A comment on Maydeu-Olivares and Coffman (2006). *Psychological Methods, 13,* 49-57.

Geiser, C., Eid, M., West, S. G., Lischetzke, T., & Nussbeck, F. W. (2012). A Comparison of Method Effects in Two Confirmatory Factor Models for Structurally Different Methods. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*, 409-436.

Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology*, *60*, 549-576.

Grigorenko, E. L., Geiser, C., Slobodskaya, H. R., & Francis, D. J. (2010). Cross-informant symptoms from CBCL, TRF, and YSR: Trait and method variance in a normative sample of Russian youths. *Psychological Assessment, 22*, 893-911.

Grigorenko, E. L., Jarvin, L., Diffley Iii, R., Goodyear, J., Shanahan, E. J., & Sternberg, R. J. (2009). Are SSATS and GPA enough? A theory-based approach to predicting academic success in secondary school. *Journal of Educational Psychology, 101*, 964-981.

Grimm, K. J., Ram, N., & Hamagami, F. (2011). Nonlinear Growth Curves in Developmental Research. *Child Development*, *82*, 1357-1371.

Hayduk, L. A. (1996). *LISREL: Issues, debates, and strategies*. Baltimore: Johns Hopkins University Press.

Hedlund, J., Wilt, J. M., Nebel, K. L., Ashford, S. J., & Sternberg, R. J. (2006). Assessing practical intelligence in business school admissions: A supplement to the graduate management admissions test. *Learning and Individual Differences, 16,* 101-127.

Jeličić, H., Phelps, E., & Lerner, R. (2010). Why Missing Data Matter in the Longitudinal Study of Adolescent Development: Using the 4-H Study to Understand the Uses of Different Missing Data Methods. *Journal of Youth and Adolescence, 39,* 816-835.

Kim, K. H. (2011). The Creativity Crisis: The Decrease in Creative Thinking Scores on the Torrance Tests of Creative Thinking. *Creativity Research Journal*, *23,* 285-295.

Kline, R. B. (2011). *Principles and Practice of Structural Equation Modeling* (3rd ed.). New York, NY: The Guilford Press.

Kobrin, J. L., Patterson, B. F., Shaw, E. J., Mattern, K. D., & Barbuti, S. M. (2008). *Validity of the SAT® for Predicting First-Year College Grade Point Average*. New York: The College Board.

Kornilov, S. A., Tan, M., Elliott, J. G., Sternberg, R. J., & Grigorenko, E. L. (2012). Gifted Identification With Aurora: Widening the Spotlight. *Journal of Psychoeducational Assessment, 30,* 117-133.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77,* 1121-1134.

Kuhn, D. (2000). Metacognitive development. *Current Directions in Psychological Science, 9,* 178-181.

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the Graduate Record Examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127,* 162-181.

Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67,* 280-296.

Maker, C. J. (1996). Identification of Gifted Minority Students: A National Problem, Needed Changes and a Promising Solution. *Gifted Child Quarterly, 40,* 41-50.

Mandelman, S. D., & Grigorenko, E. L. (2013). Questioning the Unquestionable: Reviewing the evidence for the efficacy of gifted education. *Talent Development and Excellence, 5,* 125–137.

Mandelman, S. D., Tan, M., Aljughaiman, A. M., & Grigorenko, E. L. (2010). Intellectual giftedness: Economic, political, cultural, and psychological considerations. *Learning and Individual Differences, 20*, 287-297.

Mandelman, S. D., Tan, M., Kornilov, S. A., Sternberg, R. J., & Grigorenko, E. L. (2010). The Metacognitive Component of Academic Self-Concept: The Development of a Triarchic Self-Scale. *Journal of Cognitive Education & Psychology, 9*, 73-86.

McCann, M. (2005). Our Greatest Natural Resource: Gifted Education in Australia. *Gifted Education International, 19,* 90-106.

McClain, M.-C., & Pfeiffer, S. (2012). Identification of Gifted Students in the United States Today: A Look at State Definitions, Policies, and Practices. *Journal of Applied School Psychology, 28,* 59-88.

McCoach, D. B., Kehle, T. J., Bray, M. A., & Siegle, D. (2001). Best practices in the identification of gifted students with learning disabilities. *Psychology in the Schools, 38*, 403-411.

Milgram, R. M. (2000). Identifying and enhancing talent in Israel: A high national priority. *Roeper Review, 22*, 108-110.

Muthén, L. K., & Muthén, B. O. (1998-2010). *Mplus User's Guide. Sixth Edition*. Los Angeles, CA: Muthén & Muthén.

National Association for Gifted Children. (2012). Retrieved August 21, 2012, from http://www.nagc.org

Neumeister, K. L. S., Adams, C. M., Pierce, R. L., Cassady, J. C., & Dixon, F. A. (2007). Fourth-Grade Teachers' Perceptions of Giftedness: Implications for Identifying and Serving Diverse Gifted Students. *Journal for the Education of the Gifted, 30,* 479-499.

Nielsen, M. E. (2002). Gifted Students With Learning Disabilities: Recommendations for Identification and Programming. *Exceptionality, 10*, 93-111.

Noble, J., & Sawyer, R. (2002). *Predicting different levels of academic success in college using high school GPA and ACT composite score (No. 2002–4)*. Iowa City, IA: ACT.

Pfeiffer, S. I. (2001). Professional psychology and the gifted: Emerging practice opportunities. *Professional Psychology: Research and Practice, 32*, 175-180.

Pfeiffer, S. I. (2003). Challenges and Opportunities for Students Who Are Gifted: What the Experts Say. *Gifted Child Quarterly, 47*, 161-169.

Pfeiffer, S. I. (2012). Current Perspectives on the Identification and Assessment of Gifted Students. *Journal of Psychoeducational Assessment*, *30*, 3-9.

Pfeiffer, S. I., & Blei, S. (2008). Gifted identification beyond the IQ test: Rating scales and other assessment procedures. In S. I. Pfeiffer (Ed.), *Handbook of giftedness in children: Psychoeducational theory, research, and best practices* (pp. 177-198). New York, NY, US: Springer.

Pfeiffer, S. I., & Jarosewich, T. (2007). The Gifted Rating Scales-School Form. *Gifted Child Quarterly*, *51*, 39-50.

Pfeiffer, S. I., & Petscher, Y. (2008). Identifying Young Gifted Children Using the Gifted Rating Scales—Preschool/ Kindergarten Form. *Gifted Child Quarterly*, *52*, 19-29.

Pierce, R. L., Adams, C. M., Speirs Neumeister, K. L., Cassady, J. C., Dixon, F. A., & Cross, T. L. (2006). Development of an identification procedure for a large urban school corporation: Identifying culturally diverse and academically gifted elementary students. *Roeper Review, 29*, 113-118.

Powell, W. W., & Snellman, K. (2004). The Knowledge Economy. *Annual Review of Sociology*, *30*, 199-220.

Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent growth curve modeling*. Thousand Oaks, CA: Sage Publications.

Renzulli, J. S. (1978). What Makes Giftedness? Reexamining a Definition. *The Phi Delta Kappan*, *60*, 180-184, 261.

Renzulli, J. S. (1986). The Three Ring Conception of Giftedness: A Developmental Model for Creative Productivity. In R. J. Sternberg & J. E. Davidson (Eds.), *Conceptions of giftedness* (pp. 53 – 92). New York , NY: Cambridge University Press.

Renzulli, J. S. (2005). The Three-Ring Conception of Giftedness: A Developmental Model for Promoting Creative Productivity. In R. J. Sternberg & J. E. Davidson (Eds.), *Conceptions of giftedness* (2nd ed., pp. 246-279). New York, NY: Cambridge University Press.

Shaw, E. J., Kobrin, J. L., Patterson, B. F., & Mattern, K. D. (2012). *The Validity of the SAT for Predicting Cumulative Grade Point Average by College Major*. New York: The College Board.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.

Sternberg, R. J. (1988). *The triarchic mind: A new theory of human intelligence*. New York: Viking.

Sternberg, R. J. (1996a). For whom does the bell curve toll? It tolls for you. *Journal of Quality Learning, 6*, 9-27.

Sternberg, R. J. (1996b). The sound of silence: A nation responds to its gifted. *Roeper Review, 18*, 168-172.

Sternberg, R. J. (1999a). The theory of successful intelligence. *Review of General Psychology, 3*, 292-316.

Sternberg, R. J. (1999b). A triarchic approach to the understanding and assessment of intelligence in multicultural populations. *Journal of School Psychology, 37* 145-159.

Sternberg, R. J. (2005a). Intelligence. In K. J. Holyoak & R. Morrison (Eds.), *Cambridge Handbook of Thinking and Reasoning* (pp. 751-773). New York: Cambridge University Press.

Sternberg, R. J. (2005b). The theory of successful intelligence. *Revista Interamericana de Psicología/Interamerican Journal of Psychology, 39*, 189-202.

Sternberg, R. J. (2005c). The WICS Model of Giftedness. In R. J. Sternberg & J. E. Davidson (Eds.), *Conceptions of giftedness* (2nd ed., pp. 327-342). New York, NY: Cambridge University Press.

Sternberg, R. J. (2006). The Rainbow Project: Enhancing the SAT through assessments of analytical, practical, and creative skills. *Intelligence, 34*, 321-350.

Sternberg, R. J. (2009). The Rainbow and Kaleidoscope Projects: A new psychological approach to undergraduate admissions. *European Psychologist, 14*, 279-287.

Sternberg, R. J. (2010). Assessment of gifted students for identification purposes: New techniques for a new millennium. *Learning and Individual Differences, 20*, 327-336.

Sternberg, R. J., Bonney, C. R., Gabora, L., & Merrifield, M. (2012). WICS: A Model for College and University Admissions. *Educational Psychologist*, *47*, 30-41.

Sternberg, R. J., & Grigorenko, E. L. (2004). Successful Intelligence in the Classroom. *Theory Into Practice 43*, 274-280.

Sternberg, R. J., Grigorenko, E. L., & Jarvin, L. (2006). Identification of the gifted in the new millennium: Two assessments for the broad identification of gifted students. *KEDI Journal of Educational Policy, 3*, 7-27.

Sternberg, R. J., Jarvin, L., & Grigorenko, E. L. (2011). *Explorations in Giftedness*. New York, NY: Cambridge University Press.

Sternberg, R. J., & The Rainbow Project Collaborators. (2006). The Rainbow Project: Enhancing the SAT through assessments of analytical, practical and creative skills. *Intelligence, 34* 321-350.

Sternberg, R. J., The Rainbow Project Collaborators, & The University of Michigan Business School Project Collaborators. (2004). Theory-Based University Admissions Testing for a New Millennium. *Educational Psychologist*, *39*, 185-198.

Stilwell, L. A., Dalessandro, S. P., & Reese, L. M. (2011). *Predictive Validity of the LSAT: A National Summary of the 2009 and 2010 LSAT Correlation Studies*. Newtown, PA: Law School Admission Council.

Subotnik, R. F., Edmiston, A. M., & Rayhack, K. M. (2007). Developing national policies in STEM talent development: Obstacles and opportunities. *Science Education: Models & Networking of Student Research Training under 21*, *16*, 28-38.

Talento-Miller, E., & Rudner, L. M. (2005). *GMAT® Validity Study Summary Report for 1997 to 2004*. McLean, Virginia: Graduate Management Admission Council.

Tan, M., Aljughaiman, A. M., Elliott, J. G., Kornilov, S. A., Prieto, M., Bolden, D., et al. (2009). Considering language, culture, and cognitive abilities: The international translation and adaptation of the Aurora Assessment Battery. In E. L. Grigorenko (Ed.), *Multicutural psychoeducational assessment* (pp. 443-468). New York, NY: Springer.

Tannenbaum, A. J. (1958). History of interest in the gifted. In N. B. Henry (Ed.), *Education for the gifted: The fifty-seventh yearbook for the National Society for the Study of Education, Part 2.* (pp. 21-38): Chicago, IL, US: National Society for the Study of Education; Chicago, IL, US: University of Chicago Press.

The World Bank. (2003). *Lifelong Learning in the Global Knowledge Economy: Challenges for Developing Countries*. Washington, D.C.: The World Bank.

Thompson, P. (2004). *Skating on thin ice: the knowledge economy myth*. Glasgow, UK: University of Strathclyde.

Tomlinson, S. (2008). Gifted, talented and high ability: selection for education in a one-dimensional world. *Oxford Review of Education, 34*, 59-74.

U.S. Department of Education. (1993). *National Excellence: A Case for Developing America's Talent*. Washington, DC: Author.

VanTassel-Baska, J., Feng, A. X., & De Brux, E. (2007). A Study of Identification and Achievement Profiles of Performance Task-Identified Gifted Students Over 6 Years. *Journal for the Education of the Gifted, 31*, 7-34.

Ziegler, A., Stoeger, H., & Vialle, W. (2012). Giftedness and Gifted Education: The Need for a Paradigm Change. *Gifted Child Quarterly, 56*, 194-197.

Appendix A

Aurora-*a* Examples

LIMITED METAPHORS: Words-Analytical (Aurora-*a*)

Directions:

Sometimes people compare things that seem very different. Below are sentences that compare things, but the sentences aren't finished. Finish the sentences by explaining how the first thing is like the second thing. Be creative! There are no wrong answers.

Example:

<u>*Homework*</u> is like <u>*health food*</u> because

<u>it is good for you, even though you might not like it!</u>

FIGURATIVE LANGUAGE: Words-Creative (Aurora-*a*)

Directions:

Below are some sentences from stories. The underlined parts say something in an interesting way. For example, "My mom says it's <u>raining cats and dogs!</u>" doesn't mean that cats and dogs are falling from the sky. It means that it is raining very hard.

Think about what these sentences *really* mean. Then choose the sentence that would make the *most* sense if it came next in the story.

Example:

She was nervous about entering the bike race, but her parents' support <u>built her up</u>.

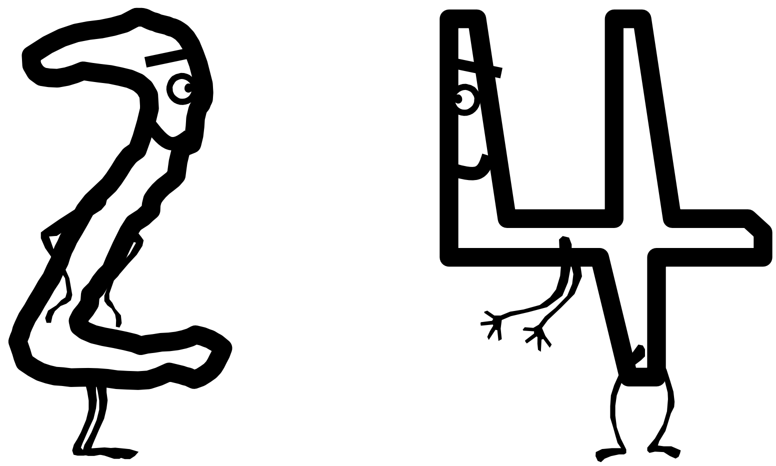
A) After she won 3rd place in her school, she was very glad she decided to compete.
B) She didn't race in the end, and was glad to just watch with her mom and dad.
C) Using the special pedals they had built for her, she was able to win the race!
D) The high heels her mother gave her helped her to pedal faster than ever.

A) is the correct answer because "built her up" means they made her feel better about herself so she decided to compete in the race.

Directions:

Here are some interesting questions using cartoon numbers. Read the questions and use your imagination to answer them in a creative way. There are no wrong answers!

NUMBER TALK: Numbers-Creative (Aurora-*a*)



Number 2 and number 4 are talking, and they are having a really good time. Why are 2 and 4 getting along so well?

2 and 4 get along so well because they have a lot in common!

They are both even numbers. 2 is happy that when he is added with another 2, they are like 4. And 4 thinks it's great that when 2 and 4 get together they make 24, the number of hours in a day!

MONEY EXCHANGE: Numbers-Practical (Aurora-*a*)

Directions:

Below are questions about kids and their money. Do your best to answer them. You may draw pictures or show your work to help you figure out the answer if you like. Write your answers in the blanks.

Example:

Abby and Ben buy ice cream together.
Abby's ice cream costs $4.00.
Ben's ice cream costs $3.00.
Abby also owes Ben $1.00 from another day.

How much should each of them pay so that Abby doesn't owe Ben any more money?

Work:

Abby should pay for some of Ben's ice cream so she won't owe him any more money.

If she pays for hers plus $1.00 of his, she will pay $4.00 + $1.00 = $5.00.

Now Ben can pay $1.00 less for his ice cream: $3.00 - $1.00 = $2.00.

Abby: $5.00 Ben: $2.00

Aurora-*g* Example

Directions for Verbal Classification (Aurora-*g*)

In each question below, there are four words given. Three of these go together in some way, and one does not belong. Figure out which word does not belong and circle it.

Example A:

Hail Snow Sun Rain

All of these words have something to do with weather. Three of them fall to the ground from the sky: Hail, Snow and Rain. The sun does not fall to the ground like this.

The answer is SUN.

Appendix B

Aurora-*r* Examples

Aurora-*r* (Teacher Rating Scale) Creative

Domain Specific	Is good at coming up with original solutions to					
	problems that involve words.	The exact opposite of my student	Not like my student	Somewhat like my student	A lot like my student	Exactly like my student
	problems that involve numbers.	The exact opposite of my student	Not like my student	Somewhat like my student	A lot like my student	Exactly like my student
	problems that involve images.	The exact opposite of my student	Not like my student	Somewhat like my student	A lot like my student	Exactly like my student
Non-Domain Specific	When posed with a "what if" situation, likes to generate lots of different possible outcomes	The exact opposite of my student	Not like my student	Somewhat like my student	A lot like my student	Exactly like my student