Validation of methods to identify and measure the underachievement of gifted students

Author:
Jackson, Rahmi

Publication Date:
2017

DOI:
https://doi.org/10.26190/unsworks/19552

VALIDATION OF METHODS TO IDENTIFY AND MEASURE THE UNDERACHIEVEMENT OF GIFTED STUDENTS

by

Rahmi Luke Jackson

A thesis submitted to

The University of New South Wales

in fulfilment of the requirements

for the degree of Doctor of Philosophy

April, 2017

ORIGINALITY STATEMENT

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree of diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by others, with whom I have worked at UNSW or elsewhere, is explicitly acknowledged in the thesis. I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentations, and linguistic expression is acknowledged.

Signed:.....

Date: 10/04/2017

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my supervisor, Dr Jae Jung, for his time, encouragement, expertise, and for not accepting mediocrity from me in a thesis on underachievement. I am also thankful to my co-supervisor, Emeritus Professor Miraca Gross, for her wisdom and encouragement. In addition, I would like to thank my ex-principal who generously donated the data used in this research and also the staff for their participation in my research, and encouragement along the way. In particular, the Head of IT, whose computer expertise was repeatedly called upon for extracting data from archives and finding efficient methods of data linkage.

There have been many people in my life for whom without their support I would be just another statistic in this thesis, which would not have been completed. Firstly, to my amazing and most excellent wife. This work was only possible with your constant support and sacrifice, because of you I hold myself supremely blest—blest beyond what language can express. You have helped me to make the choice between what is right, and what is easy, and even further, you helped share the burden of that choice. This achievement is as much yours as mine. Secondly, to my parents, you have taught me the joy of curiosity, the comfort of a hobbit-hole, the fire in the equations, and to attempt great things. Thirdly, to the Maggs family, thank you for your generosity, endless practical support, and for adopting me as part of the family. Lastly, thank you to my brothers and sisters, extended family, friends, colleagues, students, and my church family who have all shared this journey with me, your support and interest has helped keep this project alive. Particular thanks goes to Stephen McGuinness for our spontaneous meetings, Mark and Nicole Schroder for your prayers and friendship, and to the Class of 2016 for teaching me while I taught you. *Soli Deo Gloria*

ABSTRACT

Much confusion exists about gifted underachievement as a result of significant variations in how gifted underachievement has been identified and measured. From a review of the literature, five methods were found to be commonly used to *identify* gifted underachievement, two of which also *measure* the degree of gifted underachievement. In this project, the validity of the use of these methods was assessed using empirical convergence, criterion, and generalisation evidence obtained using data collected from a K–12 school in the Sydney metropolitan area (Australia). First, convergence was assessed using three different approaches to compare the results obtained using the five different identification/ measurement methods: (a) the differences in proportions of gifted students identified as exhibiting underachievement, (b) the correlation of identification/measurement results, and (c) the agreement of the identification/measurement results. The findings suggested that the different methods commonly used to identify/measure gifted underachievement cannot be considered convergent, and therefore should not be used interchangeably. Second, a latent class model was used to assess the criterion validity of the individual identification/ measurement methods. Two of these methods were found to have strong levels of criterion validity. Third, meta-analysis was carried out across 41 different combinations of expected achievement and actual achievement data obtained from the school, to indicate that generalisation was not possible across different combinations of expected/actual achievement data. A final assessment of the validity of each identification/measurement method was completed by synthesising the empirical evidence within the context of the planned uses of the methods. The final assessment demonstrated that the simple difference method may be the most valid method to use to identify and measure gifted underachievement.

Keywords: *Underachievement, Gifted education, Identification, Measurement*

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Absolute Split I (ABSI): A statistical method for identifying gifted underachievement.

Absolute Split II (ABSII): A statistical method for identifying gifted underachievement.

Australian Curriculum, Assessment and Reporting Authority (ACARA): The national authority on school curriculum, assessment, and reporting in Australia.

American Educational Research Association (AERA): A professional organisation that promotes scientific study of education.

Australian Independent Schools association (AIS): An association that represents and advises the independent school sector in Australia.

American Psychological Association (APA): A scientific and professional organisation representing psychologists in the U.S.A.

Board of Studies, Teaching and Educational Standards NSW (BOSTES): The New South Wales state government authority on teaching and education.

Concordance Correlation Coefficient (CCC): A measurement of agreement between two continuous variables.

Catholic Education Office (CEO): The governing body of Catholic schools in Australia.

Department of Education (NSWDOE): The NSW government department responsible for state education including all public schools. Previously this department has been known as the **Department of Education and Training (NSWDET)**, and the **Department of Education and Communities (NSWDEC).**

Department of Education and Training (DET): The federal government department responsible for the national education sector. Previously this department has been known as the **Department of Education, Employment and Workplace Relations (DEEWR)**, and the **Department of Education (DOE)**.

Education Resources Information Center (ERIC): An online database of education research and literature

Gifted achievement (GA): Occurs when there is a reasonable level of agreement between a student's high level of ability and their level of achievement.

Gifted Underachievement (GUA): Occurs when there is a significant discrepancy between a student's high level of ability and their level of achievement.

Higher School Certificate (HSC): The final credential for secondary education for students in NSW Australia who are finishing and exiting high school, typically used to refer to the exit examinations which determine a student's HSC marks and university matriculation.

Otis–Lennon School Ability Test (OLSAT): An instrument that measures intelligence. The instrument is administered in a group environment and does not require a psychologist to administer or interpret.

National Assessment Program—Literacy and Numeracy (NAPLAN): An annual series of national assessments that Australian students sit in Years 3, 5, 7, and 9.

National Council on Measurement in Education (NCME): An organisation of American researchers and practitioners specialising in educational measurement and testing.

Nomination (NOM): A method for identifying gifted underachievement.

Regression (REG): A statistical method for measuring and identifying gifted underachievement.

School Assessment (SA): Refers to all school-based assessment instruments.

School Certificate (SC): A former high school credential for NSW students who finish junior high school (year 10). Often used to refer to the external examinations that students sit to determine their SC marks. The SC qualification ceased in 2011.

Simple Difference (SD): A statistical method for measuring and identifying gifted underachievement.

Underachievement (UA): Occurs when there is a significant discrepancy between any student's level of ability and their level of achievement.

1 Introduction

1.1 Context

Education has long been promulgated as the best means to better oneself and advance society. Indeed, researchers have shown that education transforms lives, and is necessary for the success, health, and happiness of individuals in the modern world (Australian Government, 2014; Carnevale, Rose, & Cheah, 2011; Cutler & Lleras-Muney, 2010; Feinstein, Sabates, Anderson, Sorhaindo, & Hammon, 2006; Grossman, 2000; Jagger et al., 2007; Mishel, Bivens, Gould, & Shierholz, 2012; Salinas-Jimenez, Artes, & Salinas-Jimenez, 2011). Remarkably, the benefits of education appear to extend beyond a single individual to the lives of people across several generations (Feinstein, Duckworth, & Sabates, 2008), and even impact the wider population through reduced rates of crime, reduced inequality, and greater productivity (Carvacho, et al., 2013; Groot & van den Brink, 2010; Machin, Marie, & Vujić, 2011; Timmermans, van Lier, & Koot, 2009). Perhaps as a result, education has been recognised internationally as a basic human right and the means to abolish war, poverty, and other world problems (United Nations, 2006). Therefore, the educational underachievement of any group has been considered to be a significant concern, and government policies across the world are focused on "eliminating all forms of underachievement" (Smith, 2010, p. 38).

Surprisingly, gifted students appear to be one group of students who are underachieving. Although giftedness is commonly perceived to provide advantage in the modern world (Jung, 2014), gifted students have in fact been reported to be suffering from an "epidemic" of underachievement (Hoover-Schultz, 2005; Moon, 2009; Rimm, 2003). This disturbing claim is supported by evidence which demonstrates that: (a) up to half the population of gifted students exhibit significant academic underachievement, (b) one in five gifted students drop out of high school, and (c) forty percent of gifted students do not

complete tertiary studies (Hsieh, Sullivan, & Guerra, 2007; Rayneri, Gerber, & Wiley, 2006; Rimm, 1997, 2003). Interestingly, the disappointing educational outcomes for many gifted students do not appear to be fully explainable by the commonly recognised sources of disadvantage including socio-economic status, ethnic background, or disability. The high levels of underachievement among gifted students appear to also arise from the unique socio-emotional and learning needs of these students (Coleman, 2014; Masden, Leung, Shore, Schneider, & Udvari, 2015; Rimm, 2003; Silverman, 1997; Wellisch & Brown, 2011). Hence, giftedness may be a distinct source of disadvantage that appears separate to, and additional to, the other common sources of disadvantage.

It is noteworthy that underachievement in gifted students has often been found to be a precursor to, or an indicator of, significant social and emotional problems (Blaas, 2014). An Australian Senate enquiry into gifted education found gifted underachievers were characterised by a range of psychosomatic and psychological symptoms, ranging from stress related eczema, stomach aches, and poor self-esteem, to depression, self-harm, and mental confusion (Commonwealth of Australia, 2000). Other studies have found that many of the social and emotional problems that arise may have serious consequences that can be long-lasting and debilitating in effect (Berndt, Kaiser & van Aalst, 1982; Cross, 2013; Harrison & Van Haneghan, 2011; Harter, 2006; Hayes & Sloat, 1989; Kroesbergen, van Hooijdonk, Van Viersen, Middel-Lalleman, & Reijinders, 2016; McCall, 1994; Missett, 2013; Rimm, 2003; Zeidner & Schleyer, 1999). In addition, many cases of gifted underachievement are thought to be a result of an undiagnosed disability (Lovett & Sparks, 2013; Moon, 2009; Reis, Baum, & Burke, 2014).

The identification and measurement of gifted underachievement, which allows for the provision of appropriate educational interventions for underachieving gifted students, may be life-changing and sometimes life-saving for the affected students (Dittrich, 2014; Gross,

2010). Some of the provisions that appear to be most useful for supporting gifted students include extension beyond material covered in the regular classroom, acceleration, ability grouping, mentoring, and specific talent development programs (Reis, Burns, & Renzulli, 1992; Colangelo et al., 2010; Eddles-Hirsch, Vialle, Rogers, & McCormick, 2010; Gagné, 2011). Nevertheless, some scholars suggest that many underachieving gifted students may simultaneously require additional interventions (Gagné, 2011). For example, Wellisch and colleagues (Wellisch, 2016; Wellisch & Brown, 2012) recommend that educators may need to plan educational and therapeutic interventions for underachieving gifted students based on the individual's exceptional intellectual strengths (their giftedness) *and* on their exceptional weaknesses (which may include socio-emotional problems, areas of learning difficulty, and the presence of disabilities).

The identification and measurement of gifted underachievement is, nevertheless, a difficult enterprise. Research has suggested that teachers and peers typically do not notice underachievement in gifted students (Jones, 2005; Lau & Chan, 2001a; Merrotsy, 2013). One possible reason for the non-visibility of gifted underachievement may be related to the fact that many underachieving gifted students are able to achieve average or even high scores relative to their age peers (Matthews & McBee, 2007; McCoach & Siegle, 2008; Reis & McCoach, 2000). In contrast, when most students underachieve at school, their underachievement tends to be easily recognised, prompting immediate support to be offered from teachers and parents. As the needs of underachieving gifted students often go unnoticed and unfulfilled, it is possible that they are more vulnerable than their lower ability peers.

1.2 Statement of the Problem

When schools attempt to respond to the problem of underachievement of gifted students, they may not always find clear guidance from the research, as much confusion

exists in the published literature (Gorard & Smith, 2004). In an early review of gifted underachievement, Dowdall and Colangelo (1982) discovered that the cause of the conflicting information appeared to be the significant variability in how gifted underachievement was identified and measured. They concluded that the use of different identification and measurement methods caused researchers to be effectively studying different populations. A more recent review of the literature in gifted underachievement by Reis and McCoach (2000), came to a similar conclusion, stating that the use of multiple different methods of identifying/measuring gifted underachievement was found to "contribute to the difficulty in studying the characteristics of this population" (p. 166). Unfortunately, the problem, that different methods of identifying/measuring gifted underachievement are used interchangeably by researchers, continues to this day (Fong & Krause, 2014; Hwang et al., 2014).

The lack of a resolution to the problem has left researchers without solid guidance for selecting a valid and consistent method for identifying/measuring gifted underachievement (Dai, Swanson, & Cheng, 2011; Neihart, Reis, Robinson, & Moon, 2002; Robinson, 2006; Schober, Reimann, & Wagner, 2004; Reis & Renzulli, 2010), and may have contributed to the perception that underachievement is an ambiguous and arbitrary construct (Smith, 2010; Ziegler, Ziegler, & Stoeger, 2012). Figg, Rogers, McCormick, and Low (2012) indeed state that gifted underachievement remains an enigma. It is noteworthy that there has been a significant decline and stagnation in the number of studies investigating gifted underachievement since the year 2000 (Morisano & Shore, 2010; Ziegler, Ziegler, and Stoeger, 2012).

1.3 Purpose of the Investigation

The main purpose of this investigation is to produce guidance for researchers and educators in the selection of the best method(s) to identify and measure gifted underachievement. To this end, a rigorous review and assessment of the multiple methods that are commonly used to identify and measure gifted underachievement was undertaken, with a focus on the validity of the use of these methods (Kane, 2013). Thereafter, a valid method (or methods) for the identification and measurement of gifted underachievement will be proposed, with the expectation that such guidance will reduce the ambiguity of the underachievement construct, and enable the advancement of research that will provide a clearer understanding of how to avoid the "frustrating loss of potential" that gifted underachievement represents (Ritchotte, Matthews, & Flowers, 2014, p.183).

1.4 Significance of the Investigation

This investigation is significant for a number of reasons. From a research perspective, it is necessary to address claims that gifted underachievement is an ambiguous and arbitrary construct with questionable conceptual meaning and usefulness. Criticisms that the construct "ought to be abandoned" (Ziegler, Ziegler, & Stoeger, 2012, p. 123) or "has probably outlived its usefulness" (Smith, 2010, p. 446) may indeed be shown to be unfounded if valid and reliable methods that identify and measure gifted underachievement are demonstrated to exist. The conceptual validation of the construct is likely to contribute to advancing the field of gifted education as it is likely to promote further research on gifted underachievement (with an identical population of gifted students), including research on its causes, interventions, and strategies for prevention, to allow for the optimal support of gifted students to fulfil their potential (Reis & McCoach, 2000).

Relatedly, the findings of the investigation may address the ongoing concern expressed by scholars over a number of decades about the inconsistent use of methods to identify and measure gifted underachievement. It is noteworthy that almost every study on gifted underachievement has denounced the state of the research in the area as a result of the use of multiple contrasting methods (e.g., Phillipson & Tse, 2007; Veas, Gilar, Miñano & Castejón, 2016), while several scholars have discussed this problem from a conceptual perspective (e.g. McCall, Beach, & Lau, 2000; Reis & McCoach, 2000; Smith, 2010; Van den Broeck, 2002a, 2002b). Interestingly, Ziegler, Ziegler, and Stoeger (2012) suggest that the continued neglect of this 'elephant in the room' may be attributed to the fact that the major "points of criticism (may) have simply been forgotten" (p. 123).

From a non-research perspective, this investigation is significant because of the improved advice that educators may be offered on how to optimally identify and measure gifted underachievement. Guidance on the selection of a valid method for the identification/measurement of gifted underachievement may greatly assist schools to meet demands from a social justice perspective and allow for the fulfilment of duty of care responsibilities towards the gifted student population. In addition, governments and policy makers may be encouraged to offer greater funding to meet the special needs of gifted students.

The significance of this investigation may extend well beyond the field of gifted education. The identification and measurement of underachievement is fundamental to understanding and resolving the underachievement of *all* students. Indeed, the problems that have plagued the field of gifted education have also plagued the broader field of education, whose researchers have suggested that "underachievement is a term over which there is little consensus" (Smith, 2010, p. 41). Outside of education, it is noteworthy that researchers have used some of the methods to identify/measure underachievement in investigations that

contribute to international discussions across the fields of sociology, psychology, law, and policy on topics including disability, ethnicity, social class, delinquency, motivation, child development, resilience, and gender gaps (American Psychiatric Association, 2000; Jackson et al., 2011; Jones, 2005; Maki, Floyd, & Roberson, 2015; Maynard, Waters, & Clement, 2013; McCall, Beach, & Lau, 2000; Preckel, Holling, & Vock, 2006; Sikora & Saha, 2011; Smith, 2010; Strand, 2014; Timmermans et al., 2009; Tuss, Zimmer, & Ho, 1995; Van den Broeck, 2002a; Wood, 2003; Zirkel & Thomas, 2010).

1.5 Description of Thesis Contents

The thesis has been divided into eleven chapters. Following the introduction chapter (Chapter 1), are the chapters on the literature review (Chapter 2), the theoretical framework (Chapter 3), methodology (Chapter 4), results of the convergence evidence of validity (Chapter 5), results of the criterion evidence of validity (Chapter 6), results of the generalisability evidence of validity (Chapter 7), an overall assessment of the validity evidence (Chapter 8), conclusion (Chapter 9), references (Chapter 10), and appendices (Chapter 11).

2 Literature Review

2.1 Introduction

The underachievement of gifted students has remained an enigma despite over six decades of research (Farquhar & Payne, 1964; Obergriesser & Stoeger, 2015). It has been suggested that the lack of progress is, at least in part, due to an unresolved inconsistency in the use of methods to identify and measure gifted underachievement (Reis & McCoach, 2000). This chapter will examine in detail the major ideas that contribute to understanding this problem of inconsistent approaches used by researchers. First, this chapter discusses the nature of gifted students and how the selection of a model of giftedness impacts the identification of gifted underachievement. Second, the apparent paradox of gifted underachievement is discussed. Third, the common methods for identifying and measuring gifted underachievement are described, including variations of these methods. Fourth, some confusion raised by the non-technical use of the term underachievement is addressed, along with proposed resolutions from the literature.

2.2 Gifted Students

2.2.1 A brief history of intelligence research

The first paradigm shift in intelligence research occurred as researchers moved from an understanding of intelligence as a physical characteristic to a psychological characteristic. Galton (1869, 1883) produced one of the earliest theories of intelligence, picturing intelligence as the level of sensitivity of human senses (i.e., a physical characteristic). He argued that the greater the sensitivity, the more information one acquired, and the smarter one was. Thereafter, Binet (1903) shifted the paradigm of intelligence into a psychological construct incorporating a range of branches (e.g., memory, problem solving), each measured separately and combined into a measure of intelligence. Binet believed that intelligence was

not a fixed construct, and that it was highly influenced by environmental factors (Siegler, 1992). In comparison, Spearman (1904) provided evidence that the separate branches of intelligence were linked, and produced a unitary theory and measure of intelligence. Spearman's general construct of intelligence is often referred to as the *g* factor, which even in modern times remains perhaps the most important psychometric entity - no other measurable factor has been found that "contributes more than *g* to the understanding as well as to the prediction of human achievements" (Neubauer & Opriessnig, 2014, p. 1). The conceptualisation of intelligence as a single *g* factor has been the foundation of research that defined and identified gifted students throughout the twentieth century.

The concept of a single measure of intelligence was first challenged by Thurstone's (1938) idea of multiple intelligences, which was further developed and popularised by Gardner (1983, 2011). This theory proposes that intelligence is not just a single general ability, but contains several separate primary mental abilities. The corollary, for defining giftedness, was that a student could be gifted in one area (e.g., quantitative reasoning) and not another (e.g., visual-spatial reasoning) and that this intellectual *asynchrony* should not be unexpected. Later research, using modern statistical techniques to analyse Gardner's original data, has since found that these multiple intelligences may be correlated, and support the existence of a general intelligence (Carroll, 1993; Gottfredson, 2003; Morgan, 1996; Plucker, Callahan, & Tomchin,1996; Pyryt, 2000; Visser, Ashton, & Vernon, 2006). It is noteworthy that modern cognitive neuroscience does not support the theory of multiple intelligences (Gottfredson, 2006; Waterhouse, 2006), while Gardner himself has admitted to the lack of evidence for the validity of his theory (Gardner, 2004). Reflecting the many evidence based denouncements of Gardner's theory, some have labelled it as a "pseudoscience" (Gottfredson, 2011; Jones, 2010).

Horn and Cattell further challenged the concept of a singular intelligence construct and developed a theory of fluid and crystallised intelligence. Fluid intelligence refers to the ability to reason and identify relationships that exist, whereas crystallised intelligence is knowledge gained from previous instruction or experience (Horn & Cattell, 1967). Horn and Cattell claimed that fluid intelligence peaks at adolescence then declines into adulthood, whereas crystallised intelligence continues to grow throughout adulthood. This was a direct contradiction of Spearman's model which claimed that intelligence "appears to become fully developed in children by about their ninth year ... there normally occurs no further change even into extreme old age" (Spearman, 1904, p. 285). One impact of the work of Horn and Cattell for the measurement of intelligence, is that tests are now significantly different for different age groups and allow for the increasing inclusion of knowledge based tasks for older age groups (Kaplan & Saccuzzo, 2010).

The most popular model of intelligence at the present time is derived from the work of Horn and Cattell. The Cattell–Horn–Carroll model, which is shown in Figure 1, contains three levels of mental abilities which differ in their degree of generality. The bottom level contains narrow abilities (e.g., reading decoding, judging rhythm etc.), several of which are highly correlated and form clusters of ability referred to as broad abilities (e.g., fluid reasoning, visual processing etc.), which form the second level of the model. All of the broad abilities are highly correlated and form part of one overall general ability (Spearman's *g*), which forms the third level of the model. This model represents the most empirically validated theory of intelligence that is currently available (Australian Psychological Society, 2013; McGrew, 2009).

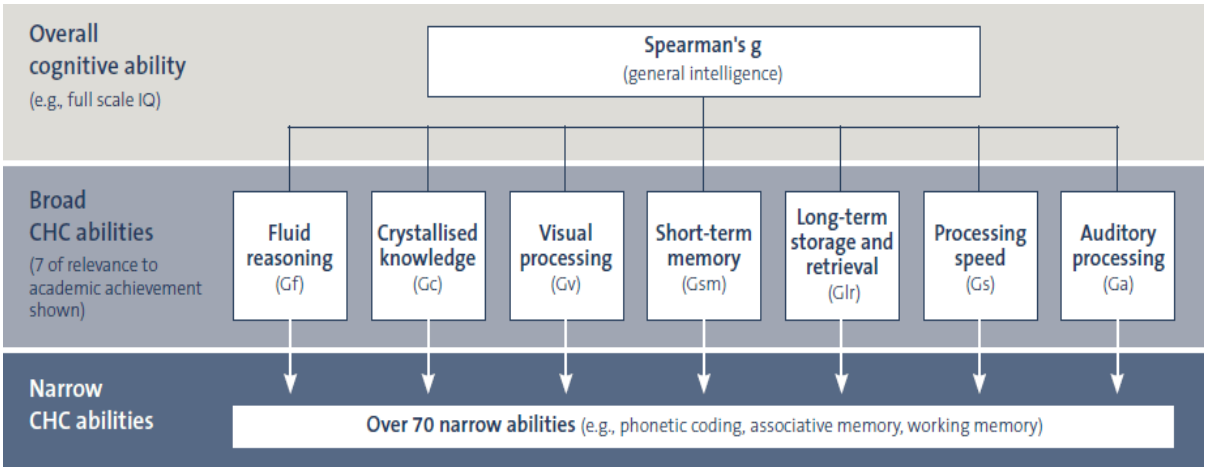


Figure 1. The Cattell–Horn–Carroll model of intelligence (Australian Psychological Society, 2013)

2.2.2 Models of giftedness

Over the last century, several significant changes to our understanding of giftedness have occurred. In one of the first major studies of giftedness, Terman (1925) considered the giftedness label to be appropriate for students with a very high score on the Stanford–Binet IQ test. Later, DeHaan and Havighurst (1957) expanded on Terman's idea of giftedness and the concept of multiple intelligences, and proposed that a student could be considered gifted if they had high ability in any one of six ability domains. It is noteworthy that while Terman (Terman & Oden, 1947) was successful in showing that gifted students are healthy, well-adjusted and sane, he also discovered that many of their talents did not actualise into adult achievements. Hence, even in the early stages of the development of the field of gifted education, a recognition existed that high ability alone did not produce high achievement.

Subsequent models of giftedness began to incorporate additional factors to better predict high achievement in adulthood. For example, both Tannenbaum's five arm star model (1983; Figure 2) and Renzulli's three ring conception of giftedness (1978, 1986; Figure 3) included non-ability factors, such as task commitment, creativity, and motivation, that

students require before they may be considered gifted. Following the dissemination of the Marland Report (1972) from the US Commissioner of Education, the focus of gifted models began to shift towards developing all students who *could* possibly become high achieving adults, rather than only those who *would* most likely become high achieving adults. With this shift in focus, some may consider the Renzulli and Tannenbaum models to be too restrictive.

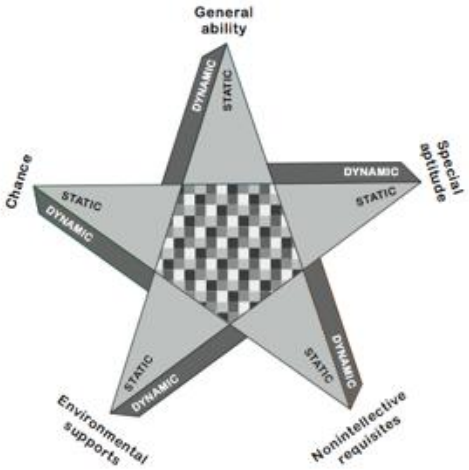


Figure 2. Five arm star of Giftedness (Tannenbaum, 1983)

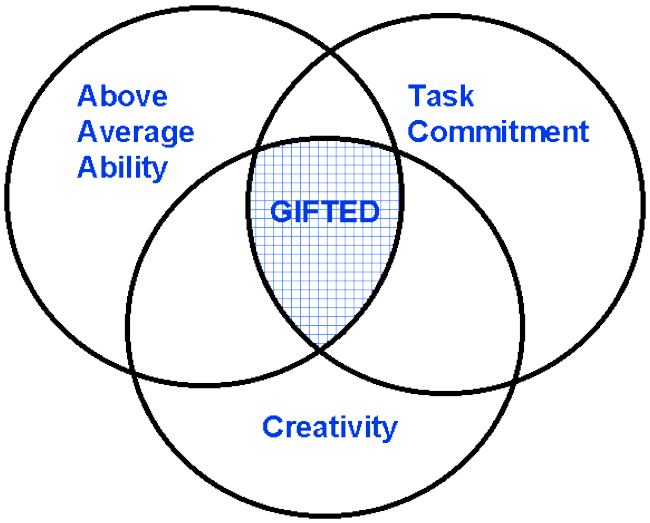


Figure 3. Three-ringed model of giftedness (Renzulli, 1986)

Modern models of giftedness hold high ability as the primary component of giftedness, but stress the importance of other factors for the development of talent. For example, the Munich model (Heller, 2004), the Universal Model of Giftedness (Jessurun, Shearer, & Weggeman, 2015; Figure 4), and Gagné's model (1985, 1995, 2003, 2009a, 2013; Figure 5) conceptualise a gifted student's performance as being moderated or catalysed by non-cognitive, environmental, or chance factors. Each of these models inherently acknowledge the existence of gifted underachievement, which may otherwise have been considered as non-giftedness in other models. Furthermore, these models suggest possible causes for gifted underachievement, including the effect of negative catalysts or moderators on the student's talent development (Gagné, 2013; Heller, 2004). These models, therefore, provide a good basis for the understanding of gifted underachievement.

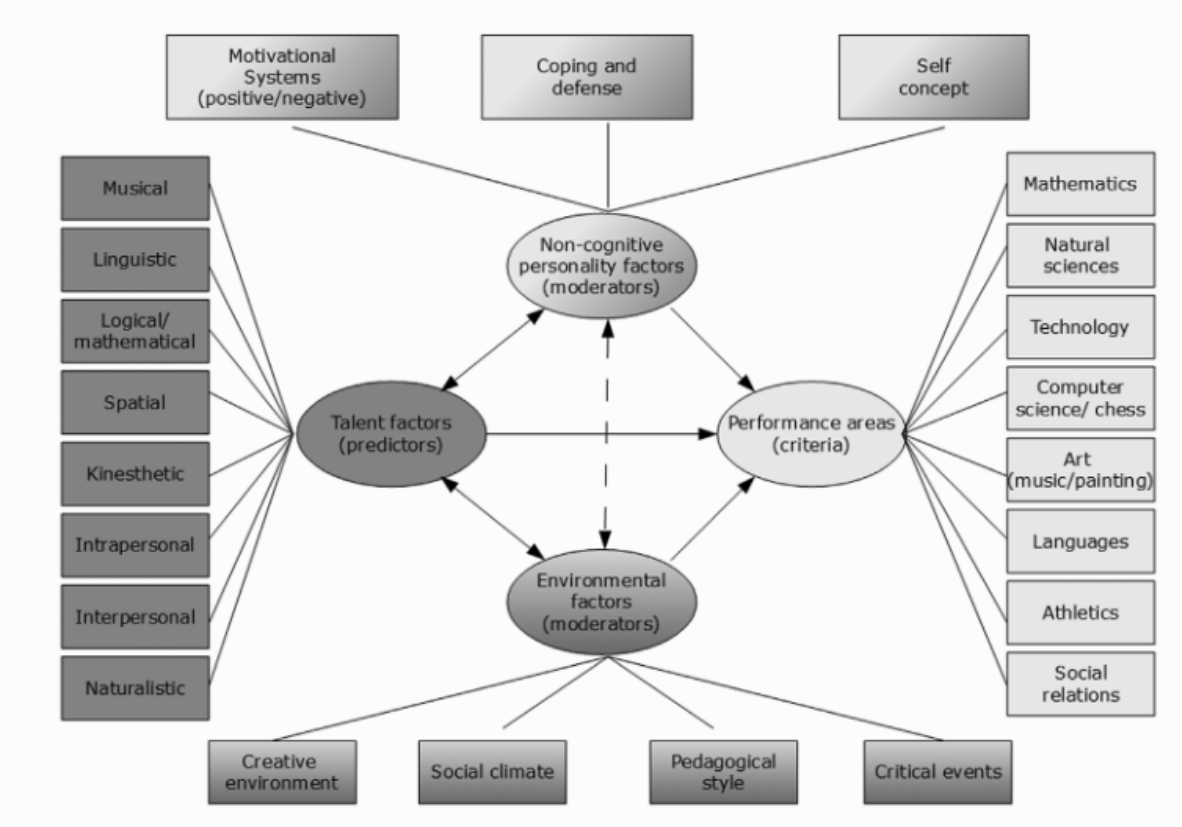


Figure 4. The Universal Model of Giftedness (Jessurun, Shearer & Weggeman, 2015)

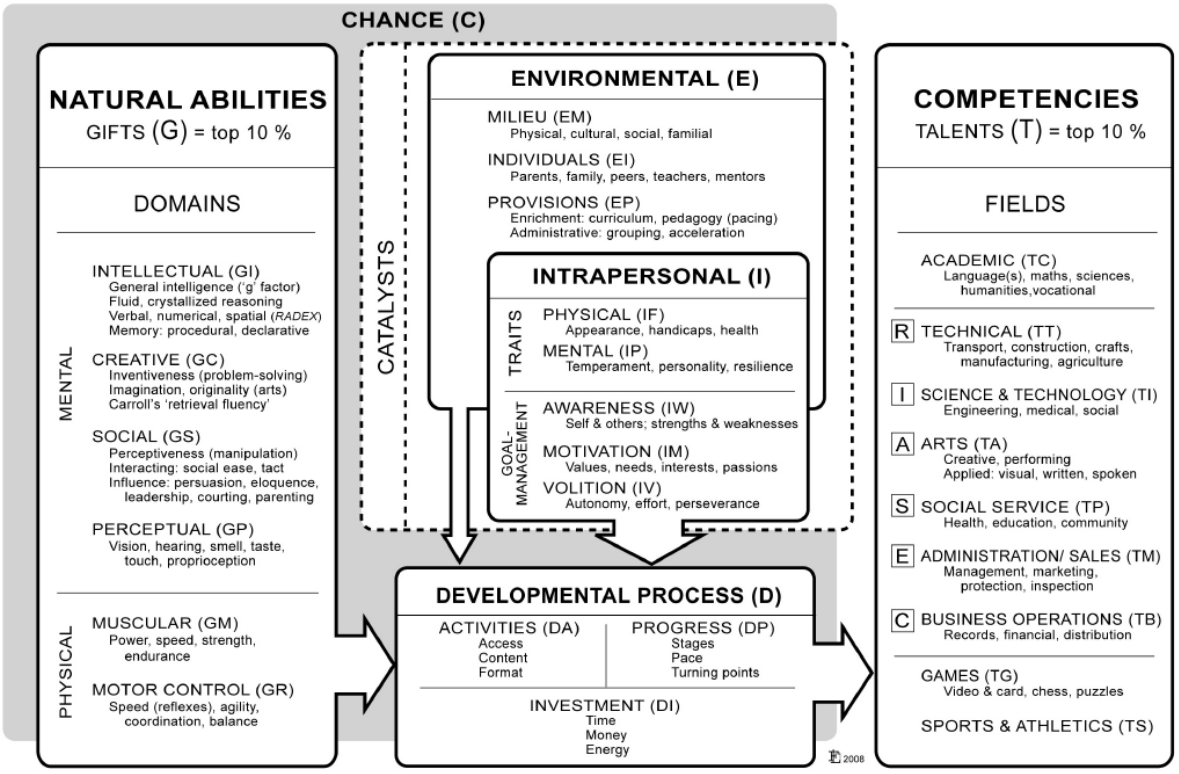


Figure 5. The Differentiated Model of Giftedness and Talent (Gagné, 2009a)

In Gagné's (2009a; 2009b; 2013) Differentiated Model of *Giftedness* and *Talent*, giftedness and talent are distinguished, as giftedness refers to the existence of high ability, while talent refers to the systematic development of this ability into high level achievements. In addition, Gagné (1998) proposed five levels of giftedness and talent based on the metric system (refer Table 1). Gagné's model suggests that all gifted students should not be expected to reach the same level of achievement, and that instead, only students at the highest levels of giftedness should be expected to be able to reach this level of achievement. According to the model, underachievement may be understood as occurring when a gifted student achieves levels of achievement that are significantly below their level of giftedness - even if their levels of achievement are still quite high relative to the general student population.

Table 1

Metric-based levels of giftedness (Gagné, 1998)

Level	Label	Prevalence	IQ equivalent	Standard deviation
1	Mildly	1:10	120	+1.3
2	Moderately	1:100	135	+2.3
3	Highly	1:1,000	145	+3.0
4	Exceptionally	1:10,000	155	+3.7
5	Extremely/Profoundly	1:100,000	165	+4.3

Among the most recent models of giftedness are Ziegler's (2005; Ziegler & Phillipson, 2012) Actiotope model (Figure 6) and Dai and Renzulli's (2008) Contextual, Emergent, and Dynamic Model, both of which propose a fundamental paradigm shift in gifted education. Rather than recognising the person to be gifted, which is argued to be "nonscientific" (Ziegler, 2005, p. 411), these newer models suggest that giftedness may be a property of the system (Harder, Vialle, & Ziegler, 2014). Among the supporters of this change in thinking are researchers who note that people who achieve excellence may be distinguishable by the quality of their learning environments and not by their abilities (Bloom, 1985; Csikszentmihalyi, 1996; Roche, 1979; Sosniak, 2006; Vaillant, 1977; Ziegler & Phillipson, 2012). While these models may become fundamental to the field of gifted education in the future, the evidence relating to, and the utility of, these new models is still being accrued (Cohen, 2012; Dai, 2012; Harder et al., 2014; Vladut, Vialle, & Ziegler, 2015). In any case, these models do not appear to be inconsistent with the processes involved in talent development proposed by scholars including Gagné (2009a, 2013), Heller (2004), and Jessurun, Shearer, and Weggeman (2015).

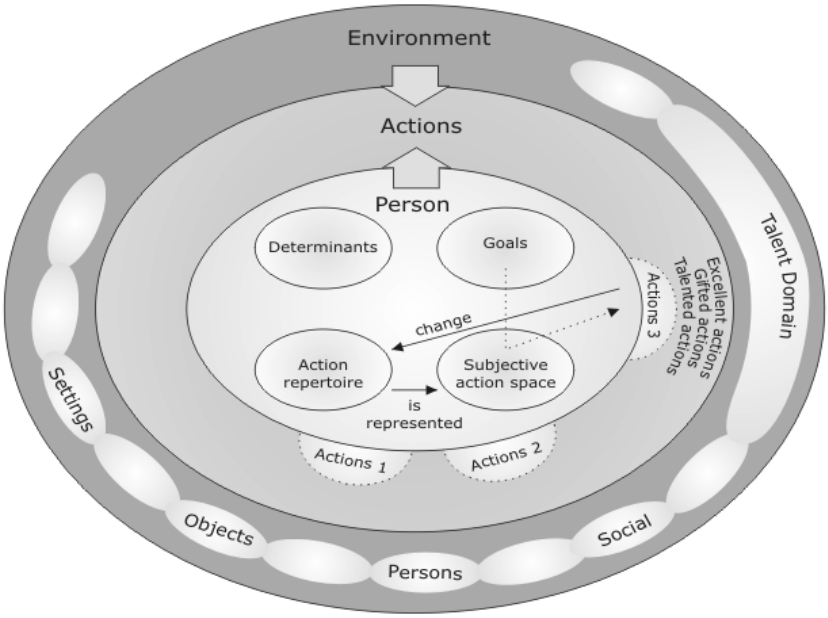


Figure 6. The Actiotope Model of Giftedness (Ziegler, 2005)

While international researchers and educators continue to debate the use of many different models of giftedness, in Australia educators seem to have reached consensus on a single model. For example, the new Australian National Curriculum has established Gagné's Differentiated Model of Giftedness and Talent (2009a, 2013) as the guiding model of giftedness in Australia (ACARA, 2015b). Furthermore, Gagné's model appears to have been broadly accepted across the peak bodies that govern the three school sectors in Australia (government, Catholic, and Independent schools) in each state/territory/diocese (e.g., the New South Wales Department of Education and Training [DET, 2004], the Sydney Catholic Education Office [CEO, 2014], and the South Australia Association of Independent Schools [AIS, 2015]).

2.2.3 Identifying gifted students

Most existing methods used to identify gifted students appear to be heavily based on measurements of intelligence (McClain & Pfeiffer, 2012; Pfeiffer, 2012). This may be

because psychometrically measured IQ scores are widely considered the single most powerful predictor of human achievement (Neubauer & Opriessnig, 2014). When intelligence tests are carried out for a large number of people, the distribution of their IQ scores forms the well known bell curve (Figure 7), which provides information about how any individual ranks in intelligence across the wider population. Knowledge of a student's IQ score has been considered "tremendously useful in both placement and programming decisions" (Assouline, 2003, p. 126).

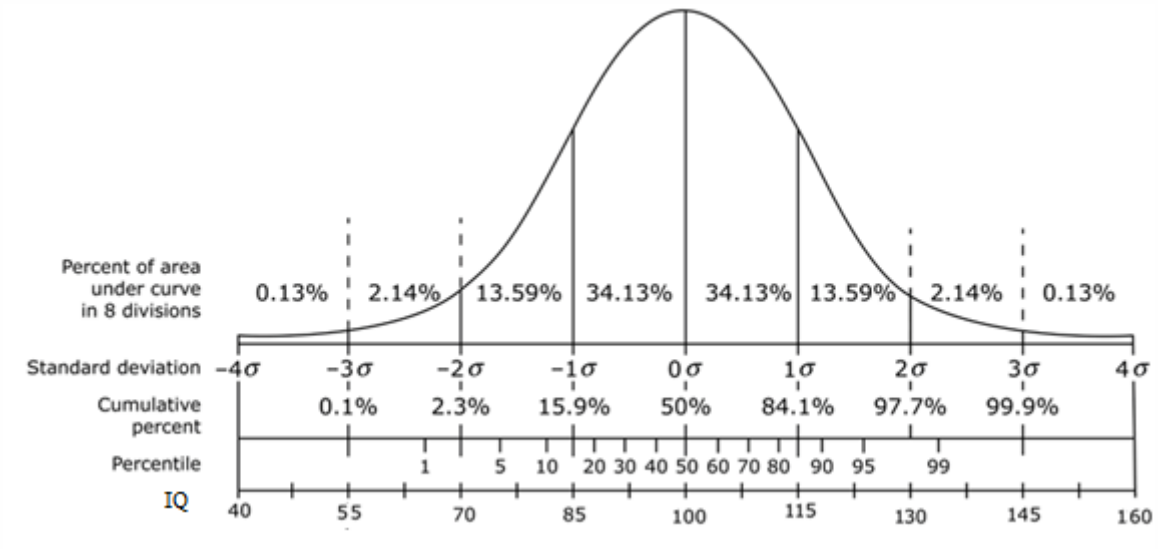


Figure 7. The Bell Curve (modified version of Ward & Murray-Ward [1999], p. 74)

Due to the usefulness of intelligence tests, a whole market comprising an estimated 3,000 commercially available intelligence tests now exist (Carlson, Geisinger, & Jonson, 2014). Among the most popular instruments are the Wechsler Intelligence Scale for Children (WISC), the Stanford–Binet (SB), Raven's Progressive Matrices, the Otis–Lennon School Ability Test (OLSAT), the Scholastic Aptitude Test (SAT), and Kaufman's Assessment Battery for Children (KABC). However, there are many differences between these

instruments, including the specific abilities and broad factors (refer Figure 1) that are tested, whether the instrument needs to be administered and interpreted by a trained psychologist, the level of language required to access the questions, the method of problem posing and student response (e.g., oral or written), whether activities are time scored (with higher scores granted for faster completion), and whether time limits are imposed. Hence, the instruments may not be considered equivalent, and the results must be interpreted with care and knowledge of the instruments. Although each instrument's measurement will be directly affected by each individual student's general intelligence and ability, the extent to which other more specific abilities additionally contribute to the measurement appears to vary.

The high costs associated with the administration of some intelligence tests present some difficulties in the use of intelligence tests. In addition to the financial cost of the test materials and the time required for analysis of results, some intelligence tests (e.g., WISC or SB) may only be administered by qualified psychologists to individual students. Consequently, the administration of individual intelligence tests to large groups of students is typically not feasible. As a result, schools often rely on other sources of information, including school grades, teacher nominations, or observations of student behaviours to select students for entry into gifted programs and provisions (McClain & Pfeiffer, 2012). One partial solution may be the use of group intelligence tests (e.g., the OLSAT). These intelligence tests may be administered to large groups of students simultaneously, do not need to be administered by qualified psychologists, and may be computer-read and analysed. Financially constrained schools often use group intelligence tests in place of individual intelligence tests, or as a screening mechanism to select a small group of students for the administration of individual intelligence tests (Worrell & Erwin, 2011).

An additional challenge to the use of intelligence tests is that, on average, students from disadvantaged backgrounds may score lower (Naglieri & Ford, 2015). Therefore, equity

concerns exist over the practice of using intelligence test results to select students for differentiated educational experiences (Ford, 2003; Herrnstein & Murray, 1994; Richert, 2003). With respect to race, research has suggested that differences in average IQ may exist across different races, and that "the average IQ differences represent real differences in the higher-order thinking skills that people have" (Gottfredson, 2003, p. 31). Debate continues over whether group differences in average IQ scores are primarily caused by environmental (Nisbett et al., 2012) or genetic factors (Gottfredson, 2013; Rushton & Jensen, 2005). Recent advances in scientific understanding suggest that environmental factors may interact with, and regulate the expression of, genes (Bakermans-Kranenburg & van Ijzendoorn, 2015).

At present, significant social and political pressures remain for the practice of identification of gifted students to be more inclusive. In response to such concerns, and their litigious nature, identification of underrepresented populations has been one of the most active research areas in the field of gifted education (Dai et al., 2011). The most common recommendations that have been proposed to date have included the avoidance of cut-off scores and the use of multiple criteria identification approaches that rely on different data sources (e.g., self, parent, peer and teacher nominations, observation, test data; Richert, 2003; Robinson, Shore, & Enersen, 2007; Worrell & Erwin, 2011). The multiple criteria identification approaches, for which there is some empirical support (Foreman & Gubbins, 2015; Rosado, Pfeiffer, & Petscher, 2015), seek to include additional insights into the student's abilities that go beyond the intelligence test. One other empirically supported recommendation is dynamic testing, which is a procedure that employs a test-intervention-retest format (with an individualised educational intervention), and identifies as gifted those students who show test scores above a predetermined threshold in the post-intervention test (Chaffey, Bailey, & Vine, 2015; Chaffey, McCluskey, & Halliwell, 2005; Merrotsy, 2016).

Some researchers have raised concerns with recommendations to reduce the reliance on scores from intelligence tests and the use of identification practices that select a demographic mix of students, possibly as the top percentiles of the different student populations have been found to "differ noticeably in their ability to handle challenging instruction" (Gottfredson, 2003, p.31). One possible outcome of using such practices is that less challenging, less effective, and less appropriate programs are offered to students identified as being gifted. Therefore, it is argued that the "democratization" of gifted education will distort gifted programs away from meeting the educational needs of gifted students into educational supplement programs for all students, which falls far short of what is needed in gifted programs (Gottfredson, 2003). The tension between these two pressures—to provide equitable opportunities to all students, and the need to meet the needs of high ability students—continues to this day (Card & Giuliano, 2015; Gallagher, 2015; Peters & Engerrand, 2016; Warne, 2016).

The presence of multiple models of giftedness, rich diversity in intelligence testing instruments, and the pressure to be more inclusive, has resulted in a plethora of different practices in the identification of giftedness. Each school, city, and country appears to have a unique policy and practice. For example, they may differ on such variables as the instruments used to measure intelligence, cut-off values for determining giftedness, whether achievement measures are also used, and whether multiple critera are used, and if so, which criteria are used and how they are weighted and combined (McClain & Pfeiffer, 2012). Therefore, one student may find himself/herself identified as gifted at one school and eligible to participate in the school's gifted program, but not at another school where he/she may be excluded from such programs. This phenomenon is referred to as *geographic giftedness* (Borland, 1989), and further complicates and restricts the ability of researchers to compare results when seeking to examine gifted students.

2.3 Gifted Underachievement

2.3.1 A paradox?

A common myth appears to exist amongst teachers, parents, students, and some researchers that gifted students do not struggle academically (Bangel, Moon, & Capobianco, 2010; Baudson & Preckel, 2013; Gallagher, Smith, & Merrotsy, 2011; Geake & Gross, 2008), and instead perform consistently at high levels in school (Clark, 2002; Smith, 2010). As underachievement is often understood to be a type of failure in learning, the concept of gifted underachievement may be considered by many to be an oxymoron (Smith, 2010; Hoover-Schultz, 2005). However, researchers argue that giftedness, by itself, does not protect a student from the many possible negative impacts on their learning (Gagné, 2009a, 2013; Gross, 2010; Subotnik, Olszewski-Kubilius, & Worrell, 2011). Furthermore, gifted students have been found to have many unique educational, social, and emotional needs that are often overlooked and not adequately provided for in schools (Blaas, 2014; Elijah, 2011; Swan et al., 2015; Yilmaz, 2015). Despite the apparent conflict in terminology, the research demonstrates that many gifted students are struggling (Salmela & Määttä, 2015) and underachieving at serious levels (Bergner & Neubauer, 2011; Rimm, 2003).

2.3.2 Definition of gifted underachievement

Similar to giftedness, a plethora of definitions exist for underachievement. Indeed, Reis and McCoach (2000) concluded after their review of the literature that the "definitions of gifted underachievement are inconsistent and sometimes incompatible" (Reis & McCoach, 2000, p. 156). Nevertheless, they found that, in general, the definitions seem to fall into one of three categories:

(a) A discrepancy between a student's ability and achievement (Baum, Renzulli, & Hébert 1995; Emerick, 2004);

(b) A discrepancy between a student's predicted achievement and actual achievement (Lupart & Pyryt, 1996); and

(c) The failure of a student to self-actualise (Rimm, 1997).

To resolve the problem of multiple definitions of gifted underachievement, Reis and McCoach (2000) proposed "an imperfect, yet workable operational definition" which is inclusive and combines the first two definition categories: "a severe discrepancy between expected achievement (as measured by cognitive/intellectual ability assessments or standardized achievement test scores) and actual achievement (as measured by class grades and teacher evaluations)" (p. 157). While the third definition category may be conceptually relevant, and consistent with the first two, it appears practically ineffective. As all of the recent research on gifted underachievement (refer Appendix 1) appears to rely on definitions of gifted underachievement that were in agreement with the Reis and McCoach definition, their definition appears to be a useful one to follow.

Ziegler, Ziegler, and Stoeger (2012) have challenged the common practice (refer Appendix 1) of using IQ based instruments for measuring expected achievement, and have instead argued that a student's previous achievement (not necessarily as measured by standardised instruments) should be used to estimate the student's expected achievement. Their argument is supported by empirical evidence which demonstrates that a student's previous achievement is the best predictor of their current/future achievement (Forget-Dubois et al., 2007; Harlen, 2005; Hecht & Greenfield, 2001; Lohman, 2005; Meisels, Bickel, Nicholson, Xue, & Atkins-Burnett, 2001; Ziegler, 2008). Therefore, while the Reis and McCoach (2000) definition has unified the conceptualisation of gifted underachievement, reasonable arguments appear to exist that support the use of non-standardised measures to measure expected achievement.

2.3.3 Utility of identifying gifted underachievement

2.3.3.1 Invisible gifted underachievement

The underachievement of most students may be immediately obvious to teachers and parents due to their low level of achievement. In contrast, the underachievement of gifted students may often remain unnoticed as they may continue to perform at an average, or even high, level relative to their age peers and grade standards. The underachievement, and giftedness, of such students is most likely to remain *invisible* (Chaffey, Bailey, & Vine, 2003). Invisible underachievement occurs when a student is performing well below their potential, but are nevertheless perceived to be succeeding and progressing normally at school. Two possible explanations for this phenomenon include a situation where a student purposefully hides their giftedness for peer acceptance (Jung, McCormick, & Gross, 2012), and the situation where gifted students are achieving highly without any effort or new learning (e.g., when they are tested on content learned previously; Tze, Daniels, & Klassen, 2016). In both cases, the discrepancy between the student's measured levels of expected achievement and actual achievement may not be determined to be severe, despite a substantial level of underachievement.

2.4 Identification of Gifted Underachievement

Any operational definition for identifying gifted underachievement must specify the following four elements to establish how underachieving gifted students may be demarcated from their peers:

(a) How expected achievement may be measured;

(b) How actual achievement may be measured;

(c) How measurements of expected achievement and actual achievement may be compared to determine the size of the discrepancy between expected and actual achievement; and

(d) How large a discrepancy should be to be considered "severe" to deem that underachievement has occurred.

The Reis and McCoach (2000) operational definition of gifted underachievement has provided guidance relating to the first two of these four elements. Nevertheless, a review of the peer-reviewed literature from 1990 to 2016 in ERIC (i.e., using the keywords *gifted underachievement* or *gifted underachiever*), indicated that while uniformity has largely been reached among scholars on the first two elements, there is substantial inconsistency with respect to the last two elements. The application of methods to identify/measure gifted underachievement used in the 52 identified studies is summarised in Appendix 1.

2.4.1 Measurement of expected achievement

In outlining an operational definition of gifted underachievement, Reis and McCoach (2000) specify ideal measurements of expected achievement to be: (a) the results of intellectual ability tests and (b) prior administrations of standardised achievement tests. In the review of the literature (Appendix 1), most studies (63%) relied on only one of these two measurements for the identification of gifted underachievement, while 77% of the studies utilised at least one of these measurements. The remaining studies used measures including previous school achievement (e.g., Emerick, 2004; Hébert, 2001), enrolment in gifted programs (typically with no information on the selection process for the gifted programs; e.g., Diaz, 1988; Hébert & Olenchak, 2000), and observations of students (e.g., Cavilla, 2015; Rafidi, 2008).

Despite the apparent consistency in the types of measurements that have been used, it is noted that intellectual ability tests and standardised achievement tests may vary significantly in the content they examine and the method of their assessment. For example, they may vary in terms of the broad ability factors that are assessed, whether time of completion is considered in scoring, whether they are completed on a one-on-one basis or in a group environment, and whether they are administered by a trained psychologist or a classroom teacher. These differences alone may produce very large discrepancies in expected achievement scores, with one study finding that two commonly used standardised intelligence tests produced mean differences greater than one standard deviation for participants (Silverman et al., 2010). Therefore, although the type of measurements used to assess expected achievement may be consistent, the diverse range of instruments that exist within each measurement type may present a significant challenge to the consistency of the identification and measurement of gifted underachievement.

2.4.2 Measurement of actual achievement

Reis and McCoach's (2000) operational definition of gifted underachievement also specified permissible measurements of actual achievement to include current class grades and teacher evaluations. As a review of the literature (Appendix 1) indicated that 73% of studies relied only on a student's school grades to determine their actual achievement, the practice appears to be widely adopted among researchers. Nevertheless, other scholars have been found to use the results of standardised achievement tests, tests of general knowledge (e.g., Figg, Rogers, McCormick, & Low 2012), and the educational pathway of students (e.g., Timmermans, van Lier, & Koot, 2009). Similar to measures of expected achievement, the diverse range of instruments that exist to measure actual achievement may present a significant challenge to the consistency of the identification and measurement of gifted underachievement.

2.4.3 Comparison of expected achievement and actual achievement

Reis and McCoach's (2000) definition did not stipulate a preferred method for comparing the measured expected and actual achievement to establish gifted underachievement. Perhaps unsurprisingly, this area remains the largest area of inconsistency among researchers in the identification of gifted underachievement. From the review of the literature (Appendix 1), it was established that four types of methods to identify underachieving gifted students are commonly used. In order of popularity, they were as follows:

(a) Absolute split (ABS) method (57%);

(b) Nomination (NOM) method (31%);

(c) Simple difference (SD) method (14%); and

(d) Regression (REG) method (10%).

Eight studies used multiple methods (either in combination, or in comparison) to identify gifted underachievement. Two further studies used a unique, alternative approach (i.e., Rasch models) in the identification of underachievement among the general student population, which may possibly be adapted to assess underachievement among gifted students (Phillipson & Tse, 2007; Van Nijlen & Janssen, 2015).

The various methods of identification of underachievement may be classified according to whether they rely on statistical techniques (that compare quantitative measurements of ability and achievement), or qualitative judgements by individuals. The methods that utilise statistical techniques are the two variations of the absolute split method, the simple difference method, and the regression method. Two of these methods, simple difference and regression methods, may also be classified as methods of *measurement* of underachievement, as they (unlike the other methods) calculate a measurement of the degree

of underachievement for each student. Figure 8 shows the different classifications of methods to identify gifted underachievement. The following sections describe each of these methods.

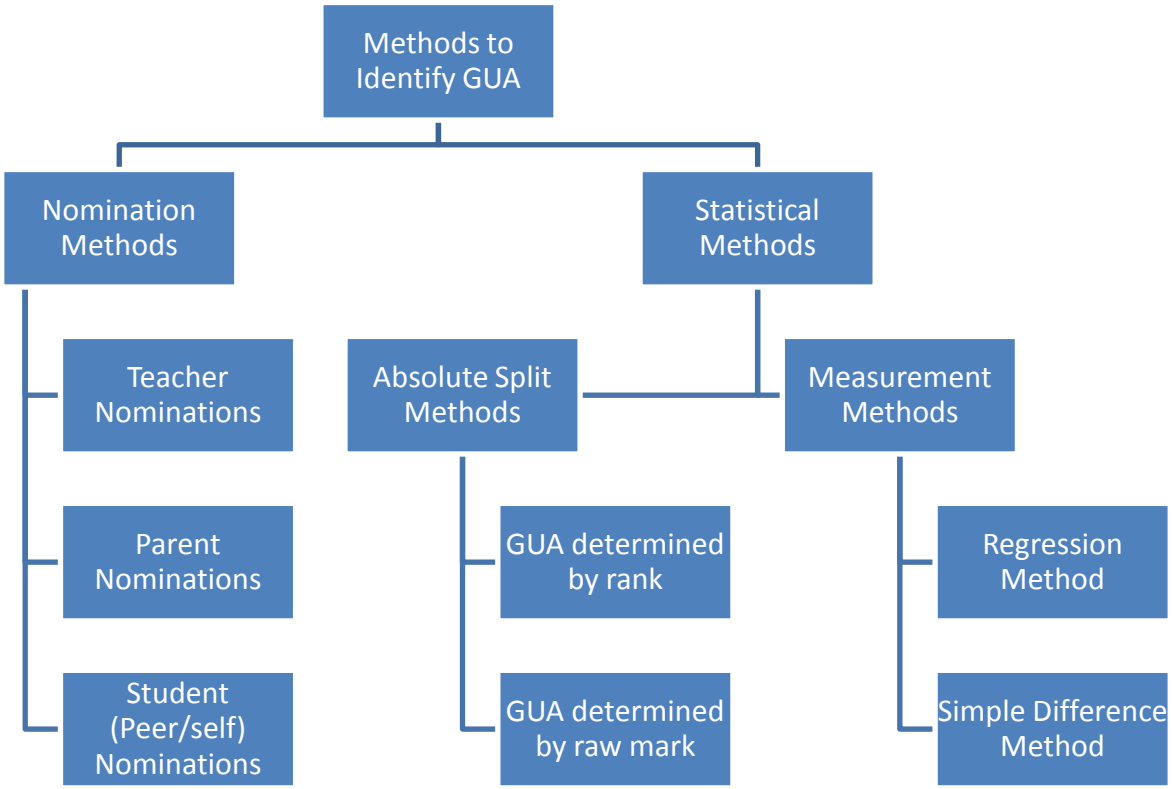


Figure 8. Classification of common methods to identify gifted underachievement (GUA).

2.4.3.1 Absolute split method

The absolute split method was found to be the most commonly used statistical technique to identify gifted underachievement (e.g., Vlahovic-Stetic, Vidovic, & Arambasic, 1999). This method requires a quantitative measure of a student's expected achievement (i.e., ability or past achievement) and their actual achievement (i.e., current achievement). Gifted underachievement is deemed when a student's expected achievement falls into a high category (e.g., top 5%) and his/her actual achievement is below some arbitrarily set threshold (e.g., bottom 50%). In the research to date, the condition of high expected achievement has

usually been determined on the basis of a student's IQ score, although occasionally a student was considered gifted on the basis other measures (Peterson & Colangelo, 1996; Peterson, 2000). The condition of low actual achievement has usually been determined by performance below a threshold achievement score (i.e., one variation of the absolute split method; Baslanti & McCoach, 2006) or a threshold rank (i.e., a second variation of the absolute split method; Hanses & Rost, 1998). It is noted that none of the other methods used to identify gifted underachievement use ranked scores. Occasionally, other indicators of low actual achievement have also been used, such as the experience of academic probation (Baker, Bridger, & Evans, 1998), the low level of difficulty of enrolled courses (Hébert, 2001; Timmermans et al., 2009), or dropping out of school (Reis et al., 2005; Sikora & Saha, 2011). Figure 9 shows the students who would be selected by an application of the absolute split method with the threshold for expected achievement chosen to be the top 10% of ability (in accordance with Gagné's model) and the threshold for actual achievement chosen to be below the first quartile (i.e. below the 75^{th} percentile). It is noted that unlike the simple difference and regression methods, the two variations of the absolute split method do not produce a measurement of the degree of underachievement.

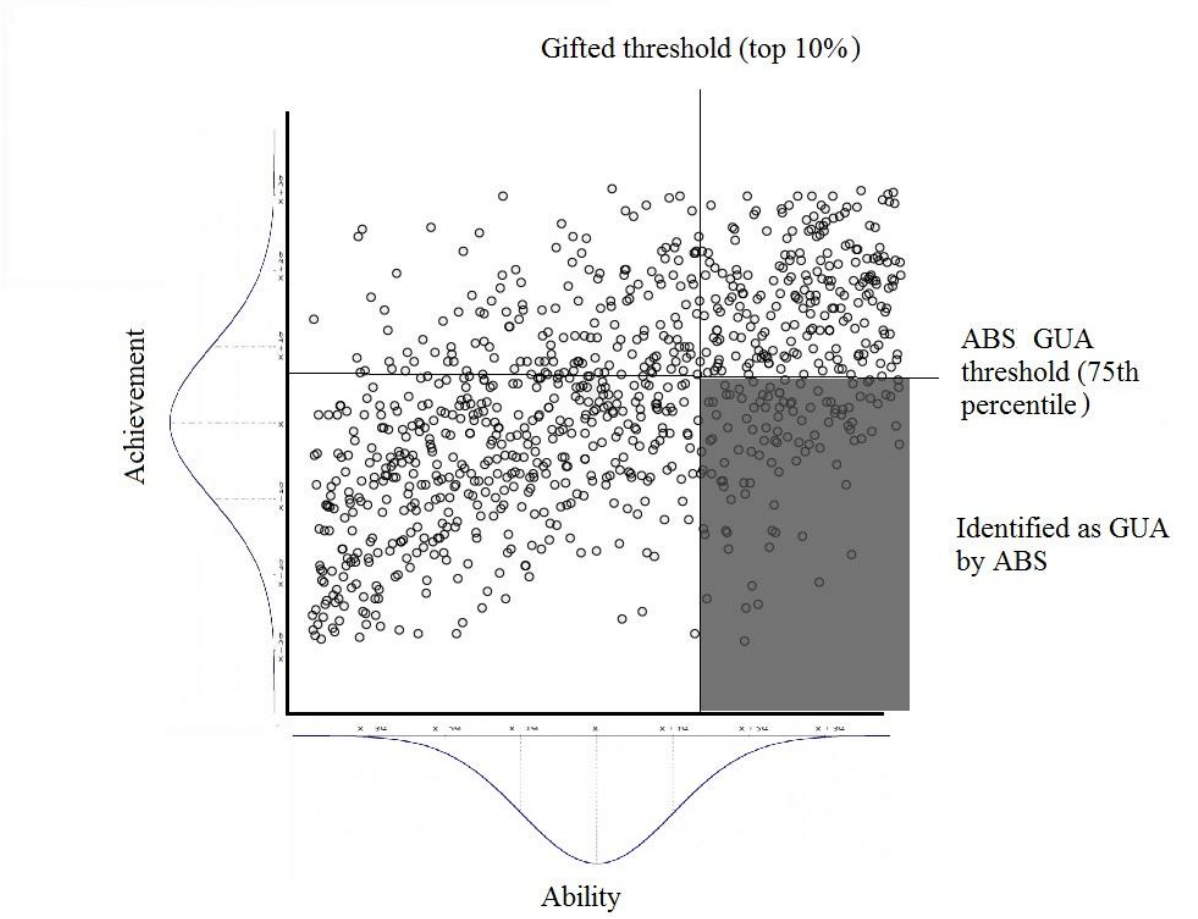


Figure 9. Absolute Split (ABS) method of identifying gifted underachievement (GUA)

2.4.3.2 Nomination method

Nomination is another commonly used method to identify underachieving gifted students. Most often, this method is used to identify participants for intervention programs and counselling services (Davis & Rimm 1998; Rimm, 1991), or for qualitative research case studies (e.g., Cavilla, 2015). The nomination method is not a statistical method, nor does it provide a measure of gifted underachievement. Although there is significant variation in how the nomination method is used, a typical implementation procedure would involve a judgement being made about whether a student is underachieving on the basis of their impressions of the student. Nominations may be made by different people, including teachers, parents, peers, and the students themselves (i.e., self-nominations). Nevertheless,

teacher nominations appear to be used almost exclusively by researchers (for some exceptions, see Lau and Chan, 2001a; Flint, 2002).

In the research to date, the processes that teachers have followed to make their selections have varied substantially. Nevertheless, teachers appear to commonly use the following information in completing nominations: student's participation in the classroom, attendance, behaviour, student records, profiles of gifted underachievers, degree of assistance provided in class, student background, and home environment (Fisher, 2005; Pipkin, Winters, & James, 2007; Rafidi, 2008). On occasion, researchers provided teachers with guidance for selection by providing an identification checklist, or trained teachers to understand and recognise underachievement (Donnelly, 2010). It is noteworthy that Lau and Chan (2001a) required teachers to provide reasons for their selections, which they later screened.

2.4.3.3 Regression method

The regression method is a complex statistical method for the identification and measurement of gifted underachievement. This method requires quantitative measurements of expected achievement and actual achievement for a group of students (e.g., all students of a certain year group at a school). After the data for all students in the group are plotted on a graph (e.g., expected achievement on the *x* axis and actual achievement on the *y* axis), a straight line that best fits the data is plotted. This reference line, sometimes called the *line of best fit*, but more technically known as the *regression line*, describes the average relationship between expected achievement and actual achievement for the assessed group. A particular student's position on the graph relative to the regression line is used to make determinations about gifted underachievement. It is noted that the regression method is simultaneously considered to be an identification method and a measurement method, as a measurement of the degree of gifted underachievement is possible by examining how far a student performs below the regression line (refer Appendix 12).

Among those students who qualify as gifted (i.e., students whose expected achievement is greater than a threshold such as the 90th percentile), those who are considered to be underachieving are those who fall at least one standard error below the regression line (Bouffard, Roy, & Vezeay, 2005; Gorard & Smith, 2004; McCall et al., 2000; Preckel et al., 2006), or occasionally, two standard errors below the regression line (Peters & van Boxtel, 1999; Plewis, 1991). Standard error is a measure of the accuracy of the regression line as a representation of the plotted data, and assesses the deviation of observed measurements from this line. Figure 10 shows the students who would be selected as exhibiting gifted underachievement by an application of the regression method with a threshold of one standard error.

The regression method is a comparative method of identifying and measuring gifted underachievement that relies on the *average* relationship between expected and actual achievement of students in the assessed group. Gifted underachievement is considered to be exhibited by those students who: (a) qualify as being gifted due to their level of expected achievement, *and* (b) have a low level of actual achievement relative to other students in the group with the same level of expected achievement. Some researchers (e.g., McCall, Beach, & Lau, 2000) argue strongly for the exclusive use of the regression method as: (a) it may be easily adapted and used with any combination of expected and actual achievement measurements, (b) it is able to identify underachieving students of all levels of ability, and (c) it consistently identifies the same proportion of underachieving students (regardless of the data combination, or ability range of the students). As a result, the regression method is much more popular in studies of underachieving students in the general student population (i.e., not restricted to gifted students; e.g., Bouffard, Roy, & Vezeau, 2005; Preckel, Holling, & Vock, 2006). However, other researchers have argued that these properties are precisely why the method is invalid (Van den Broeck, 2002a, 2002b).

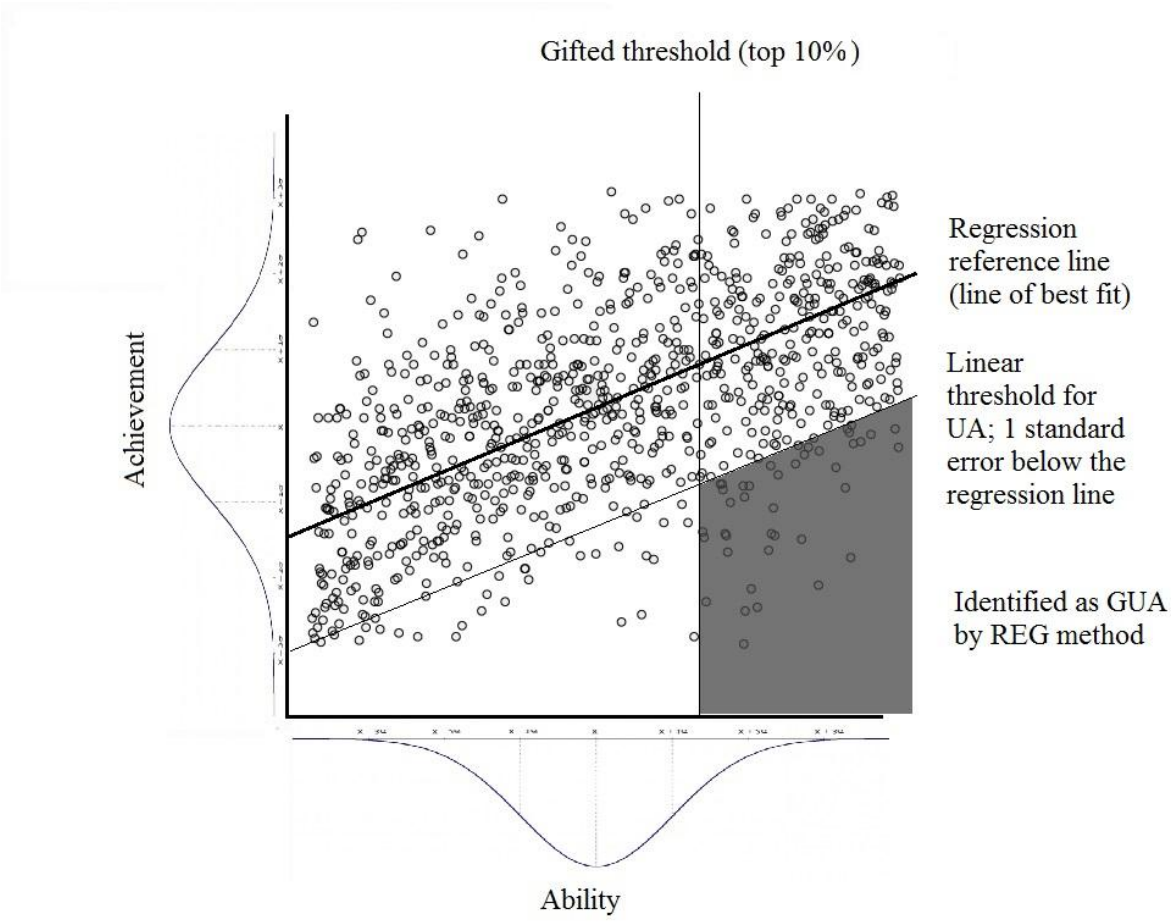


Figure 10. Regression (REG) method of identifying gifted underachievement (GUA).

2.4.3.4 Simple difference method

The simple difference method is a statistical method for the identification and measurement of gifted underachievement (Lau & Chan, 2001a). Similar to the regression method, the simple difference method requires a quantitative measurement of a student's expected achievement and actual achievement. Nevertheless, unlike in the regression method, both measurements are converted into standardised units using the standard deviation for each measurement instrument used. Thereafter, the degree of underachievement for each student is calculated by subtracting each student's standardised actual achievement score

from the student's standardised expected achievement score (refer Appendix 12). Gifted underachievement is deemed when a student's standardised expected achievement score is at least one standard deviation greater than his/her standardised actual achievement score (Carr, Borkowski, & Maxwell, 1991; Nurmi, Onatsu, & Haavisto, 1995; Tuss et al., 1995; Ziegler & Stoeger, 2010). Variations in the choice of threshold value from one standard deviation appear to be rare, and are typically small (e.g. 1.1 standard deviations, Bergner & Beubauer, 2011), in the literature.

Graphically, the simple difference method (refer Figure 11) appears similar to the regression method. However, there are two key differences. First, the reference line used is not the line of best fit, but rather a diagonal line at 45º to the axes (i.e., the line where a student's expected achievement is equal to his/her actual achievement in standardised units). Second, the units of measurement are different for the two methods. In the simple difference method, underachievement is measured using standard deviation units, whereas in the regression method, underachievement is measured using standard errors.

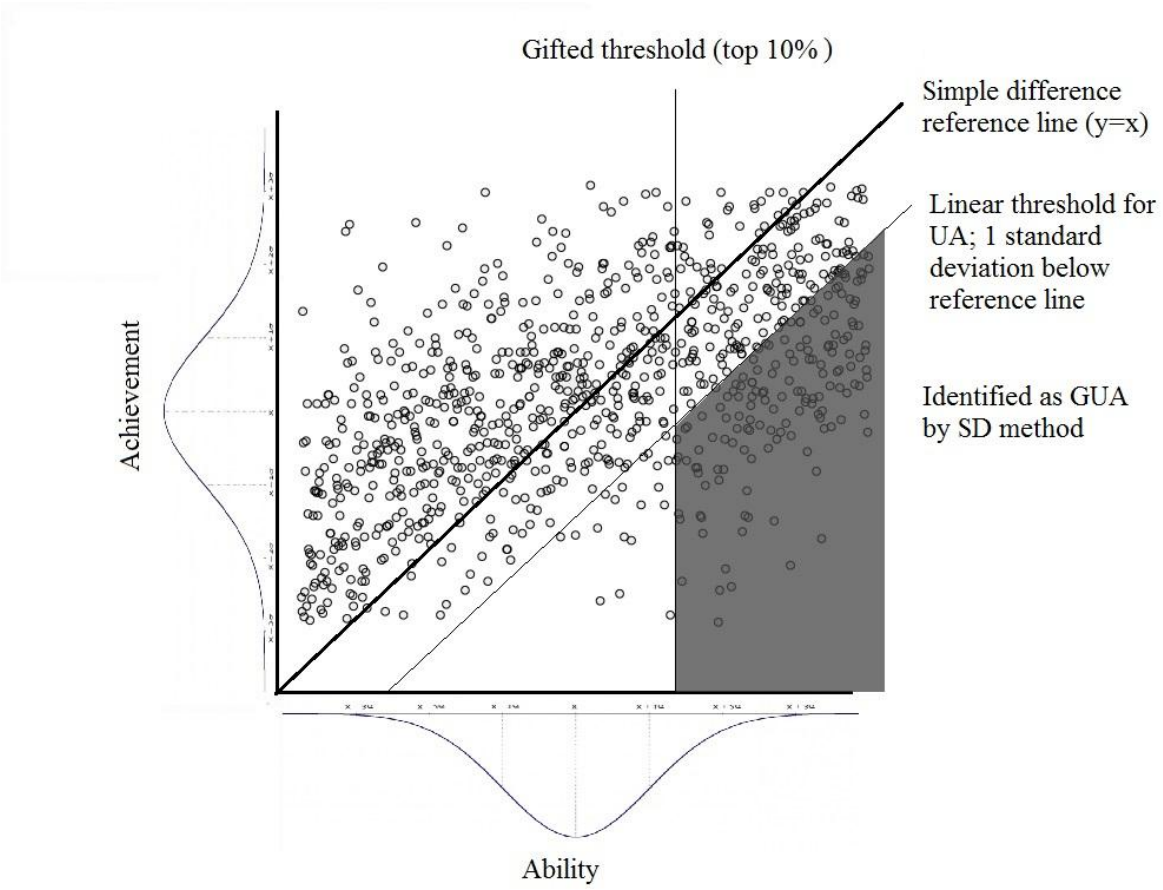


Figure 11. Simple difference (SD) method of identifying gifted underachievement (GUA).

2.4.4 "Severe" discrepancy

The Reis and McCoach (2000) definition of gifted underachievement did not prescribe a recommended threshold for determining what would constitute a "severe" discrepancy between expected and actual achievement to deem gifted underachievement. Nevertheless, all of the statistical methods for identifying gifted underachievement (i.e., absolute split, regression, and simple difference methods) rely on an arbitrarily selected threshold value to determine when a discrepancy between expected achievement and actual achievement becomes severe. In the majority (73%) of the identified studies using either the simple difference or regression methods, one standard deviation or one standard error of estimate was selected as the threshold value. Only a very few studies utilised larger

thresholds that were closer to two standard deviations or errors (e.g., Peters & van Boxtel, 1999).

The selection of the appropriate threshold has been described by Ziegler et al. (2012) as effecting a balance between the types of false classifications. A low threshold may increase the rate of incorrect classifications of gifted underachievement (type I error), while a high threshold may increase the rate of incorrect classifications of gifted achievement (type II error), which increases the number of cases of invisible underachievement. It may be the case that the use of an overly high threshold may greatly reduce the usefulness of the identification methods for gifted underachievement, as part of their purpose would be to overcome the problem of invisible underachievement.

It is noteworthy that the process for diagnosing learning disabilities appears similar to the process for identifying gifted underachievement. Specifically, a commonly used method for the identification of students with learning disabilities uses an IQ-achievement discrepancy criterion (Reschly & Hosp, 2004), which fits the definition of gifted underachievement proposed by Reis and McCoach (2000). Under this criterion, the standard psychological diagnostic manual (DSM-IV) requires "a discrepancy more than 2 standard deviations between achievement and IQ" (American Psychiatric Association, 2000, pp.49–50) to diagnose a learning disability, and that lower discrepancies of between one and two may be sufficient in some cases. While the new psychological diagnostic manual (DSM-V; American Psychiatric Association, 2013) has adopted broader diagnostic criteria, the many recent updates to the manual have been heavily criticised (British Psychological Society, 2011; Demazeux & Singy, 2015; Robbins, 2011) and the discrepancy approach still appears to be the most commonly used approach (Maki et al., 2015).

While it may be reasonable to conclude that all gifted students with a learning disability may also be identified as exhibiting underachievement, it is not reasonable to expect that all gifted students who underachieve have a learning disability. Indeed, Reis and McCoach's (2000) definition of gifted underachievement claims that "to be classified as a (gifted) underachiever, the discrepancy between expected and actual achievement must not be the direct result of a diagnosed learning disability" (p. 157). This would suggest that the larger threshold value of two standard deviations, which is used to diagnose learning disabilities, may be too extreme to be used for the identification of gifted underachievement. If the threshold of two standard deviations was adopted, the operational definitions of underachievement and learning disability may be so similar that they could be considered identical phenomenon. The lower threshold of one standard deviation therefore appears to be more appropriate.

In contrast to the simple difference and regression methods, the thresholds used with the absolute split method appear to have much more variation. Most notably, two clear variants of thresholds used with the absolute split method appear to exist in the literature: (a) a threshold of student rank, and (b) a threshold of student raw score. In addition to these different types of thresholds, the threshold values also appear to vary considerably (e.g. a GPA below 2.0 in Baslanti, 2008; a GPA below 3.49 in Matthews & McBee, 2007; below the 85th percentile [i.e., rank] in Figg et al., 2012; below the average in Guldemond, Bosker, Kuyper, & van der Werf, 2007). Reflecting these differences in the thresholds used, two variants of the absolute split method (i.e., absolute split I and absolute split II) will be examined in this investigation as two separate identification methods. One of these will use a threshold of student rank (i.e., below the 75^{th} percentile), in recognition of the "extremeness" of using the most commonly used student rank threshold of below the 50^{th} percentile (i.e., the 50^{th} percentile may be as large as two standard deviations below the level gifted students are

expected to achieve; Uttl, 2005), while the other will utilise the most commonly used student raw score threshold of 80% (Appendix 1).

2.4.5 Impact of variations in methods for identifying gifted underachievement

A number of problems have arisen from the existence of multiple methods to identify and measure gifted underachievement. One significant impact may be the wide range of prevalence estimates of gifted underachievement. This problem was noted as early as 1964 when Farquhar and Payne claimed that "there is an extreme range in the absolute number of individuals identified as under- and over-achievers, depending upon the particular technique used" (pp. 882–883). The problem continues to this day. At one extreme, Hsieh et al. (2007) found that as many as 54% of gifted students were underachieving, while at the other, Colangelo, Kerr, Christensen, and Maxey (1993) suggested that less than 1% were underachieving. Other researchers have found the prevalence of gifted underachievement to fall somewhere in between (e.g., 9% by Matthews & McBee, 2007; 16% by Guldemond, Bosker, Kuyper, & van der Werf, 2007). This may lead to confusion as various stakeholders are unable to assess the precise prevalence, and therefore the seriousness, of the problem. One possible consequence is that the problem may be dismissed.

Relatedly, the research into interventions to address the needs of gifted students who underachieve is hampered by contradictory results when trials are repeated using different identification methods to select participants. Unfortunately, researchers have still not reached a consensus on the optimal interventions to assist gifted students who are underachieving (Ryan & Coneybeare, 2013). It is not clear if this reflects a problem with the interventions that have been studied, or the significant variability in how students were identified as underachieving. Indeed, every review on the topic of gifted underachievement has denounced the use of different methods to identify and measure gifted underachievement because it

limits the ability of research findings to be compared (Dowdell & Colangelo, 1982; Reis & McCoach, 2000).

Of note to this investigation, two prior studies have attempted to empirically compare the methods for identification of gifted underachievement. Annesley, Odhner, Madoff, and Chansky (1970) investigated four identification methods: (a) simple difference with a one standard deviation threshold, (b) simple difference with a one standard error of measurement threshold, (c) regression with a one standard error of estimate threshold, and (d) teacher nomination. They used all four methods with the same group of 157 first grade students and concluded that the four methods do not identify the same group of gifted students. More recently, Lau and Chan (2001a) investigated four different identification methods: (a) simple difference with a one standard deviation threshold, (b) regression with a one standard error of estimate threshold, (c) absolute split using a rank-based threshold of below 25th percentile (bottom quartile), and (d) a combination of teacher and peer nominations (as a single nomination method). They applied all four methods to 126 Grade 7 students, and reached the conclusion that the three statistical methods were in agreement, while the nomination methods identified different students to the statistical methods. A possible reason for the disagreement between the two studies may lie in the limitations of these studies (e.g., relatively small sample sizes, participants of a single grade, a single combination of ability and achievement measurements, and the non-use of statistical methods that measure and assess agreement among the methods).

It is unfortunate that, despite the efforts of researchers, the concern from six decades ago – "it is obvious that a dire need exists to adopt standard definitions of the procedures for identifying discrepant achievers" (Farquhar & Payne, 1964, p. 883) – remains unresolved.

2.5 Semantics

In addition to the confusion arising from the inconsistent use of identification methods, there is also considerable confusion arising from the use of the term underachievement with different technical or non-technical meanings. As a result of the ambiguous use of the term, Smith (2010) has suggested that it "has probably outlived its usefulness" (p. 46), while Plewis (1991) has proposed that it be abandoned altogether. Nevertheless, both scholars simultaneously acknowledge that there remains a crucial concept underlying underachievement that is important to educational research. Faced with this problem, both Smith (2010) and Plewis (1991) recommend a general conception of underachievement as an individual phenomenon, that occurs when a student's performance is not commensurate with their level of ability (which is consistent with Reis and McCoach's [2000] operational definition of gifted underachievement). This section attempts to elucidate some of the common uses of the term underachievement that differ from the recommended definition.

2.5.1 Relative position, achievement gap, or underachievement?

One of the most common uses of the term underachievement is to indicate the relative achievement of two groups of students. In this use, one group is said to underachieve if the mean achievement within that group is below the mean achievement from at least one other group. The most common bases of grouping include ethnicity, gender, age, country, or social class. Hence, a researcher may claim that boys are underachieving compared to girls (as is commonly reported in reading and writing subjects; e.g. Bush, 2005), or that one country is underachieving compared to others (e.g., on the basis of the Trends in Mathematics and Science Study [TIMSS] results). Plewis (1991) refers to such comparisons as the *relative position* of one group with respect to another, while Smith (2010) uses the expression *achievement gaps* between different groups. It is noted that even though achievement gaps

exist between groups and are often well-documented, the individual students within any "underachieving" group may each still be achieving to their full potential depending on their level of ability (Gottfredson, 2000; Smith, 2010; Swann, 1985).

2.5.2 Low achievement and underachievement

The term underachievement is also sometimes used synonymously with the term *low achievement* (e.g., Jarjoura, Tayeh, & Zgheib, 2015). Faced with this problem, Smith (2007), and separately Landis and Reschly (2013), proposed a differential definition for low achievement, as achievement well below the average for a broad sample of age peers. Smith (2010) has suggested that students identified as low achievers under this definition may be a homogenous group with large numbers of students from low socioeconomic status backgrounds. It is noted that it will be possible for gifted students who achieve above the average of their age peers (i.e., not exhibiting low achievement), to nevertheless be underachieving as they may not be achieving to their level of expected achievement.

2.5.3 Underachievement, underachievers, or underachieving?

The terms underachiever and underachievement are often used interchangeably by researchers. Underachievement refers to the actual achievement of a student being significantly below their level of expected achievement. Hence, a single case of poor actual achievement may be recognised as underachievement. In contrast, for a student to be labelled as an underachiever, Reis and McCoach (2000) suggest that the underachievement "must persist over an extended period of time" (p. 157). Nevertheless, in practice the term underachiever appears to be used quite readily, and without evidence of persisting underachievement (refer Appendix 1). Such an application of the label underachiever is perhaps inappropriate.

In addition, when making reference to the phenomenon of a severe discrepancy between expected and actual achievement, it may be ideal to follow the guidelines of the American Psychological Association (APA) relating to disability. The APA mandates that in scientific and professional communication, authors use person-first language and therefore refer to "people with a disability", rather than "disabled people" (APA, 2010). This language highlights that each individual's experience of the disability is unique and that the disability does not become a defining feature of their identity (Dunn & Andrews, 2015). Consequently, rather than labelling an individual as an underachiever, it may be preferable for researchers to instead refer to an individual who is (or was) underachieving, or simply to the underachievement itself.

2.6 Concerns with the Gifted Underachievement Construct

It is noted that the construct of gifted underachievement has been plagued with the longstanding problem of inconsistent results from investigations into the issue, which has kept gifted underachievement as a seemingly perpetual enigma (Dowdall & Colangelo, 1982; Farquhar & Payne, 1964; Reis & McCoach, 2000; Obergriesser & Stoeger, 2015). Some researchers have therefore expressed doubts about the usefulness and the validity of the gifted underachievement construct, going so far as to suggest that it may perhaps be an ambiguous and arbitrary construct (Smith, 2010; Ziegler, Ziegler, & Stoeger, 2012). Despite the criticisms, many researchers, including some critics (Smith, 2010; Ziegler, Ziegler, & Stoeger, 2012), defend the necessity and importance of the construct.

2.6.1 IQ-based measures

Ziegler, Ziegler and Stoeger (2012) provide a significant critique of the Reis and McCoach (2000) version of the gifted underachievement construct which promotes the use of IQ-based measures of expected achievement. Specifically, Ziegler et al. (2012) argue that if

the gifted underachievement construct is assessed using IQ-based measures, the construct may have limited validity. They provide three criticisms from three different perspectives: theoretical, methodological and empirical. The theoretical criticism is that the reliance on IQ measures is fallacious as any identified gifted underachievement demonstrates not the "failure" of the student to achieve, but rather a failure of the IQ-based model of giftedness used to predict the student's achievement. Their empirical criticism, which appears to be the basis of the theoretical criticism, presents evidence that a variety of research has established IQ as a poor predictor of *adult* achievement. However, the gifted underachievement construct finds its utility primarily in the identification of *children* who may benefit from interventions rather than adults, and many scholars strongly argue that IQ is the best predictor of a student's future achievement (Assouline, 2003; Neubauer & Opriessnig, 2014).

The third criticism from Ziegler, Ziegler, and Stoeger (2012) is methodological in nature. This criticism relates to the problem that when two instruments are combined, their errors in measurement are also combined, resulting in an unsatisfactorily high level of error in the identification of gifted underachievement. They present significant statistical arguments which demonstrate that in a situation where no underachievement actually exists, approximately 10% of all students would still be identified as underachieving given typical estimates of reliability for IQ instruments (.85) and school marks (.55). Nevertheless, if the instruments used are restricted to those that are the most accurate among those available, the false identification rate is reduced significantly.

To resolve these three issues Ziegler et al. (2012) suggested that IQ instruments should be excluded from the assessment of the underachievement construct, in favour of a student's previous achievement as the measure of expected achievement. Unfortunately, such a recommendation may be problematic. From a theoretical perspective, the underachievement construct may be considered to find its utility in identifying those students whose needs are

not being met and may therefore benefit from school-based interventions. However, if the construct is constrained only to past student achievement, students who exhibit chronic underachievement may never be identified or supported. Furthermore, the use of IQ-based measures to determine a student's level of expected achievement is more likely to detect otherwise invisible gifted students, who may be among the most disadvantaged and vulnerable students in schools (Blaas, 2014; Funk-Werblo, 2003; Silverman, 1997).

2.6.2 Adaptability to different models of giftedness

Some scholars have expressed concerns relating to the adaptability of the gifted underachievement construct to the diverse conceptions of giftedness that exist (Ziegler, 2008; Ziegler, Ziegler, & Stoeger, 2012). In response, an alternative, delphic, conceptualisation of gifted underachievement has been proposed, which suggests that "underachievers are talented persons whose current achievement is below experts' expectations. Without intervention, this will result in unfavourable prognoses for the achievement of excellence" (Ziegler, Ziegler, & Stoeger, 2012, p.124). Unfortunately, the increasing adaptability of the definition may lead to greater inconsistency. For example, in comparison with the conceptualisation of Reis and McCoach (2000), this conceptualisation removes direction on suitable instruments for identification. Moreover, it makes reference to "underachievers", which may be inappropriate for reasons of ethics and empirical validity in this thesis.

2.6.3 Arbitrary threshold values

The selection of a threshold to distinguish between what may be deemed as underachievement and what may be considered a typical fluctuation in achievement will necessarily be arbitrary. While it does appear to be common consensus that the threshold should be one standard deviation (cf. Appendix 1; Reis & McCoach, 2000; Ziegler, Ziegler, & Stoeger, 2012) there has been no theoretical or empirical argument to establish that this threshold is appropriate. Furthermore, this common consensus only appears to exist for the

simple difference and regression methods of identification. In contrast, the absolute split methods have a much greater degree of diversity in thresholds utilised. Moreover, the threshold of one standard *deviation* for the simple difference method is not equivalent to the threshold of one standard *error* for the regression method. One possible solution may arise if instead of utilising these thresholds researchers and practitioners refer to the *degree* of underachievement.

2.6.4 Underachieving environments

Some scholars have proposed a paradigm shift in the field of gifted education whereby giftedness is not considered a construct of an individual student, but rather a construct of the environment that students are in (Dai & Renzulli, 2008; Harder et al., 2014; Ziegler, 2005; Ziegler & Phillipson 2012). Relatedly, Funk-Werblo (2003) have proposed that gifted underachievement should be assessed at the level of the school rather than the individual student. Nevertheless, such a proposal may be at odds with the primary utility of the gifted underachievement construct, which is to identify students whose fundamental needs are not being met so that appropriate interventions may be provided. For example, some may argue that it is unreasonable to expect no students attending any single school to be exhibiting gifted underachievement, regardless of the high performance of the school as a whole, as many factors beyond the school may influence an individual student's achievement. Nevertheless, the *additional* use of the gifted underachievement construct to assess the effectiveness of a school (or even individual teachers) may have potential, as the current practice of measuring and therefore ranking schools on the basis of their student's achievement may be subject to manipulation (e.g., the exclusive enrolment of high achievers). It may indeed be the case that many schools appear to provide excellent environments due to the high level of results that are produced, when in fact many students are exhibiting underachievement and the school is having a negative impact.

2.7 Summary

In this chapter, the relevant literature on giftedness, gifted underachievement, and the identification/measurement of gifted underachievement was discussed. There appear to be several different methods that are commonly used to identify and measure gifted underachievement, even though the practice of using these methods appears to be causing considerable confusion and conflict among researchers and practitioners. Furthermore, no firm basis or guidance appears to exist on how to make an optimal choice among these methods. The next chapter describes the theoretical framework that was used to guide this investigation.

3 Theoretical Framework

3.1 Introduction

The purpose of this chapter is to outline the theoretical framework that guided the investigation. This theoretical framework is expected to allow for the collection of the most important and useful evidence to contribute to the assessment of the methods that identify and measure the underachievement of gifted students.

3.2 Validity Theory and a Validation Framework

Validity is the "most fundamental consideration" (AERA, APA & NCME, 1999, p.9) when assessing the use of measurement and identification instruments, and is therefore characterised as "the reigning deity" (Cizek, Rosenberg & Koons, 2008, p. 397) of psychometrics. Nevertheless, modern validation appears to be one of the most misunderstood and misused concepts, and therefore a clear guide for how to carry out validation is needed (Frisbie, 2005). This section attempts to outline the basic concepts of validity, the changes that modern validity theory has brought, and a guiding framework for validation for this investigation.

3.2.1 Validity theory

Often validity is used to refer to whether an instrument measures what it claims to measure (Field, 2013). In contrast to this traditional view, modern validity theory suggests that it is not the instrument itself that is validated, but rather the degree to which actions and interpretations are supported by empirical evidence and theoretical rationale (AERA, APA & NCME, 1999; Borsboom, 2012; Cizek et al., 2008; Cizek, Bowen & Church, 2010; Cronbach, 1971; Fraenkel & Wallen, 2006; Kane, 2006; 2013; Messick, 1989; 1995; Wolming & Wikström, 2010). An illustrative example is that under modern validity theory, rather than making an assessment of whether a ruler is a valid instrument, the focus should be

on whether the measurements from a ruler are being used appropriately (e.g., the use of height measurements to restrict shorter children from going on certain rides).

Traditionally, three types of validity have been assessed to validate an instrument: criterion, content and construct validity (Fraenkel & Wallen, 2006). *Criterion validity* examines the relationship between scores obtained with the instrument being examined and scores obtained from another instrument which is known to be valid, while *content validity* examines the content and format of the instrument to determine whether it is logical, comprehensible, adequate and appropriate. *Construct validity* examines the validity of the psychological constructs (i.e., a hypothesised variable which is not directly observable but explains some aspect of human behaviour, such as intelligence) being measured by the instrument. The validity of the construct (and thus of its measurement) relies on how clearly the construct is defined, whether testable hypotheses are able to be made based on the construct, and whether empirical evidence supports the testable hypotheses (Fraenkel & Wallen, 2006). Kane (2006, 2013) suggests that while the assessment of these three types of validity are important in modern validity theory, they are not sufficient in themselves to justify the validity of interpretations made from the measurements.

Three major changes from the traditional approach have been introduced in modern validity theory. First, and as noted previously, there has been a shift away from validating the instrument itself and onto validating the interpretation and use of data from the instrument. Second, with the focus on the appropriateness of the interpretation and use of the data, criterion, content, and construct validity are now considered to be different sources of evidence of validity. Third, there is now an increased focus on the context and consequences of the uses of a measurement as another source of evidence of validity. For example, the appropriateness of the decisions made on the basis of a measurement to achieve positive consequences or avoid negative consequences may be considered evidence of validity.

However, it is noted that very few researchers attempt to include consequential evidence to support validity, leading some to conclude that such evidence is too difficult or impossible to locate, or that such evidence does not exist (Cizek, Bowen & Church, 2010; Fraenkel & Wallen, 2006; Reckase, 1998).

3.2.2 Kane's validation framework

Due to the complex nature of modern validity theory, a guiding framework is helpful to ensure that evidence is collected and used appropriately. Kane (2006, 2013) provides such a guiding framework, which appears to be well received by scholars (Brennan, 2013; Davies, 2012; Moss, 2013; Newton, 2013; Sireci, 2013) and utilised across a variety of disciplines (Aryadoust, 2011; Bell et al., 2012; Hill et al., 2012; Kumazawa, 2013; McGaghie, Cohen, & Wayne, 2011;Wang, Choi, Schmidgall & Bachman, 2012). It is noteworthy that Chapelle, Enright, and Jamieson (2010) consider Kane's approach to be "a clear improvement" (p.12) to existing practices when it was evaluated in comparison to the standards outlined in AERA, APA, and NCME (1999). Indeed, a consensus appears to exist that Kane provides effective and simple guidelines for validation studies (Chapelle, 2012).

Kane's (2006, 2013) framework requires researchers to use empirical evidence and sound reasoning to make arguments for the validity of their interpretations and uses of instruments. To guide the researcher, Kane suggests that, first, an outline of the formal logical steps involved in the interpretation and use of instruments be prepared. Thereafter, the researcher is required to determine which of the formal logical steps are the most contestable, to focus the collection of evidence to determine whether these contestable steps may be empirically demonstrated. Kane adopts a specific set of terms to refer to each part of this validation process:

(a) the outline of formal logical steps involved in the interpretation and use of instruments is called the *interpretation/use argument*;

(b) each logical step is referred to as an *inference;* and

(c) the concluding discussion of evidence supporting or opposing the validity of the interpretation/use of instruments is called the *validity argument*.

3.2.3 The interpretation/use argument

The interpretation/use argument is an outline of each of the formal logical steps (i.e., the inferences) that are made between the observation of performance to the eventual interpretation/use of instrument data. Each inference will rely on assumptions, which are also stated as part of the interpretation/use argument. Kane (2006, 2013) notes that a good interpretation/use argument is one that is clear, coherent and complete. Of interest to this investigation, Kane identified five common inferences (i.e., scoring inference, generalisation inference, extrapolation inference, decision inference, and implication inference) that arise when using test instruments to reach conclusions about people. These are outlined in the following sections and further elucidated using an example (interpretation/use arguments typically include some or even all of these inferences):

3.2.3.1 Scoring inference

Scoring is the first inference in the use of any instrument. Scoring is the process where a person's observed performance (e.g., written answer in a test, artwork, actions during a wrestling match etc.) is assigned a numerical value based on a set of scoring rules. The two assumptions associated with this inference are that an appropriate scoring rule was used and that the scoring rule was applied by the scorers as specified. While scoring rules are typically assessed when a test is developed, their appropriate application may be evaluated.

3.2.3.2 Generalisation inference

Generalisation is commonly the second inference. Generalisation is required to allow the results of an instrument from different occasions and different people to be compared.

Kane (2013) explains that two types of generalisation need to be considered when using measurements from an instrument. First, there is an expectation that the measurements are generalisable across the different occasions on which the instrument was administered. This type of generalisation requires that the different conditions under which the instrument was used did not significantly affect the measurements. Some degree of generalisability across occasions is almost always necessary for the interpretation of measurements to be meaningful (Brennan, 2001; see Moss, 1994 for counter examples).

The second type of generalisation is the expectation that an instrument makes consistent measurements. This is more commonly known as reliability (Punch, 2005). While there are several methods to estimate reliability (e.g., test-retest, split-half, Cronbach's alpha), each is fundamentally based on the concept of taking multiple measurements and calculating the variation between the measurements (Field, 2013; Fraenkell & Wallen, 2006; Trochim & Donnelly, 2007). Such generalisability over tasks is always a necessary, but not a sufficient, condition of validity (Kane, 2013). It is noted that reliability evidence is probably the most common type of validity evidence reported for educational and psychological measurements (Cizek et al., 2008).

As with the scoring inference, the generalisability inference is typically assessed during the development of the instrument through trials of the instrument with a large sample. There are two common assumptions for the generalisation inference. These are that the sample of observations taken to validate the instrument is representative of the population of interest, and that this population is large enough to control sampling errors (Kane, 2006). An instrument may be biased if these assumptions are not satisfied.

3.2.3.3 Extrapolation inference

Extrapolation is commonly the third inference. Extrapolation occurs when the measurement from an instrument is used to make a conclusion about a person (Kane, 2006; 2013). Wools, Eggen, and Sanders (2010) suggest that such extrapolations may take one of two forms. First, as no test instrument can exhaustively measure the constructs being examined, a person's performance on the items in the instrument must be extrapolated to estimate their level of competence on the construct. For example, performance on a short written test of Spanish vocabulary (i.e., the instrument) may be used to estimate the person's proficiency in Spanish (i.e., the construct being examined). Kane refers to this type of extrapolation as theory-based interpretation, and describes it as an interpretation of a person's score on the instrument to provide information about their knowledge or skill on the construct of interest.

A second extrapolation may also be involved if the test result is used to make claims about how the person might perform on similar tasks. For example, having scored highly on the Spanish language test, the interpretation may be that this person will succeed in communicating for simple interactions (e.g., for navigation, or transactions) when visiting Spain. Kane describes this extrapolation as an interpretation of a person's level of skill to provide information on how they may perform in practice or other real-life situations.

3.2.3.4 Decision inference

The *decision* inference is made when a measurement from an instrument is used as the basis for the selection of a particular action (Kane, 2006; 2013; Wools et al., 2010). For example, according to the performance of students on an instrument, decisions could be made about class placement, the award of a scholarship or prize, or the provision of educational interventions. As some interpretations and uses of instruments do not require any decisions to be made, this inference is not always included in the interpretation/use argument. The

assumptions associated with the decision inference are that the meaning of a person's scores are easily interpretable by those making the decision, and that the use of the instrument scores will lead to an appropriate decision outcome.

3.2.3.5 Implication inference

An *implication* inference is made when conclusions are reached about a person, which results in several traits becoming linked to the person. This inference is similar to the extrapolation inference, although the nature of traits being linked to the person are not directly measured by the test instrument itself. While in the extrapolation inference the person's score on an instrument is used to imply their skill or ability in a larger domain than was tested (e.g., a broader knowledge of the language than was tested), in the implication inference, other constructs, which are only theoretically connected, may also be implied. For example, if a researcher concludes that a person is gifted on the basis of performance on an intelligence test, some traits of gifted people including the desire for challenging content and high motivation may also become associated with the person, even though they were not specifically tested. Not all interpretations and uses of instruments require implications to be made, and therefore this inference is not always included in the interpretation/use argument.

3.2.3.6 Exemplar interpretation/use argument

While Kane's framework is designed to assist researchers to understand and follow the requirements of modern validity theory, its complexity may mean that it is best demonstrated with the aid of an example - a mathematics test, performance on which is used to place students into classes, with high scores being interpreted to mean that a student should be placed into a high level class. To make this interpretation, several inferences must be made. In the *scoring* inference, the student's written responses to each test item are marked. The *generalisation* inference allows these marks to be converted into percentile ranks, while *extrapolation* occurs when the percentile rank of each student is used to make conclusions

about the relative level of skill of each student. Finally, a *decision* is made using the level of skill for each student to place them into a specific class. In this example, four logical inferences were made between the student's original raw responses in the mathematics test and the interpretation/use of the mathematics test.

The interpretation/use argument has been summarised in Table 2, which incorporates the assumptions associated with each inference.

Table 2

Interpretation/use argument for a placement testing system (Kane, 2006, p. 24)

I1: ***Scoring***: Marking or scoring the individual's performance	
A1.1.	The scoring rule is appropriate
A1.2.	The scoring rule is applied as specified
I2: ***Generalisation***: A comparison of an individual's score to others' scores	
A2.1.	The observations made in testing are representative of the individual's level of skill
A2.2.	The number of scores obtained is large enough to control for sampling error
I3: ***Extrapolation***: A conclusion about the individual's level of skill	
A3.1.	The successful completion of test tasks require competencies developed in the courses undertaken by the individual and competencies required in subsequent courses to be undertaken by the individual
A3.2.	There are no test tasks that are irrelevant to the skills being assessed to seriously bias the interpretation of scores
I4: ***Decision***: Placement in a specific course based on their measured level of skill	
A4.1.	Performance in courses, beyond the initial course, depends on the level of skill in the competencies developed in earlier courses in the sequence
A4.2.	Students with a low level of skill in the prerequisites for a course are not likely to succeed in the course
A4.3.	Students with a high level of skill in the competencies taught in a particular course would not substantially benefit from taking the course

3.2.4 The validity argument

The second part of Kane's argument-based framework to assess validation is the validity argument. The validity argument systematically examines each of the inferences and assumptions outlined in the first part of the framework (i.e., the interpretation/use argument) to show that they are reasonable and appropriate. Nevertheless, Kane (2006, 2013) suggests that in the conduct of a validation study, evidence should be acquired for only the most critical and problematic inferences, due to the limitless nature of the evidence that may be gathered and the reasonableness of many inferences and assumptions that do not require further evidence. Where empirical support is necessary, Kane provides some guidance. The following sections outline in detail the types of evidence Kane suggests may be relevant for the validation of inferences in the interpretation/use argument (adapted from Kane, 2006, pp. 34–38, 51–56).

3.2.4.1 Evidence for the scoring inference

Typically, the assessment of whether an instrument is being scored appropriately relies on the judgement of a panel of experts. Experts may be requested to assess the scoring criteria, whether the scoring criteria have been implemented correctly, and the procedures for selecting and training scorers (Clauser, Harik, & Clyman, 2000). The accuracy and consistency of the scorers may be analysed by comparing the scores of the panel of experts and trained scorers through measures such as inter-rater reliability (Kane, 2006).

3.2.4.2 Evidence for the generalisation inference

Evidence to support generalisation typically comes from reliability studies and generalisability studies (Kane, 2006). In a reliability study, the ability of an instrument to reproduce the same measurements under specific conditions is assessed. In contrast, a generalisability study seeks to measure whether an instrument produces the same measurements when a specific condition (or conditions) are allowed to change. For example,

if generalising across different groups of people, a generalisability study will measure how much variation in measurement occurs due to a change in the groups that are assessed. If either a reliability or generalisability study indicates that significant sampling errors exist, the generalisation inference is not supported (Kane, 2006). While the design of a reliability and generalisability study is different, both attempt to determine whether the measurements from an instrument are homogeneous (the same) across a series of repetitions of measurements. Therefore, evidence for the generalisation inference may also be referred to as homogeneity evidence.

3.2.4.3 Evidence for the extrapolation inference

There are two types of empirical evidence relevant to the extrapolation inference, convergence evidence and criterion evidence. Convergence evidence is typically gathered by assessing the correlation between the scores of two instruments that claim to measure the same trait or skill (Campbell & Fiske, 1959; Cronbach, 1971; Santelices & Taut, 2011), with convergence being deemed to be established when a high correlation is found between the instrument scores. Some scholars (Altman & Bland, 2002; Zaki, Bulgiba, Ismail & Ismail, 2012) recommend that, if possible, instead of a sole reliance on correlation coefficients, a number of different measures of convergence should be simultaneously used to acquire a more complete assessment of the degree of convergence between instruments.

Criterion evidence, which is evidence obtained by comparing the observed scores in a test with a criterion score from another source with established validity (Cronbach, 1971; Messick, 1989; Kane, 2006), is also used as a source of evidence to support the extrapolation inference. An example of a situation where criterion evidence is gathered is when scores on a short written test designed to make conclusions about an individual's intelligence are similar to his/her scores on an IQ test, which has a known and acceptable degree of validity, and may more thoroughly test the variable in question (i.e., intelligence). As the measurement of a

criterion score may be very time consuming and financially onerous, it is typically carried out with only a small number of participants (Kane, 2006).

3.2.4.4 Evidence for the decision inference

The evidence that may be presented to support the decision inference is highly dependent on the nature of the decision to be made. Furthermore, while most inferences that form part of the interpretation/use argument are evaluated in terms of their plausibility, the decision inference may be evaluated in terms of its consequences (Kane, 2006). Typically, to evaluate a decision, an assessment is often made as to how well it achieves its goals, whether positive outcomes are reached, or negative outcomes are avoided (Kane, 2006; Shepard, 1993, 1997). For example, a school's decision to stream its classes may be evaluated in terms of the resulting improvements to student learning. Nevertheless, significant debate continues regarding how specifically to acquire consequential evidence and whether it is indeed necessary (Cizek et al., 2010; Cizek et al., 2008; Haertel, 2013; Hubley & Zumbo, 2011; Kane, 2013; Lane, 2012, 2013; Nichols & Williams, 2009). Despite the ongoing debate, Kane (2006) notes that "in education, as in medicine, there is an obligation to avoid doing harm if it can be avoided" (p. 56).

3.2.4.5 Evidence for the implication inference

Similar to the decision inference, the relevant evidence to support an implication inference will depend on the specific implications themselves. As the implication inference is made when traits or skills are linked to a person without the direct testing of such traits or skills, evidence to support the implication inference involves the presentation of evidence relating to the tests of such traits or skills. For example, if a student scores highly in a mathematics test, two possible implications may be that the student thinks logically and enjoys mathematical tasks. Both of these implications may be directly tested, using

instruments that are specifically designed to assess logical thinking and enjoyment of mathematical tasks, for the provision of evidence to support the implication inference.

3.3 The Application of Kane's Validation Framework to Gifted Underachievement: An Interpretation/Use Argument

In this section, an interpretation/use argument relating to the current use of methods which seek to identify and measure the underachievement of gifted students will be developed. Generally, the purpose of identification and measurement of gifted underachievement is to determine the appropriateness of offering educational intervention programs for particular students (Reis & McCoach, 2000). Unlike Kane's exemplar of a mathematics test used to place students into appropriate classes, the identification of gifted students who are underachieving for placement into intervention programs rely on a combination of measurements from two different instruments (i.e., one measuring expected achievement and one measuring actual achievement). Wools, Eggen, and Béguin (2016) demonstrate that interpretation/use arguments may incorporate measurements from multiple instruments by including the inferences for each instrument in parallel but separate inferences, until the information from the multiple instruments is combined. Consequently, the interpretation/use argument for the identification and measurement of gifted underachievement must include the scoring, generalisation, and extrapolation inferences separately for both of the instruments used to measure expected achievement and actual achievement. It is noted that these inferences are specific to the instruments used, and not to the various methods (i.e., absolute split I, absolute split II, nomination, simple difference, and regression methods) of identification and measurement of gifted underachievement.

The information from the instruments designed to assess expected achievement and actual achievement should be combined only *after* extrapolations of student scores to their

levels of expected achievement and actual achievement. At this point, the levels of expected achievement and actual achievement of students may be compared to assess their level of gifted underachievement. This step should be considered to be an additional *extrapolation*, as it goes beyond the scoring of individual levels of expected or actual achievement and the making of comparisons of these scores with other students, to reaching a conclusion about levels of gifted underachievement. Therefore, an additional extrapolation inference that combines information from the instruments that measure expected achievement and actual achievement should be included in the interpretation/use argument. As the extrapolation inference will vary for each of the different methods (i.e., absolute split I, absolute split II, nomination, simple difference, and regression methods) used to identify and measure gifted underachievement, the empirical evidence that assesses this inference will be useful to establish the individual validity of these methods.

In recognition of the different combinations of instruments available to assess gifted underachievement (refer Appendix 1), and the need for the conclusion about an individual's gifted underachievement (arrived at under the second extrapolation inference) to be *independent* of the data combination used, an additional *generalisation* inference is necessary in the interpretation/use argument. This second generalisation inference will be that any individual method used to identify and measure gifted underachievement will identify and measure gifted underachievement independently of the different combinations of expected achievement and actual achievement data used.

The final inference relates to the decision that is made following the identification and measurement of gifted underachievement, which will essentially be whether to place each individual student into an intervention program. This decision inference will be specific to the intervention program that is utilised by the researcher or educator. Figure 12 and Table 3 provides greater details on the interpretation/use argument that guided this project.

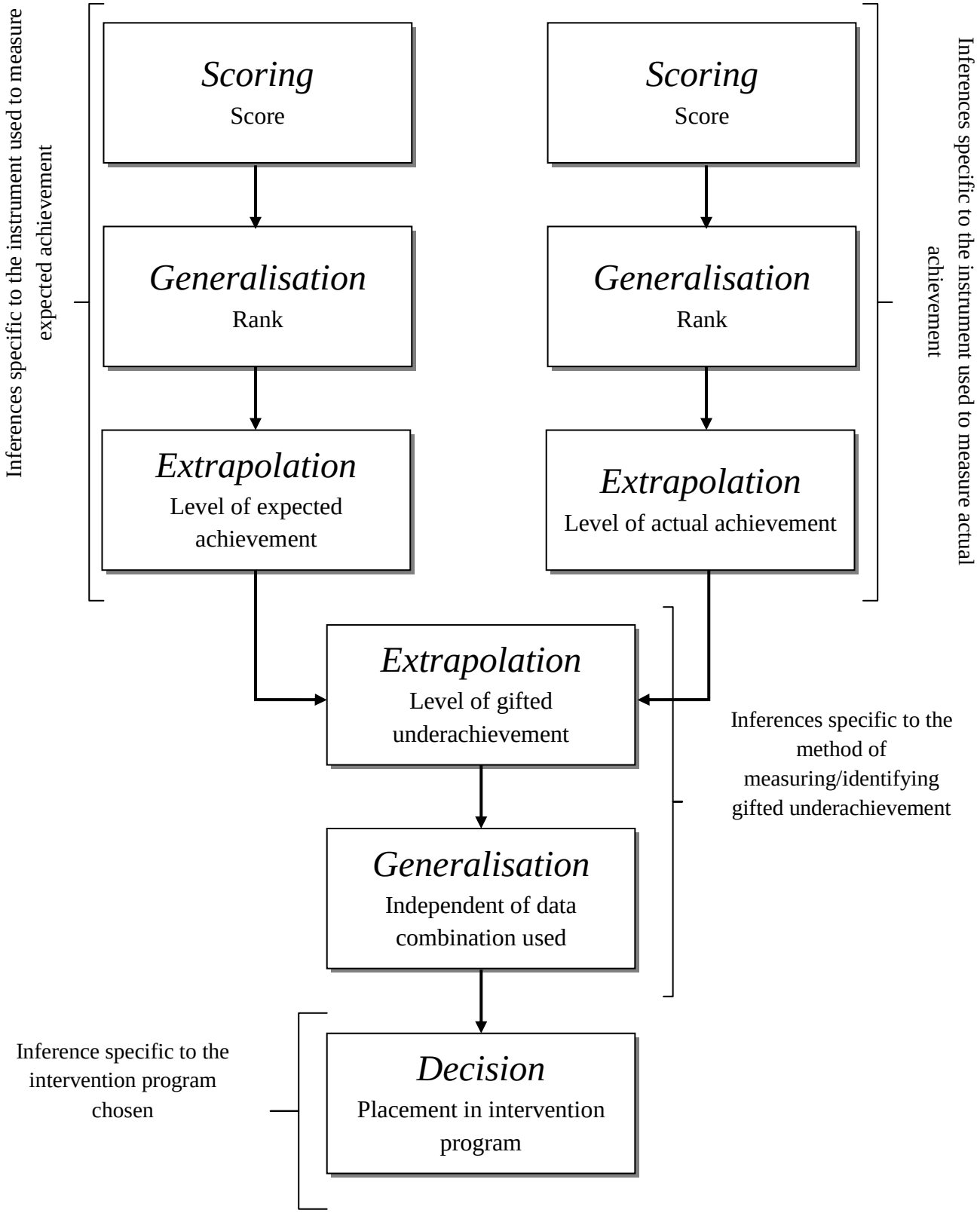


Figure 12. Interpretation argument for methods that identify/measure gifted underachievement

Table 3

Interpretation argument for methods that identify/measure gifted underachievement.

I1: ***Scoring***: Scoring the individual's performance to determine scores for expected achievement and actual achievement	
A1.1.	The scoring rules are appropriate
A1.2.	The scoring rules are applied as specified
I2: ***Generalisation***: A comparison of the individual's score to others' scores	
A2.1.	The observations made in testing are representative of the individual's level of expected achievement or actual achievement with respect to the expected achievement or actual achievement of others
A2.2.	The number of scores obtained is large enough to control for sampling error
I3: ***Extrapolation***: A conclusion about the individual's expected achievement and actual achievement	
A3.1.	The test tasks appropriately assess the individual's expected achievement or actual achievement
A3.2.	There are no test tasks that are irrelevant to the assessment of expected achievement or actual achievement to seriously bias the interpretation of scores
I4: ***Extrapolation***: A conclusion about the individual's gifted underachievement	
A4.1.	The expected achievement and actual achievement scores are used together to appropriately assess the individual student's gifted underachievement
A4.2.	There are no irrelevant sources of variability that would seriously bias the interpretation of scores as measures of gifted underachievement
I5: ***Generalisation***: A comparison of gifted underachievement across different data combinations	
A5.1.	The identification/measurement of gifted underachievement using this data combination is representative of the individual student's gifted underachievement across other data combinations
A5.2.	The number of observations is large enough to control for sampling error
I6: ***Decision***: Placement in a specific intervention program based on the individual's gifted underachievement	
A6.1.	The intervention program positively impacts gifted students who underachieve
A6.2.	The intervention program does not negatively impact gifted students who underachieve

3.4 The Application of Kane's Validation Framework to Gifted Underachievement: A Validity Argument

This section will analyse the proposed interpretation/use argument for gifted underachievement to identify the most critical and problematic inferences that will become the focus of this investigation. Separate analyses were undertaken for all the inferences that comprise the interpretation/use argument, with the exception of the inferences relating to the measurement of expected and actual achievement (which were instead separated into analyses for standardised and non-standardised instruments that assess achievement, in recognition of the substantial differences in the issues affecting the validity of standardised and non-standardised instruments that assess expected and actual achievement).

3.4.1 Assessment of the validity of using standardised instruments to measure expected/actual achievement

There are several thousand commercially available standardised instruments that may be used to measure expected or actual achievement (Carlson et al., 2014), with varying levels of validity. For the most commonly used instruments in Australia (e.g., Wechsler Intelligence Scale for Children-fifth edition [WISC-V], Wechsler Individual Achievement Test-third edition [WIAT], Stanford–Binet-fifth edition [SB-V], Woodcock–Johnson Tests of Cognitive Ability and Tests of Achievement-second edition [WJ-II]), validity evidence is reported in a technical manual (e.g., Harcourt Assessment, 2007; Pearson, 2014), while independent reviews of these instruments carried out by psychometric experts are also regularly reported in the *Mental Measurements Yearbook* (Carlson et al., 2014). Of these instruments, the WISC-V, SB-V may be used to measure expected achievement, WIAT-II may be used to measure actual achievement, and the WJ-II contains separate tests that may be used to measure either expected or actual achievement. The independent reviews of standardised instruments that assess expected or actual achievement including the WISC-V (Benson, in

press; Keith, in press), the WIAT-III (Miller, 2010; Willse, 2010), the SB-V (Johnson & D'Amato, 2005; Kush, 2005; Sink & Eppler, 2007; Vacca, 2007), and the WJ-III (Cummings, 1995; Lee & Stefany, 1995) formed the basis of the following sections.

3.4.1.1 Validity of the scoring inference for achievement scores

The necessary evidence to support the scoring inference include expert reviews (of test items, scoring criteria, implementation of scoring criteria, and procedures for selecting and training scorers) and measures of inter-scorer agreement. The independent expert reviews published in the *Mental Measurements Yearbooks* provide most of this evidence. For each of the selected commonly used standardised instruments, the experts reviewed the test items and their scoring criteria and supported the validity of the scoring inference (Benson, in press; Cummings, 1995; Johnson & D'Amato, 2005; Keith, in press; Kush, 2005; Lee & Stefany, 1995; Miller, 2010; Sink & Eppler, 2007; Willse, 2010; Vacca, 2007). These instruments have also been found to have high levels of inter-scorer agreement. For example, the SB-V was identified to have a median correlation between different scorers of 0.90 (Sink & Eppler, 2007; Johnson & D'Amato, 2005; Kush, 2005; Vacca, 2007), while the inter-scorer coefficients for the WISC-V appear to be 0.97 or greater (Pearson, 2014). No evidence could be found that assessed the implementation of the scoring criteria, or the procedures for selecting and training scorers for the commonly used standardised instruments.

3.4.1.2 Validity of the generalisation inference for achievement ranks

The necessary evidence for supporting the generalisation inference is provided by reliability and generalisability studies. Two main types of reliability evidence were commonly reported for the achievement instruments: split-half and test-retest reliability. The reliability evidence reported by the instrument publishers was extensive and all of the reliability estimates were considered acceptable by the independent reviewers (Benson, in press; Cummings, 1995; Johnson & D'Amato, 2005; Keith, in press; Kush, 2005; Lee &

Stefany, 1995; Miller, 2010; Sink & Eppler, 2007; Willse, 2010; Vacca, 2007). For example, a test-retest stability coefficient of 0.92 was reported for the full-scale IQ construct of the WISC-V (Pearson, 2014), and a composite stability coefficient of 0.97 was reported for the WIAT-II (Harcourt Assessment, 2007). Almost all of the split-half and test-retest reliability coefficients reported for the selected commonly used instruments ranged from 0.75 to 0.95, which may be considered ideal values (Tavakol & Dennick, 2011).

In addition, the standardisation procedure used in the development of each instrument was often reported on meticulously to show that the samples used during the test development were representative of the wider population with respect to age, sex, race/ethnicity, parent education level, and geographic region. The independent expert reviewers noted that the processes used to select normative samples during the development of these tests were "most impressive" and concluded that the samples used are representative of the wider population (Johnson & D'Amato, 2005). While no generalisability study could be found, the independent expert reviewers were of the view that the generalisability of samples used was sufficient to support the generalisability of test interpretations (Benson, in press; Cummings, 1995; Johnson & D'Amato, 2005; Keith, in press; Kush, 2005; Lee & Stefany, 1995; Miller, 2010; Sink & Eppler, 2007; Willse, 2010; Vacca, 2007).

3.4.1.3 Validity of the extrapolation inference for levels of achievement

The necessary empirical evidence to assess the extrapolation inference includes convergence evidence and criterion evidence. The technical manuals for each of the selected instruments reported convergence evidence with a series of other commonly used instruments. For example, the WIAT-II technical manual reported correlations of the WIAT-II with the WIAT-I, Wide Range Achievement Test-third edition (WRAT-III), and the Differential Ability Scales (DAS), to be 0.84, 0.76, and 0.63 respectively (Harcourt Assessment, 2007). The independent reviewers for each of the commonly used instruments

concluded that the convergence statistics tended to be within the acceptable to ideal range (Benson, in press; Cummings, 1995; Johnson & D'Amato, 2005; Keith, in press; Kush, 2005; Lee & Stefany, 1995; Miller, 2010; Sink & Eppler, 2007; Willse, 2010; Vacca, 2007). Therefore, the convergence between the commonly used standardised measures of achievement generally appear to be well supported. It is nevertheless noteworthy that one study (Silverman et al., 2010) found that the average difference in scores between the common instruments to be greater than one standard deviation. This highlights the need for the convergence evidence to be interpreted cautiously, as the different instruments may not be interchangeable, even though each appears valid.

Criterion evidence was provided using two different methods. First, for the expected achievement instruments (e.g., WISC-V, SB-V, and WJ-II), some actual achievement instruments (e.g., WIAT-II and Kaufman Assessment Battery for Children [KABC-II]) were used as criterion measures. The consistently high correlations between the various instruments that measure a student's expected achievement and their respective criterion measures (e.g., .81 between the WISC-V and WIAT-II, and .81 between the WISC-V and KABC-II was 0.81; Pearson, 2014) provided evidence that supported the predictive validity of these instruments (Benson, in press; Cummings, 1995; Johnson & D'Amato, 2005; Keith, in press; Kush, 2005; Lee & Stefany, 1995; Sink & Eppler, 2007; Vacca, 2007). These results were also supported by the many large meta-analyses of independent studies that have found standardised intelligence tests to be among the most useful instruments in the prediction of human behaviour (refer Neubauer & Opriessnig, 2014). Nonetheless, some have expressed concerns about the validity of some of these instruments due to factors including the Flynn effect (Pietschnig & Voracek, 2015), the use of inappropriate ratios rather than normalised scores (Thurstone, 1926), impact of disadvantage on measurement (Merrotsy, 2013) and the possible vested financial interest of scorers (Hertwig & Ortmann, 2001).

The second type of criterion evidence produced, for both the expected and actual achievement instruments, has been special population studies. These studies (reported in the technical manuals) directly measure the usefulness of the commonly used standardised instruments in measuring differences between special groups (e.g., gifted, intellectual disability, specific learning disorders, autism, etc.) and a matched control group. Multiple independent reviews of the selected commonly used instruments found that the differences between the measurements taken from these special groups and the control group were as expected (Benson, in press; Cummings, 1995; Johnson & D'Amato, 2005; Keith, in press; Kush, 2005; Lee & Stefany, 1995; Miller, 2010; Sink & Eppler, 2007; Willse, 2010; Vacca, 2007). For example, on the WIAT-II, individuals with reading disorders scored an average of 16 points lower on the word reading subtests than the control group, while individuals with an intellectual disability tended to score more than 30 points lower on each subtest than the control group. As these differences matched expectations, evidence is provided of the validity of the instruments.

3.4.2 Assessment of the validity of non-standardised instruments to measure expected/actual achievement

A major problem in making an assessment of the validity of non-standardised measures of achievement is that because they are not standardised, there may be substantial variation in the application and design of instruments, and therefore substantial variation in the validity of their interpretation. Even though researchers of gifted underachievement almost all report a reliance on school grades (refer Appendix 1), it is important to note that the actual test items completed may be entirely different for potentially each student in each study. Nevertheless, such variation may be necessary to correctly reflect the unique content, learning environment, and skills learned by each student (Haladyna, 2006; Reis & McCoach, 2000; Rock & Stenner, 2005; William, 2001, 2003). As there may be no commonly used non-

standardised instruments, an assessment of the validity of non-standardised instruments outlined in this section may only represent a rough guide to the relevant issues.

3.4.2.1 Validity of the scoring inference for achievement scores

The scoring inference appears to be supported when a number of key practices are followed. For example, scholars have found that high levels of inter-rater agreement are possible when: (a) assessment tasks incorporate clear instructions for students, (b) scoring criteria specifies clear indicators of success, avoids vague terms, and are detailed, (c) scoring criteria are constructed by the scorer, or scorers have been provided with adequate training and exemplar samples, and (d) moderation activities have been conducted among scorers to ensure consistency in scoring (Davis, 2016; Harlen, 2005). Further evidence to support the scoring inference with non-standardised instruments may be obtained through an examination of the cognitive process that scorers follow while making an assessment of student performance (Bejar, 2012).

3.4.2.2 Validity of the generalisation inference for achievement ranks

Despite the widespread use of teacher judgements in the assessment of student achievement, their reliability and generalisability is not always guaranteed (MacCann & Stanley, 2010). For example, many researchers have found that teacher assessments of students may be influenced by the non-academic traits of students, including gender, attractiveness, ethnicity, behaviour, motivation, socioeconomic status, and social skills (Bennett, Gottesman, Rock, & Cerullo, 1993; Beswick, Willms, & Sloat, 2005; Glock, Korlak-Schwerdt, Klapproth, & Böhmer, 2013; Hammes, Bigras, & Crepaldi, 2014; Harlen, 2005; Hurwitz, Elliot, & Braden, 2007; Kaiser, Retelsdorf, Südkamp, & Möller, 2013; Ritts, Patterson, & Tubbs, 1992; Strand, 2012; Tiedemann, 2002), or the teacher's level of education and experience (Beswick et al., 2005; Davis, 2016; Mashburn & Henry, 2004). Others argue that the requirement for teachers to provide final grades for students on the basis

of broad descriptive criteria (e.g., the common grade scale; BOSTES, 2016) may be subjective and unreliable (Johnson, 2013). To overcome these problems, some suggest that the average of multiple assessment tasks should be used instead of a single assessment score (Messick, 1996; Goldschmidt, Martinez, Niemi, & Baker, 2007). It is noteworthy that the common practice of researchers of using a grade point average (GPA) may be seen as reflecting this advice (refer Appendix 1).

3.4.2.3 Validity of the extrapolation inference for achievement

Multiple analytical/theoretical arguments have been made to support the use of non-standardised instruments to measure achievement. These include the fact that such instruments, unlike standardised instruments: (a) may involve multiple types of tasks, (b) may involve the acquisition of information over a longer period of time, (c) generally provide students with greater opportunities to demonstrate their capabilities, (d) may be directly linked to experiences in the classroom, and (e) may provide an assessment environment similar to the learning environment (Bagnato, 2005; Haladyna, 2006; Pellegrini, 2001; Reis & McCoach, 2000; Rock & Stenner, 2005; William, 2001, 2003). Furthermore, due to the provision of multiple opportunities for students to demonstrate their capabilities, their scores may be less affected by factors such as student stress, lack of sleep, or sickness (Haladyna, 2006; Reis & McCoach, 2000; Rock & Stenner, 2005; William, 2001, 2003).

Some empirical evidence exists to support the extrapolation of non-standardised instrument scores to a measure of a student's level of achievement. Several studies have found a moderate to strong correlation between teacher assessment scores and standardised test scores. For example, Begeny, Eckert, Monterello, and Storie (2008) noted a correlation of .79, Hinnant, O'Brien, and Ghazarian (2009) identified a correlation of .67, while Martin and Shapiro (2011) reported a correlation of .81. In comparison, a recent meta-analysis found the overall mean correlation between teacher judgements and student performance on

standardised achievement tests to be 0.63 (Südkamp, Kaiser, & Möller, 2012). Moreover, criterion evidence obtained by multiple scholars suggests that teacher assessment may be at least as accurate as standardised tests in predicting future performance (Forget-Dubois et al., 2007; Harlen, 2005; Hecht & Greenfield, 2001; Lohman, 2005; Meisels et al., 2001; Ziegler, 2008), with some researchers claiming that past teacher assessments of achievement may be superior to standardised measurements of *g* in predicting future performance (Ziegler et al., 2012).

3.4.3 Validity of the extrapolation inference for gifted underachievement

The validity of the measures for gifted underachievement (i.e., the absolute split I, absolute split II, nomination, simple difference, and regression methods) is yet to be fully established by researchers. The sections below summarise the existing contributions to the assessment of the validity of the extrapolation inference for gifted underachievement.

3.4.3.1 Regression to the mean

Regression to the mean is a statistical artefact that may affect any series of test results, if there is an element of chance in arriving at a correct response to any of the test questions. In the situation where a student scores highly in one test due to such chance factors, he or she is likely to achieve a score that is closer to the mean on a subsequent test. While the difference in the two test scores may be considered to reflect different levels of achievement by educators, a more appropriate interpretation may be that it is a statistical artefact that arose from chance factors (Barnett, van der Pols, & Dobson, 2005). In practice, many tests are likely to have some elements of chance, particularly if students are in a position to guess answers (e.g., multiple choice questions).

Some scholars have noted that since there is some degree of chance in many tests of expected and actual achievement, a regression to the mean effect, which may be interpreted

as underachievement, may be quite common. Therefore, it is possible that the measurements of gifted underachievement are inflated, and that a greater proportion of gifted students are identified as underachieving than is actually the case (Cone & Wilson, 1981; Wilson & Reynolds, 1985; McCall, Evahn, & Kratzer, 1992). This argument has been used to discredit the use of some identification/measurement methods (i.e., absolute split I, absolute split II, and simple difference methods) and support the use of the regression method, which inherently corrects for the regression to the mean artefact (McCall et al., 2000).

3.4.3.2 Compounding errors in measurement

Some scholars have argued (Smith, 2003, 2010; Ziegler et al., 2012) that when using statistical techniques to identify gifted underachievement, the majority of identified cases may be due to measurement error. Specifically, as the measurements of both expected achievement and actual achievement are prone to errors in measurement, any combination of these two measurements (needed to measure gifted underachievement) must also have a degree of error. It is noteworthy that Smith (2003, 2010) in fact suggested that rather than being a real phenomenon, underachievement may reflect a statistical artefact of compounded errors (Smith, 2003, 2010). Ziegler et al. (2012) investigated the matter for a hypothetical situation where there is no underachievement, assuming a normal distribution of expected achievement and actual achievement measurements, and typical estimates of errors, and concluded that "approximately 10% of the total (not only of the gifted) students would be considered underachievers, even if underachievement were a nonexistent phenomenon" (Ziegler et al., 2012, p. 126) due to compounding errors in measurement. The issue was further highlighted by Silverman et al. (2010), who demonstrated that when two of the most commonly used intelligence tests (i.e., WAIS and SB) were administered to the same sample, the average discrepancy in scores was larger than the typical threshold used to establish underachievement.

3.4.3.3 Subjectivity of nominations

Teachers, peers, and parents may have many interactions with an individual child that may allow each of them to develop a unique view of the child. Despite their rich knowledge, however, research suggests that these groups largely fail to correctly identify cases of gifted underachievement (Dunne & Gazeley, 2008; Jones & Myhill, 2004; Lau & Chan, 2001a). One possible reason for this may be that teacher, peer, and parent nominations require highly subjective judgements from untrained individuals, based on incomplete information, and skewed personal bias (Richert, 2003). For example, researchers have found that teachers and students "tend to rate as more desirable and successful those students who are most similar to themselves" (Dowdall & Colangelo, 1982, p. 181). While the provision of appropriate training for nominators may be helpful, the issue of personal bias may be difficult, if not impossible, to resolve without also relying on more objective methods of identification.

3.4.3.4 Convergence

Two studies have been conducted to date to determine whether the results from the different identification methods converge. The first study (Annesley et al., 1970) examined three of the statistical methods (i.e., absolute split I, simple difference, and regression) and found no evidence of convergence. The second study (Lau & Chan, 2001a), which also included the nomination method, indicated that while the statistical methods (i.e., absolute split I, simple difference, and regression) were convergent, the nomination method was not. As these findings are inconsistent, some ambiguity exists in the convergence of the various methods used to identify and measure gifted underachievement.

3.4.4 Validity of the generalisation inference for gifted underachievement

Although generalisation of the use of methods to identify/measure gifted underachievement across different combinations of ability and achievement measures may be ideal and often implicitly assumed (refer Appendix 1), such a generalisation may not be

theoretically supported. One reason why generalisation across different data combinations may not be possible is that some data combinations are clearly more valid than others. For example, past achievement in English may not be an appropriate measure of expected achievement for comparison with a student's actual achievement in mathematics. A related concern may be the difference in the point of time when the two measurements were obtained. For example, it may be questionable to compare a student's expected achievement with his/her actual achievement, if the assessment of expected achievement was made two years prior to the assessment of his/her actual achievement (Deary, 2006; Weinert & Schneider, 1999; Ziegler et al., 2012).

Furthermore, the different instruments that are available to measure expected and actual achievement may have varying levels of difficulty, or may be standardised with different populations. For example, when two (or more) courses in the same field are offered at different levels of difficulty, they may attract different groups of students (i.e., higher ability students may be more likely to take the most difficult course, while lower ability students may be more likely to take the easiest course). Therefore, a student's actual achievement in one course, as measured by their position on the bell curve for that student group, may not be equivalent to another student's achievement in another course. Consequently, any differences in the selection of achievement measurements, and their combinations, may have an impact on the identification and measurement of gifted underachievement.

Finally, as the degree of measurement error is unique to each expected or actual achievement measurement, the degree of error in the level of underachievement may also be unique. The varying degrees of measurement error in each data combination may result in different proportions of false identifications and different levels of bias in the measurements of gifted underachievement (Ziegler et al., 2012). Such variations between the different

measurements of achievement, which are referred to by some as the problem of heterogeneity (Reis & McCoach, 2000), may mean that the generalisation of gifted underachievement across different data combinations may be difficult.

3.4.5 Validity of the decision inference

The validity of the decision to place particular students into intervention programs is also yet to be established by researchers. Ritchotte, Rubenstein, and Murry (2015) note that the various existing interventions designed to reverse gifted underachievement have produced conflicting results to date. Some propose that the lack of consistent and conclusive evidence for the positive effects of such interventions may be related to the inconsistent use of methods to identify and measure gifted underachievement by scholars (Dowadell & Colangelo, 1982; Reis & McCoach, 2000). It is therefore possible that until the validity of methods to identify/measure gifted underachievement is fully established, any attempts to validate intervention programs may be futile.

3.5 Inferences Chosen for Further Investigation

The analysis of the proposed interpretation/use argument has identified three inferences whose validity has not yet been established: (a) the extrapolation inference for gifted underachievement, (b) the generalisation inference for gifted underachievement across different expected/actual achievement data combinations, and (c) the decision inference of placement of students into specific intervention programs. As the validity of the placement of students into specific intervention programs may not be assessed until after the validity of the methods used to identify and measure gifted underachievement is established, the chosen inferences for this investigation are the two inferences that are specific to the methods for the identification and measurement of gifted underachievement – the extrapolation inference for

gifted underachievement and the generalisation inference for gifted underachievement across different expected/actual achievement data combinations.

3.6 Summary

Many researchers have raised concerns about the validity of each of the methods used to identify and measure gifted underachievement, despite their widespread use (Barnett, et al., 2005; Dowdall & Colangelo, 1982; Richert, 2003; Ziegler et al., 2012). As a result, it is not clear whether any method designed to identify and measure gifted underachievement is valid. In this chapter, Kane's validation framework was discussed and used to construct a series of logical inferences relating to the identification and measurement of gifted underachievement. Following an evaluation of each inference, two inferences that have yet to be fully supported were selected for further investigation. The next chapter describes how the investigation was designed to gather the evidence recommended by Kane to establish the validity of both of the inferences for each of the methods commonly used to identify and measure gifted underachievement.

4 Methodology

4.1 Introduction

The purpose of this chapter is to discuss the methodology used in the investigation. The discussion outlines the research problem, research questions, research design, population, sample selection, and data sources that relate to this investigation.

4.2 The Research Problem

As outlined in the previous chapters, a current and significant problem for research into gifted underachievement is that the validity of methods to identify and measure gifted underachievement is not yet established. At present, researchers interchangeably use multiple methods (i.e., absolute split I, absolute split II, nomination, simple difference, and regression), typically without any justification.

4.3 The Research Questions

Kane (2006) noted that finding a resolution to the problem of validity may require an extremely large investigation due to the range of evidence that may need to be gathered. To provide reasonable restrictions to the current project, the most problematic inferences from the interpretation/use argument (i.e., the extrapolation and generalisation inferences relating to the identification and measurement of gifted underachievement) were selected for empirical investigation. Hence, the research questions that guided the project are as follows:

1. *Is the extrapolation inference reasonable for each of the methods used to identify and measure gifted underachievement?*
2. *Is the generalisation inference reasonable for each of the methods used to identify and measure gifted underachievement?*

To support an extrapolation inference, Kane (2006) recommended that two sources of empirical evidence, convergence evidence and criterion evidence, may be provided. To support a generalisation inference, he recommended that a generalisability study, which determines the degree of variation that occurs due to specific changes in how a research method is applied across different occasions of use, be conducted. In the context of this investigation, separate generalisability studies were needed for the statistical methods (i.e., generalisation across different combinations of expected and actual achievement data) and the nomination method (generalisation across the different nominators). Figure 13 below illustrates the structure of this investigation.

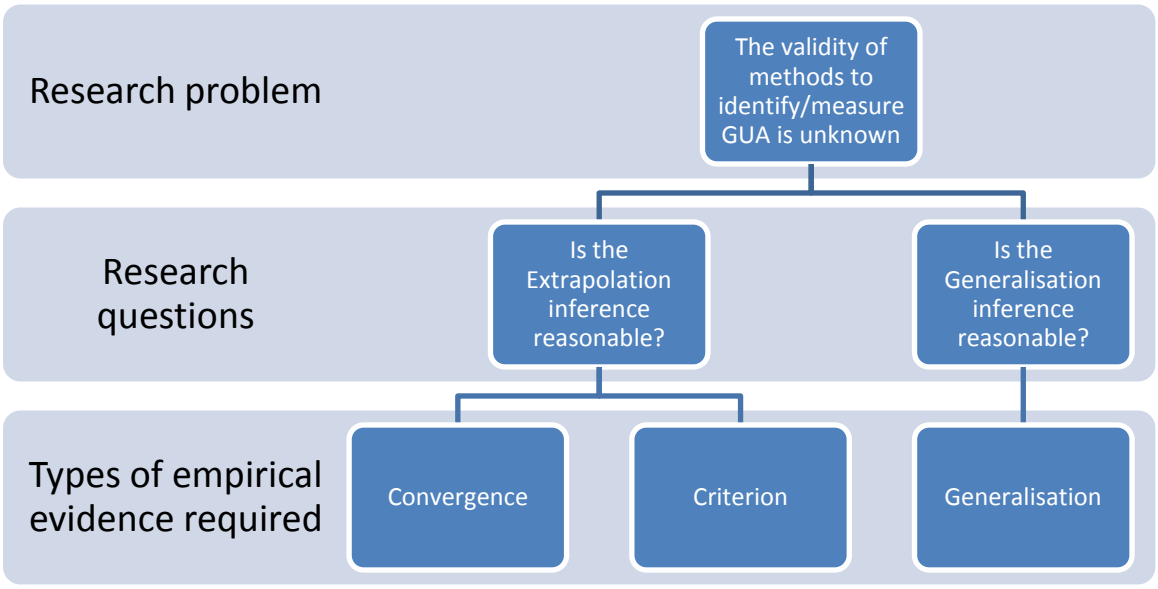


Figure 13. Structure of investigation

4.4 Selection of Methods Used to Identify/Measure Gifted Underachievement

A review of the research has found five commonly used methods to identify gifted underachievement: the two common variations of the absolute split method, the nomination method, the regression method, and the simple difference method. Two of these methods

(i.e., the regression and simple difference methods), also provide a measurement of the degree of gifted underachievement, while four of these methods (i.e., the two variations of absolute split, regression, and simple difference) may be considered statistical methods due to their reliance on a statistical comparison of measurements of expected achievement and actual achievement. The only non-statistical method, the nomination method, relies on a subjective comparison of a student's perceived expected achievement and actual achievement. Among the various forms of nomination, data were collected on the most popular form of the nomination method (i.e., teacher nominations; Appendix 1). Table 4 summarises all of the methods used for the identification/measurement of gifted underachievement that were included in this investigation.

The two variants of the absolute split method have been nominally labelled as "absolute split I" and "absolute split II". The absolute split I method identifies gifted underachievement when a gifted student achieves an actual achievement rank below the first quartile (i.e., below the 75^{th} percentile) for the group under investigation. The absolute split II method identifies gifted underachievement when a gifted student achieves a raw achievement score of below 80%. The two variants of the absolute split method differ in terms of whether the threshold for gifted underachievement is a lower than expected *rank* or a lower than expected *score*.

In all applications of the methods used to identify/measure gifted underachievement, giftedness was defined using Gagné's (2009a, 2013) model. Therefore, the selection of gifted students for investigation was made on the basis of whether the student's expected achievement score placed them within the top 10% of age peers in the ability domain that was measured. As the various methods used to identify/measure gifted underachievement were applied to many combinations of expected achievement and actual achievement data, it is possible that a student may be considered gifted according to one expected achievement

measurement, but not according to another. This variation in the identification of students as gifted due to the use of different instruments reflects the experience of researchers when using multiple instruments (Borland, 1989).

Table 4

Summary of methods used to identify/measure gifted underachievement

Method of identification	Threshold for significant discrepancy	Information on the degree of gifted underachievement (i.e., measure of gifted underachievement)	Type of method
Absolute split I	Below 75th percentile	No	Statistical
Absolute split II	Below 80% raw score	No	Statistical
Simple difference	1 standard deviation	Yes	Statistical
Regression	1 standard error of estimate	Yes	Statistical
Nomination	Personal judgement of the nominator	No	Nomination

4.5 Research Design

A research design is important to ensure that appropriate data are collected to answer the research questions (Bryman, 2004; Fraenkel & Wallen, 2006). The selection of a particular research design will largely direct the methods used by the researcher to collect and analyse data (Bryman, 2004). Of the many different types of research designs, a correlational research design was chosen for this project, as the types of evidence required to answer the research questions as suggested by an application of Kane's validation framework (i.e., convergent, criterion, and generalisation evidence) require measurements of correlation between different variables.

4.6 Sample Selection

Due to the contradictory findings of previous correlation studies that compared the different methods used to identify/measure gifted underachievement with relatively small

sample sizes (Annesley et al., 1970; Lau & Chan, 2001a), a large sample size was targeted for the current project. Specifically, archive data that has already been collected by a school were chosen to overcome the difficulty of collecting data relating to large numbers of participants from a small population of gifted students (i.e., approximately 10% of the student population).

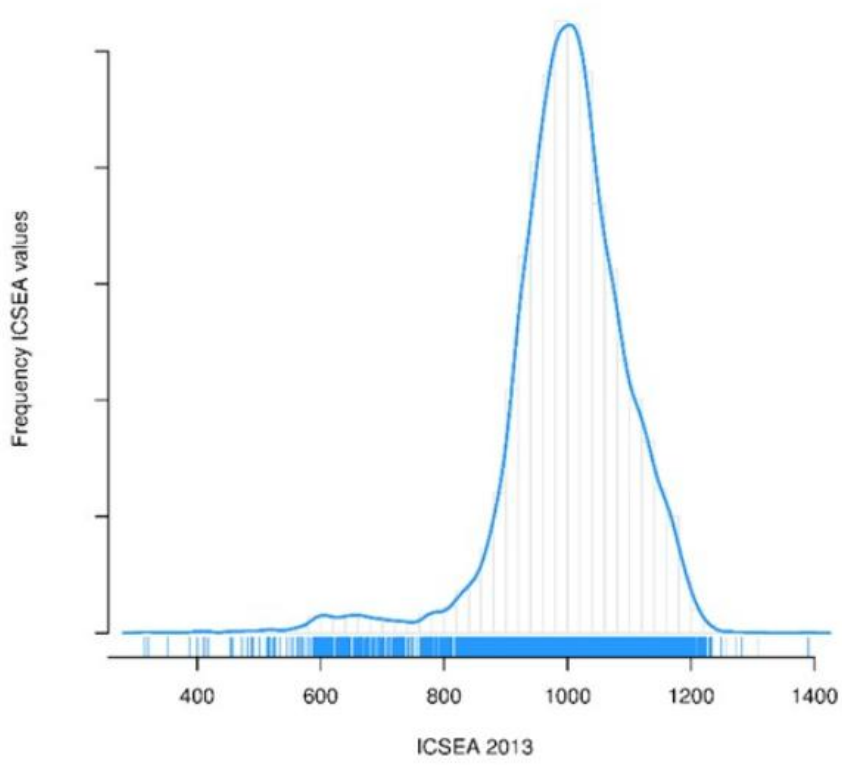


Figure 14. National distribution of ICSEA values 2013 (ACARA, 2015a)

The data originated from a co-educational K–12 independent Christian school located in the south-western suburbs of Sydney, Australia. The school, which has been in operation for over three decades, has a total enrolment of over 1,300 students, of which 41% have a language background other than English, and 53% are male. Students from the school have diverse language and religious backgrounds. The school's Index of Community Socio-Educational Advantage (ICSEA), which is a measure of the social and economic conditions

of the households served by the schools is 1066, which is only slightly above the Australian national average (ACARA, 2015a). The national distribution of ICSEA values for 2013 is shown in Figure 14.

The archive data comprised all of the Grade 7 to 12 expected achievement and actual achievement data available at the school for more than the past 10 years. While the school has systematically tested all students with an expected achievement measure upon entrance, and has maintained consistent records of multiple measurements of achievement for every student, there are some limitations to the data. For example, each instrument was not administered to all students, some instruments were only administered over a limited number of years (e.g., 2003–2009), while other instruments were only administered to certain year groups (e.g., Grades 7–9). Consequently, there are differing sample sizes for each combination of expected achievement and actual achievement instruments. Despite these limitations, the comparison of these different data sets accurately reflect the common practice of researchers attempting to combine or compare results from different studies in gifted underachievement (Dowdall & Colangelo, 1982). As the archival data did not include any teacher nomination data, teacher nominations were newly collected by requesting teachers at the school to complete a qualitative survey.

4.7 Instruments from the Archive Data

Over the history of the school from which the data were obtained, many different instruments have been used to assess expected and actual achievement, at different stages, and with varying degrees of overlap. Historically, separate electronic databases were maintained by different staff members for each of the instruments that have been used at the school. It is noted that some instruments could not be included in this project due to a lack of compatibility in the databases (e.g., student data were not stored with unique identifiers that

allowed matching to student data in other databases). In the following sections, the nature of the instruments that were chosen for inclusion in this study are discussed. Each of these instruments have been used for multiple years, and as a result have provided large amounts of student data.

4.7.1 The Otis–Lennon School Ability Test (OLSAT)

The Otis–Lennon School Ability Test (OLSAT) is a commonly used instrument to measure "cognitive abilities that relate to a student's academic success in school" (Pearson, 2015). It is typically administered to groups of students in a classroom setting by the classroom teacher to ensure the student's performance in the test is representative of their ability in the classroom. The multiple choice instrument focuses on the verbal comprehension, verbal reasoning, pictorial reasoning, figural reasoning, and quantitative reasoning aspects of intelligence, and provides both verbal and non-verbal ability scores as well as a total score called the School Ability Index (SAI). The scores are normalised to a mean of 100 and a standard deviation of 16 (Harcourt Educational Measurement, 2003).

The OLSAT appears to be a well-supported instrument for the measurement of expected achievement. It appears to have a high level of reliability (.92; Johny, Lukose & Magno, 2012) and accuracy (i.e., standard errors of measurement of between 5.5 and 5.8 points; Harcourt Educational Measurement, 2003). Furthermore, many studies have independently confirmed that it has a suitable level of convergence with individually administered IQ tests that are better known and more established (i.e., WISC and SB): .62 (Guilmette, Kennedy, & Queally, 2001), .67 (Tyler-Wood & Carri, 1991), .71 (Ryan, 2007), .73 (Weschler, 1991), .76 (Duncan, 2009), .85 (Dyer, 1985; Oakland, 1985), and .89 (Swets, 1988). Multiple independent tests have also provided criterion evidence using school achievement (Calaguas, 2012; Maddux, 2010; Magno, 2009; Medallon & Cataquis, 2011; Morse, 2010). Nevertheless, independent reviewers have also identified some weaknesses of

the OLSAT, including: (a) its reliance on an early conception of intelligence, (b) the lack of evidence of test-retest reliability, (c) the lack of the conduct of confirmatory factor analysis to demonstrate the validity of the subscores (verbal/non-verbal), and (d) the lack of guidance on the appropriate use of the instrument in the technical manual (Maddux, 2010; Morse, 2010).

4.7.2 The Higher School Certificate (HSC)

The Higher School Certificate (HSC) is the highest school-based qualification that a student receives in the state of New South Wales, Australia (NSW) after the completion of thirteen years of education. Students choose subjects that have been developed or endorsed by the Board of Studies, Teaching and Educational Standards NSW (BOSTES) and complete these subjects over a two year period from Grade 11. For each subject, students must complete a range of assessment tasks that are set and marked by the school, along with a final external examination, which is set and marked under standardised conditions with quality assurance protocols (BOSTES, 2010, 2013b). Students receive an overall HSC mark for each subject they complete, which combines their school assessment marks and the external examination marks (BOSTES, 2011, 2013a). In 2013, the BOSTES reported that 74,276 students received HSC results (BOSTES, 2013a). The results of the HSC are used by universities as the primary basis for admission, via the Universities Admissions Centre. In NSW, the HSC results are perhaps the most important indicator of school achievement.

4.7.3 The School Certificate (SC)

The School Certificate (SC) is a recently retired qualification given to NSW students after the completion of Grade 10 and prior to the commencement of HSC studies. The SC consisted of five mandatory external exams that assessed students on their knowledge of content taught over a two year period in English, mathematics, science, computing skills, and humanities (history and geography) from Grade 9. The SC credential was designed to ensure that students leaving high school without completing Grades 11 and 12 (the HSC) would

have a credential that measures their achievement by state standards, and to allow for ease of comparison by employers and educational institutions. From 2012, the School Certificate was retired (Patty, 2011), as it had become redundant due to the raising of the legal age at which students could leave school to 17. While in use, it was considered to be the best instrument for "marking the end of junior secondary schooling" (NSW Department of Training and Education Co-ordination, 1997, p. 33). In 2011, the BOSTES reported that 90,491 students completed the SC (BOSTES, 2011).

4.7.4 National Assessment Program—Literacy and Numeracy (NAPLAN)

The National Assessment Program—Literacy and Numeracy (NAPLAN) was introduced in 2008 to monitor student progress in reading, writing, language (spelling, grammar and punctuation), numbers, patterns, algebra, measurement, data, space, and geometry compared to a set of national minimum standards. According to the Australian Curriculum Assessment and Reporting Authority (ACARA), the administering body, "NAPLAN is not a test of content. Instead, it tests skills in literacy and numeracy that are developed over time through the school curriculum" (ACARA, 2016). All students in Australia sit the external NAPLAN examinations under standardised conditions in Grades 3, 5, 7, and 9. ACARA follows strict procedures to ensure consistent marking and examination conditions (ACARA, 2016).

4.7.5 School Assessment (SA)

School Assessment (SA) is perhaps the most common method used to evaluate a student's achievement (Reis & McCoach, 2000). While school assessments may have low and unknown levels of reliability, and varying levels of reliability for each task (MacCann & Stanley, 2010), they are considered useful as they: (a) may involve multiple types of tasks (not exclusively examinations), (b) are acquired over a long period of time, (c) generally provide students with opportunities to demonstrate their developed skills and knowledge, and

(d) are directly linked to the courses and experiences in the classroom (Reis & McCoach, 2000; William, 2001, 2003). Hence, using a student's final school assessment results may produce a much more valid, reliable, and accurate measure of a student's current achievements than data obtained from other instruments (William 2001). Furthermore, it is possible that the benefits of using school assessment results may outweigh the benefits obtained by standardisation (Kane, 2006).

4.8 Description of the Archive Data

The statistical methods that identify and measure gifted underachievement combine the measurements of expected achievement and actual achievement into a single variable. As there are several different instruments that produce expected achievement and actual achievement measurements, a large number of unique combinations of expected achievement and actual achievement measurements are possible. This section provides some key descriptive information on the data included in this project, including: (a) an outline of the time period that each instrument was used at the school, (b) the groups that were administered each instrument, and (c) descriptive statistics on the data obtained (i.e., mean, range, sample size, histogram). At the end of this section, the number of gifted students for each combination of expected achievement/actual achievement data is reported.

4.8.1 The Otis–Lennon School Ability Test (OLSAT)

The OLSAT has been used at the school from 2006 until the present time. The instrument is primarily used to assist with class placement of students as they enter high school. Hence, every student sits the OLSAT when they enter high school (either during Grade 6, or during the enrolment process). The school currently has records for 2,501 students, who have an average School Ability Index (SAI) of 102.1, with a standard deviation of 14.7, and range from 50 to 149. The only cases where a student would not complete the

OLSAT, is when the results from an individual intelligence test (e.g., the WISC or SB) are available for that student, or if enrolment occurred during the senior years (Grades 11 or 12). Figure 15 outlines a histogram showing the distribution of OLSAT scores collected from the archive data.

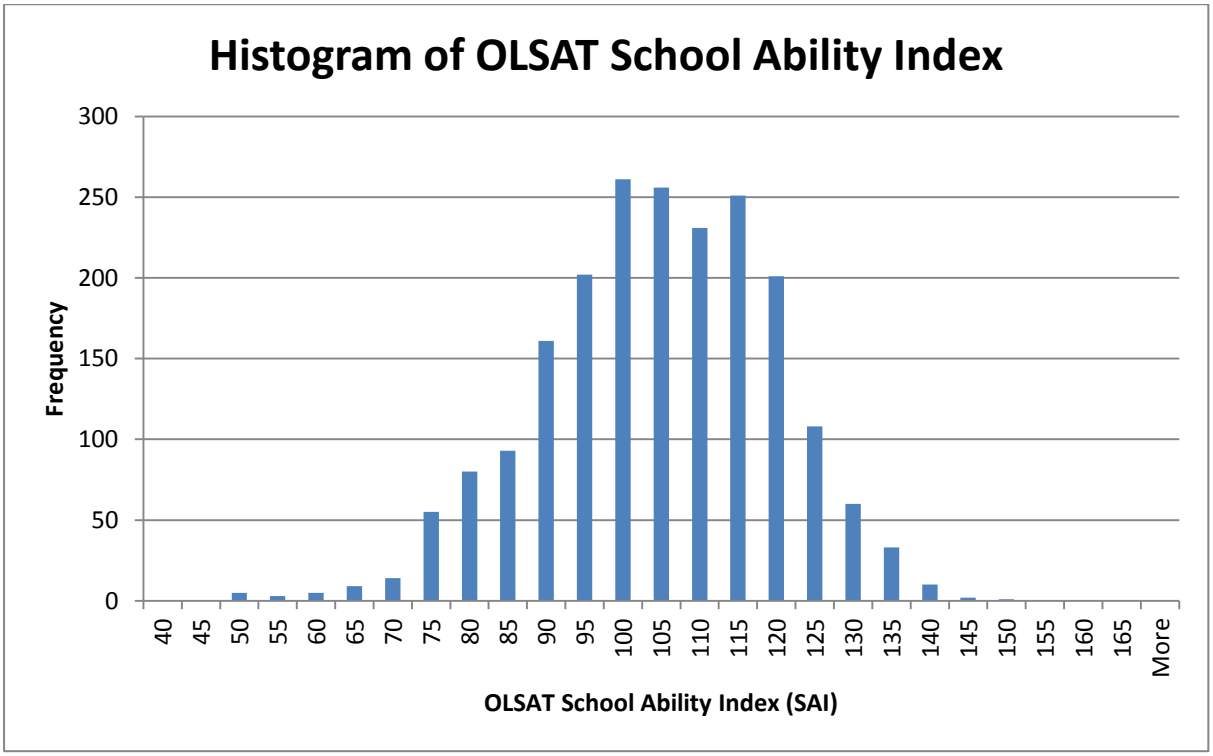


Figure 15. Histogram of all student OLSAT School Ability Index scores

4.8.2 The Higher School Certificate (HSC)

Although the school has always participated in the HSC examinations, its electronic archives only contain student results from 2002. For the period that records exist, 1,267 students completed HSC examinations at the school. While students who complete the HSC are able to choose all of their subjects from a wide selection, English must be chosen by students who intend to pursue tertiary education and the school has historically required that mathematics be completed. The focus of this investigation was therefore on English and mathematics subjects, as they consistently had high enrolment rates, which ensured an

adequate sample size. The descriptive statistics for one English subject and two mathematics subjects are provided in Table 5. Subjects that gifted students are least likely to select (i.e., Standard English and General Mathematics) were excluded from the analysis. Figures 16 and 17 show histograms of the distribution of English and mathematics HSC scores collected from the archive data.

Table 5

Descriptive statistics for HSC data

Course	Mean (%)	Range (%)	Sample size
Advanced English	75	37–92	613
Mathematics	73	19–97	412
Extension I Mathematics	71	10–98	215

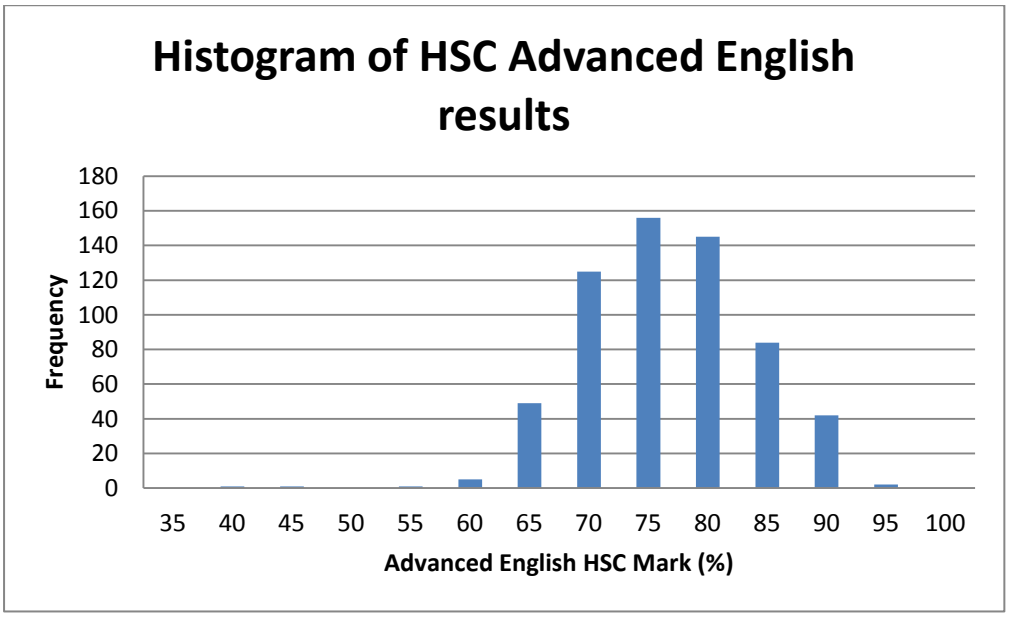


Figure 16. Histogram of all student HSC Advanced English marks

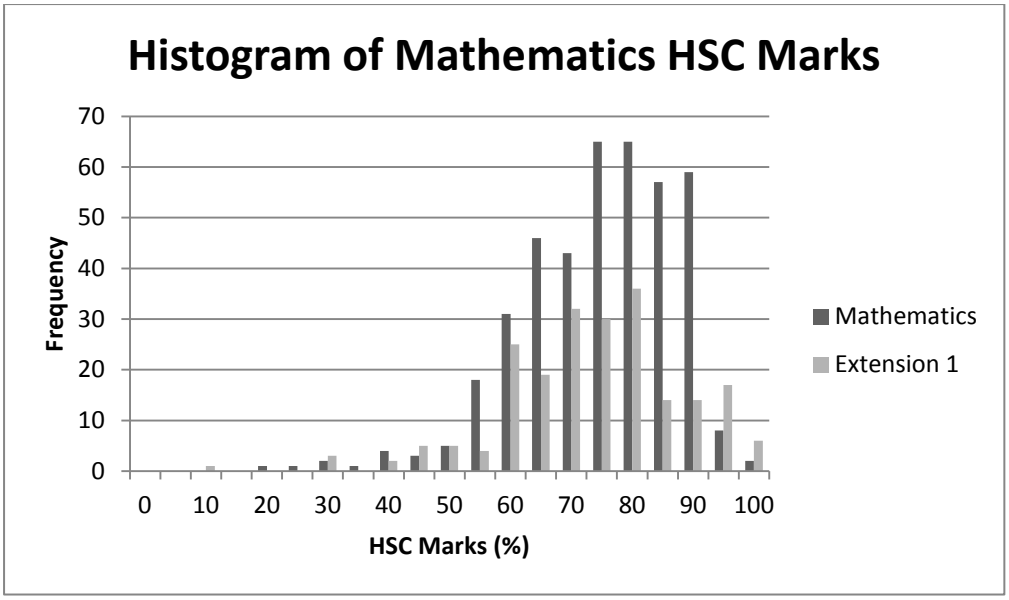


Figure 17. Histogram of all student HSC Mathematics and Extension I Mathematics marks

4.8.3 The School Certificate (SC)

The school archives contained student SC results from 2001 to 2011. In this time period, 1,221 students completed SC examinations at the school. The descriptive statistics for the English and mathematics results are shown in Table 6, while Figures 18 and 19 show histograms of this data.

Table 6

Descriptive statistics for SC data

Course	Mean (%)	Range (%)	Sample size
English	77	23–95	1207
Mathematics	74	40–100	1202

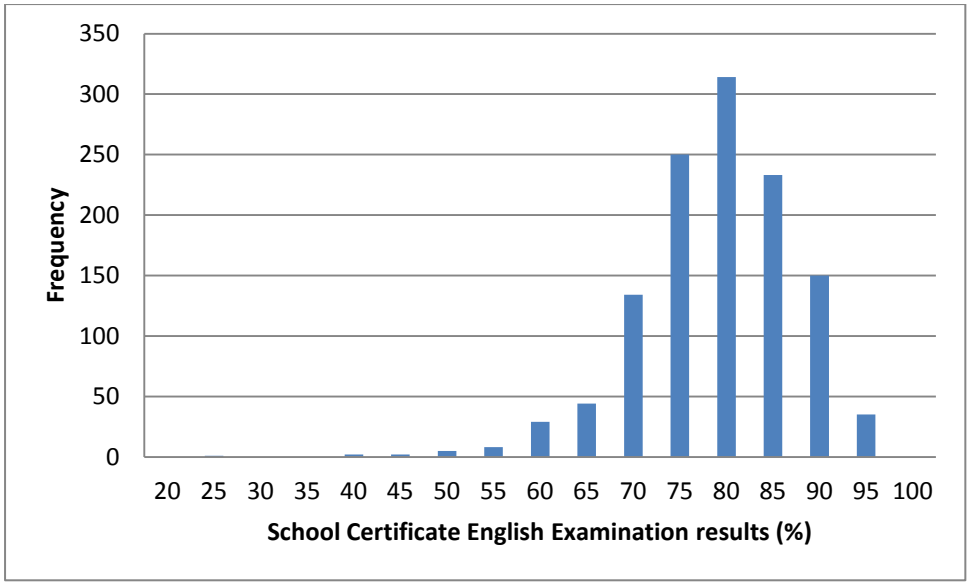


Figure 18. Histogram of SC English data

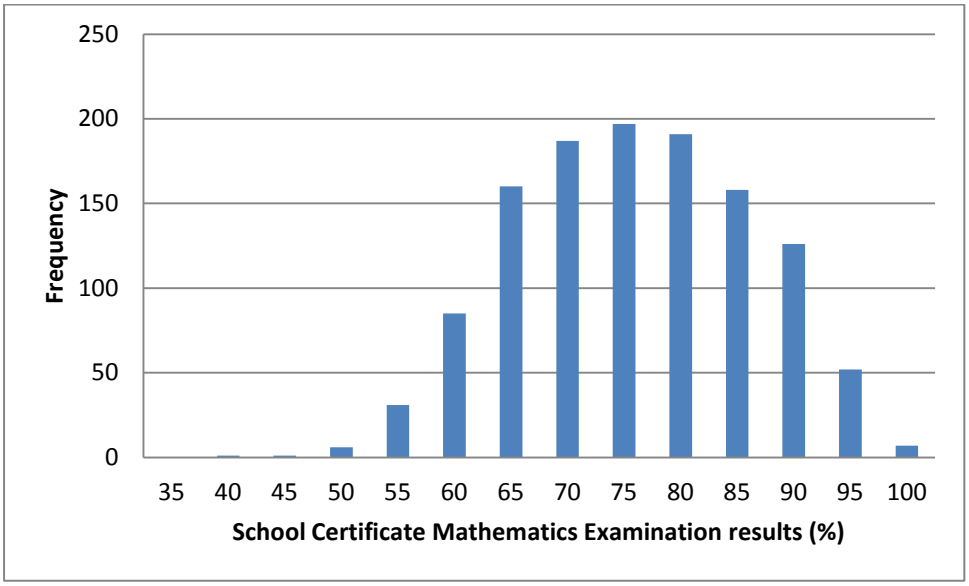


Figure 19. Histogram of SC Mathematics data

4.8.4 National Assessment Program—Literacy and Numeracy (NAPLAN)

The school's records include NAPLAN results from 2008. For the period that records exist, 2,599 students completed literacy examinations and 2,564 students completed

numeracy examinations. The descriptive statistics of the literacy and numeracy results are noted in Table 7, while cumulative histograms for each set of data are shown in Figures 20 and 21.

Table 7

Descriptive statistics for NAPLAN data

Grade	Mean (score)		Range (score)		Sample size	
	Literacy	Numeracy	Literacy	Numeracy	Literacy	Numeracy
Year 3	437	424	226–592	214–666	598	593
Year 5	512	509	327–672	340–748	620	611
Year 7	552	558	360–741	361–786	708	702
Year 9	594	612	386–743	416–874	673	658

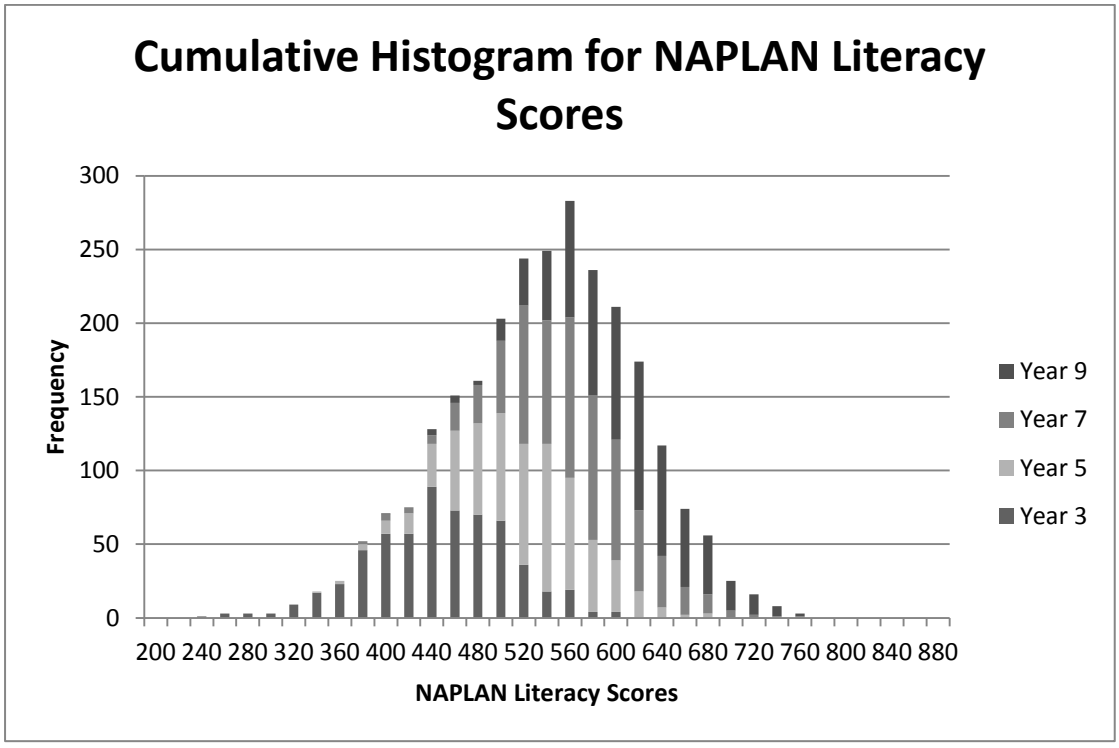


Figure 20. Cumulative histogram for NAPLAN Literacy scores

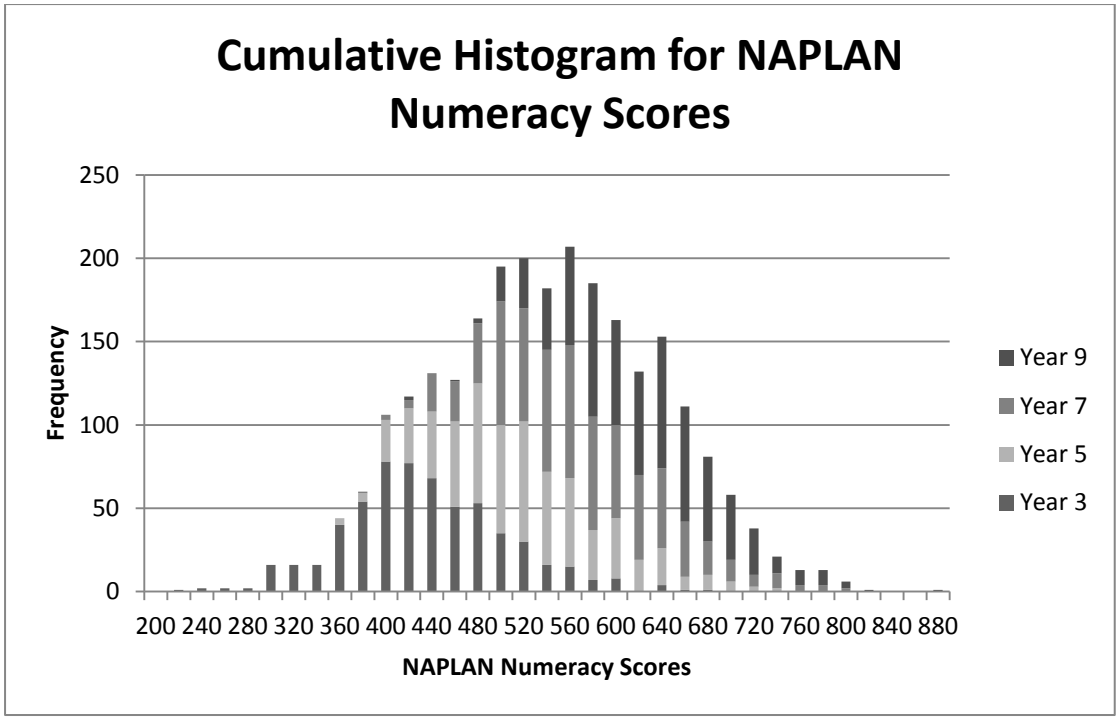


Figure 21. Cumulative histogram for NAPLAN Numeracy scores

4.8.5 School Assessment (SA)

Data for school assessments were available for the period from 2002 to 2012. While the data comprised a total of 367,567 individual assessment task marks, they were combined using school-specified weights to form weighted average marks in English and mathematics for each semester. This reduced the total number of records to 56,448. Thereafter, following the recommendation by Goldschmidt et al. (2007) and Messick (1996) to increase the validity of these measurements by taking the average, these marks were averaged over the junior high school years (Grades 7–10) and the senior high school years (Grades 11–12). This further reduced the total number of records to 6,511. The descriptive statistics for these data are included in Table 8 and a histogram for each set of data is shown in Figures 22 to 25.

Table 8

Descriptive statistics for School Assessment data

Course	Mean (%)	Range (%)	Sample size
Junior English	59	5–96	2,054
Junior Mathematics	66	0–98	1,934
Senior English	59	7–95	1,268
Senior Mathematics	59	4–96	1,255

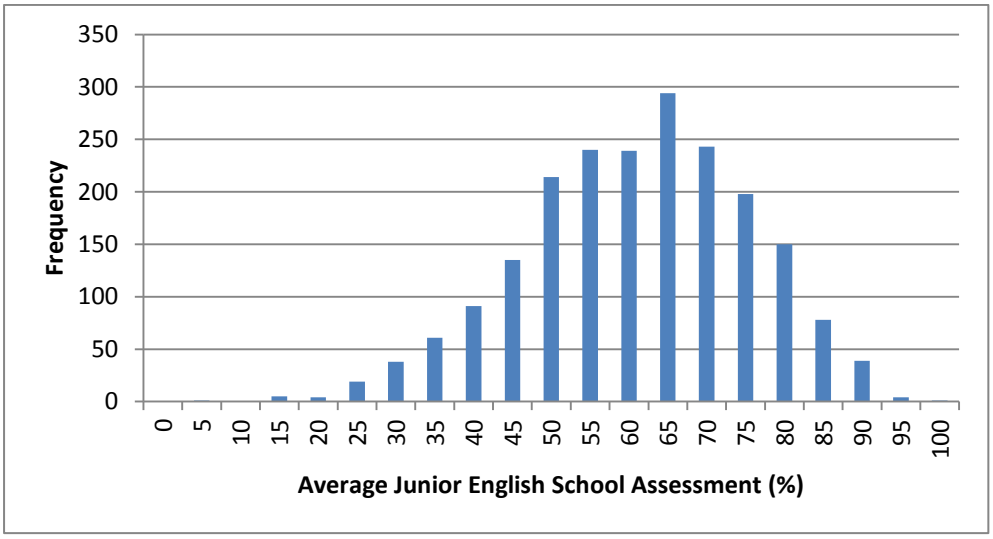


Figure 22. Histogram of Junior English SA data

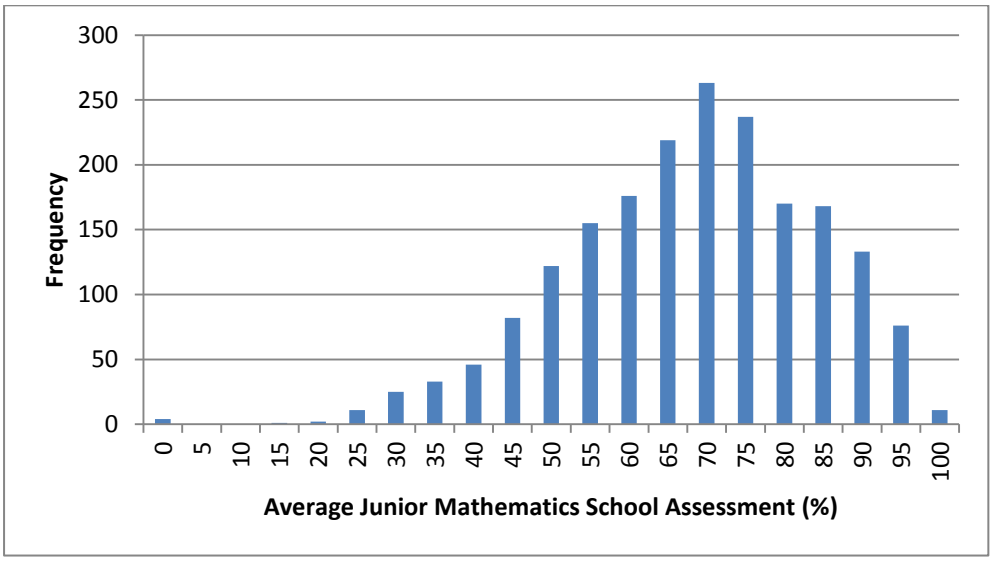


Figure 23. Histogram of Junior Mathematics SA data

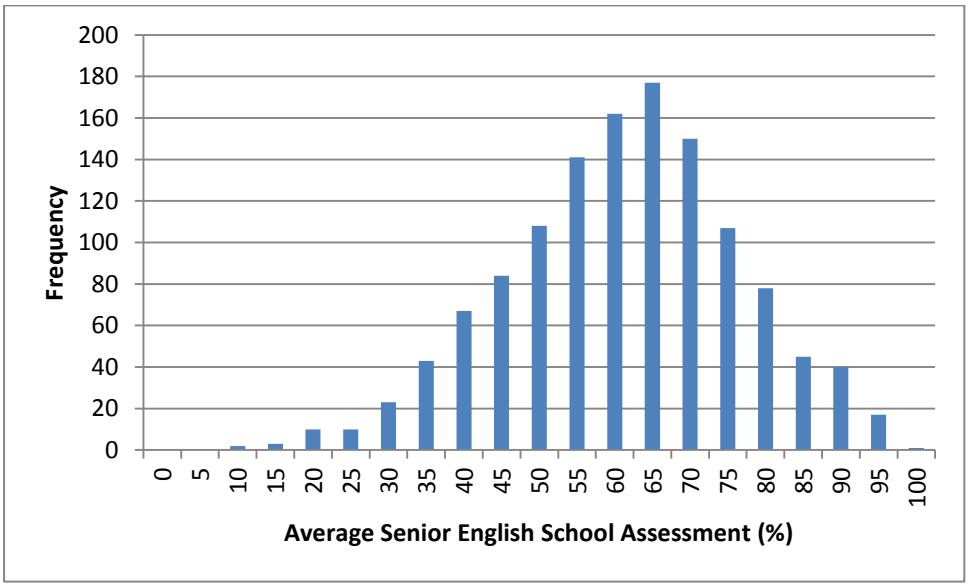


Figure 24. Histogram of Senior English SA data

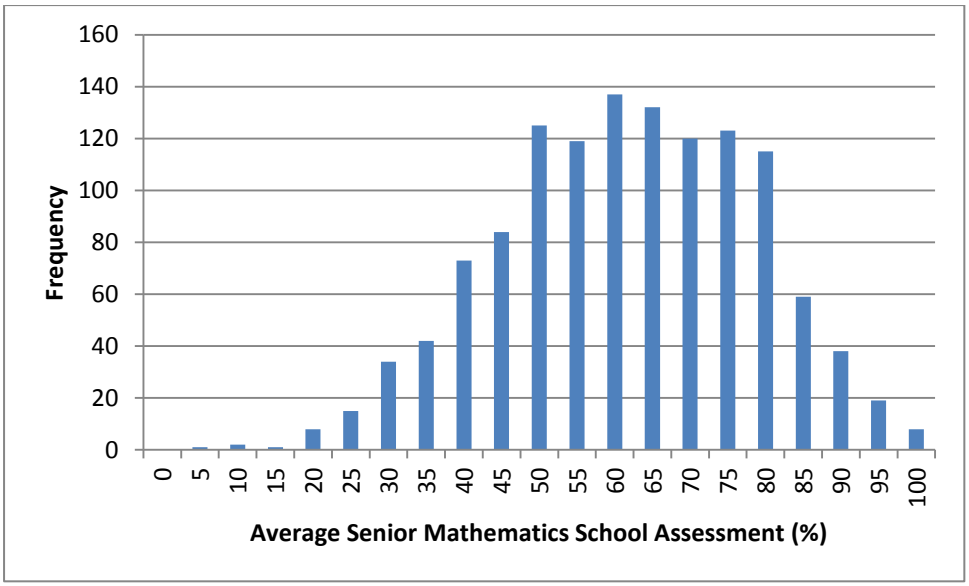


Figure 25. Histogram of Senior Mathematics SA data

4.9 Research Instrument

The data in the school archives did not include any nominations of gifted underachievement. Therefore, a survey instrument was needed to collect appropriate nomination data. A survey instrument was chosen as it allowed for a small amount of information to be collected in an effective and efficient manner from all high school teachers at the school.

4.9.1 Development of the survey

In a review of the literature, no published instruments that showed how researchers requested teachers to nominate gifted students who were underachieving could be identified. Nevertheless, most researchers indicated that teachers were asked to classify each student on their class list as achieving or underachieving (Annesley, 1970; Carr et al., 1991; Jones & Myhill, 2004; Lee-Corbin & Evans, 1996), while some required the teachers to provide reasons for their nominations (Dunne & Gazeley, 2008; Lau & Chan, 2001a; Sharp, Kendall, & Schagen, 2003). These two elements of the existing literature were incorporated into the survey instrument that was developed for this project, which was essentially designed to reflect common practices used in the literature with respect to teacher nominations of gifted underachievers (i.e., the use of simple instruments that comprise few elements).

4.9.2 Presentation of the survey

The survey was presented as a self-administered online form. The questions in the survey asked teachers to: (a) nominate a gifted student who had achieved significantly below their potential in the past semester, (b) the class the student was in, and (c) the reasons for their nomination. To encourage a high response rate, the survey was designed to be clear, quick, and simple for each teacher to complete (Fraenkel & Wallen, 2006). Teachers were invited to complete the survey multiple times, as necessary, to allow all gifted students they

taught, and had exhibited underachievement in the past semester, to be nominated. The survey may be found in Appendix 7.

4.9.3 Data collection

Multiple scholars have questioned the validity of nominations from teachers who have no training to distinguish between gifted achievement and gifted underachievement (Dunne & Gazeley, 2008; Jones, 2005; Schultz, 2002). Hence, some researchers (Fisher, 2005; Kanevsky & Keighley, 2003) have made a practice of providing guidance to teachers on gifted underachievement before they are asked to classify students. To increase the validity of the teacher nominations in this project, the investigator provided an accredited professional development presentation on gifted underachievement to all high school staff at the selected school. This presentation, which occurred on the 14th of July, 2015 during a professional development week, was timely as the staff had recently completed their half yearly reports for students in their classes. The presentation outlined Gagné's differentiated model of giftedness and talent (Gagné, 1998; 2009a; 2009b; 2013), the nature and problem of gifted underachievement, and the specific ways in which to identify gifted students who exhibit underachievement (Reis & McCoach, 2000). At the conclusion of the presentation, staff were invited to participate in the research project. The participant information statement and the email invitation sent to each staff member may be found in Appendices 6 and 8, respectively.

To supplement the nomination data, additional archive data sets were collected from the school assessment records in 2015. These data were necessary to allow the results of the teacher nominations (which were collected in 2015) to be compared to the results from the statistical methods. Moreover, to ensure that a suitable sample size was achieved, high school teachers of all subjects were invited to participate. Therefore, further school assessment data were also collected from each of these subjects to compare the teacher nomination data to the

results of the statistical methods. The additional data included 1,483 student scores, with a mean of 67% and a range of 6% to 100%. A histogram of these data is shown in Figure 26.

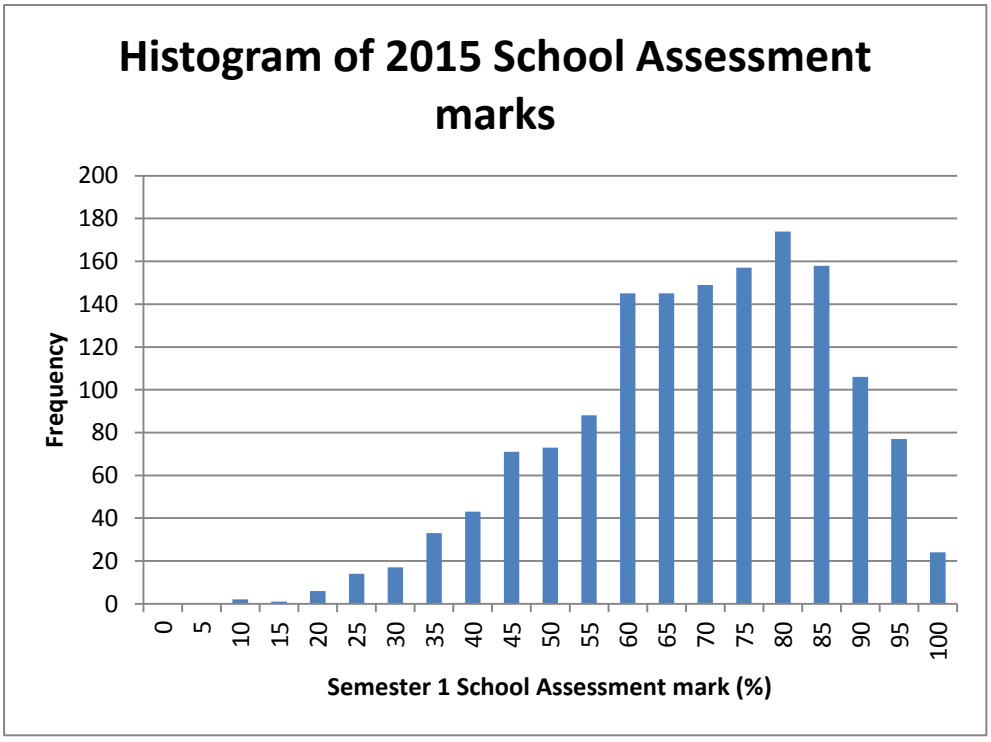


Figure 26. Histogram of additional school assessment marks

4.9.4 Screening of data

Following the practice of researchers who screen teacher nominations by excluding any nominations that did not include reasons, or included inappropriate reasons (Dunne & Gazeley, 2008; Lau & Chan, 2001a; Sharp et al., 2003), a screening process was applied to the 122 teacher nominations of gifted underachievement that were received. Not one of the received nominations was discarded after a thorough inspection of all reasons provided for the nominations. Generally, the participating teachers cited observations of common characteristics of underachieving gifted students (79% of responses) and/or made comparisons of measured ability with observed or measured achievement of the students (29% of responses). The teachers nominated students in each grade across 79 different high

school classes in each of the teaching areas in the school. On average, the participating teachers nominated 1.4 gifted students as underachieving in each class. As the remaining students in each class were assumed to be classified as “not underachieving”, the survey administration resulted in a total of 1,505 teacher classifications of students.

4.10 Data Preparation

4.10.1 Standardisation

The first step that was carried out in the preparation of the data for analysis was to convert each measured score into standardised units, with a mean of zero and a standard deviation of one. Standardisation was possible for many instruments as they have published means and standard deviations across large populations. These publications include the technical manual for the OLSAT (Harcourt Educational Measurement, 2003), *Report on the Scaling of the HSC* (UAC, 2001–2014), the BOSTES *Results Analysis Package* for SC results (BOSTES, 2001–2011), and the *National Reports* on NAPLAN results (ACARA, 2008–2013). For each of these instruments (except OLSAT), standardisation was applied to each subgroup with a published mean and standard deviation (e.g., for each year, grade, subject, etc.).

The school assessments were the only set of data that did not have published means and standard deviations. The school reports database that was used to extract school assessments included 445 unique courses (i.e., unique by year, grade, subject, and semester). The mean scores across these courses varied considerably, with a range of 34–93%, an average of 61%, and a standard deviation of 8.5%. As a result, a decision was made to standardise scores for each course based on the mean and standard deviation specific to each course. As noted previously, the scores for each student were then averaged across the

multiple grades and semesters to produce an average score in English and mathematics for both junior (Grades 7–10) and senior (Grades 11–12) years.

4.10.2 Selecting combinations of expected and actual achievement

Each of the instruments selected to measure expected achievement or actual achievement provided multiple sets of measurements. For example, the OLSAT provides three sets of measurements (i.e., School Ability Index, Verbal score, and Non-Verbal score), while the NAPLAN provides two sets of measurements (i.e., Literacy score and Numeracy score). In total, eleven sets of measures exist to assess expected achievement and ten sets of measures exist to assess actual achievement (refer Table 9). An instrument was only considered able to measure expected achievement if it was measured before the actual achievement was measured. For example, the Higher School Certificate (HSC) could not be used as a measure of expected achievement as it is the final assessment of a student's achievement before they leave high school.

Table 9

All sets of measurements of expected achievement and actual achievement from the school archives

Expected achievement measurements		Actual achievement measurements	
Instrument	Subcomponents	Instrument	Subcomponents
OLSAT	SAI	NAPLAN	Literacy
	VS		Numeracy
	NV		
Prior NAPLAN	Literacy	HSC	English
	Numeracy		Mathematics
Prior SC	English	SC	English
	Mathematics		Mathematics
Prior Junior SA	English	Junior SA	English
	Mathematics		Mathematics
Prior Senior SA	English	Senior SA	English
	Mathematics		Mathematics

Note. SAI = School Ability Index; VS =Verbal score; NV = Non-verbal score

From all of the measures of expected and actual achievement, a total of 110 combinations of expected achievement and actual achievement measurements were possible. To reduce the total number of combinations studied to a more manageable size, only those combinations relating to the same type of content were included in the investigation. For example, combinations of expected achievement in numeracy and actual achievement in mathematics were included, but combinations of expected achievement in numeracy with actual achievement in English were excluded. With respect to analyses relating to the nomination method, as the additional school assessment results collected in 2015 related to multiple subjects, only the general School Ability Index (SAI) subcomponent from the OLSAT was used as the expected achievement measurement. Furthermore, the averaging of school assessments across multiple grades (i.e., across 7–10 to form the Junior School Assessment score) was not appropriate for comparison with the nomination data, as nomination data were only collected for a single semester (instead, a school assessment score was calculated as the weighted total score across all of assessments in the first semester of 2015). The result was a total of 41 sets of combinations of expected achievement and actual achievement measurements, which are shown in Table 10 along with the sample size of gifted students examined in each data combination.

4.10.3 Data-linkage across separate databases

As the data from each instrument were historically stored in separate databases, a process of data-linkage occurred to match data in the separate databases that came from the same students. As the data provided were de-identified, data-linkage was carried out using student ID numbers that were unique to the student and common across the different databases. Unfortunately, three older databases (circa 1990–2000, containing Ravens Progressive Matrices, Slossons Intelligence Tests, and state academic competitions results) needed to be excluded from this research as the student ID numbers used were not common

Table 10

Sample size for all combinations of expected achievement and actual achievement measurements studied

Instrument pair (i.e., expected achievement – actual achievement)	Combination	n
OLSAT–NAPLAN	SAI–Lit	390
	SAI–Num	387
	VS–Lit	328
	NV–Num	496
OLSAT–SC	SAI–E	39
	SAI–M	39
	VS–E	23
	NV–M	46
OLSAT–HSC	SAI–E	55
	SAI–M	24
	VS–E	36
	NV–M	29
NAPLAN–SC	Lit–E	39
	Num–M	76
NAPLAN–HSC	Lit–E	41
	Num–M	45
SC–HSC	E–E	69
	M–M	106
OLSAT–Junior SA	SAI–E	158
	SAI–M	158
	VS–E	123
	NV–M	214
OLSAT–Senior SA	SAI–E	75
	SAI–M	75
	VS–E	55
	NV–M	92
NAPLAN–Junior SA	Lit–E	164
	Num–M	318
NAPLAN–Senior SA	Lit–E	75
	Num–M	131
Junior SA–SC	E–E	75
	M–M	71
SC–Senior SA	E–E	76
	M–M	191
Junior SA–HSC	E–E	79
	M–M	49
Senior SA–HSC	E–E	46
	M–M	39
Junior SA–Senior SA	E–E	89
	M–M	77
OLSAT–SA	SAI–SA	290

Note. SAI = School Ability Index; VS =Verbal Score; NV = Non-verbal score; E = English; M = Mathematics; Lit = Literacy; Num = Numeracy

across the databases. The matching of student ID numbers was undertaken in *Excel* using the *VLOOKUP* function. Due to differences in the grades and years that the various instruments were administered, there was not a complete set of matched pairs of student results for each of the combinations of expected and actual achievement measurements.

4.10.4 Selection of gifted students

The model of giftedness adopted for this investigation was Gagné's (2009a, 2013) *Differentiated Model of Giftedness and Talent*. Reflecting the definition of giftedness in this model, gifted students were identified as those students who scored in the top 10% of their age peers in each of the specific measures of expected achievement used for each data combination. It is noteworthy that Gagné supports the use of intelligence tests, school grades, and standardised achievement tests for the identification of gifted students (Gagné, 2007). The sample size of gifted students identified in each data combination is provided in Table 10.

4.11 Meta-Analysis

The statistical tools used in this investigation produced results for each of the 41 combinations of expected achievement and actual achievement measurements, and for each of the methods (i.e., absolute split I, absolute split II, nomination, regression, and simple difference) used to identify and measure gifted underachievement (i.e., up to 410 results for each statistical method). Such a large volume of results may make it difficult to extract a single meaningful conclusion. To address this problem, it may be inappropriate to examine the simple averages of the results, as true patterns may disappear and false patterns based on pooling artefacts may appear (i.e., Simpson's paradox, Blyth, 1972; Tang, He & Tu, 2012). Thus, other, more sophisticated, methods may be needed to combine and simplify the large number of results.

A possible solution is meta-analysis, which may be considered to be an appropriate method for combining information from multiple sources (Hunter & Schmidt, 2004; Rosenthal & DiMatteo, 2001). Meta-analysis may be used to: (a) produce an overall estimate of an effect size by combining multiple studies, and (b) test the significance of an effect with a level of statistical power related to the combined sample size. For example, if several different researchers were examining the impact of breastfeeding on intelligence, a meta-analysis could combine these results to produce a single measurement of the effect of breastfeeding on intelligence and establish whether the effect is statistically significant (Franke, 2001). It is noted that other methods of combining information from multiple studies (e.g., making a comparison of the count of the number of studies where an effect was found to be significant, with those studies that found the effect to be insignificant), may actually reduce the statistical power and may be less accurate than the individual studies (Borenstein, Hedges, Higgins, & Rothstein, 2009). Moreover, it is noted that researchers in the field of gifted education have recently been encouraged to use meta-analysis techniques to increase the precision and reliability of research findings and to answer unique questions (Steenbergen-Hu & Olszewski-Kubilius, 2016). Therefore, in this investigation, meta-analytical methods were used.

The calculations required in a meta-analytical approach to determine an overall effect size from multiple studies are completed over multiple steps (Borenstein, 2009). First, the weighting of each study must be determined. In a meta-analysis, each study is weighted according to the precision and sample size of the study, with the weighting being equal to the variance of the effect size (Borenstein et al., 2009). Second, the weighting is applied by multiplying the effect size from each study by its variance. Third, all of the weighted effect sizes are summed. Fourth, the variances from each study are added to calculate the total weighting of the multiple studies. Last, the sum of weighted effect sizes is divided by the

total weighting of the multiple studies to produce the overall weighted average effect size for the multiple studies. The calculations relating to each of these steps are outlined in Appendix 12.

4.12 Application of the Methods to Identify/Measure Gifted Underachievement

Each of the five methods for the identification/measurement of gifted underachievement were applied to all 41 combinations of expected achievement and actual achievement measurements. Nevertheless, it is noted that the absolute split II method could not be applied to the OLSAT-NAPLAN instrument pair, as the NAPLAN instrument does not publish its maximum score to allow for the calculation of a percentage score. Furthermore, as previously discussed, the nomination method could only be assessed with newly collected teacher nomination data, as the school archives did not contain any teacher nomination data. The total number of cases of gifted underachievement identified for each method for each combination of measurements, and over all data combinations, is summarised in Table 11.

The lines of best fit that were used for the regression method for each of the combinations of expected achievement and actual achievement measurements are described in Table 12. Each line of best fit is described by the slope of the line (the gradient, m), the y-intercept (b), and a measure of how well the line of best fit actually fits the data (correlation, r). It is noteworthy that the correlation between the expected achievement and actual achievement measurements varies from small ($r = 0.17$) to very large ($r = 0.83$), although the average correlation ($r = 0.51$) may be considered large (Field, 2013). Furthermore, almost all slopes of the line of best fit (m) are positive and less than one, and most (71%) of the y-intercepts (b) are negative. These observations indicate that the students, on average, have lower levels of actual achievement than expected achievement.

Table 11
Summary of the total number of cases of gifted underachievement identified for each method of identification applied to each pair of measurements

Instrument pair	Combination	ABSI	ABSII	SD	REG	NOM
OLSAT–NAPLAN	SAI–Lit	107	-	109	63	-
	SAI–Num	36	-	43	50	-
	VS–Lit	82	-	65	44	-
	NV–Num	70	-	82	63	-
OLSAT–SC	SAI–E	9	4	6	5	-
	SAI–M	5	7	4	5	-
	VS–E	2	1	3	2	-
	NV–M	7	8	8	5	-
OLSAT–HSC	SAI–E	50	33	45	9	-
	SAI–M	21	14	21	3	-
	VS–E	31	20	29	7	-
	NV–M	26	16	28	2	-
NAPLAN–SC	Lit–E	4	2	5	7	-
	Num–M	4	4	8	7	-
NAPLAN–HSC	Lit–E	33	18	35	4	-
	Num–M	41	28	42	6	-
SC–HSC	E–E	50	30	47	12	-
	M–M	92	65	96	14	-
OLSAT–Junior SA	SAI–E	50	120	59	21	-
	SAI–M	58	56	58	24	-
	VS–E	31	86	35	14	-
	NV–M	86	87	94	29	-
OLSAT–Senior SA	SAI–E	45	62	46	12	-
	SAI–M	38	47	43	12	-
	VS–E	31	34	29	7	-
	NV–M	49	63	55	16	-
NAPLAN–Junior SA	Lit–E	24	93	22	21	-
	Num–M	104	107	147	51	-
NAPLAN–Senior SA	Lit–E	35	51	30	6	-
	Num–M	67	80	82	21	-
Junior SA–SC	E–E	4	1	0	8	-
	M–M	2	4	1	4	-
SC–Senior SA	E–E	37	38	29	12	-
	M–M	89	123	87	36	-
Junior SA–HSC	E–E	57	23	49	6	-
	M–M	39	25	39	4	-
Senior SA–HSC	E–E	17	5	17	8	-
	M–M	24	12	21	4	-
Junior SA–Senior SA	E–E	23	44	12	4	-
	M–M	17	36	11	4	-
OLSAT–SA	SAI–SA	106	116	116	46	122
Total		1703	1563	1758	678	122

Note. SAI = School Ability Index; VS =Verbal Score; NV = Non-verbal score; E = English; M = Mathematics; Lit = Literacy; Num = Numeracy; ABSI = Absolute Split I; ABSII = Absolute Split II; SD = Simple Difference; REG = Regression; NOM = Nomination

Table 12
Regression fits used for the regression method of identification/measurement of gifted underachievement

Instrument pair	Combination	Regression Fit		
		m	*b*	*r*
OLSAT–NAPLAN	SAI–Lit	0.52	0.20	0.62
	SAI–Num	0.77	0.21	0.69
	VS–Lit	0.51	0.24	0.59
	NV–Num	0.69	0.18	0.66
OLSAT–SC	SAI–E	0.48	0.27	0.60
	SAI–M	0.74	0.21	0.72
	VS–E	0.55	0.29	0.65
	NV–M	0.66	0.19	0.70
OLSAT–HSC	SAI–E	0.32	–0.85	0.25
	SAI–M	0.28	–1.00	0.19
	VS–E	0.33	–0.83	0.25
	NV–M	0.29	–1.07	0.23
NAPLAN–SC	Lit–E	0.74	0.17	0.78
	Num–M	0.74	0.09	0.83
NAPLAN–HSC	Lit–E	0.52	–1.02	0.35
	Num–M	0.18	–0.90	0.17
SC–HSC	E–E	0.74	–1.17	0.44
	M–M	0.22	–0.82	0.20
OLSAT–Junior SA	SAI–E	0.57	–0.09	0.56
	SAI–M	0.61	–0.10	0.62
	VS–E	0.57	–0.04	0.54
	NV–M	0.55	–0.14	0.60
OLSAT–Senior SA	SAI–E	0.22	–0.11	0.20
	SAI–M	0.35	–0.04	0.34
	VS–E	0.24	–0.10	0.21
	NV–M	0.32	–0.05	0.33
NAPLAN–Junior SA	Lit–E	0.84	–0.15	0.73
	Num–M	0.38	–0.02	0.72
NAPLAN–Senior SA	Lit–E	0.40	–0.10	0.31
	Num–M	0.35	–0.04	0.37
Junior SA–SC	E–E	0.69	0.21	0.79
	M–M	0.92	0.19	0.72
SC–Senior SA	E–E	0.53	–0.19	0.42
	M–M	0.47	–0.15	0.51
Junior SA–HSC	E–E	0.97	–1.29	0.60
	M–M	0.94	–1.66	0.44
Senior SA–HSC	E–E	0.86	–0.68	0.82
	M–M	1.21	–1.54	0.68
Junior SA–Senior SA	E–E	0.58	–0.11	0.54
	M–M	0.53	–0.04	0.47
OLSAT–SA	SAI–SA	0.45	–0.24	0.36

Note. SAI = School Ability Index; VS =Verbal Score; NV = Non-verbal score; E = English; M = Mathematics; Lit = Literacy; Num = Numeracy

Table 13

Mean measurements of gifted underachievement

Instrument pair	Data	SD	REG
OLSAT–NAPLAN	SAI–Lit	0.66	0.07
	SAI–Num	0.09	–0.12
	VS–Lit	0.53	–0.04
	NV–Num	0.28	–0.11
OLSAT–SC	SAI–E	0.55	0.00
	SAI–M	0.32	0.17
	VS–E	0.48	0.03
	NV–M	0.51	0.18
OLSAT–HSC	SAI–E	2.05	0.08
	SAI–M	2.17	0.03
	VS–E	2.00	0.09
	NV–M	2.28	–0.08
NAPLAN–SC	Lit–E	0.36	0.21
	Num–M	0.44	0.10
NAPLAN–HSC	Lit–E	1.73	–0.12
	Num–M	2.51	–0.02
SC–HSC	E–E	1.50	–0.09
	M–M	2.18	–0.12
OLSAT–Junior SA	SAI–E	0.88	0.08
	SAI–M	0.88	0.16
	VS–E	0.72	–0.02
	NV–M	1.00	0.11
OLSAT–Senior SA	SAI–E	1.40	–0.04
	SAI–M	1.21	0.08
	VS–E	1.33	–0.03
	NV–M	1.35	0.10
NAPLAN–Junior	Lit–E	0.50	0.14
	Num–M	0.96	0.14
NAPLAN–Senior	Lit–E	0.96	–0.18
	Num–M	1.28	0.04
Junior SA–SC	E–E	0.24	–0.05
	M–M	–0.16	–0.13
SC–Senior SA	E–E	0.90	–0.01
	M–M	1.13	0.05
Junior SA–HSC	E–E	1.24	–0.13
	M–M	1.64	–0.13
Senior SA–HSC	E–E	0.83	–0.13
	M–M	1.17	–0.07
Junior SA–Senior	E–E	0.51	–0.33
	M–M	0.50	–0.31
OLSAT–SA	SAI–SA	1.29	0.25
Weighted average		0.82	0.00

Note. SAI = School Ability Index; VS =Verbal Score; NV = Non-verbal score; E = English; M = Mathematics; Lit = Literacy; Num = Numeracy

Table 13 reports the average measurements of gifted underachievement, for each of the 41 data combinations, using the two methods (i.e., regression and simple difference methods) that produce measurements of the degree of gifted underachievement. A weighted average measurement of gifted underachievement was calculated for both methods across all of the data combinations following the procedure discussed and outlined in Appendix 12. Interestingly, the weighted average measurement of gifted underachievement from the simple difference method is close to the threshold used for identifying gifted underachievement (one standard deviation), while the weighted average measurement from the regression method is zero. These differences will be more formally investigated in the following chapter.

4.13 Summary

This chapter has outlined the methodology followed for this investigation. It has discussed how the research was designed to answer the research questions using data collected from a sample of the target population and how the data was prepared for analysis. The following chapters outline the data analyses and the results of these analyses.

5 Convergence Evidence

5.1 Introduction

This chapter begins to answer the first research question: "*Is the extrapolation inference reasonable for each of the methods to identify and measure gifted underachievement?"* It was determined in Chapter 4 that to answer this question, convergence evidence and criterion evidence should be provided. This chapter outlines the data analyses and results that form the convergence evidence.

5.2 Outline of Chapter

There are three major approaches that researchers use to assess convergence:

- Proportions: An examination of whether similar proportions of classifications are made between two sets of results (Ho et al., 2014). If the difference between proportions identified are large and statistically significant, convergence is not supported.
- Association or correlation: A measurement of the degree to which two variables are related (Lau & Chan, 2001a). If the association or correlation between variables is small or not significant, convergence is not supported.
- Agreement: A measurement of the degree to which two variables are equal (Bland & Altman, 1999; Agresti, 2013; Hanneman, 2008). If the agreement is weak, convergence is not supported.

Following Zaki et al. (2012), who propose that comparisons of proportions, correlation, and agreement are *all* used as convergence evidence, all three approaches will be used to gather convergence evidence in this investigation.

To provide convergence evidence to assess whether the extrapolation inference is reasonable for any particular method used to identify and measure gifted underachievement, the degree of convergence between all of the methods used to identify and measure gifted underachievement (i.e., the two variations of absolute split, nomination, regression, and simple difference) will need to be examined. It is noted that, due to differences in data type, the statistical analyses required to assess the convergence of the *identifications* of gifted underachievement will be different to those required to assess the convergence of the *measurements* of the degree gifted underachievement from the different methods (Agresti, 2013; Hanneman, 2008). Specifically, the identification results are considered to be a set of classifications (i.e., gifted underachievement or gifted achievement) that take the form of dichotomous categorical data (also called binary or boolean data), while the measurement results take the form of continuous data that may take on any value.

While the two sets of analyses will be conducted in relation to each of the three approaches to providing convergence evidence for all of the methods used to identify/measure gifted underachievement, as the proportion approach is only applicable to identification data (and not to measurement data), this chapter will provide a total of five sets of statistical results to assess convergence. Each of these five sets of results will include statistical measurements and a statistical test of significance where relevant.

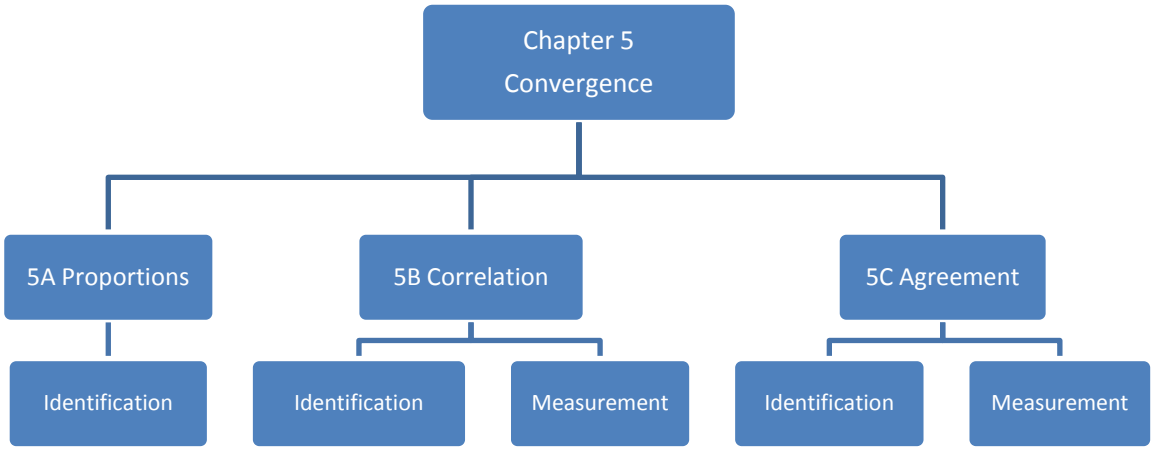


Figure 27. Outline of Chapter 5

Figure 27 provides details on how the chapter is organised. First, the chapter is divided into three major sections according to the three approaches for the assessment of convergence evidence (i.e., proportions of identification, association and correlation, and agreement). Second, each major section, except proportions of identification, will comprise separate analyses relating to convergence evidence for the identification results and the measurement results. A holistic discussion incorporating all of the convergence evidence occurs at the end of the chapter.

5.3 A Note on the Interpretation of Probability Values

As this investigation uses many statistical measurements and statistical tests to answer the research questions, correct interpretation of these statistical analyses is of paramount importance to this investigation. Recently, concern has been raised with the validity of current practices that primarily rely on probability values (*p*-values) to make decisions (Gelman & Loken, 2014; Goodman, 2008; Johnson, 2013; Lew, 2012; Nuzzo, 2014; Peng, 2015; Trafimow & Marks, 2015; Ziliak, 2010). In response, the American Statistical Association (ASA) has released a statement to address key issues surrounding the

misconception and misuse of probability values (*p*-values). Specifically, they provided the following six principles for the correct use and understanding of probability values:

1. *P-values can indicate how incompatible the data are with a specified statistical model.*
2. *P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.*
3. *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*
4. *Proper inference requires full reporting and transparency.*
5. *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.*
6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis. (Wasserstein & Lazar, 2016)*

Importantly, they also noted that "statistical significance is not equivalent to scientific, human, or economic significance " (Wasserstein & Lazar, 2016, p. 10). This is because statistical significance is highly dependent on the sample size and measurement precision achieved in a study (Field, 2013). Therefore, very small effects, of no practical significance, can be found to be statistically significant with a large enough sample size. Consequently, in the following chapters, where many statistical tests are used, in some cases with large sample sizes, an emphasis has been made on making decisions on the basis of the size of the measured effect, rather than primarily on the probability values. In addition, to provide full reporting and transparency, a comprehensive guide to equations used and additional results produced are also provided in the appendices.

5.4 Proportions

One approach to establishing that the different methods of identifying gifted underachievement are convergent is to assess whether they identify similar proportions of cases of gifted underachievement. This section examines the differences in proportions of gifted underachievement identified by the different methods, and determines whether the measured differences are large and statistically significant.

5.4.1 Analysis of proportions

For each method of identification, the proportion of identifications was calculated as the number of identified cases of gifted underachievement divided by the number of gifted students examined. In addition, the weighted average proportion of identification was calculated for each identification method across all data combinations following the meta-analysis methods previously described (and shown in Appendix 12). The proportions of identification are reported in Table 14, while the weighted average values are included in the final rows of Table 14 and in Figure 28. The proportion of cases identified as gifted underachievement varied from 0% to 91% of gifted students across the different data combinations and methods. The broad range of results possibly demonstrates why researchers to date have had difficulty in ascertaining the true rate of underachievement for gifted students. The weighted average proportions for each identification method had a narrower range (i.e., 13% for the regression method to 44% for the absolute split II method).

Table 14

Proportions of gifted students identified as underachieving

Instrument pair	Combination	ABSI	ABSII	SD	REG	Nom
OLSAT–NAPLAN	SAI–Lit	0.27	-	0.28	0.16	-
	SAI–Num	0.09	-	0.11	0.13	-
	VS–Lit	0.25	-	0.20	0.13	-
	NV–Num	0.14	-	0.18	0.13	-
OLSAT–SC	SAI–E	0.23	0.10	0.15	0.13	-
	SAI–M	0.13	0.18	0.10	0.13	-
	VS–E	0.09	0.04	0.13	0.09	-
	NV–M	0.15	0.17	0.17	0.11	-
OLSAT–HSC	SAI–E	0.91	0.60	0.82	0.16	-
	SAI–M	0.38	0.25	0.38	0.05	-
	VS–E	0.86	0.56	0.81	0.19	-
	NV–M	0.41	0.25	0.44	0.03	-
NAPLAN–SC	Lit–E	0.10	0.05	0.13	0.18	-
	Num–M	0.05	0.05	0.11	0.09	-
NAPLAN–HSC	Lit–E	0.80	0.44	0.85	0.10	-
	Num–M	0.54	0.37	0.55	0.08	-
SC–HSC	E–E	0.71	0.43	0.67	0.17	-
	M–M	0.55	0.39	0.57	0.08	-
OLSAT–Junior SA	SAI–E	0.32	0.76	0.37	0.13	-
	SAI–M	0.37	0.35	0.37	0.15	-
	VS–E	0.25	0.70	0.28	0.11	-
	NV–M	0.40	0.40	0.44	0.13	-
OLSAT–Senior SA	SAI–E	0.60	0.83	0.61	0.16	-
	SAI–M	0.51	0.63	0.57	0.16	-
	VS–E	0.56	0.80	0.53	0.13	-
	NV–M	0.53	0.68	0.60	0.17	-
NAPLAN–Junior SA	Lit–E	0.15	0.57	0.13	0.13	-
	Num–M	0.33	0.34	0.46	0.16	-
NAPLAN–Senior SA	Lit–E	0.47	0.69	0.40	0.08	-
	Num–M	0.50	0.60	0.62	0.16	-
Junior SA–SC	E–E	0.05	0.01	0.00	0.10	-
	M–M	0.03	0.06	0.01	0.06	-
SC–Senior SA	E–E	0.49	0.50	0.38	0.16	-
	M–M	0.47	0.64	0.46	0.19	-
Junior SA–HSC	E–E	0.70	0.28	0.60	0.07	-
	M–M	0.56	0.36	0.56	0.06	-
Senior SA–HSC	E–E	0.34	0.10	0.34	0.16	-
	M–M	0.37	0.18	0.32	0.06	-
Junior SA–Senior SA	E–E	0.26	0.49	0.13	0.04	-
	M–M	0.22	0.47	0.14	0.05	-
OLSAT–SA	SAI–SA	0.37	0.40	0.40	0.16	0.42
Weighted Average		0.32	0.44	0.33	0.13	0.42

Note. SAI = School Ability Index; VS =Verbal Score; NV = Non-verbal score; E = English; M = Mathematics; Lit = Literacy; Num = Numeracy; ABSI = Absolute Split I; ABSII = Absolute Split II; SD = Simple Difference; REG = Regression; NOM = Nomination

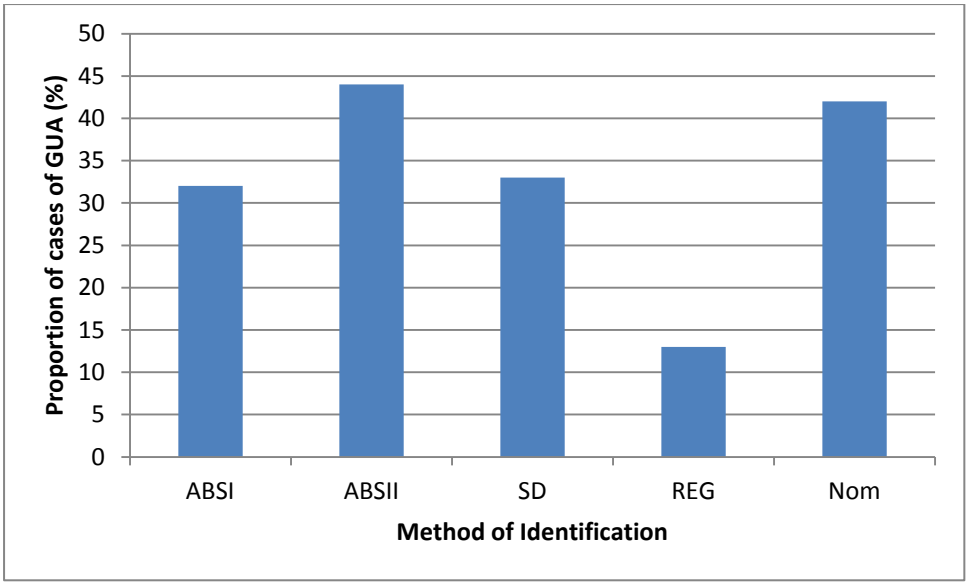


Figure 28. Proportion of cases of gifted underachievement from each method of identification

5.4.2 Comparisons of proportions

5.4.2.1 Cochran's Q test

Cochran's Q test (Cochran, 1950) may be used to statistically test whether the average proportion of gifted underachievement cases identified by each identification method is indeed different. The test is commonly used to compare the proportions of diagnoses from multiple diagnostic tests which are each applied to the same sample (Cohen et al., 2015; Fleiss, Levin, & Paik, 2003), and formally assesses the null hypothesis that the proportion of gifted underachievement cases identified is equal for each of the identification methods (i.e., H_0: $p_1=p_2=p_3=p_4=p_5$). Cochran's test requires the calculation of a Q test statistic (following the method outlined in Appendix 12), which measures the amount of variation between the different proportions. The Q test statistic is compared to a chi-square distribution to determine whether any differences in proportions measured are statistically significant.

The Q test statistic of 86 ($p < 0.001$) for the data suggested that a statistically significant difference does indeed exist between the weighted average proportions of

identifications from the five methods that identify gifted underachievement. Therefore, the conclusion that the average proportion of identifications from the five methods is *not* equal is supported. Nevertheless, as the difference in proportions of identification may only be due to the proportions from one or two methods, further investigations are necessary to determine if any pairs of the identification methods may be considered to have equal proportions of identification.

5.4.2.2 McNemar's test

McNemar's (1947) test is used to determine whether there is a statistically significant difference between two sets of dichotomous classifications made of the same group of people (Agresti, 2013; Nussbaum, 2015; Tang et al., 2012), and is therefore an appropriate test to determine whether any pairs of identification methods have different or equal proportions of identification (Roberts, Sheffield, McIntire, & Alexander, 2011). To allow a deeper analysis of the differences in proportions of gifted underachievement classifications across the different identification methods, the McNemar test was applied to every possible pairing of identification methods for each data combination.

The results of the McNemar test are reported in Table 15 and Table 16. Table 15 reports the comparisons of the statistical identification methods to each other for each of the 41 data combinations, along with the weighted averages of the difference in proportions across different data combinations for each pair of statistical identification methods. Table 16 reports the comparisons of the statistical identification methods to the nomination method for the only data combination on which comparisons were possible (i.e., OLSAT as the measure of expected achievement and school assessment as the measure of actual achievement).

In Table 15, each column of values represents the difference in proportion of cases identified as gifted underachievement between two methods of identification. A positive

value indicates that Method 1 identified a lower proportion of gifted underachievement cases than Method 2, while a negative value indicates that Method 2 identified a lower proportion of gifted underachievement cases than Method 1. As the weighted averages were calculated across both positive and negative values, it is possible that the calculation of the weighted averages may have "artificially" reduced the size of the difference in proportions. Therefore, the weighted average of the absolute values of the difference in proportions were also calculated (i.e., the measure reported in the final row of Table 15).

The observed difference in proportion of gifted underachievement classifications varied from –0.90 to 0.45. The McNemar test results showed that most (62%) of these individual differences in proportions for each data combination were statistically significant ($p < 0.05$) and therefore do not support convergence. Nevertheless, it is noted that only one of the four statistical methods of identification (i.e., the regression method) produced a significant difference in proportion when compared with the nomination method. Hence, the results appear to support the possibility of convergence between the nomination method and each of the other three statistical methods (i.e., the absolute split I, absolute split II, and simple difference methods).

The weighted average of the differences in the proportions of gifted underachievement identifications for each pair of statistical identification methods ranged from –0.34 to 0.05. The regression method appeared to be particularly different to the other methods as, on average, it identified between 20% to 34% less cases of gifted underachievement than the other three statistical methods. The weighted average difference in proportions between the other pairs of statistical methods had a narrower range (±0.05).

A comparison of the weighted averages of the differences in the proportions of gifted underachievement with the weighted averages of the absolute values of the differences in the

proportions of gifted underachievement for each identification method, showed that the latter statistic was larger. This suggested that the weighted averages of the differences in the proportions of gifted underachievement may be reduced, if there are substantial inconsistencies across the various data combinations in terms of which of each pair of investigated identification methods identified the larger proportion of gifted underachievement. Interestingly, the weighted averages of the absolute values of the differences in the proportions of gifted underachievement revealed that the absolute split II method may also be substantially different to the other statistical methods (e.g., on average, it identified between 17% and 35% more or less cases of gifted underachievement than the other statistical methods). Only the absolute split I and simple difference methods appeared to have a small difference in the weighted average of the absolute values of the differences in proportions of gifted underachievement.

When the McNemar test was carried out on the weighted average (and weighted average of the absolute) values, it is noteworthy that *all* of the values were statistically significant to indicate that convergence was not supported between any of the pairs of statistical methods. Nevertheless, these probability values must be interpreted cautiously (Wasserstein & Lazar, 2016). For example, despite the McNemar test showing that all of the weighted average differences in proportions of gifted underachievement were statistically significant ($p < 0.05$), the size of some of the differences may not have practical significance (Field, 2013; Wasserstein & Lazar, 2016). Therefore, the very small difference in the proportion of cases of gifted underachievement identified in the absolute split I and simple difference methods may possibly be suggestive of practical convergence.

Table 15

Difference in proportion and McNemar test results

Comparison of	Method 1:	ABSI			ABSII		SD
with	Method 2:	ABSII	SD	REG	SD	REG	REG
OLSAT–NAPLAN	SAI–Lit	-	0.01	–0.11*	-	-	–0.12*
	SAI–Num	-	0.02	0.04*	-	-	0.02*
	VS–Lit	-	–0.05*	–0.12*	-	-	–0.06*
	NV–Num	-	0.04*	–0.01	-	-	–0.06*
OLSAT–SC	SAI–E	0.13*	0.08	0.10	–0.05	–0.03	0.03
	SAI–M	–0.05	0.03	0.00	0.08	0.05	–0.03
	VS–E	0.04	–0.04	0.00	–0.09	–0.04	–0.04
	NV–M	–0.02	–0.02	0.04	0.00	0.07	0.07
OLSAT–HSC	SAI–E	–0.31*	–0.09	–0.75*	0.22*	–0.44*	–0.65*
	SAI–M	–0.29*	0.00	–0.75*	0.29*	–0.46*	–0.75*
	VS–E	–0.31*	–0.06	–0.67*	0.25*	–0.36*	–0.61*
	NV–M	–0.34*	0.07	–0.83*	0.41*	–0.48*	–0.90*
NAPLAN–SC	Lit–E	–0.05	0.03	0.08	0.08	0.13*	0.05
	Num–M	0.00	0.05	0.04	0.05	0.04	–0.01
NAPLAN–HSC	Lit–E	–0.37*	0.05	–0.71*	0.41*	–0.34*	–0.76*
	Num–M	–0.29*	0.02	–0.78*	0.31*	–0.49	–0.80*
SC–HSC	E–E	–0.29	–0.04	–0.55*	0.25*	–0.26*	–0.51*
	M–M	–0.25	0.04	–0.74*	0.29*	–0.48*	–0.77*
OLSAT–Junior SA	SAI–E	0.44*	0.06	–0.18*	–0.39*	–0.63*	–0.24*
	SAI–M	–0.01	0.00	–0.22*	0.01	–0.20*	–0.22*
	VS–E	0.45*	0.03	–0.14*	–0.41*	–0.59*	–0.17*
	NV–M	0.00	0.04	–0.27*	0.03	–0.27*	–0.30*
OLSAT–Senior SA	SAI–E	0.23*	0.01	–0.44*	–0.21*	–0.67*	–0.45*
	SAI–M	0.12*	0.07	–0.35*	–0.05	–0.47*	–0.41*
	VS–E	0.24*	–0.04	–0.44*	–0.27*	–0.67*	–0.40*
	NV–M	0.15*	0.07	–0.36*	–0.09	–0.51*	–0.42*
NAPLAN–Junior SA	Lit–E	0.42	–0.01	–0.02	–0.43*	–0.44*	–0.01
	Num–M	0.01	0.14*	–0.17*	0.13*	–0.18*	–0.30*
NAPLAN–Senior SA	Lit–E	0.21*	–0.07	–0.39*	–0.28*	–0.60*	–0.32*
	Num–M	0.10*	0.11*	–0.35*	0.02	–0.45*	–0.47*
Junior SA–SC	E–E	–0.04	–0.05	0.05	–0.01	0.09*	0.11*
	M–M	0.03	–0.01	0.03	–0.04	0.00	0.04
SC–Senior SA	E–E	0.01	–0.11*	–0.33*	–0.12*	–0.34*	–0.22*
	M–M	0.18*	–0.01	–0.28*	–0.19*	–0.46*	–0.27*
Junior SA–HSC	E–E	–0.43*	–0.10*	–0.65*	0.33*	–0.22*	–0.54*
	M–M	–0.29*	0.00	–0.71*	0.29*	–0.43*	–0.71*
Senior SA–HSC	E–E	–0.26*	0.00	–0.20*	0.26*	0.07	–0.20*
	M–M	–0.31*	–0.08	–0.51*	0.23*	–0.21*	–0.44*
Junior SA–Senior SA	E–E	0.24*	–0.12*	–0.21*	–0.36*	–0.45*	–0.09*
	M–M	0.25*	–0.08*	–0.17*	–0.32*	–0.42*	–0.09*
OLSAT–SA	SAI–SA	0.04	–0.02	–0.30*	0.01	–0.34*	–0.32*
Weighted average		0.05*	0.01 *	–0.20*	–0.04*	–0.34*	–0.22*
Weighted average of absolute values		0.17*	0.05*	0.22*	0.18*	0.35*	0.23*

Note. SAI = School Ability Index; VS =Verbal Score; NV = Non-verbal score; E = English; M = Mathematics; Lit = Literacy; Num = Numeracy; ABSI = Absolute Split I; ABSII = Absolute Split II; SD = Simple Difference; REG = Regression; NOM = Nomination; *$p < 0.05$

Overall, the analysis of proportions has suggested that the four statistical methods used to identify gifted underachievement (i.e., absolute split I, absolute split II, regression, and simple difference methods) may not be convergent in terms of the proportions of gifted underachievement that are identified. In particular, the regression method and the absolute split II method appeared to identify different proportions of gifted underachievement. Nevertheless, among the statistical identification methods, marginal evidence for convergence was identified between the absolute split I and the simple difference methods (i.e., although no statistical convergence was demonstrated, the difference in the proportions of gifted underachievement cases identified using the two identification methods across different data combinations was small). When comparisons were made between the proportions of gifted underachievement identified using the nomination method and the statistical identification methods, some convergence evidence was found between the nomination method and the absolute split I, absolute split II, and the simple difference methods.

Table 16

Difference in proportion and McNemar test results for nomination

Method	Difference in proportion
Absolute split I	–0.05
Absolute split II	0.02
Regression	–0.26*
Simple difference	–0.03

*$p < 0.05$

5.5 Association and Correlation

This section provides evidence using the second approach to establish the convergence of the various methods that identify and measure gifted underachievement. Specifically, statistical measures were used to assess how strongly the results obtained from the various methods used to identify and measure gifted underachievement are related to one another, and whether the relationship is statistically significant. Tests of association were conducted on data from the methods that identify gifted underachievement (i.e., the absolute split I, absolute split II, simple difference, regression, and nomination methods), while additional tests of correlation were conducted on data from the methods that measure the degree of gifted underachievement (i.e., the simple difference and regression methods).

5.5.1 Association evidence

Association refers to the strength of the relationship between nominal, ordinal, or dichotomous variables. Two variables may be considered to be strongly associated with one another when a change in one variable is consistently met with a change in the other variable. The strength of the association between variables may be measured using association coefficients which have a maximum value of +1 and a minimum value of –1. A strong association between variables, that is also statistically significant, may form a part of the evidence that suggests convergence between variables. The following subsections outline the statistical analyses to measure and test association.

5.5.1.1 Contingency tables

When examining the association between two dichotomous variables, it is common to analyse these variables using a contingency table (Nussbaum, 2015). Table 17 provides an example of a simple 2 x 2 contingency table that compares the results of two methods for identifying gifted underachievement. Each letter (i.e., a, b, c, or d) represents the counts of

the number of cases of gifted underachievement or gifted achievement identified by the two methods. The letters "a" and "d" represent the number of cases where the two methods produce the same classifications, while the letters "b" and "c" represent the number of cases where the two methods produce different classifications. The analysis of association evidence relies on these four values, with larger values for a and d, and smaller values for b and c, indicative of a higher degree of association. While larger contingency tables are possible to allow for the comparison of variables with more than two values (non-dichotomous variables), all contingency tables used in this chapter are 2 x 2 contingency tables. Approximately 400 contingency tables were created for this chapter and are reported in Appendix 9.

Table 17

A generic contingency table comparing the classifications given by two methods (Nussbaum, 2015)

		Method 1		Row
		GUA	GA	Total
Method 2	GUA	a	b	a + b
	GA	c	d	c + d
Column total		a + c	b + d	TOTAL

Note. GUA = Gifted Underachievement; GA = Gifted Achievement

5.5.1.2 The Phi coefficient

The phi coefficient (ϕ) is a measure of association between two categorical variables (Field, 2013) that is equivalent to the commonly used Pearson correlation coefficient calculated with continuous variables (Cernovsky, 2002; Garson, 2012; Nussbaum, 2015). The phi coefficient is calculated using the values from the contingency table (refer to Appendix 12 for the relevant formula). Pett (1997) suggests that phi coefficient values from 0.00 to 0.29 are indicative of a *weak* level of association, values from 0.30 to 0.49 are indicative of a *low* level of association, values from 0.50 to 0.69 are indicative of a *moderate* level of

association, values from 0.70 to 0.89 are indicative of a *strong* level of association, and values from 0.90 to 1.00 are indicative of a *very strong* level of association. In addition, Park, Riddle, and Tekian (2014) suggest that a phi coefficient value of greater than 0.70 is required to establish that a relationship is at "sufficient levels" (p. 618) for convergence.

5.5.1.3 Chi-Square (χ^2) test of independence

The chi-square test is used to determine whether the association between two variables is statistically significant, and is commonly used in combination with the phi coefficient to determine whether convergence between variables is supported (Lau & Chan, 2001a). Convergence is deemed to be supported when the chi-square test result indicates that the probability of no association is very low.

It is noted that two conditions need to be met for the results of the chi-square test to be meaningful:

(a) No more than 20% of the cells in the contingency table should contain values less than five (Fisher's exact test should be used instead when this assumption is violated; Park et al., 2014); and
(b) The two variables should not be from the same participants (the McNemar test should be used instead when this assumption is violated; Elwood, 2007; Glantz, 2011; Rothman, Lash & Greenland, 2008).

In this project, the variables being examined for association (the results from the different methods of identifying gifted underachievement) are from the same participants, and hence the second assumption is violated. Nevertheless, to allow for comparability to the findings of Lau and Chan (2001a), who also compared methods that identified gifted underachievement and used the chi-square test to assess the statistical significance of association with the same participants, the chi-square test results are included here. The more

appropriate McNemar test has already been carried out and reported. The calculations required for carrying out the chi-square test and Fisher's exact test are outlined in Appendix 12.

5.5.1.4 Association results

The phi coefficient values and the chi-square test results between the nomination and each of the statistical methods (i.e., absolute split I, absolute split II, regression, and simple difference) are summarised in Table 19, while the phi coefficient values and chi-square test results showing the associations between the statistical methods are reported in Table 18. These tables together contain a total of 238 measurements of association and tests of significance. It is noted that as the simple difference method classified all cases from one data combination (i.e., Junior school assessment in English – School Certificate in English) as gifted achievement, one of the conditions for the conduct of the chi-square test was violated, and the Fisher's exact test was used instead in this instance.

Only 50 of the 238 measurements of association (i.e., 21% of all measurements of association) could be classified as *strong* or *very strong* associations according to Pett's (1997) criteria. A further 73 (31%) could be classified as *moderately strong*, 76 (32%) could be classified as of *low strength*, and 35 (15%) could be classified as of *weak strength*. If Park et al.'s (2014) criteria are used, only 21% of the associations appeared to be strong enough to indicate that convergence may be possible.

As the association between every possible pair of statistical identification methods was measured for each of the data combinations, it was possible to calculate a weighted average association for each possible pairing of the statistical methods. These values were calculated using meta-analysis techniques as previously described (refer Appendix 12) and are reported in Table 18 and Figure 29. All of the weighted average associations may be

classified as being of *low* or *moderate* strength according to Pett's (1997) guidelines, and none met Park et al.'s (2014) criteria for convergence. Nevertheless, two pairs of identification methods (i.e., absolute split I – absolute split II, and absolute split I – simple difference) with the highest weighted average phi coefficient values (i.e., 0.65 and 0.62, respectively), were only slightly below Park et al.'s (2014) threshold. The associations between the nomination method and the statistical identification methods were weaker (i.e., phi coefficient values ranged from 0.13 to 0.34) than the associations between the statistical identification methods.

The chi-square tests showed that 185 (78%) of the 238 measured associations, along with all of the weighted average associations had a level of association that was statistically significant ($p < 0.05$). It is noted that Lau and Chan (2001a) obtained similar results and concluded that the methods were convergent. Nevertheless, as an assumption of the chi-square test was violated (i.e., the data was obtained from the same participants), a similar conclusion about convergence may be inappropriate. The more appropriate McNemar test, as previously reported, evinced that the differences in proportions of gifted underachievement identified were statistically significant.

Generally, the association results reported in this section do not appear to support the convergence of the methods that identify gifted underachievement. No pairing of the identification methods produced a weighted average association (or an association in the case of pairings with the nomination method) that was large enough to suggest a strong or very strong association according to Pett (1997), or a level of association that is supportive of convergence proposed by Park et al. (2014). Furthermore, the chi-square test results do not appear to be meaningful, while the McNemar test results were generally non-supportive of the associations between the statistical identification methods. The nomination method,

Table 18

Summary of the phi coefficients of association and chi-square test of significance

Association of with Instrument pair	Method 1: Method 2: combination	ABSI			ABSII		SD
		ABSII	SD	REG	SD	REG	REG
OLSAT–NAPLAN	SAI–Lit	-	0.64*	0.70*	-	-	0.70*
	SAI–Num	-	0.65*	0.73*	-	-	0.92*
	VS–Lit	-	0.61*	0.64*	-	-	0.79*
	NV–Num	-	0.54*	0.70*	-	-	0.78*
OLSAT–SC	SAI–E	0.62*	0.61*	0.70*	0.79*	0.88*	0.90*
	SAI–M	0.82*	0.88*	1.00*	0.72*	0.82*	0.88*
	VS–E	0.69*	0.34	1.00*	0.55	0.69*	0.34
	NV–M	0.92*	0.44*	0.82*	0.39	0.76*	0.58*
OLSAT–HSC	SAI–E	0.39*	0.51*	0.14	0.58*	0.36	0.21
	SAI–M	0.45*	1.00*	0.14	0.45	0.32	0.14
	VS–E	0.45	0.61*	0.20	0.55*	0.44	0.24
	NV–M	0.38	0.56*	0.09	0.21	0.25	0.05
NAPLAN–SC	Lit–E	0.69*	0.88*	0.72*	0.61*	0.50*	0.82*
	Num–M	1.00*	0.50*	0.54*	0.50*	0.54*	0.93*
NAPLAN–HSC	Lit–E	0.44	0.84*	0.16	0.37	0.37	0.14
	Num–M	0.40	0.23	0.12	0.34	0.31	0.10
SC–HSC	E–E	0.54*	0.83*	0.28	0.60*	0.52*	0.31
	M–M	0.49*	0.64*	0.15	0.41*	0.31*	0.13
OLSAT–Junior SA	SAI–E	0.38*	0.57*	0.58*	0.43*	0.22	0.51*
	SAI–M	0.92*	0.70*	0.56*	0.64*	0.53*	0.56*
	VS–E	0.38*	0.59*	0.62*	0.37*	0.24	0.57*
	NV–M	0.93*	0.54*	0.48*	0.51*	0.48*	0.45*
OLSAT–Senior SA	SAI–E	0.56*	0.75*	0.36*	0.43*	0.20	0.35*
	SAI–M	0.78*	0.71*	0.43*	0.56*	0.34*	0.38*
	VS–E	0.57*	0.78*	0.34	0.44*	0.19	0.36
	NV–M	0.72*	0.65*	0.43*	0.40*	0.31*	0.38*
NAPLAN–Junior	Lit–E	0.36*	0.55*	0.62*	0.34*	0.33*	0.92*
	Num–M	0.94*	0.52*	0.50*	0.49*	0.49*	0.47*
NAPLAN–Senior	Lit–E	0.58*	0.65*	0.32	0.33*	0.20	0.36*
	Num–M	0.82*	0.63*	0.43*	0.45*	0.35*	0.34*
Junior SA–SC	E–E	0.49*	0.00[a]	0.69*	0.00[a]	0.34*	0.00[a]
	M–M	0.70*	0.70*	0.70*	0.49*	0.47*	0.49*
SC–Senior SA	E–E	0.71*	0.81*	0.44*	0.68*	0.43*	0.55*
	M–M	0.65*	0.60*	0.44*	0.35*	0.27*	0.53*
Junior SA–HSC	E–E	0.40*	0.50*	0.18	0.50*	0.45*	0.22
	M–M	0.52*	0.62*	0.15	0.52*	0.29	0.15
Senior SA–HSC	E–E	0.46*	0.81*	0.60*	0.46*	0.58*	0.60*
	M–M	0.53*	0.85*	0.27	0.62*	0.51*	0.31
Junior SA–Senior	E–E	0.49*	0.67*	0.37*	0.40*	0.22	0.55*
	M–M	0.57*	0.77*	0.44*	0.44*	0.25	0.57*
OLSAT–SA	SAI–SA	0.75*	0.71*	0.52*	0.67*	0.48*	0.49*
Weighted average		0.65*	0.62*	0.54*	0.48*	0.39*	0.58*

Note. SAI = School Ability Index; VS =Verbal Score; NV = Non-verbal score; E = English; M = Mathematics; Lit = Literacy; Num = Numeracy; *$p<0.05$; [a]Fisher's exact test used

which showed phi coefficient values ranging from 0.13 to 0.34 with the other identification methods, appeared to be the most non-convergent with the other identification methods.

Table 19

Summary of the phi coefficients of association and chi-square test of significance comparing nomination and statistical identification methods

Method	Phi
Absolute split I	0.27*
Absolute split II	0.33*
Simple difference	0.34*
Regression	0.13

*$p < 0.05$

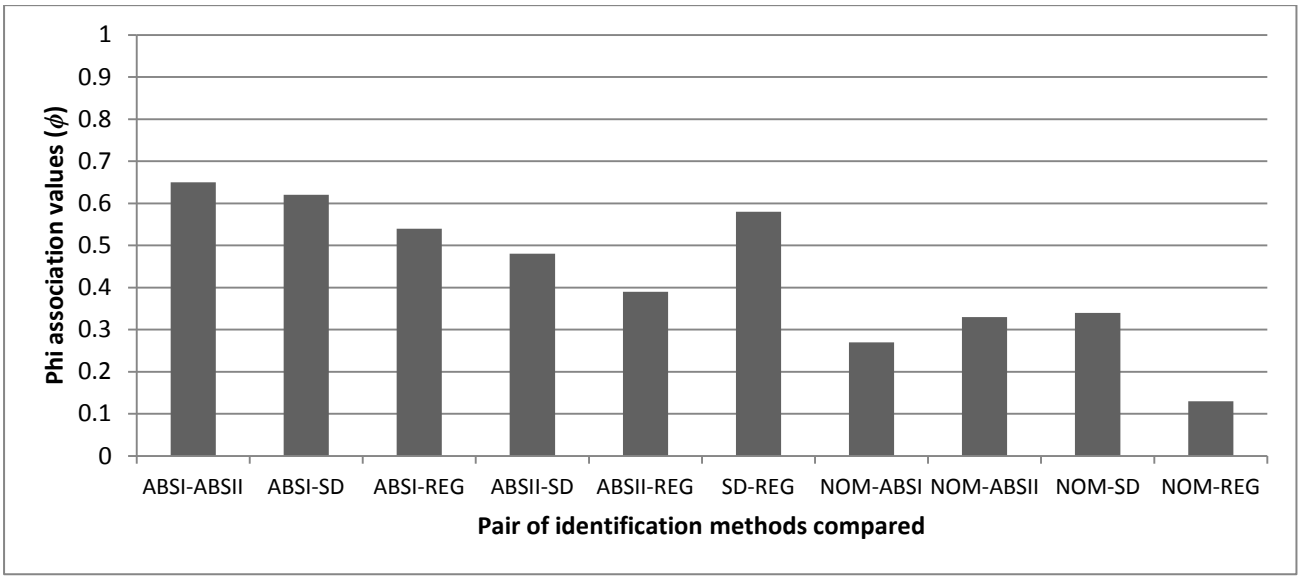


Figure 29. Weighted average phi coefficient values

5.5.2 Correlation evidence

Correlation refers to the strength of the relationship between two continuous variables. As two of the methods used to identify gifted underachievement (i.e., the simple difference method and the regression method) calculate continuous variables that may be

used to measure the degree of gifted underachievement, a correlation analysis will be possible to assess the strength of the relationship between these variables. In comparison to association analysis, correlation analysis has greater statistical power, which may mean that it provides stronger evidence to assess the possibility of convergence. The following sections outline the statistical analyses required to measure and test correlations between continuous variables.

5.5.2.1 Pearson's correlation coefficient

The Pearson correlation coefficient (*r*, Field, 2013; Hair, Anderson, Black, Babin, & Black, 2010) is commonly used to assess the strength of the relationship between two continuous variables. Johnson and Wichern (2007) suggest that Pearson correlation coefficient values of 0.00 to 0.29 represent a *very weak* linear relationship, values of 0.30 to 0.49 represent a *weak* linear relationship, values of 0.50 to 0.69 represent a *moderate* linear relationship, values of 0.70 to 0.89 represent a *strong* linear relationship, and 0.90 to 1.00 represent a *near perfect* linear relationship between the two variables. In addition, a *t*-test is commonly used to determine whether the measured correlation is statistically significant (Field, 2013). The test statistic, the formula for which is provided in Appendix 12, is calculated on the basis of the size of the correlation and the sample.

5.5.2.2 Correlation results

Table 20 provides details of the Pearson correlation coefficient (*r*) calculations and the results of the *t*-tests. These results show that across almost all data combinations, the measurements of gifted underachievement obtained using the simple difference and regression methods were *nearly perfectly* related to one another. Furthermore, across all data combinations, the relationship was found to be statistically significant. When the weighted average correlation between the simple difference and regression measurements of gifted underachievement was calculated using the meta-analysis methods outlined in Appendix 12,

Table 20

Results for Pearson correlation coefficients

Instrument pair	Data	*r*
OLSAT–NAPLAN	SAI–Lit	0.97*
	SAI–Num	1.00*
	VS–Lit	0.97*
	NV–Num	0.99*
OLSAT–SC	SAI–E	0.95*
	SAI–M	0.99*
	VS–E	0.91*
	NV–M	0.97*
OLSAT–HSC	SAI–E	0.98*
	SAI–M	0.99*
	VS–E	0.98*
	NV–M	0.97*
NAPLAN–SC	Lit–E	0.99*
	Num–M	0.98*
NAPLAN–HSC	Lit–E	0.98*
	Num–M	0.88*
SC–HSC	E–E	1.00*
	M–M	0.89*
OLSAT–Junior SA	SAI–E	0.98*
	SAI–M	0.99*
	VS–E	0.98*
	NV–M	0.95*
OLSAT–Senior SA	SAI–E	0.97*
	SAI–M	0.97*
	VS–E	0.97*
	NV–M	0.93*
NAPLAN–Junior SA	Lit–E	1.00*
	Num–M	0.94*
NAPLAN–Senior SA	Lit–E	0.97*
	Num–M	0.92*
Junior SA–SC	E–E	0.98*
	M–M	1.00*
SC–Senior SA	E–E	0.99*
	M–M	0.94*
Junior SA–HSC	E–E	1.00*
	M–M	1.00*
Senior SA–HSC	E–E	1.00*
	M–M	1.00*
Junior SA–Senior SA	E–E	0.98*
	M–M	0.99*
OLSAT–SA	SAI–SA	0.89*
Weighted average		0.97*

Note. SAI = School Ability Index; VS =Verbal Score; NV = Non-verbal score; E = English; M = Mathematics; Lit = Literacy; Num = Numeracy; *$p<0.05$

a *nearly perfectly* linear relationship was again suggested. The correlation results support the possibility of convergence between the simple difference and regression methods of measuring gifted underachievement.

Overall, contrasting information was provided from the association and correlation analyses. While the association analyses largely indicated non-convergence between the identification methods (with the absolute split I and absolute split II methods, and the absolute split I and simple difference methods having the strongest, albeit marginal, claims for convergence), the correlation analyses indicated a high level of convergence between the measurements of the degree of gifted underachievement obtained using simple difference and regression methods.

5.6 Agreement

This section provides evidence from the third approach (i.e., agreement) that may be used to assess the convergence of methods that identify and measure gifted underachievement. Agreement may be considered to occur when the various methods used to identify gifted underachievement identify the same gifted students as underachieving, or when the two methods used to measure gifted underachievement measure the same degree of gifted underachievement for each student. As for the previous sections, different statistical analyses will be used to investigate the level of agreement between the methods that identify and measure gifted underachievement.

5.6.1 Usefulness of agreement analyses

Altman and Bland (Altman & Bland, 1983; Bland & Altman, 1986; 1999) argue that the other two approaches used to assess convergence may be limited because their results may suggest convergence, even though no true convergence exists. For example, measurements of the differences in the proportions of classifications of gifted

underachievement may be problematic, because even if the proportions obtained using the different identification methods are similar, they may in fact identify different students as exhibiting gifted underachievement. Similarly, measures of correlation may suggest that two measurements of gifted underachievement have a perfect linear relationship, which only suggests that if the value of one measurement is known, the other may be predicted (i.e., a high correlation may not be used to infer that the two measurements are the same, or even similar). To address such issues, an examination of agreement statistics, which use a method of quantifying the differences between two sets of data, may be necessary (Altman & Bland, 2002).

5.6.2 Agreement of identification methods

Percentage agreement and Cohen's kappa are two commonly used measures of agreement between categorical variables, and may therefore be appropriate to assess the level of agreement between the methods that identify gifted underachievement (McHugh, 2012). Fleiss et al. (2003) suggest that percentage agreement is the simplest method used to measure the degree of agreement between two categorical variables (e.g., gifted achievement and gifted underachievement), while Viera and Garrett (2005) suggest that the kappa agreement measure may be the most commonly used agreement statistic.

5.6.2.1 Percentage agreement

The percentage agreement measurement provides information on the total percentage of identical classifications obtained from two different methods. However, there are two reported problems with this statistic as a measure of agreement (Hoffmann & Ninonuevo, 1994):

(a) The percentage agreement may be greatly inflated when one classification category has a much larger identification rate. For example, when comparing two tests that

diagnose an uncommon medical condition, the large rate of negative results may suggest a high level of agreement, even if the tests disagree on every positive result.

(b) Some agreement is expected to occur by chance alone. Sim and Wright (2005) suggest that if the measurements agree "purely by chance, they are not really 'agreeing' at all; only agreement beyond that expected by chance can be considered 'true' agreement" (p. 258).

Due to these problems, the percentage agreement statistic was not used in this investigation.

5.6.2.2 Cohen's kappa statistic

The Cohen's kappa (κ) statistic may be used to assess the degree of agreement between any classification tasks (Agresti, 2013; Brennan & Prediger, 1981; Cohen, 1960; Czodrowski, 2014; Sim & Wright, 2005; Warrens, 2011), and is commonly used in the health fields to assess the agreement between multiple methods to diagnose a health condition (Correia et al., 2011; Ewe et al., 2013; Ghanizadeh, 2013; Lindsley et al., 2011). The procedure in the calculation of the statistic is shown in Appendix 12. The Cohen's kappa statistic is preferred over the percentage of agreement because it attempts to take into account the agreement that would be expected purely by chance.

Values for kappa range from –1 to 1, with values of 1 indicative of perfect agreement, values of zero indicative of no agreement greater than that expected by chance, and a negative value indicative of less agreement than may be expected by chance. The interpretation of specific kappa values within this range may vary according to the context in which it is being used (Kundel & Polansky, 2003; Landis & Koch, 1977). Nevertheless, several different scholars have proposed interpretations for values which are summarised in Table 21. It is noted that while the choice of these thresholds and interpretations may be somewhat arbitrary (Sim & Wright, 2005), the Fleiss et al. (2003) and Walts, et al. (2011)

guidelines appear to be an appropriate intermediate position. Therefore, a value of 0.75 was deemed necessary for convergence to be supported. Tests of significance for Cohen's kappa values are generally not used, as it is rare for these values to not be greater than expected by chance (Agresti, 2013; Bakeman, & Gottman, 1997).

Table 21

Common interpretations of kappa values

Kundel & Polansky, 2003; Landis & Koch, 1977		Fleiss et al., 2003; Walts et al., 2011		McHugh, 2012; Rettew et al., 2009; Tang et al., 2012	
kappa value	Agreement	kappa value	Agreement	kappa value	Agreement
< 0.00	Less than chance	< 0.00	Less than chance	< 0.00	Less than chance
0.01–0.20	Slight	0.01–0.39	Poor		
0.21–0.40	Fair			< 0.80	Unacceptable
0.41–0.60	Moderate	0.40–0.74	Fair–good		
0.61–0.80	Substantial				
0.81–1.00	Almost perfect	0.75–1.00	Excellent	0.81–1.00	Acceptable

5.6.2.3 Kappa agreement results

This section reports the kappa agreement values (κ) between each pair of methods used to identify gifted underachievement for each data combination (refer Tables 22 and 23). The kappa agreement statistics ranged from –0.32 to 1.00. Approximately half (54%) of the kappa values were found to be greater than the Walts et al. (2011) threshold for fair agreement ($\kappa = 0.40$), while only 14% were above their threshold for excellent agreement ($\kappa = 0.75$). Therefore, strong support could not be found for convergence between the various methods that identify gifted underachievement. In particular, the level of agreement between the nomination method and the other statistical methods was consistently less than expected by chance alone ($\kappa < 0.00$), suggesting that the nomination method, consistent with the

findings of Lau and Chan (2001a), may be the least convergent with the other methods used to identify gifted underachievement.

When the weighted average kappa values were calculated using the meta-analysis methods (as outlined in Appendix 12; refer Figure 30), *none* of the weighted average kappa values was above the Walts et al. (2011) threshold for excellent agreement ($\kappa = 0.75$). Nevertheless, two pairs of methods (i.e., absolute split I – absolute split II, and absolute split I – simple difference) had weighted average kappa values that were close (0.69 and 0.64 respectively) to the Walts et al. (2011) threshold. These results tentatively and weakly support the possibility that among the methods used to identify gifted underachievement, *some* convergence may exist between the absolute split I and absolute split II methods, and between the absolute split I and simple difference methods. None of the remaining weighted average kappa agreement values were above the threshold for fair agreement ($\kappa = 0.40$).

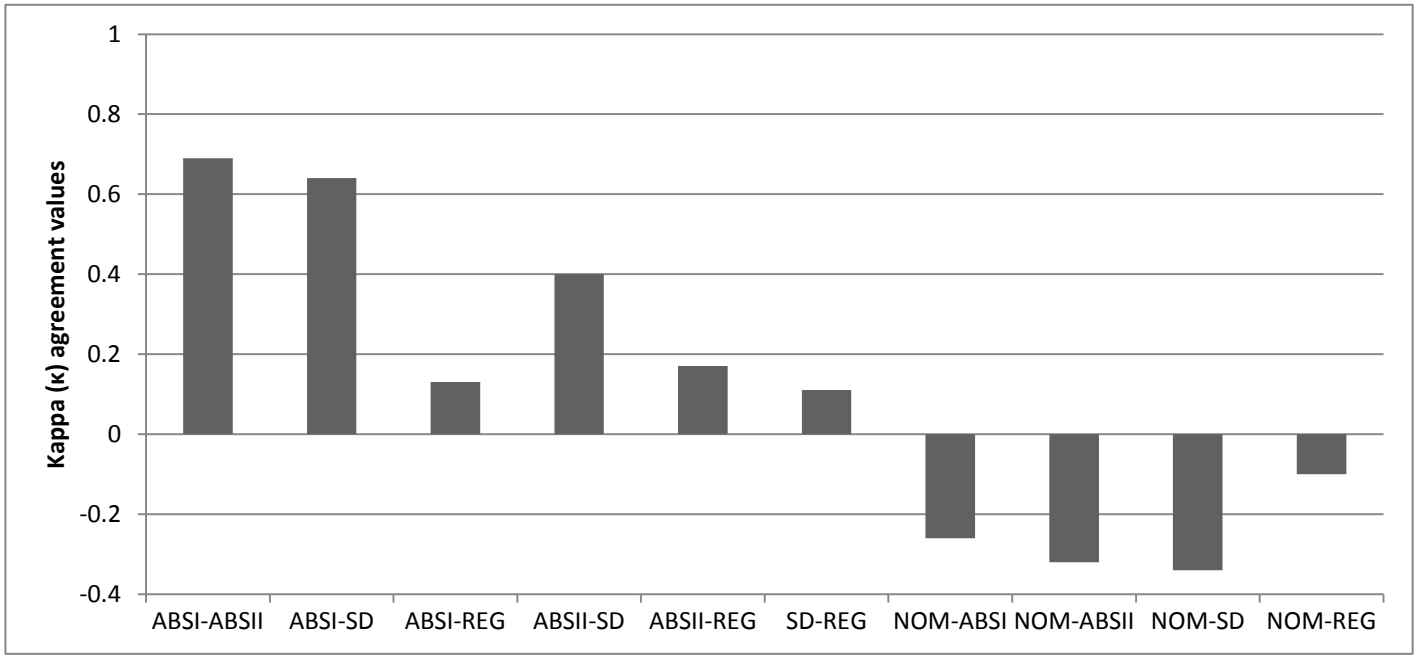


Figure 30. Kappa agreement for pairs of identification methods

Table 22

Kappa agreement results between statistical methods for identification

Instrument pair	Data pair	ABSI			ABSII		SD
		ABSII	SD	REG	SD	REG	REG
OLSAT–NAPLAN	SAI–Lit	-	0.64	0.66	-	-	0.66
	SAI–Num	-	0.65	0.71	-	-	0.91
	VS–Lit	-	0.61	0.60	-	-	0.77
	NV–Num	-	0.53	0.70	-	-	0.76
OLSAT–SC	SAI–E	0.55	0.59	0.66	0.77	0.87	0.89
	SAI–M	0.80	0.87	1.00	0.69	0.80	0.87
	VS–E	0.65	0.33	1.00	0.47	0.65	0.33
	NV–M	0.92	0.44	0.81	0.39	0.73	0.56
OLSAT–HSC	SAI–E	0.26	0.47	0.04	0.50	0.23	0.08
	SAI–M	0.33	1.00	0.04	0.33	0.19	0.04
	VS–E	0.34	0.60	0.07	0.46	0.32	0.11
	NV–M	0.25	0.47	0.02	0.08	0.11	0.01
NAPLAN–SC	Lit–E	0.64	0.87	0.69	0.54	0.40	0.80
	Num–M	1.00	0.46	0.51	0.46	0.51	0.93
NAPLAN–HSC	Lit–E	0.32	0.83	0.05	0.24	0.24	0.04
	Num–M	0.28	0.23	0.03	0.21	0.17	0.02
SC–HSC	E–E	0.45	0.83	0.15	0.53	0.43	0.18
	M–M	0.39	0.63	0.05	0.28	0.18	0.03
OLSAT–Junior SA	SAI–E	0.26	0.57	0.50	0.32	0.09	0.41
	SAI–M	0.92	0.70	0.47	0.64	0.46	0.47
	VS–E	0.25	0.59	0.55	0.26	0.11	0.49
	NV–M	0.93	0.54	0.38	0.51	0.37	0.33
OLSAT–Senior SA	SAI–E	0.48	0.75	0.23	0.37	0.08	0.21
	SAI–M	0.76	0.71	0.31	0.56	0.20	0.25
	VS–E	0.49	0.78	0.20	0.36	0.07	0.23
	NV–M	0.69	0.65	0.31	0.39	0.18	0.25
NAPLAN–Junior SA	Lit–E	0.23	0.55	0.61	0.21	0.20	0.92
	Num–M	0.94	0.50	0.45	0.47	0.43	0.36
NAPLAN–Senior SA	Lit–E	0.53	0.65	0.18	0.28	0.08	0.23
	Num–M	0.80	0.62	0.31	0.45	0.22	0.20
Junior SA–SC	E–E	0.39	0.00	0.64	0.00	0.20	0.00
	M–M	0.65	0.66	0.65	0.39	0.47	0.39
SC–Senior SA	E–E	0.71	0.79	0.33	0.66	0.32	0.47
	M–M	0.61	0.60	0.35	0.33	0.17	0.43
Junior SA–HSC	E–E	0.27	0.49	0.06	0.40	0.33	0.10
	M–M	0.42	0.62	0.04	0.42	0.16	0.04
Senior SA–HSC	E–E	0.34	0.81	0.53	0.34	0.56	0.53
	M–M	0.43	0.84	0.13	0.55	0.41	0.18
Junior SA–Senior SA	E–E	0.44	0.62	0.24	0.27	0.09	0.46
	M–M	0.49	0.74	0.32	0.32	0.12	0.49
OLSAT–SA	SAI–SA	0.74	0.71	0.42	0.67	0.37	0.36
Weighted average		0.69	0.64	0.13	0.40	0.17	0.11

Note. SAI = School Ability Index; VS =Verbal Score; NV = Non-verbal score; E = English; M = Mathematics; Lit = Literacy; Num = Numeracy

Table 23

Kappa agreement results between nomination and statistical methods for identification

Method	kappa agreement
Absolute split I	–0.26
Absolute split II	–0.32
Regression	–0.10
Simple difference	–0.34

5.6.3 Agreement of measurement methods

For the two methods that not only identify gifted underachievement, but also measure the degree of gifted underachievement (i.e., the simple difference and regression methods), additional agreement analyses were possible. The greater statistical power of such analyses mean that stronger evidence may be provided to assess the possibility of convergence between the simple difference and regression methods.

5.6.3.1 Concordance correlation coefficient

The concordance correlation coefficient (CCC) is a measure of the agreement between continuous variables (Lin 1989, 2000) that is equivalent to the Cohen's kappa statistic used with categorical data (Lin, Hedayat & Wu, 2007; Robieson, 1999). As for the Cohen's kappa statistic, the CCC already takes into consideration the level of agreement that may be attributed to chance, and no tests are usually performed to determine whether the size of the CCC value is significant. Appendix 12 provides details on how the statistic is calculated. Table 25 provides CCC threshold values that are commonly used by scholars to interpret the strength of agreement between variables (Johnson, Chumlea, Czerwinski, & Demearath, 2012; McBride, 2005; Panter , Costa, Dalton, Jones, & Ogilvie, 2014; Stepnowsky, Zamora, Barker, Liu, & Sarmiento, 2013; Wang & Hui, 2015).

Table 24

Concordance correlation coefficients

Instrument pair	Data pair	SD–REG
OLSAT–NAPLAN	SAI–Lit	0.70
	SAI–Num	0.92
	VS–Lit	0.70
	NV–Num	0.89
OLSAT–SC	SAI–E	0.61
	SAI–M	0.87
	VS–E	0.64
	NV–M	0.82
OLSAT–HSC	SAI–E	0.40
	SAI–M	0.42
	VS–E	0.46
	NV–M	0.31
NAPLAN–SC	Lit–E	0.74
	Num–M	0.77
NAPLAN–HSC	Lit–E	0.26
	Num–M	0.23
SC–HSC	E–E	0.43
	M–M	0.26
OLSAT–Junior SA	SAI–E	0.65
	SAI–M	0.68
	VS–E	0.66
	NV–M	0.56
OLSAT–Senior SA	SAI–E	0.50
	SAI–M	0.55
	VS–E	0.57
	NV–M	0.49
NAPLAN–Junior SA	Lit–E	0.79
	Num–M	0.56
NAPLAN–Senior SA	Lit–E	0.53
	Num–M	0.48
Junior SA–SC	E–E	0.70
	M–M	0.95
SC–Senior SA	E–E	0.63
	M–M	0.54
Junior SA–HSC	E–E	0.31
	M–M	0.26
Senior SA–HSC	E–E	0.47
	M–M	0.35
Junior SA–Senior SA	E–E	0.44
	M–M	0.49
OLSAT–SA	SAI–SA	0.84
Weighted Average		0.66

Note. SAI = School Ability Index; VS =Verbal Score; NV = Non-verbal score; E = English; M = Mathematics; Lit = Literacy; Num = Numeracy

Table 25

Interpretation of Concordance Correlation Coefficient (CCC) values (McBride, 2005)

CCC value	Strength of agreement
<0.90	Poor
0.90–0.95	Moderate
0.95–0.99	Substantial
>0.99	Almost perfect

Table 24 provides the results of the CCC calculations used to measure the degree of agreement between the simple difference and regression measurements of gifted underachievement for each data combination. It is noted that the CCC values ranged from 0.23 to 0.95, with 39 of the 41 data combinations considered to have a "poor" level of agreement and only two considered to have a "moderate" level of agreement (McBride, 2005). The weighted average concordance correlation coefficient, which was calculated following meta-analysis methods (refer Appendix 12), was found to be 0.66, which is well below the threshold for a "moderate" or a higher level of agreement. These results do not support the convergence of the simple difference and regression methods for measuring gifted underachievement.

5.6.3.2 Paired t-test

As the simple difference and regression methods were used to produce measurements of the degree of gifted underachievement with the same group of students, the measurements may be considered *paired*. The mean of the differences between each pair of measurements, calculated in standard deviation units, may be analysed to assess the level of agreement between these measurements (Hair et al., 2010). Baguley (2009) provides some guidelines on the interpretation of the size of the mean differences (i.e., less than 0.2 is considered a *negligible* difference, greater than 0.2 is considered a *small* difference, greater than 0.5 is considered a *medium* difference, and greater than 0.8 is considered a *large* difference). A

difference of one standard deviation units is equivalent to the adopted threshold for a *severe* discrepancy needed to identify gifted underachievement. In addition to an examination of the mean differences, a paired *t*-test (which is often called a dependent *t*-test) may be used to determine whether the mean differences are statistically significant.

The differences between the means of the two sets of measurements (shown in Table 26) demonstrate that the simple difference and regression methods may not be considered to produce similar measurements of gifted underachievement. Following Baguley's (2009) guidelines, 15 (37%) of the mean differences may be classified as *large* (including 14 that were larger than one standard deviation), 6 (15%) may be classified as *medium*, 10 (24%) may be classified as *small*, and 10 (24%) may be classified as *negligible*. Furthermore, the weighted average difference between means (0.59) was classifiable as being of *medium* size (Baguley, 2009). It is noteworthy that for almost all (38 out of 41) data combinations, the paired *t*-test demonstrated that the measurements of gifted underachievement by the simple difference and regression methods were statistically significantly different.

Table 26

Paired t-test results

Instrument pair	Data	*d*
OLSAT–NAPLAN	SAI–Lit	0.33*
	SAI–Num	0.07*
	VS–Lit	0.36*
	NV–Num	0.24*
OLSAT–SC	SAI–E	0.42*
	SAI–M	0.04
	VS–E	0.51*
	NV–M	0.18*
OLSAT–HSC	SAI–E	1.63*
	SAI–M	1.60*
	VS–E	1.44*
	NV–M	2.06*
NAPLAN–SC	Lit–E	0.04
	Num–M	0.13*
NAPLAN–HSC	Lit–E	2.14*
	Num–M	2.37*
SC–HSC	E–E	0.34*
	M–M	2.18*
OLSAT–Junior SA	SAI–E	0.56*
	SAI–M	0.39*
	VS–E	0.63*
	NV–M	0.91*
OLSAT–Senior SA	SAI–E	1.35*
	SAI–M	1.09*
	VS–E	1.13*
	NV–M	1.30*
NAPLAN–Junior	Lit–E	0.09*
	Num–M	0.75*
NAPLAN–Senior	Lit–E	1.02*
	Num–M	1.27*
Junior SA–SC	E–E	0.13*
	M–M	0.00
SC–Senior SA	E–E	0.40*
	M–M	1.13*
Junior SA–HSC	E–E	0.07*
	M–M	0.24*
Senior SA–HSC	E–E	0.10*
	M–M	0.36*
Junior SA–Senior	E–E	0.77*
	M–M	0.61*
OLSAT–SA	SAI–SA	0.34*
Weighted average		0.59*

Note. SAI = School Ability Index; VS =Verbal Score; NV = Non-verbal score; E = English; M = Mathematics; Lit = Literacy; Num = Numeracy; *$p<0.05$

5.6.3.3 Bland–Altman plots

In addition to the concordance correlation coefficient and paired *t* tests, Bland–Altman plots may be used to evaluate the level of agreement between two methods (Bland & Altman, 1986, 1995, 1999; Cabrera et al., 2002). Singhal and Siddhu (2011) suggest that these plots may be considered to be "the best method for comparing the measurements obtained by two methods when the true value is unknown" (p.1598), while Preiss and Fisher (2008) note that in medical research, "Bland–Altman analysis has largely replaced the correlation coefficient as the predominant tool for evaluating the interchangeability of two methods for clinical measurement" (p. 257). A typical Bland–Altman plot is shown in Figure 31.

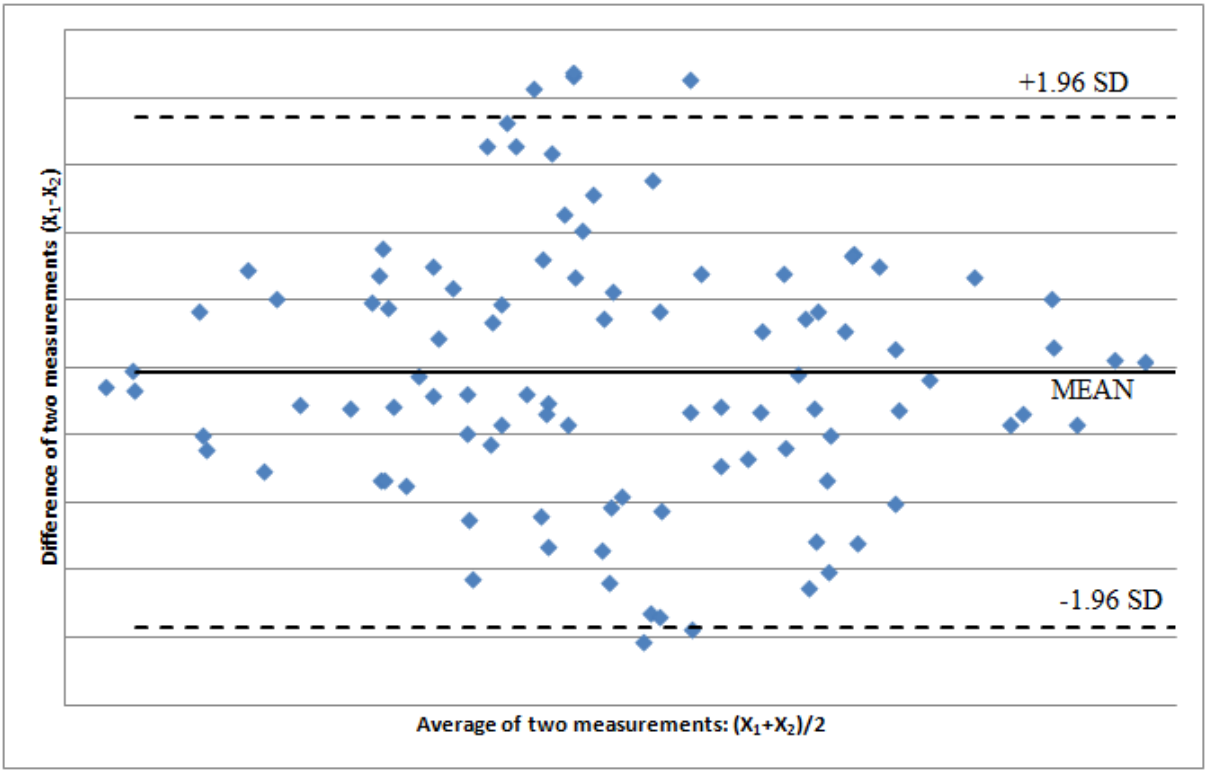


Figure 31. A sample Bland–Altman plot

In the construction of the Bland–Altman plot, two new variables are calculated from the original two measurements of each participant. The first variable is the average of the two measurements for each participant, while the second variable is the difference between the two measurements for each participant. These variables are then plotted on a graph (i.e., the Bland–Altman plot), with the average of the two measurements on the horizontal axis and the difference between the two measurements on the vertical axis. Bland–Altman plots typically specify the mean difference between the two measurements and the 95% confidence intervals for the mean difference (these are shown as a solid horizontal line and as dashed horizontal lines respectively in Figure 31).

Prior to the use of a Bland–Altman plot to determine whether two sets of measurements agree, multiple scholars (Bland & Altman, 1986, 1995, 1999; Giavarina, 2015) recommend that an assessment be made of whether a linear pattern, which may be indicative of systematic bias (e.g., the level of agreement between the two measurements may increase or decrease as the degree of underachievement increases or decreases) exists. Any such bias should be removed before the Bland–Altman plot may be interpreted (Bland & Altman, 1986, 1995, 1999; Giavarina, 2015). The interpretation itself is based on two main factors: (a) the size of the mean difference of measurements (previously analysed when conducting paired *t*-tests), and (b) the size of the confidence interval for the mean differences. For agreement to occur, the Bland–Altman plot should demonstrate a small mean difference *and* narrow confidence intervals (Connelly, 2008). As the size of the mean difference was previously analysed, the analysis of Bland–Altman plots for this investigation focused on the size of the confidence interval for the mean difference of measurements.

Two steps need to be taken to determine an acceptable threshold for the size of the confidence interval for the mean difference of measurements. First, as no standard guidelines exist, the researcher must establish how large a difference in the measurements should be to

be considered significant (Bland & Altman, 1986, 1995, 1999; Dewitte, Fierens, Stockl, & Thienpont, 2002). Such a threshold should be determined a priori, and will depend on the units of measurement, the planned use of the instrument, and whether the difference may have any significance for identification (Connelly, 2008; Dewitte et al., 2002; Giavarina, 2015; Hanneman, 2008). In this investigation, a threshold of one standard deviation may be appropriate, as such a threshold has been typically adopted to establish "severe discrepancy" in gifted underachievement (refer Appendix 1).

The second step requires the researcher to establish the percentage of scores that assess difference in measurement that are in excess of the threshold, which may be considered small enough to constitute an acceptable level of agreement between the measurements obtained using the two methods (Bland & Altman, 1986, 1995, 1999; Hanneman, 2008). As no study in the field of education was found that used Bland–Altman plots, guidance was sought from other fields where these plots are used regularly. One relevant study may be Hanneman (2008), which compared two medical instruments and used a threshold for the confidence interval that was equivalent to ±0.42 standard deviations. While the measurements being examined in the current investigation may not be used to diagnose medical conditions, they may nevertheless be used to diagnose related conditions (Reschly & Hosp, 2004). Therefore, a confidence interval of ±0.5 standard deviations, which was broadly similar to that used in Hanneman (2008) was selected. Such a confidence interval indicates that a maximum of 5% of measurements may be different by at least one standard deviation for agreement to be established between the measurements obtained using the two methods.

A common feature of the Bland–Altman plots that were developed for this investigation was an asymmetry of data points above and below the line of best fit. Additionally, a large section of each plot below the line of best fit tended to be devoid of data

points (an example of such a plot is provided in Figure 32). The unique pattern of the data distribution in the Bland–Altman plots may be attributed to the collection of data on underachievement from gifted students only. Indeed, more symmetrical data distributions (refer Figure 33) were achieved when measurements from both gifted and typical students were included.

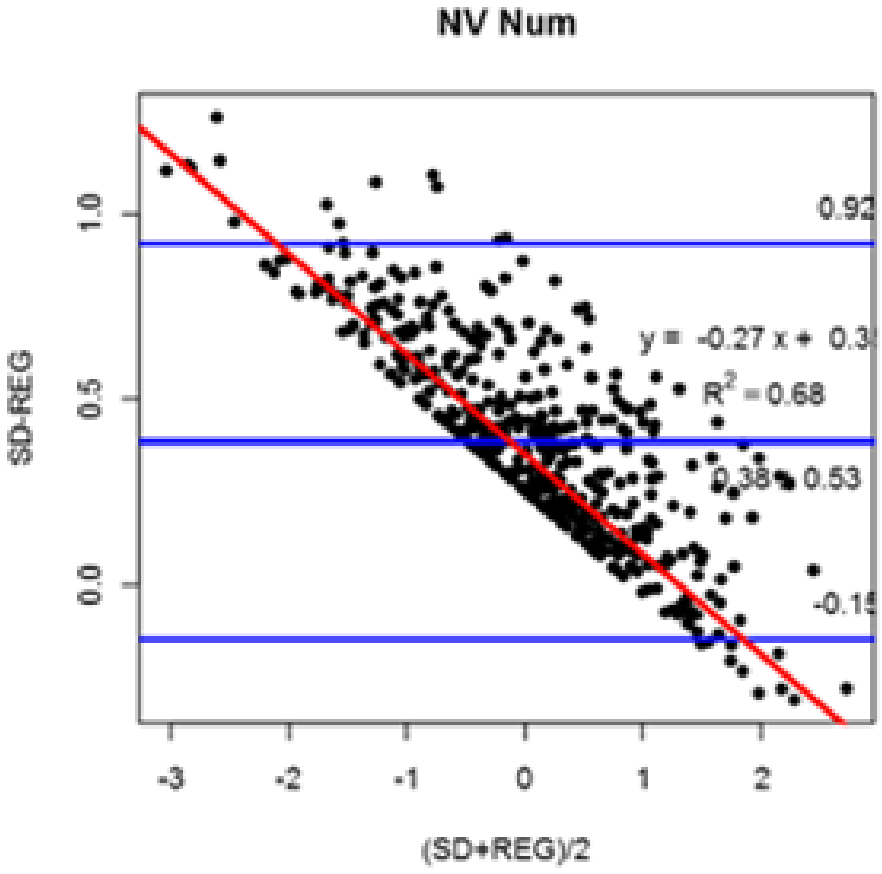


Figure 32. Bland–Altman plot for data showing asymmetric distribution

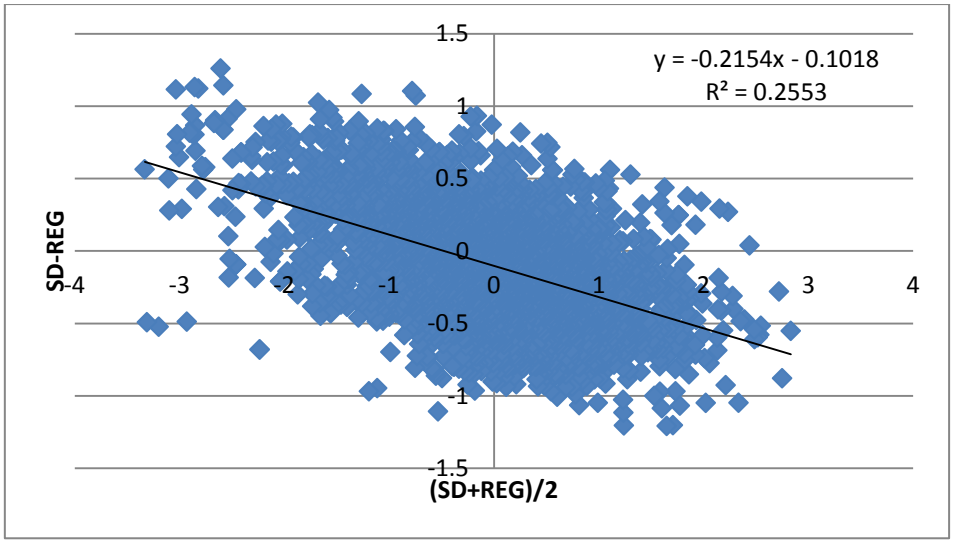


Figure 33. A Bland–Altman plot for OLSAT-NAPLAN (NV–NUM) data including typical students

The initial interpretation of the Bland–Altman plots that included data from both gifted and typical students suggested a tendency for a large linear systematic pattern between the two measurements of underachievement. Indeed, more than half (i.e., 25 out of 40) of the Bland–Altman plots exhibited such a pattern. Figure 34 shows how the Bland–Altman plot is altered after the removal of this linear pattern and the removal of data from typical students (Bland & Altman, 1999; Giavarina, 2015). The measured confidence intervals from the final Bland–Altman plots (that included data only from gifted students) for all data combinations, and after adjustment for systematic bias, are shown in Table 27. Appendix 10 contains all of the Bland–Altman plots.

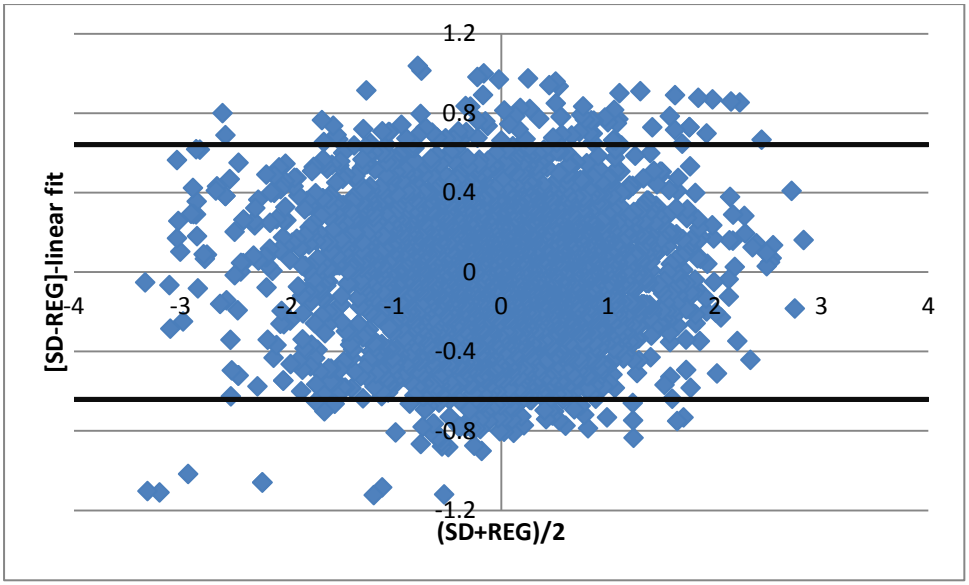


Figure 34. Bland–Altman plot for OLSAT-NAPLAN (NV–NUM) data after removing the linear bias and showing 95% CI for data

The adjusted Bland–Altman plots indicated that only ten (24%) of the confidence intervals were smaller than Hanneman's (2008) limit for agreement (i.e., ±0.42), while only twelve (29%) were smaller than the predetermined acceptable limit for agreement for this project (i.e., ±0.50), leaving 29 (71%) that did not meet the predetermined acceptable limit for agreement. Furthermore, the weighted average of the confidence intervals, as calculated using meta-analysis methods (refer Appendix 12) was found to be 0.65, which is larger than the predetermined acceptable limit for agreement. Therefore, the Bland–Altman plot results (as for the results of the analyses using concordance correlation coefficients and paired *t* tests) do not generally support the convergence of the simple difference and regression methods for measuring gifted underachievement.

Table 27

Confidence Intervals (CI) from the Bland–Altman plots

Instrument	Data pair	CI
OLSAT–NAPLAN	SAI–Lit	1.03
	SAI–Num	0.47
	VS–Lit	1.02
	NV–Num	0.64
OLSAT–SC	SAI–E	1.04
	SAI–M	0.54
	VS–E	0.87
	NV–M	0.74
OLSAT–HSC	SAI–E	0.99
	SAI–M	0.94
	VS–E	0.91
	NV–M	1.05
NAPLAN–SC	Lit–E	0.52
	Num–M	0.64
NAPLAN–HSC	Lit–E	0.63
	Num–M	1.45
SC–HSC	E–E	0.29
	M–M	1.24
OLSAT–Junior SA	SAI–E	0.39
	SAI–M	0.36
	VS–E	0.53
	NV–M	0.53
OLSAT–Senior SA	SAI–E	0.64
	SAI–M	0.55
	VS–E	0.84
	NV–M	0.71
NAPLAN–Junior	Lit–E	0.22
	Num–M	0.58
NAPLAN–Senior	Lit–E	0.81
	Num–M	0.90
Junior SA–SC	E–E	0.41
	M–M	0.08
SC–Senior SA	E–E	0.56
	M–M	0.74
Junior SA–HSC	E–E	0.03
	M–M	0.06
Senior SA–HSC	E–E	0.30
	M–M	0.26
Junior SA–Senior	E–E	0.46
	M–M	0.52
OLSAT–SA	SAI–Sem1	0.66
Weighted Average		0.65

5.7 Discussion

This chapter described the convergence evidence that was gathered to assist in the evaluation of the validity of the extrapolation inference (i.e., the use of information about a student's expected achievement and actual achievement to determine the student's gifted underachievement) in the interpretation/use argument. Specifically, evidence relating to the convergence of the five commonly used methods to identify and measure gifted underachievement (i.e., absolute split I, absolute split II, regression, simple difference, and nomination) was assessed across 41 different combinations of expected achievement and actual achievement measurements for students attending a school in the Sydney metropolitan area. Three different approaches were used to contribute evidence towards the assessment of convergence.

The first approach involved making comparisons of the proportions of identification of gifted underachievement. Consistent with the findings of Annesley et al. (1970), which used a similar approach and determined that the identification methods differed significantly, the results from this investigation suggested that the five identification methods do not appear to identify the same proportions of students exhibiting gifted underachievement. Two methods in particular, the absolute split II method and the regression method, appeared to be particularly non-convergent with the other methods. Nevertheless, when individual pairs of identification methods were compared, some support for convergence was found between the nomination method and the two absolute split methods, and between the nomination method and the simple difference method. Furthermore, it was also noted that despite a lack of evidence of statistical convergence, the difference in proportions of gifted underachievement identified using the simple difference and absolute split I methods was very small.

The second approach involved an examination of the degree of association and correlation between the results of the methods used to identify and measure gifted underachievement. Lau and Chan (2001a) previously assessed the association of results from a number of identification methods and determined that only the statistical methods were closely related. In the current investigation, the strength of the relationships between the statistical identification methods were found to be moderately strong (with the strongest being between the two absolute split methods, and between absolute split I and the simple difference method), but below the level necessary for convergence, while the relationships between the nomination method and each of the statistical methods were consistently *weak* (Park et al., 2014). Therefore, while the results of this investigation appear to be broadly similar to that of Lau and Chan (2001a), they were nevertheless interpreted as being non-supportive of convergence between the five identification methods. It is noted that when a correlation analysis was undertaken of measurements obtained from the two methods that also measure the degree of gifted underachievement, a *nearly perfect* linear relationship, suggesting a high level of convergence between the simple difference and regression methods, was found.

The third approach used to collect convergence evidence was an examination of the degree of agreement between the results of the methods used to identify and measure gifted underachievement. While agreement appeared to exist occasionally for some pairs of identification methods for a limited number of data combinations, the weighted average agreement levels for each pair of identification methods was below the recommended thresholds for convergence (nevertheless, the size of the weighted average agreement between the absolute split I and absolute split II methods, and between the absolute split I and simple difference methods, was close to this threshold). In particular, the regression method and the nomination method appeared to have the lowest levels of convergence with the other

methods. Similarly, assessments of the degree of agreement between the measurements of gifted underachievement from the simple difference and regression methods, using concordance correlation coefficients, mean differences in measurements, and Bland–Altman plots, failed to identify substantial support for convergence between these methods.

While convergence did appear to exist for some of the identification/measurement methods with some specific data combinations, generally the evidence collected in this chapter indicated that convergence was not supported. Indeed, consistent evidence across the multiple approaches used to assess convergence could not be found for *any* pair of identification/measurement methods. For example, the proportional evidence showing convergence between the nomination method and three of the statistical identification methods ($\delta \leq 0.05$ with the McNemar's test results showing no significant difference), contrasted with the evidence obtained from the association and agreement approaches, which suggested non-convergence ($\phi < 0.35$ and $\kappa < 0.00$ respectively). Similarly, correlational evidence indicating a high level of convergence between the simple difference and regression measurements of gifted underachievement ($\bar{r} = 0.97$) contrasted with multiple sources of evidence using the agreement approach, which uniformly suggested non-convergence ($CCC = 0.66$; $\bar{d} = 0.59$ which was found to be significantly different using a *t*-test; $CI = 0.65$). Nevertheless, among the various identification methods, the strongest level of convergence appeared to exist between the absolute split I and absolute split II methods ($\overline{|\delta|} = 0.05$; $\bar{\phi} = 0.65$; $\bar{\kappa} = 0.69$), and between the absolute split I and simple difference methods ($\overline{|\delta|} = 0.05$; $\bar{\phi} = 0.62$; $\bar{\kappa} = 0.64$). Generally, the measurements obtained from these pairs of methods tended to be close to the required thresholds for statistical convergence.

5.8 Summary

This chapter has collected evidence to assess the convergence of the different methods that identify and measure gifted underachievement. The evidence shows that convergence between the various methods used to identify and measure gifted underachievement is unsupported. The next chapter will discuss the collection of criterion evidence to assess the validity of the extrapolation inference of the interpretation/use argument.

6 Criterion Evidence

6.1 Introduction

This chapter provides criterion evidence to answer the first research question that guided the project: *'Is the extrapolation inference reasonable for each of the methods to identify and measure gifted underachievement?'*. In Chapter 4, a determination was made that an assessment of the extrapolation inference (i.e., the use of information about a student's expected achievement and actual achievement to determine the student's gifted underachievement) required the collection of both convergence evidence (presented in Chapter 5) and criterion evidence. Criterion evidence may be collected by comparing the measurements obtained using the various identification methods to criterion values (i.e., true values, or at least well-accepted estimates of the true values; Kane, 2006; 2013). This chapter outlines how a criterion value may be established for the identification of gifted underachievement, and uses that criterion value to assess the validity of the extrapolation inference.

6.2 Accuracy

The process of obtaining criterion evidence by making comparisons of measurements to criterion values may be referred to as the process of making assessments of *accuracy*. Accuracy is a statistical measure of the correctness of classifications and is commonly used to assess the usefulness of new diagnostic tests (Alonzo & Pepe, 1999). Accuracy may be considered equivalent to the convergence of a measurement with the *true* value. Hence, the calculations described in Chapter 5 to assess convergence may be used to assess accuracy if comparisons are made with criterion values.

The concept of accuracy is sometimes broken down into two sub-components called sensitivity and specificity. *Sensitivity* refers to the number of correct classifications of

positive identifications (e.g., gifted underachievement) divided by the total number of positive identifications, and provides an indication of how well the classification method avoids making false positive identifications (Type I errors). In contrast, *specificity* refers to the number of correct classifications of false identifications (e.g., gifted achievement) divided by the total number of false identifications, and provides an indication of how well the classification method under examination avoids making false negative identifications (Type II errors). Table 28 is a contingency table that outlines the possible classifications of conditions to demonstrate the concepts of sensitivity and specificity.

Table 28

Demonstration of accuracy concepts

		True classification	
		Condition present	Condition not present
Classification from method	Condition present	True positive	False Positive (Type I error)
	Condition not present	False negative (Type II error)	True negative

6.3 Latent Class Analysis

Kane (2006, 2013) has noted that the construction of a suitable criterion value may be difficult when the variable being measured is not directly observable. A commonly adopted solution to the problem is the use of latent class analysis (Christensen et al. 1992; Collins & Huynh, 2014; Delaney, Holder, Allan, Kenkre, & Hobbs, 2003; Espeland & Handelman, 1989; Goldberg & Wittes, 1978; Hadgu, Dendukuri, & Hilden, 2005; Walter, Frommer, & Cook, 1991). Mammadov, Ward, and Riedl (2016) have argued that the field of gifted education would benefit from using this methodological tool.

Latent class analysis is useful in situations where multiple classification methods are used to classify the same variable, the true classification of which is not directly observable. Specifically, it allows the classifications from the multiple methods to be combined to calculate an estimate of the most likely true classification for each participant. The latent class analysis approach assumes that a true underlying condition (e.g., gifted underachievement) exists, and that each of the different classification methods are influenced by this underlying condition. Estimations of accuracy may then proceed by a comparison of the results of each individual classification method to the estimated true classification. Although latent class analysis may be carried out on many statistical software packages, the *poLCA* statistical package within the *R* software environment (Linzer & Lewis, 2011) was used in this investigation.

For the latent class approach to be useful, there must be more variables that are input (i.e., the number of methods of identification of gifted underachievement) than the number of classifications (i.e., gifted achievement and gifted underachievement; Agresti, 2013). This requirement was satisfied with respect to the identification of gifted underachievement, as five identification methods (i.e., absolute split I, absolute split II, nomination, regression, and simple difference) are used to estimate two true classifications (i.e., gifted achievement and gifted underachievement). Nevertheless, it was not satisfied with respect to the measurement of gifted underachievement, as only two measurement methods (i.e., regression and simple difference) are used to estimate two true classifications (i.e., gifted achievement and gifted underachievement). Hence, criterion values could only be estimated for the identification of gifted underachievement, and latent class analysis was not carried out to assess the accuracy of the measurement of gifted underachievement.

6.4 Results and Discussion

Latent class analysis was undertaken on the entire data set to create a single latent class model that provided criterion value estimates. Table 29 contains:

(a) Contingency tables relating to each of the five identification methods that show the proportion of cases that were in agreement or disagreement with the latent class model; and

(b) Measurements of accuracy of the results obtained from each of the five identification methods with the latent class model. The following calculations were made: Cohen's kappa (κ) agreement, phi (ϕ) association, and the difference between proportions (δ) identified.

Table 29

Accuracy results using a latent class model

Method		Latent Class		Accuracy		
		GUA	GA	κ	ϕ	δ
SD	GUA	0.89	0.09			
	GA	0.11	0.91	0.80	0.80	–0.01*
REG	GUA	0.54	0.03			
	GA	0.46	0.97	0.51	0.56	–0.22*
ABSI	GUA	0.95	0.05			
	GA	0.05	0.95	0.90	0.90	0
ABSII	GUA	0.87	0.20			
	GA	0.13	0.80	0.67	0.67	0.04*
NOM	GUA	0.46	0.79			
	GA	0.54	0.21	–0.33	–0.34	0.13*

Note. GUA = Gifted Underachievement; GA = Gifted Achievement; SD = Simple difference; REG = Regression; ABSI = Absolute split I; ABSII= Absolute split II; NOM = Nomination
*$p < 0.05$ from McNemar's test

The results demonstrated that the absolute split I method had the highest degree of accuracy out of the five identification methods. Among the other methods, the simple

difference method also had a high level of accuracy, while the absolute split II and regression methods appeared to have more moderate levels of accuracy. It is noteworthy that both the absolute split I and simple difference methods had: (a) kappa agreement values classifiable as *almost perfect* agreement with the latent class model (Kundel & Polansky, 2003), (b) association levels with the latent class model that were above the required threshold for convergence ($\phi > 0.70$; Park et al., 2014), and (c) a difference in identified proportions of gifted underachievement with the latent class model of 0.01 or less. Nevertheless, of these two methods, McNemar's test indicated that only the absolute split I method produced a proportion of identification of gifted underachievement that was *not* statistically significantly different to the latent class model.

The analyses generally suggested that the nomination method may be the least accurate. For example, a negative Cohen's kappa statistic suggested that the nomination method had less agreement with the latent class model than may be expected by chance, while the negative phi coefficient indicated that the students who were *not* identified by teachers as exhibiting gifted underachievement were more likely to be exhibiting gifted underachievement according to the latent class model. Moreover, the nomination method identified 13% more students as exhibiting gifted underachievement than the latent class model, which was a statistically significant difference.

These results strongly suggest that the validity of the nomination method is not supported. As the inclusion of any method in the latent class analysis influences the latent class model that is constructed, the analyses were repeated without the nomination method. Table 30 provides details of the analyses relating to the refined latent class model. In this model, all methods, except the regression method, had a very small difference in the proportion of students identified as exhibiting gifted underachievement, and two methods (i.e., the simple difference and absolute split I methods) had associations and agreements

greater than the threshold for convergence (ϕ > 0.70, Park et al., 2014; $\kappa = 0.75$, Walts et al., 2011). However, a substantial change to the Cohen's kappa statistic for the regression method was noted. As the new results indicated that the regression method may have only poor agreement ($\kappa < 0.40$) with the criterion model and also identified 30% less cases of gifted underachievement than the latent class model it was decided that the inclusion of the regression method may distort the model. Accordingly, the analyses were repeated once again without the regression method.

Table 30

Accuracy results using a latent class model with nomination method removed

Method		Latent Class		Accuracy		
		GUA	GA	κ	ϕ	δ
SD	GUA	0.89	0.09			
	GA	0.11	0.91	0.80	0.80	–0.01*
REG	GUA	0.40	0.01			
	GA	0.60	0.99	0.39	0.50	–0.30*
ABSI	GUA	0.94	0.05			
	GA	0.06	0.95	0.89	0.89	–0.01*
ABSII	GUA	0.85	0.21			
	GA	0.15	0.79	0.64	0.64	0.03*

Note. GUA = Gifted Underachievement; GA = Gifted Achievement
*p < 0.05 from McNemar's test

The final results indicated that the absolute split I ($\kappa = 0.97$; $\phi = 0.97$; $\delta = 0.01$) and simple difference ($\kappa = 0.75$; $\phi = 0.75$; $\delta = -0.02$) methods continue to have a high level of accuracy, according to measures of agreement and association with the latent class model that satisfy widely adopted thresholds for convergence (ϕ > 0.70, Park et al., 2014; $\kappa = 0.75$, Walts et al., 2011). These results (refer Table 31) suggested that both the absolute split I and simple difference methods may be used interchangeably with the criterion variable. The third identification method (i.e., the absolute split II method) was only slightly below the

relevant thresholds. All three methods appeared to have small differences in the proportions of gifted underachievement cases identified in comparison to the latent model (i.e., absolute value of 0.01 to 0.03). Although the McNemar test found that these differences in proportions were statistically significant, this may be a consequence of the large sample size (i.e., 4,988) used in the analyses (Field, 2013; Hair et al., 2010).

Table 31

Accuracy results using a latent class model with nomination and regression methods removed

Method		Latent Class		Accuracy		
		GUA	GA	κ	ϕ	δ
SD	GUA	0.86	0.11			
	GA	0.14	0.89	0.75	0.75	–0.02*
ABSI	GUA	0.99	0.02			
	GA	0.01	0.98	0.97	0.97	0.01*
ABSII	GUA	0.86	0.20			
	GA	0.14	0.80	0.66	0.66	0.03*

Note. GUA = Gifted Underachievement; GA = Gifted Achievement
*$p < 0.05$ from McNemar's test

The results show that the validity of the extrapolation inference with respect to two of the methods used to identify gifted underachievement (i.e., absolute split I and simple difference), is supported. The support, however, is not equal, with greater support for the absolute split I method than the simple difference method. Simultaneously, the validity of the extrapolation inference for the nomination and regression methods was not supported by the results. Hence, the validity of both the nomination and regression methods to identify gifted underachievement remains questionable.

6.5 Summary

This chapter has collected criterion evidence to assess the validity of the different methods that identify gifted underachievement. The evidence demonstrated that criterion

validity is only supported for two of the statistical methods used to identify gifted underachievement (i.e., the absolute split I and simple difference methods). In the next chapter, evidence was collected to assess the validity of the generalisation inference of the interpretation/use argument for the project.

7 Generalisation Evidence

7.1 Introduction

This chapter provides empirical evidence to answer the second research question that guided the project: *'Is the generalisation inference reasonable for each of the methods to identify and measure gifted underachievement?'*. The generalisation inference refers to the identification or measurement of gifted underachievement for any individual identification/ measurement method, irrespective of: (a) the data combinations used to assess expected achievement and actual achievement (for the absolute split I, absolute split II, simple difference, and regression methods), or (b) the nominator (for the nomination method). Therefore, this chapter attempts to ascertain whether the results from the statistical methods are generalisable across different data combinations, and whether the results from the nomination method are generalisable across different nominators. If there is no significant variation, the results are considered to be *homogeneous*. This chapter outlines the data analysis and results gathered for such homogeneity evidence.

7.2 Homogeneity

Homogeneity occurs when a group of studies investigating the same effect are found to have consistent effect sizes (the opposite of homogeneity is *heterogeneity*, which occurs when these studies are found to have inconsistent effect sizes). As this investigation gathered evidence across 41 different combinations of expected achievement and actual achievement measurements, the analyses relating to each data combination may be treated as a separate study for the conduct of meta-analyses to determine whether the results of the identification or measurement of gifted underachievement for each individual method are homogeneous. Unfortunately, the nomination method may not be assessed for generalisation using this approach as it was only examined in one study.

7.3 Meta-Analysis

Meta-analytical methods are typically required to assess the homogeneity of effect sizes across a series of studies that investigate the same issue (Borenstein et al., 2009). It is a very powerful statistical approach to analyse and combine information from multiple studies, and is usually carried out to: (a) determine an overall effect size from a number of studies, (b) determine whether the effect is consistent (homogeneous), and if not, (c) determine whether there are any patterns in the inconsistency (heterogeneous) of the effect. In this project, the overall effect sizes across the 41 different data combinations for each identification/ measurement method have already been calculated and reported as weighted average statistics (refer Chapter 5). Therefore, this chapter is focused on an examination of the homogeneity of the effect sizes across the 41 different data combinations for each individual identification/measurement method. When homogeneity was not supported, further follow up investigations were undertaken to determine whether the inconsistency may be explained by the differences in the data combinations.

7.4 Test of Homogeneity

The general process to carry out a homogeneity test requires multiple steps:

(a) The effect size is calculated for each study;

(b) A weighted average of the effect sizes is calculated across all studies (Borenstein et al., 2009; Appendix 12);

(c) A test statistic, *Q*, is calculated to measure the total amount of variation in the effect size across the different studies (Appendix 12); and

(d) A statistical test is carried out by comparing the measured amount of variance in the effect size, *Q*, to the amount of expected variance (which is based on the number of studies examined and a chi-square distribution).

Additionally, two further statistics may be calculated to summarise the results of a homogeneity test. The first of these is the standard deviation of effect sizes across the investigated studies (i.e., *T*) which may be used to determine 95% confidence intervals for the weighted average effect sizes. The second statistic (i.e., I^2), calculates the proportion of observed variance that is real. Therefore, *T* is a measure of the amount of heterogeneity, and I^2 is an indication of how much variation in the effect size is due to this heterogeneity. This section will report the homogeneity test results and statistics to examine the homogeneity of the individual methods that identify or measure gifted underachievement.

7.4.1 Homogeneity results and discussion

A test of homogeneity was carried out for each of the statistical methods that identify/measure gifted underachievement across the 41 data combinations that were reported in Chapter 5. Specifically, tests were carried out for the following:

(a) the proportion of cases identified as gifted underachievement (% GUA) for the absolute split I, absolute split II, simple difference, and regression methods; and
(b) the mean measurement of the degree of gifted underachievement ($\bar{\mu}$) for the simple difference and regression methods.

Table 32 summarises the results from these tests, and reports the weighted average effect size for each statistic across the 41 data combinations, the 95% confidence interval for each weighted average effect size, the standard deviation of the effect sizes (*T*), the proportion of observed variance that is real (I^2), the homogeneity test statistic (*Q*), and the probability that the test statistic comes from a homogeneous group of effect sizes (*p*-value).

Table 32

Homogeneity test results

Statistic	Weighted average value (± 95% CI)	T	I^2	Q	p-value for homogeneity
% GUA: ABSI	0.32 ± 0.32	0.17	0.76	165	$p < 0.001$
% GUA: ABSII	0.45 ± 0.29	0.15	0.74	139	$p < 0.001$
% GUA: SD	0.33 ± 0.33	0.17	0.77	170	$p < 0.001$
% GUA: REG	0.13 ± 0.00	0.00	0.00	6	1.00
$\bar{\mu}$: SD	0.82 ± 1.06	0.54	0.97	1366	$p < 0.001$
$\bar{\mu}$: REG	0.00 ± 0.18	0.09	0.48	77	$p < 0.001$

Note. GUA = Gifted Underachievement; % GUA = proportion of gifted underachievement identified; $\bar{\mu}$ = mean measurement

The results from the tests of homogeneity suggest that only one of the studied statistics (i.e., the proportion of cases of gifted underachievement identified by the regression method) show homogeneous results across the different data combinations. The probability of the other statistics being homogeneous was found to be very low ($p < 0.001$). The homogeneity of the identified proportions of gifted underachievement obtained using the regression method is not unexpected due to the nature of the method which uses a line of best fit between the expected and actual achievement measures to assess underachievement, and therefore adjusts for variations in the expected and actual achievement measures in different data combinations (McCall et al., 2000). All the other tested statistics were found to show heterogeneity across the different data combinations.

While most of these statistics were found to have similar levels of heterogeneity (T) and percentages of real variation (I^2), one exception was the mean measurement of gifted underachievement from the regression method ($\bar{\mu}$). In this case, the degree of heterogeneity was small (which was measured by the T statistic, in standard deviation units, to be 0.09) and the percentage of variance found not to be due to errors of measurement was much lower (measured by I^2 to be 48%) than the other variables tested for homogeneity. The small

amount of heterogeneity and low percentage of real variance for this statistic suggests that the heterogeneity, while statistically significant, may perhaps not be of practical significance.

Overall, the results of the homogeneity tests suggested that, with the exception of the regression method to identify gifted underachievement, the statistical methods used to identify and measure gifted underachievement may not be homogeneous across different combinations of expected achievement and actual achievement data.

7.5 Meta-Regression

As the results of the homogeneity tests suggested that homogeneity may not exist for almost all of the statistics obtained from the statistical methods used to identify and measure gifted underachievement, a meta-regression analysis was carried out to explore possible explanations for the heterogeneity. Meta-regression analysis is a procedure that may be used to determine the percentage of variance in heterogeneous statistics that is explainable by the variance of other variables (Field, 2013; Hair et al., 2010). While meta-regression analysis follows the principles of regular regression analysis, it extends regression analysis by summarising a pattern across multiple studies.

To carry out a meta-regression analysis, possible variables that may explain a portion of the variance in the heterogeneous statistics need to be selected. In consideration of the fact that the regression method was found to identify a homogenous proportion of gifted underachievement across different data combinations, it was hypothesised that the properties of the line of best fit between the expected and actual achievement measurements (i.e., the regression line) may explain some of the heterogeneity in the measurements obtained from the other statistical identification/measurement methods. Three variables associated with the regression line (i.e., the gradient [m], the intercept [b], and a measure of how well the line fits the data, typically provided by the Pearson correlation value [r]) were therefore selected as

three possible predictor variables. The values of these three variables for the different data combinations were reported in Table 12.

7.5.1 Meta-regression results and discussion

The results from the meta-regression analyses are summarised in Table 33. They indicated that the three predictor variables individually explain large amounts of the variance of the heterogeneous statistics, with the Pearson correlation coefficient (*r*) explaining over 50% of the variance in three (i.e., $\bar{\mu}$:SD, % GUA:ABSI, and %GUA:SD) out of the five heterogeneous statistics. The findings suggested that variables associated with the relationship between the expected achievement and actual achievement measurements may individually explain substantial portions of the heterogeneity in the measurements obtained from the various statistical identification/measurement methods. Nevertheless, it is noted that none of the three predictor variables accounted for a substantial portion of the variance of the mean measurement of gifted underachievement from the regression method ($\bar{\mu}$: REG).

Table 33

Meta-regression results

Statistic	Percentage of variance of statistic explained by selected predictor variable (R^2)		
	m	*b*	*R*
% GUA: ABSI	0.16	0.44	0.54
% GUA: ABSII	0.25	0.00	0.37
% GUA: SD	0.20	0.38	0.52
$\bar{\mu}$: SD	0.30	0.55	0.69
$\bar{\mu}$: REG	0.02	0.07	0.02

Note. GUA = Gifted Underachievement; % GUA = proportion of gifted underachievement identified; $\bar{\mu}$ = mean measurement

7.6 Multiple Regression

In consideration of the fact that variables associated with the relationship between the expected achievement and actual achievement measurements may individually have an impact on the identification and measurement of gifted underachievement, and the *collective* impact of these variables may be *different* to the impact of individual variables, additional analyses were conducted using multiple regression analyses. Multiple regression analyses may be useful as it is able to examine the combined impact of multiple predictor variables associated with the relationship between the expected achievement and actual achievement measurements, to determine the total amount of variance in the heterogeneous statistics that is explainable by the relationship between the expected achievement and actual achievement measurements. It is noted that the sole reliance on the individual findings of the meta-regression analysis may lead to incorrect conclusions, as some of the variance explained by each individual predictor variables may be shared.

In this investigation, multiple regression models were developed for each of the heterogeneous statistics (i.e., % GUA:ABSI, %GUA:ABSII, %GUA:SD, $\bar{\mu}$:SD, and $\bar{\mu}$:REG) using the same three predictor variables (the gradient [m], the intercept [b], and the correlation [r] between the expected achievement and actual achievement measurements). The *SPSS* software package (version 22) was used to carry out the multiple regression analyses with the standard *Enter* method (Field, 2013).

7.6.1 Assumptions of multiple regression analysis

Multiple regression analysis makes several assumptions about the data being examined that need to be confirmed prior to the analysis (Field, 2013; Hair et al., 2010):

(a) The dependent variable is measured on a continuous scale;

(b) Two or more predictor variables exist;

(c) The independence of residuals (i.e., size of the error of the model for predicting the values of the dependent variable), which may be confirmed with a Durbin–Watson statistic that is close to 2;

(d) A normal distribution of the residuals, which may be visually confirmed using a histogram of residuals and a probability-probability plot (P–P plot);

(e) A linear relationship between the dependent and predictor variables, which may be visually confirmed using a scatter plot of each individual predictor variable and dependent variable, and a scatter plot of the combined predictor variable and the dependent variable;

(f) Homoscedasticity (i.e., same level of variance along the line of best fit), which may be visually confirmed using a scatter plot of each individual predictor variable and dependent variable, and a scatter plot of the combined predictor variable and the dependent variable (Figure 35 shows the patterns of the scatter plots that indicate homoscedasticity and heteroscedasticity);

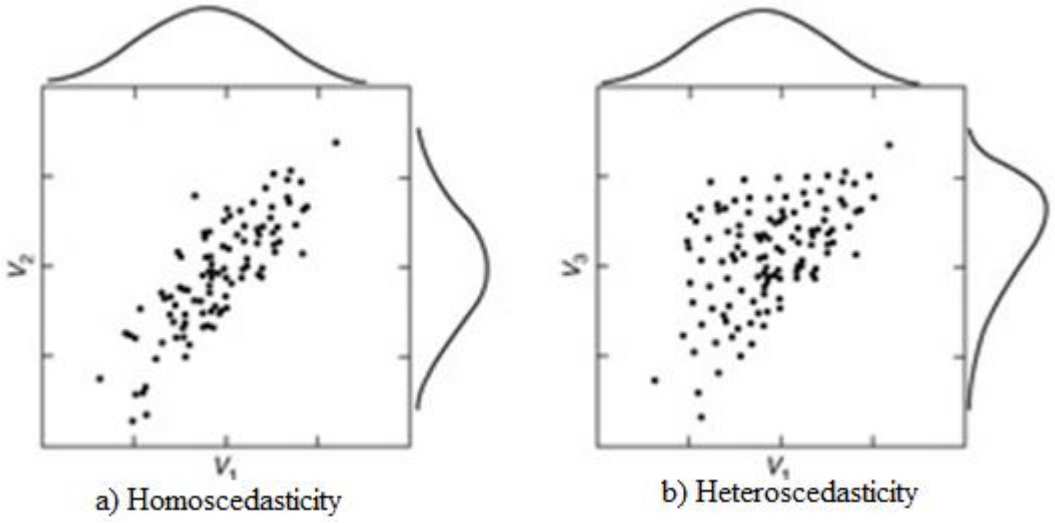


Figure 35. Comparison of homoscedasticity and heteroscedasticity (Stamatis, 2002)

(g) No perfect multicollinearity (i.e., the predictor variables must not be highly correlated), which may be confirmed by correlations of less than 0.9, or average variance inflation factor values that are not substantially greater than one; and

(h) No significant outliers, which may substantially influence the multiple regression model developed and bias the results. The presence of problematic data points may be detected using a range of statistical measures including the size of standardised residuals, Cook's distance values, leverage values, Mahalanobis distance values, covariance ratio values, and the size of change in regression coefficients when a case is deleted from the analysis.

The multiple tests of the assumptions of multiple regression analysis are presented in Appendix 11 and summarised in Table 34. The results show that all of the assumptions were consistently met for each statistic under investigation.

Table 34

Summary of multiple regression analysis assumption test results

	Assumptions of Multiple Regression Analysis met?					
Statistic	Durbin – Watson	Normality	Linear	Homoscedasticity	Collinearity	No. of influential cases removed
% GUA: ABSI	Yes	Yes	Yes	Yes	Yes	2
% GUA: ABSII	Yes	Yes	Yes	Yes	Yes	1
% GUA: SD	Yes	Yes	Yes	Yes	Yes	2
$\bar{\mu}$: SD	Yes	Yes	Yes	Yes	Yes	2
$\bar{\mu}$: REG	Yes	Yes	Yes	Yes	Yes	3

Note. GUA = Gifted Underachievement; % GUA = proportion of gifted underachievement identified; $\bar{\mu}$ = mean measurement

7.6.2 Multiple regression results and discussion

The results of the multiple regression analyses, with the heterogeneous statistics as the dependent variable and the variables describing the relationship between the expected and actual achievement measurements (i.e., gradient [*m*], *y*-intercept [*b*], and correlation [*r*]) as independent variables, are summarised in Table 35. The results demonstrate that the relationship between the expected achievement and actual achievement measurements explain a large percentage of the variation in almost all of the heterogeneous statistics. Specifically, the multiple regression models were able to explain more than 45% of the variation in four out of five heterogeneous statistics (i.e., %GUA:ABSI, %GUA:ABSII, %GUA:SD, and $\bar{\mu}$: SD), and an average, 59% of the variance in all heterogeneous statistics (70% when the regression method is excluded). Hence, the relationship between the expected achievement and actual achievement measurements appears to have a substantial impact on the identification and measurement of gifted underachievement under the various identification and measurement methods.

Table 35

Multiple regression results

Statistic	Multi-Regression model Coefficient of predictor variable				
	Model Constant	*m*	*b*	*R*	Model R^2
% GUA: ABSI	0.63*	–0.18	–0.22*	–0.42	0.70
% GUA: ABSII	0.88*	0.39	0.21*	–1.26*	0.46
% GUA: SD	0.61*	–0.39	–0.25*	–0.19	0.67
$\bar{\mu}$: SD	1.71*	–1.97*	–1.03*	0.24	0.97
$\bar{\mu}$: REG	–0.03	–0.37	–0.03	0.45	0.16

Note. GUA = Gifted Underachievement; % GUA = proportion of gifted underachievement identified; $\bar{\mu}$ = mean measurement

* $p < 0.05$

Interestingly, of the parameters of the multiple regression models, (i.e., *m*, *b*, *r*, and the model constant), the model constant and y intercept (*b*) were consistently identified to be statistically significant contributors of the identification/measurement of gifted underachievement across the different identification/measurement methods. Only two other parameters (i.e., *m* for the mean measurement of gifted underachievement from the simple difference method, and *r* for the proportion of identifications from the absolute split II method) made statistically significant contributions for individual identification/measurement methods. Of note, none of the parameters of the multiple regression model were found to make a statistically significant contribution to the measurement of gifted underachievement using the regression method.

7.7 Generalisation Evidence for the Nomination Method: Inter-Rater Agreement

To assess the generalisation inference of the interpretation/use argument for the nomination method, an investigation was made of the level of agreement among different nominators. Specifically, assessments were made using Kane's guidelines on inter-rater agreement (Kane, 2013). The use of meta-analytical procedures to assess the generalisation inference, as for the other methods of identification and measurement, was inappropriate, as the nomination method was only examined in one study.

The nomination data for this project was provided from teachers (i.e., between one and six teachers) who classified each student as exhibiting gifted achievement or gifted underachievement. To assess the level of agreement among the different teachers, a variation of the Cohen's kappa agreement statistic for multiple observers, as described by Agresti (2013) and Fleiss et al. (2003), was used. Typically, the Cohen's kappa statistic describes the agreement in classifications between two observers. In the situation where three or more

observers exist, a measure of the overall agreement in classifications may be calculated as the average of the Cohen's kappa agreement statistics for each possible pairing of observers. This overall multi-observer agreement statistic provides a single measurement of agreement, with the same scale and interpretation as the kappa (κ) statistic.

The overall multi-observer kappa agreement statistic for this project was found to be 0.31, which is below the adopted threshold for *fair* agreement (Fleiss et al., 2003; Walts et al., 2011) and indicative of a poor level of agreement among the nominators. The result was suggestive of only *weak* support for generalisation of the nomination method across different nominators. It is nevertheless noted that some portion of the disagreement may reflect differences in a student's level of performance in different subject areas (e.g., a student underachieving in English may not necessarily be underachieving in mathematics), rather than a lack of consistency in the nominations.

7.8 Summary

This chapter has collected evidence to assess the generalisability of the different methods that identify and measure gifted underachievement across the different data combinations used to assess expected and actual achievement. The evidence shows that, except for the regression identification method, generalisability is not supported. It was apparent that the relationship between the expected and actual achievement measurements may account for a substantial proportion of the heterogeneity in the different methods used to identify and measure gifted underachievement. The next chapter will make an overall assessment of the validity of each of the methods used to identify and measure gifted underachievement by considering all of the empirical evidence that was collected in this investigation.

8 Discussion

8.1 Introduction

Kane's framework suggests that the validity of the interpretation of the results obtained from the various methods used to identify/measure gifted underachievement may be determined by empirical evidence within the context of the use of these methods. The context therefore needs to be outlined, and should inform the validity arguments needed to evaluate the validity of each identification/measurement method for each inference in the planned interpretation/use argument. If any inferences are not found to be substantive, then the planned interpretation/use argument may need to be revised. This chapter presents validity arguments that combine the empirical evidence gathered in Chapters 5, 6 and 7, with the planned context of the use of these methods, to allow for a final assessment of the validity of the various methods used to identify/measure gifted underachievement.

8.2 Context Revisited

There are multiple uses of the various methods that identify/measure gifted underachievement:

(a) The selection of gifted students by educators and counsellors for participation in intervention programs which aim to reverse underachievement (Reis & McCoach, 2000; Rimm, 2003). Nevertheless, as "students underachieve for so many different reasons" (Reis & McCoach, 2000, p. 152), an automatic placement of an underachieving gifted student into any intervention program may not be appropriate. Furthermore, the adequate provision of such interventions may be difficult, due to the large numbers of students who may be involved (Hsieh et al., 2007), and the diversity of the interventions that are likely to be necessary (Figg et al., 2012).

(b) The selection of gifted students for participation in research studies which aim to advance understanding of gifted underachievement, and advance understanding of the appropriate interventions for gifted underachievement. The findings of such studies may allow for the provision of advice to parents, schools, policymakers, and students on how to optimally prevent or ameliorate gifted underachievement (Jovanović, Teovanović, Mentus, & Petrović, 2010).

(c) The identification and measurement of underachievement in all students, including those in the general student population and students with learning disabilities (Jones, 2005; Lau & Chan, 2001a; Maki et al., 2015; Zirkel & Thomas, 2010).

The contexts of the planned uses of the methods that identify and measure gifted underachievement appear to highlight the importance of accurate identification and measurement of underachievement, to enable the provision of appropriate educational interventions for these students and the subsequent realisation of their potential.

8.3 Interpretation/Use Argument Revisited

To assess the validity of the interpretation of the results obtained from the various methods used to identify/measure gifted underachievement, an outline of the logical steps in the interpretation/use argument for this project was presented in Chapter 3 (refer Figure 12). The proposed interpretation/use argument included the scoring, generalisation, and extrapolation inferences separately for both of the instruments used to measure expected achievement and actual achievement. Thereafter, it included three additional inferences that had yet to be empirically established:

(a) Extrapolation: The information from the instruments designed to assess expected achievement and actual achievement are combined for an assessment of an individual's gifted underachievement; and

(b) Generalisation: Any individual method used to identify/measure gifted underachievement will identify/measure gifted underachievement independently of the data combinations (or nominators) used to assess expected achievement and actual achievement; and

(c) Decision: The placement of an underachieving gifted student into a specific intervention program.

Kane (2013) suggested that convergence and criterion evidence may be collected to assess the extrapolation inference, and that a generalisability study may be conducted to assess the validity of the generalisation inference. The convergence, criterion, and generalisation evidence to support these two inferences were reported and discussed in Chapter 5, Chapter 6, and Chapter 7, respectively. A determination was made (refer Chapter 3) that the decision inference should not be evaluated until after both the extrapolation and generalisation inferences were supported, and the accuracy of at least one method of identifying/measuring gifted underachievement was established.

8.4 Validity Arguments

This investigation gathered empirical evidence to assess the appropriateness of the extrapolation and generalisation inferences from the proposed interpretation/use argument for each of the identification/measurement methods. This section will discuss the empirical evidence for the validity of each of these inferences for each identification/measurement method.

8.4.1 Extrapolation inference

8.4.1.1 Absolute split I

The extrapolation inference relating to the absolute split I method for the identification of gifted underachievement appeared to have the most empirical support of all

the methods studied in this investigation. Strongly supportive criterion evidence was identified for the method when the results of the absolute split I method were assessed with the criterion estimates provided through latent class analysis. A weaker level of support was found for the convergence of the results of the absolute split I method with two of the other identification methods (i.e., the absolute split II and simple difference methods). Overall, the empirical evidence is supportive of the validity of the extrapolation inference for the absolute split I method.

Despite the empirical support, the absolute split I method has a number of limitations. One major limitation is that many gifted students who are underachieving may not be identified as such under this method (and therefore remain as invisible gifted underachievers), due to the fixed expected achievement thresholds that are used to assess the underachievement of gifted students of different ability levels. That is, as the expected achievement of both highly and moderately gifted students are considered to be the same under the absolute split I method, highly gifted students who demonstrate actual achievement at the level of moderately gifted students may *not* be classified as exhibiting gifted underachievement. A possible solution to the problem may be to introduce a variation to the absolute split I method that utilises multiple combinations of expected achievement/actual achievement thresholds (e.g., 99th percentile expected achievement/below 85th percentile actual achievement and 90th percentile expected achievement/below 75th percentile actual achievement etc.). Nevertheless, such an approach will require the setting of a number of arbitrary thresholds, which may lead to other problems.

A second limitation of the absolute split I method is that it only classifies students as exhibiting gifted underachievement or gifted achievement. Further information, including information on the degree of gifted underachievement, is not provided by the method, although such information may be useful in the identification of students who may have the

greatest need for such interventions, and the formulation of appropriate educational intervention programs for the identified students.

8.4.1.2 Absolute split II

Only a weak level of empirical support was found for the extrapolation inference relating to the absolute split II method. While the convergence evidence suggested that the results of the absolute split II method converge best with the results of the absolute split I method, the level of convergence was not enough to conclude that the two methods may be used interchangeably. Furthermore, the level of convergence with the other identification methods was typically poor. Similarly, the criterion evidence, obtained by assessing the convergence of the criterion estimates from latent class analysis with the results of the absolute split II method, was only marginally supportive of the method.

The limitations of the absolute split I method also apply to the absolute split II method. Nevertheless, an additional limitation is that the absolute split II method uses raw achievement scores as its threshold for actual achievement, rather than standardised ranks. The use of raw scores may be problematic as it may not take into consideration the difficulty of the achievement instrument used, and therefore may not be easily interpreted to indicate a student's level of actual achievement.

8.4.1.3 Nomination

The nomination method appeared to have the least empirical support as a method for the identification of gifted underachievement. The convergence evidence indicated that the results of the nomination method were some of the most divergent from the results of the other identification methods, while the criterion evidence suggested that the nomination results had less agreement with the criterion estimates than may be expected with a random classification of students (i.e., students *not* identified by the nomination method were more

likely to be exhibiting underachievement). The findings do not support the extrapolation inference relating to the nomination method.

8.4.1.4 Regression

As for the nomination method, the regression method for the identification and measurement of gifted underachievement appears to have minimal empirical support. The low level of convergence of the results of the regression method with the results of all the other identification methods was exemplified by: (a) the large differences in the proportions of gifted underachievement identified using the method in comparison to the other methods, and (b) a level of agreement with the other identification methods that was only slightly greater than that may be expected by chance. Similar conclusions about convergence were reached when the regression method was examined as a measure of the degree of gifted underachievement, and compared to the simple difference measure of gifted underachievement. As for the convergence evidence, the criterion evidence indicated that the results of the regression method had inadequate levels of agreement and association with the criterion estimates obtained using latent class analysis.

It is noteworthy that Van den Broeck claims that the regression method "is logically inconsistent with the concept of underachievement" (2002a, p.197) and that the "regression adjustment… is the direct source of the lack of validity" (2002b, p. 209). Unlike the other methods which compare a student's expected achievement and actual achievement, the regression method compares a student's actual achievement to patterns of achievement from the other students in the group. As arguably "almost all students" do not work at the level they are capable of (Reis & McCoach, 2000, p. 157), the regression method is likely to underestimate how much achievement may be expected from each student, and hence, underestimate the level of underachievement. This expectation was consistently supported by the results in this investigation which showed that the regression method always identified a

substantially lower proportion of cases of gifted underachievement than any of the other methods. Essentially, the regression method appears only to identify cases of *local* underachievement (i.e., students who are underachieving in comparison to the sample studied).

Although both the absolute split I and simple difference methods may, arguably, also identify cases of local underachievement (i.e., due to their reliance on the distribution of student scores during the standardisation procedure), the same problem of under-identification does not appear to exist for these methods. This may be so, as unlike for the regression method, their reliance on the distribution of student scores for the expected and actual achievement measurements are *independent* (i.e., rather than placing reliance on the local relationship between expected and actual achievements, the absolute split I and simple difference methods place reliance separately on the local expected achievement and local actual achievement). In addition, the problem may be overcome by the use of instruments that have been standardised in large populations (i.e., not restricted to a single school), such as the NAPLAN, OLSAT, HSC, and SC.

8.4.1.5 Simple difference

The simple difference method for the identification of gifted underachievement received varying levels of positive empirical support from the multiple tests that were undertaken. Specifically, the criterion evidence was supportive of the accuracy of the results obtained from the method when comparisons were made with criterion estimates obtained using latent class analysis, while the convergence evidence with the absolute split I and II methods (but not with the regression or nomination methods) was at a weaker level. Collectively, the empirical evidence appeared to support the validity of the extrapolation inference for the simple difference method.

It is noteworthy that when the simple difference method was assessed as a measure of the degree of gifted underachievement, and compared to the regression method as a measure of the gifted underachievement, the two measurement methods were found to be divergent (i.e., the tests of association and agreement between these methods provided inconsistent results).

8.4.2 Generalisation inference

Two variations of the generalisation inference were tested in this investigation. The generalisation inference for the absolute split I, absolute split II, regression, and the simple difference methods related to the identification or measurement of gifted underachievement for any individual identification/measurement method, irrespective of the combinations of data that were used to assess expected achievement and actual achievement. For the nomination method, which does not explicitly utilise data obtained from any instruments, the generalisation inference related to the generalisability of the results obtained using the nomination method across different nominators. This investigation gathered evidence to determine whether the generalisation inference for the different identification/measurement methods was supported.

8.4.2.1 Absolute split I

The empirical evidence suggested that the generalisation of the results from the absolute split I method for the identification of gifted underachievement was not supported. Specifically, the proportion of cases identified as gifted underachievement by the absolute split I method varied significantly when different data combinations were used to assess the expected achievement and actual achievement of gifted students. Follow-up analyses showed that 70% of these variations may be explained by the relationship between the expected achievement and actual achievement measurements. Overall, the results indicated that the relationship between the expected achievement and actual achievement measurements, in

each data combination, may have a substantial impact on the results obtained using the absolute split I method to identify gifted underachievement.

8.4.2.2 Absolute split II

As for the absolute split I method, the empirical evidence suggested that the generalisation of results from the absolute split II method was not supported. That is, the proportion of cases of gifted underachievement identified by the absolute split II method was found to vary substantially as different data combinations were used to assess expected achievement and actual achievement. The follow-up analyses determined that 46% of the variations observed in the proportion of gifted underachievement identified using the absolute split II method was explainable by the relationship between the expected achievement and actual achievement measurements. Consequently, and as for the absolute split I method, the relationship between the expected achievement and actual achievement measurements, in each data combination, may have a major impact on the results obtained using the absolute split II method to identify gifted underachievement.

8.4.2.3 Nomination

As for the two absolute split methods, the empirical evidence was not supportive of the generalisation of the results obtained from the nomination method. Specifically, when the level of agreement between nominators in the classification of the same students as exhibiting gifted achievement or gifted underachievement was examined, a poor level of agreement was established. It therefore appeared that the individual teachers making the nomination may differ substantially in the identification of gifted underachievement. The finding was supportive of the concerns, expressed by many scholars, that teacher-based nominations may be highly subjective (Dunne & Gazeley, 2008; Jones & Myhill, 2004; Lau & Chan, 2001a).

8.4.2.4 Regression

Unlike the other methods used to identify gifted underachievement, the empirical evidence suggested that the generalisation inference relating to the results of the regression method (as a method of identification of gifted underachievement) was supported. Specifically, the regression method was found to consistently identify the same proportion of cases of gifted underachievement (13%) regardless of the data combinations used to assess expected achievement and actual achievement. In contrast, support was not found for the generalisability of the regression method as a method of measurement of the degree of gifted underachievement. Follow up analyses determined that 16% of the variation in the mean measurement of the degree of gifted underachievement may be explained by the relationship between the expected achievement and actual achievement measurements in each data combination.

8.4.2.5 Simple difference

The generalisation inference relating to the simple difference method was not supported. That is, the proportion of cases of gifted underachievement identified by the simple difference method was found to vary substantially as different data combinations were used to assess expected achievement and actual achievement. Follow-up analyses suggested that 67% of the variations may be explained by the relationship between the expected achievement and actual achievement measurements in each data combination. Similarly, the simple difference method, when used as a method to measure the degree of gifted underachievement, was found to vary substantially across the different data combinations. The relationship between the expected and actual achievement measurements in each data combination was able to explain 97% of the variation in the mean measurement of gifted underachievement. The results indicate that the relationship between the expected achievement and actual achievement measurements, for each data combination, may have a

substantial impact on the identification and measurement of gifted underachievement using the simple difference method.

8.5 Revised Interpretation/Use Argument

The empirical evidence suggested that while the extrapolation inference may only be considered valid for the absolute split I and simple difference methods (as methods for the identification of gifted underachievement), the generalisation inference may not be reasonable for any of the identification and measurement methods (except for identifications from the regression method). Therefore, the original interpretation/use argument was revised to demonstrate the possible valid uses of these methods. Specifically, the generalisation inference was removed from the previous interpretation/use argument (refer Figure 36).

An implication of this change is that the choice of instruments used to assess expected achievement and actual achievement may be important in the interpretation of the results obtained using the identification/measurement methods. Furthermore, as the extrapolation inference was not supported for the nomination or regression methods, and only weakly supported for the absolute split II method, the revised interpretation/use argument should only be considered valid for the absolute split I and the simple difference method as methods of identification of gifted underachievement.

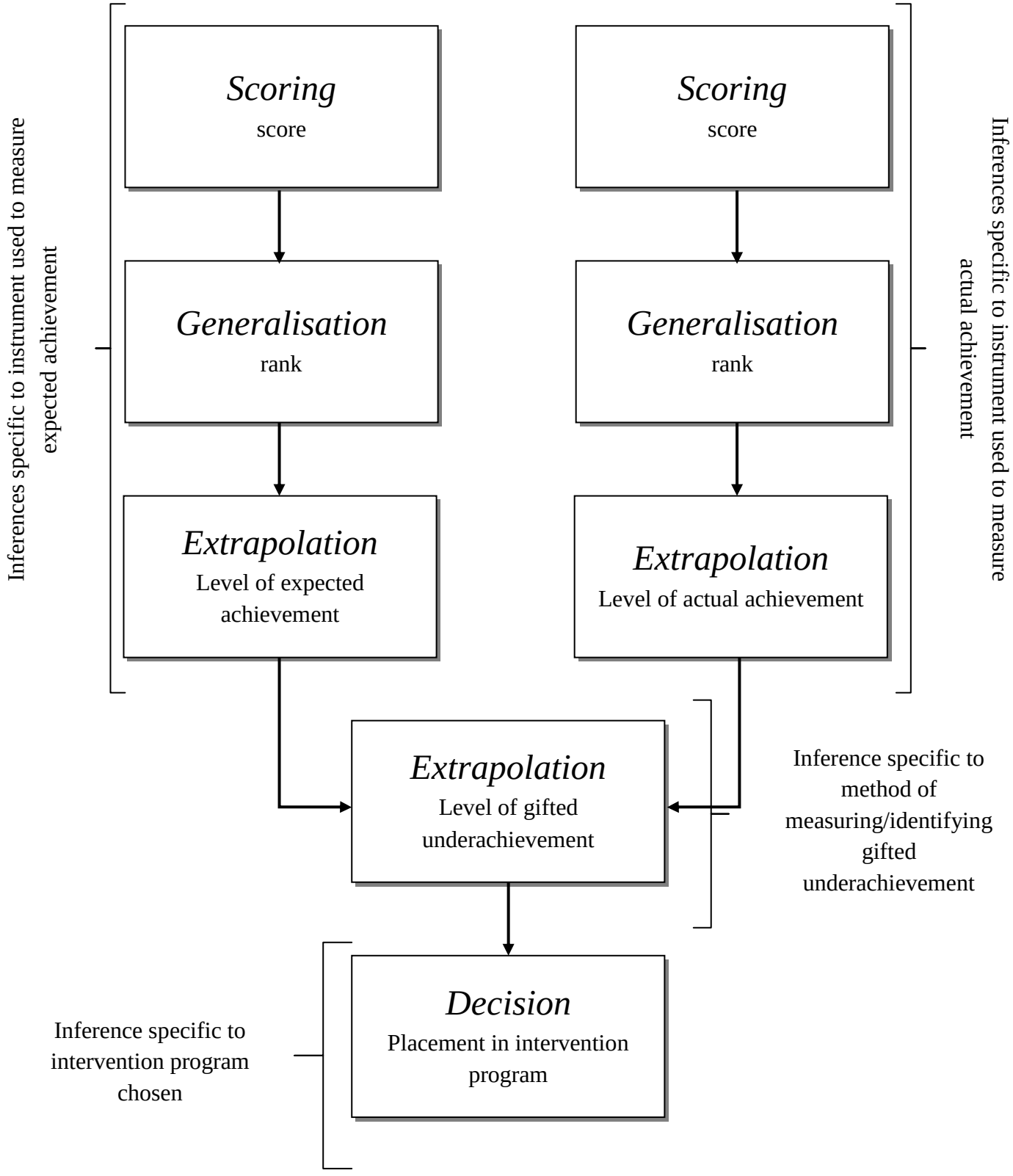


Figure 36. Revised network of logical steps for the interpretation of ability and achievement scores

8.6 An Overall Assessment

The convergence and criterion evidence gathered in this investigation suggested that the commonly used methods to identify (i.e., the absolute split I, absolute split II, nomination, regression, and simple difference methods) and measure (i.e., the regression and simple difference methods) gifted underachievement may not *all* be used interchangeably. Importantly, these methods appeared to have different levels of validity when used to reach a conclusion about gifted underachievement. It is noteworthy that none of these methods (with the exception of the regression method as a method of identification of gifted underachievement, for which strong empirical validity support was not found) produced consistent results across the different data combinations that may be used to assess expected/actual achievement, or across different nominators. Indeed the final generalisation inference was removed in the revised interpretation/use argument.

Among the identification/measurement methods, the nomination and regression methods appeared to have the least amount of empirical validity support in the assessment of gifted underachievement. For example, the nomination method produced the lowest levels of association with the other identification methods, while its level of agreement with the other methods was consistently less than may be expected by chance. Similarly, the regression method regularly identified a much lower proportion of cases of gifted underachievement than the other methods, had a low level of agreement with the other methods, and had a poor level of accuracy as assessed using latent class criterion modelling. While the absolute split II method had a higher level of empirical support in comparison to either of these methods, the level of support tended to be only moderate. Furthermore, the absolute split II method may have some potentially problematic issues, including a reliance on non-standardised achievement scores (which may leave the method vulnerable to variations in the level of difficulty of the achievement instruments used), and a reliance on a fixed expected

achievement threshold (which may mean that any differences in the level of ability of gifted students are ignored, leading to a possible failure to recognise underachievement in gifted students at the highest ability levels).

The two methods with the strongest empirical support were the absolute split I and simple difference methods. Both methods demonstrated reasonable levels of convergence with the other identification/measurement methods, and had sufficiently high levels of agreement with the latent class criterion. Nevertheless, the absolute split I method also appears to have a number of weaknesses. First, as for the absolute split II method, the reliance of the absolute split I method on a fixed expected achievement threshold may mean that all gifted students will be treated as having the same level of ability, possibly leading to a systematic non-identification of underachievement among gifted students at the highest ability levels. Second, the absolute split I method is unable to produce a measurement of the degree of underachievement, which may represent valuable additional information that informs decision-making on the most appropriate intervention programs and provisions for the identified students. The simple difference method, which is free from both of these limitations, therefore appears to be the more valid and versatile method that is recommended for use by researchers and educators.

The overall final assessment of the validity of the commonly used methods that identify and measure gifted underachievement, as guided by the revised interpretation/use argument (which has rejected the generalisation inference), the context of their use, and the gathered empirical evidence, indicated that the simple difference method is the only supported method to identify and measure gifted underachievement. Not only is the method accurate in the identification and measurement of gifted underachievement, it is also appropriate to the context of the provision of appropriate educational interventions for identified students to support the realisation of their potential.

8.7 Validity of the Gifted Underachievement Construct

As this investigation has empirically demonstrated that all of the commonly used methods to identify/measure gifted underachievement may have some limitations, many researchers may question whether the concept of gifted underachievement is indeed valid. This is particularly the case as some scholars have already remarked on the "inherent" problems of the construct (Reis & McCoach, 2000). Nonetheless, the utility and importance of gifted underachievement has been consistently defended (Reis & McCoach, 2000; Smith, 2010; Zielger, Ziegler, & Stoeger, 2012) and recognised as one of the most active areas of research in gifted education (Dai et al., 2011). The findings of the study are now discussed with respect to the validity of the gifted underachievement construct.

8.7.1 Confidence in extrapolation

The findings of this investigation have demonstrated that the extrapolation from levels of expected and actual achievement to a level of underachievement is valid when the simple difference method of identification/measurement is used. As a result, researchers and practitioners may have some level of have confidence that the identification and measurement of underachievement using the simple difference method will produce meaningful results that are useful and valid.

8.7.2 A lack of generalisability

In contrast, the lack of empirical support for the validity of the generalisation inference implies that different students may be identified as exhibiting gifted underachievement when different combinations of instruments designed to assess expected and actual achievement are used. One factor that may contribute to the substantial variability in the results obtained from different instruments may relate to differences in content, scoring guides, presentation, test administration procedures, pedagogy, and areas of curricular

emphases by teachers (Reis & McCoach, 2000). The problem, which is referred to as the *heterogeneity of criterion*, has been recognised as resulting in the identification/measurement of different "types" of underachievement with limited comparability (Land & Sher, 2015; Reis & McCoach, 2000). Indeed, as noted by Coe (2010), when instruments "differ in terms of their content or style of assessment... it is less clear whether and how they may legitimately be compared" (p. 279). Nonetheless, a solution to this problem is proposed.

Multiple researchers have argued that when it is valid to consider a student's average score across a set of different instruments, it will also be valid to compare scores from these instruments (Coe, 2008, 2010; Newton, 2005). However, the interpretation of such average scores must relate to a shared construct, which may provide a basis for comparison of the different instruments. For example, the average of a student's scores across different school subjects may be considered to be a measure of the student's "general achievement", and a comparison of the student's scores across different school subjects may be considered to represent a comparison of the student's achievement in different subjects to the student's general achievement. While some may argue that the calculation of an average score using scores obtained in different subjects is meaningless (Murphy, 2007), Coe (2008) successfully validated the existence of a general achievement construct when he showed that across 34 different subjects, a single latent trait was able to explain 83% of the observed variation with a reliability of .95. Therefore, it is plausible that a refinement to the gifted underachievement construct to one that utilises an average across multiple measurements of a student's expected and actual achievement may increase its generalisability.

8.7.3 Additional considerations

In addition to the findings from this study, other theoretical issues may need to be considered in any refinement to the gifted underachievement construct, including the removal of arbitrary thresholds that are commonly used in the identification/measurement of gifted

underachievement, how a suitable combination of instruments that assess expected and actual achievement may be chosen, and the implications of any refinements of the construct for invisible gifted students.

8.7.3.1 Removal of arbitrary thresholds

One approach to address the arbitrariness of the commonly used thresholds to assess gifted underachievement may be to refrain from their use and to focus on making assessments of the *degree* of underachievement of gifted students. The degree of gifted underachievement may be easily assessed by calculating the size of the discrepancy between expected and actual achievement using the simple difference method. A system similar to Gagné's (1998) metric levels of giftedness may be useful to guide the interpretation of these measurements (refer Table 36). While the labels and ranges used in such guidelines may also be arbitrary, they would no longer form a part of the gifted underachievement construct itself, similar to guidelines that are currently employed to support the interpretation of other statistical measurements, including Cronbach's alpha (Field, 2013).

Table 36

Possible levels of gifted underachievement

Level	Label	Prevalence	Underachievement Range
0	Mild underachievement	5:10	Below +1
1	Moderate underachievement	3:10	+1 to +1.27
2	High underachievement	1:10	+1.28 to +2.29
3	Exceptional underachievement	1:100	+2.3 to +2.99
4	Extreme/Profound underachievement	1:1000	+3 or above

8.7.3.2 Choice of instrument combinations

Unfortunately, this investigation did not explicitly examine the question of which combinations of instruments to assess expected and actual achievement may be the most valid. Nevertheless, the follow-up meta-regression and multiple regression analyses

demonstrated that the relationship between the expected and actual achievement measurements may be very important. Of note, the multiple regression analyses demonstrated that 97% of the variation in the mean measurement of gifted underachievement by the simple difference method may be explained by the relationship between the expected and actual achievement measurements (and 70% of the variation may be explained by the strength of the relationship alone). Therefore, it appears reasonable to require any choice of the combination of instruments designed to assess expected and actual achievement to produce measurements that are at least moderately related to one another.

Additionally, Harder et al. (2014) have suggested that the use of assessments of specific abilities to be "much more valid" (p. 98) than assessments of general intelligence in the prediction of specific achievements. Therefore, when an assessment is being made of underachievement in a specific field (e.g., mathematics), it may be reasonable to encourage the use of measurements of expected achievement relating to the corresponding domains of ability (e.g., as outlined in the Cattell–Horn–Carroll model of intelligence, Figure 1). Similarly, measurements of general intelligence may be more appropriate to assess expected achievement when measurements of general achievement (e.g., the previously described average of school assessments across multiple subjects) are used to assess actual achievement.

8.7.3.3 Invisible gifted students

One limitation of the underachievement construct may relate to its lack of acknowledgement of invisible gifted students, who underachieve on both instruments designed to assess expected and actual achievement, and tend to be from highly disadvantaged backgrounds (Merrotsy, 2013, 2016). As these students are likely to underachieve on the measures of expected achievement, they may not be identified using any of the traditional identification/measurement methods. Nevertheless, scholars have suggested

that these students may benefit from targeted educational interventions *prior to* the administration of any identification/measurement instruments that should then be used as measures of expected achievement (Chaffey, Bailey & Vine, 2003; 2015; Chaffey, McCluskey, Halliwell, 2005; Merrotsy, 2013; 2016).

8.8 Summary

This chapter synthesised the collected empirical evidence within the context of the planned uses of the identification/measurement methods to reach a conclusion that the simple difference method may be the most valid and versatile method for use by researchers and practitioners. The next chapter will provide a final summary for this investigation, discuss its implications, and suggest areas for further research.

9 Conclusions

9.1 Research Purpose Revisited

The purpose of this research was to investigate the validity of the different methods that identify and measure gifted underachievement. Following recommendations from Kane's framework of validation, three types of empirical evidence (i.e., convergence, criterion, and generalisation evidence) were collected to assist in the assessment of validity.

The purpose of this chapter is to present the conclusions of the research by synthesising the key findings. First, this chapter will summarise the major findings from the investigation and answer the research questions. Second, the implications and recommendations are discussed, including a proposed refinement to the operational definition of gifted underachievement. Third, the limitations of the research and possible areas for future investigation will be outlined.

9.2 Summary of Major Findings

A number of major findings were made in this research. The first of these is that the different methods used for the identification and measurement of gifted underachievement may not be equivalent, do not identify the same groups of students, and therefore should not be used interchangeably. Second, one method, the simple difference method, was demonstrated to have the most validity and versatility for the identification and measurement of gifted underachievement. Third, the methods for the identification and measurement of gifted underachievement do not generally appear to be generalisable across different combinations of expected and actual achievement measures (or across different nominators).

The convergence of the methods for the identification and measurement of gifted underachievement was assessed using five different statistical measures and tests, across 41 different combinations of expected and actual achievement data. The results strongly

suggested that all of these methods were not adequately convergent for them be used interchangeably. The regression and nomination methods were noted as particularly divergent from the other methods.

As the various methods for the identification and measurement of gifted underachievement were not found to be equivalent, a criterion was required to determine the relative degree of validity of the methods. An appropriate criterion was constructed using latent class analysis. Using this criterion, only two methods, the absolute split I and the simple difference methods, were found to have sufficient evidence to support the validity of their use in the identification of gifted underachievement. It is noteworthy that the criterion evidence allowed for a ranking of the different methods by their degree of agreement with the criterion variable. The resulting order of validity, from highest to lowest, was: the absolute split I method, the simple difference method, the absolute split II method, the regression method, and the nomination method.

Thereafter, a generalisability study was conducted to examine whether each of the identification/measurement methods produced homogeneous results across 41 different data combinations of expected and actual achievement. The results of this study, which did not support generalisability (the only exception was the proportion of gifted underachievement identified by the regression method, which was consistently found to identify 13% of cases), suggested that the data combinations strongly affected the results obtained for each of the identification/measurement methods. Furthermore, it was established that the relationship between the expected and actual achievement measurements had a large impact on the results. The separate generalisability study to examine whether the results from the nomination method were generalisable across different nominators similarly suggested that generalisability was not supported.

A final assessment of validity of the identification/measurement methods was undertaken through an examination of the empirical evidence within the context of the planned uses of the methods. The empirical evidence clearly did not support the validity of the regression and nomination methods and only weakly supported the absolute split II method. Of the two remaining methods with the strongest empirical support, the absolute split I method had a number of limitations in the context of the planned uses of the method. Therefore, a conclusion was reached that the simple difference method may be the best method for the identification and measurement of gifted underachievement.

9.3 Answer to Research Questions

The findings of the study were evaluated to answer the research questions.

1. *Is the extrapolation inference reasonable for each of the methods used to identify and measure gifted underachievement?*

   The extrapolation inference appeared to be reasonable only for the simple difference method.

2. *Is the generalisation inference reasonable for each of the methods used to identify and measure gifted underachievement?*

   The generalisation inference appeared to be reasonable only for one method (i.e., the regression method as a method of identification of gifted underachievement) for which the extrapolation inference did not appear to be reasonable.

9.4 Refinements to the Gifted Underachievement Construct

Throughout this investigation, many arguments have been made from theoretical, methodological, and empirical perspectives to refine the construct of gifted underachievement. The following is an outline of these refinements to the original conceptualisation from Reis and McCoach (2000).

Refinement supported by the new empirical findings from this study:

(a) Only the simple difference method is valid for comparing expected and actual achievement scores (refer section 8.6);
(b) The nomination and regression methods are substantially different from the other methods used to identify/measure gifted underachievement, to suggest that these methods may identify/measure a related but different construct (refer section 6.4);
(c) The commonly used methods to identify/measure gifted underachievement are not interchangeable (refer section 5.7);
(d) The choice of the combination of instruments used to assess expected and actual achievement is pivotal to the identification/measurement of gifted underachievement (refer section 7.4.1); and
(e) The combination of instruments designed to assess expected and actual achievement should produce scores that are at least moderately related to one another (refer section 8.7.3.2);

Refinements supported by the theoretical or methodological arguments in this study:

(f) The two absolute split methods have significant areas of bias, and therefore, limited validity (refer sections 8.4.1.1 & 8.4.1.2);
(g) The use of measurements of the *degree* of underachievement is preferable to the use of arbitrary thresholds to determine the existence of underachievement (refer section 8.7.3.1);
(h) The use of measures of intelligence are preferred to other measures to assess expected achievement, although the use of measures of past achievement is feasible (refer section 2.6.1); and
(i) The term ‘underachiever’ is undesirable (refer section 2.5.3).

Refinements based on the empirical findings and theoretical arguments from others:

(j) The matching of scores on specific sub-components of intelligence (i.e., expected achievement) to scores on specific sub-components of achievement (i.e., actual achievement) is more valid than the matching of scores on general intelligence to scores on specific sub-components of achievement (Harder et al., 2014);

(k) The use of average achievement across school subjects to measure general achievement is a more generalisable conception of the construct of underachievement than the use of specific achievement scores in individual school subjects (Cole, 2008, 2010);

(l) The selection of instruments for both expected and actual achievement should have a high degree of accuracy (Ziegler & Ziegler, 2012); and

(m) The assessment of underachievement may have additional utility in the assessment of the educational environment (e.g., 'underachieving schools', or 'underachieving teachers' may be identified when large groups of students in a particular environment exhibit underachievement; Funk-Werblo, 2003).

9.5 Recommendations for Researchers

This investigation has identified a number of implications and recommendations that researchers may consider to increase the validity of their research. First, and most importantly, this investigation demonstrated that the commonly used methods of identification and measurement of gifted underachievement may not all be equivalent and, therefore, should not be used interchangeably. Attempts by Reis and McCoach (2000) appear to have largely increased the uniformity in operational definitions of gifted underachievement (refer Appendix 1), although different methods continue to be used in the identification and measurement of gifted underachievement. It is expected that this investigation provides

sufficient evidence to convince researchers of: (a) the problems associated with the inconsistent selection of methods to identify and measure gifted underachievement, and (b) the inappropriateness of making comparisons or combining the results obtained from these different methods.

For over sixty years, researchers have used various methods for the identification and measurement of gifted underachievement, with minimal, if any, justification of the selected methods (Dowdall & Colangelo, 1982; Gowan, 1955; Ryan & Coneybeare, 2013). Over this time, no attempt has been found to establish the validity of any of these methods. This investigation has determined that the simple difference method is the most valid and useful method for the identification and measurement of gifted underachievement. Therefore, researchers are advised to adopt the simple difference method over the other commonly used current identification/measurement methods. If this recommendation is widely adopted, it will help resolve one of the longest standing issues in the field of gifted education by improving the comparability and validity of studies on gifted underachievement.

The findings of this investigation further demonstrated that the combination of expected and actual achievement measurements used for the identification and measurement of gifted underachievement may have a substantial impact on the validity of the results. Researchers are therefore encouraged to ensure that the instruments used to produce measures of expected achievement and actual achievement each have high levels of accuracy, and that the measure of expected achievement relates to a specific domain that is expected to contribute to the measure of actual achievement in a compatible domain. Alternatively, researchers should adopt measures relating to non-specific general domains for both expected and actual achievement (e.g., general intelligence and general achievement). The finding highlights the fundamental importance of the validity of the measurements of expected and actual achievement in the identification and measurement of gifted underachievement.

9.6 Recommendations for Practice

In terms of practice, this investigation has identified a number of implications and recommendations that schools and education policies should consider to increase the educational and welfare outcomes for gifted students, and possibly for all students. First, from the literature reviewed, it is clear that "to be gifted is to be vulnerable" (Silverman, 1997, p. 37). The weighted average proportion of cases of gifted underachievement identified by the simple difference method across all 41 data combinations in this investigation suggested that, on average, one-third of gifted students may be underachieving (refer Table 14). In consideration of the large numbers of affected students, educators, schools, and policymakers are encouraged to take active steps to address the specific needs of gifted students. Indeed, multiple studies have repeatedly demonstrated the benefits of investing in specialist programs for gifted students (e.g., Colangelo, Assouline, & Gross, 2004; Dare & Nowicki, 2015; Feldhusen & Moon, 1992; Gross, 1995, 2006; Hertzog & Chung, 2015; Kulik & Kulik, 1992; Noble et al., 2007; Olszewski-Kubilius, 1995; Park, Lubinski, & Benbow, 2013; Plucker & Callahan, 2014; Vogl & Preckel, 2014).

Relatedly, the findings of the study provide clear guidelines on *how specifically* underachievement should be identified in gifted students by educators and other practitioners. This investigation, amongst others, has demonstrated that relying on teachers to nominate gifted students who exhibit underachievement may not be a particularly valid or tenable approach. Instead, a systematic approach to identification and measurement, that makes use of the simple difference method and involves the screening of all gifted students during their schooling, may be necessary. Specifically, in recognition of the increasing use of sophisticated software packages (e.g., *Edumate*) to manage large electronic databases in modern schools, it may be useful to apply the simple difference method to student achievement and related data (e.g., HSC, NAPLAN, school assessment, intelligence test

results) to identify and measure gifted underachievement. The resulting information may be useful as triggers for investigation, and the provision of appropriate educational and related interventions (e.g., more challenging programs, enrichment, extension, mentors, and counselling) for the affected students. Further, it is suggested that the intervention programs or provisions provided are appropriately match the measured degree of underachievement of the affected students (perhaps following a similar approach to the response to intervention model, where a higher degree of underachievement links a student to a more significant interventions).

One of the major findings of this investigation is that the validity of the identification/ measurement of gifted underachievement may depend on the specific combination of expected and actual achievement data utilised. It is suggested that schools ideally utilise relevant sub-components of an intelligence or cognitive ability test as measures of expected achievement in combination with a compatible measure of actual achievement. For example, a school could use the non-verbal component of the OLSAT to assess expected achievement and the results of a student's achievement test in mathematics to assess actual achievement. Alternatively, if underachievement is being assessed across subject areas, it may be appropriate to use assessments of general intelligence to assess expected achievement and averages of student achievement across subject areas to assess actual achievement. Schools should also preferentially select instruments that have a higher degree of accuracy, for example, aggregated school marks (e.g., a weighted mark that combines multiple tasks) or externally administered and standardised tests. It may not be too difficult for arrangements to be made so that the existing data management software packages are programmed to automatically perform the necessary calculations to produce an easy-to-interpret report on the underachievement of students enrolled in each school.

It is noted that while this study was specifically designed to investigate the identification and measurement of gifted underachievement, the findings may also have application to the general student population. Indeed, a significant discrepancy between expected and actual achievement is likely to be a major problem that affects all students. Schools with data management software programs may have minimal difficulty in making arrangements for periodic assessments of underachievement across all of its students using the simple difference method, with the available data on expected and actual achievement. Nevertheless, further research may be needed to determine whether any modifications to the identification process may be necessary for students who are not gifted.

A final recommendation for practice is for the underachievement statistics derived from the simple difference method to be used by external authorities to assess the educational "health" of schools and the effectiveness of teachers at schools. The currently utilised public statistics to evaluate schools (e.g., HSC results and NAPLAN results) may be easily impacted by socioeconomic factors and by (written and unwritten) school enrolment policies. For example, an academically selective high school may consistently obtain the highest HSC scores, as it only enrols the most highly achieving students, although these students may be exhibiting significant levels of underachievement. In comparison, statistics that describe the degree of underachievement among students in a school (e.g., the percentage of students who exhibit underachievement, how the percentage changes over time, the average degree of underachievement, and the breakdown of underachievement among various sub-groups including gender, age, ethnicity etc.) may provide more appropriate information about the quality of education that is being provided. Such statistics could be reported by schools in their public annual reports and in publically funded websites such as *My School*.

9.7 Limitations of the Research

The limitations in the research should be taken into consideration when the findings are interpreted. The main limitation of this investigation is that the empirical evidence to assess the validity of the commonly used methods to identify and measure gifted underachievement was obtained using data from high school students attending a single school. Additionally, the data used was archival, and the archives did not contain uniform records of the same sets of instruments for all students across all calendar years and grades. Consequently, additional testing with new samples from other schools, across a range of school sectors, geographical locations including rural areas, and school levels (e.g., primary school students), may be desirable. Moreover, the results of the study should not necessarily be generalised beyond the population used for this investigation.

9.8 Areas for Further Investigation

Last, the findings of this study point to a multitude of areas for further investigation:

(a) This investigation determined that the combination of expected and actual achievement measurements may have an impact on the validity and generalisability of the different methods that identify and measure gifted underachievement. Therefore, further investigations into the ideal combination of expected and actual achievement measures may be beneficial. It may be the case that no one single solution exists, and that multiple data combinations will need to be used simultaneously. If so, further research may be necessary to establish the optimal manner in which such multiple data combinations should be used.

(b) Only teacher nominations were investigated in this study. Therefore, a more thorough examination of the nomination method, using data from a number of different types of nominators (e.g., parents, peers, self, counsellors, tutor, pastoral carers, mentors etc.),

and a greater number of nominations, may be useful to provide a clearer understanding of the validity of the nomination method. It is possible that the validity of the nomination method may be improved through the use of alternative nomination instruments, or through the specific training that may be given to nominators to prevent subjectivity in judgements.

(c) One area that was not investigated in this study was the discrimination of underachievement from normal fluctuations in achievement. While the commonly adopted threshold of one standard deviation between expected and actual achievement levels to deem underachievement does appear to be widely accepted, there is no strong argument for the optimality of this somewhat arbitrary value. Further research may therefore be needed to identify a more appropriate threshold, if one exists. In practice, as schools may not have the resources to assist all individual cases of gifted underachievement as defined using the current threshold (particularly as this investigation demonstrated that approximately one third of such students may be exhibiting underachievement), pragmatics may demand that a larger threshold be reached before any interventions are provided. To address this problem, it may be useful for schools to use measurements of the degree of underachievement, rather than making dichotomous classifications of gifted students as exhibiting or not exhibiting underachievement. One approach may be to use a system similar to Gagné's metric levels of giftedness, where levels of gifted underachievement rather than levels of giftedness are outlined. A possible example is shown in Table 36. Research may be necessary to determine how such a system could be established and how it may be used to optimally provide educational interventions for affected students.

(d) The findings of this study may be used in the identification and measurement of underachievement in all students. Nevertheless, as data were collected from gifted high school students in this investigation, a replication study may be necessary with a representative sample from the general student population.

(e) A new method for the identification of gifted underachievement was developed at the concluding stages of this investigation, and as a result, was not included. The inclusion of this method in a future validation study is recommended (Veas, Gilar, Miñano, & Castejón, 2016).

(f) Finally, much is already known about what may be done to prevent gifted underachievement. Nevertheless, there continues to be an urgent need for schools, teachers, parents, and educational authorities to change practices and policies to better meet the specific educational needs of gifted students. Hence, the final emphasis of this investigation is to restate and expand upon a conclusion of Reis and McCoach (2000, p. 167) in their review of gifted underachievement ("researchers must move beyond describing this educational dilemma and instead strive to find solutions") and, further, convince policy makers, schools, teachers, and parents to adopt and use them. It is expected that this investigation has helped to provide the necessary tools for researchers to do so.

9.9 Summary

This chapter has summarised the main findings of the investigation, acknowledged its limitations, discussed important implications for research and practice, and suggested possible areas where further investigation may be helpful. It is hoped that the present research will be useful to both researchers and practitioners as they seek to understand and serve students, including those whose struggles so often go unnoticed.

10 REFERENCES

Abelman, R. (2007). Fighting the war on indecency: Mediating TV, internet, and videogame usage among achieving and underachieving gifted children. *Roeper Review, 29*(2), 100-112.

Agresti, A. (2013). *Categorical Data Analysis* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Albaili, M. A. (2003). Motivational goal orientations of intellectually gifted achieving and underachiving students in the United Arab Emirates. *Social Behaviour and Personality, 31*(2), 107-120.

Alonzo, T. A., & Pepe, M. S. (1999). Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Statistics in Medicine, 18*(22), 2987-3003. doi:10.1002/(SICI)1097-0258(19991130)18:22<2987::AID-SIM205>3.0.CO;2-B

Altman, D. G., & Bland, J. M. (1983). Measurement in medicine: The analysis of method comparison studies. *The Statistician, 32*, 307-317.

Altman, D. G., & Bland, J. M. (2002). Commentary on quantifying agreement between two methods of measurement. *Clinical Chemistry, 48*(5), 801-802.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [AERA, APA, and NCME]. (1999). *Standards for Educational and Psychological Testing.* Washington, WA: American Psychological Association.

American Psychiatric Association. (2000). *Diagnostic and Statistical Manual of Mental Disorders (DSM-IV)* (4th ed.). Washington, WA: American Psychiatric Association. doi:10.1176/appi.books.9780890423349

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders (DSM-V)* (5th ed.). Arlington, VA: American Psychiatric Association Publishing.

American Psychological Association. (2010). *Publication Manual of the American Psychological Association* (6th ed.). Wasington, WA: American Psychological Association.

Annesley, F., Odhner, F., Madoff, E., & Chansky, N. (1970). Identifying the first grade underachiever. *The Journal of Educational Research, 63*(10), 459-462.

Aryadoust, V. (2011). Validity arguments of the speaking and listening modules of international English Language testing system: A synthesis of existing research. *Asian English for Specific Purposes Journal, 7*(2), 28-54.

Association of Independent Schools (AIS). (2015). *Supporting your gifted and talented child's achievement and well-being: A resource for parents.* Retrieved December 12, 2015, from http://www.ais.sa.edu.au/__files/f/201292/2014%20Parent%20Booklet%20May%203.pdf

Assouline, S. G. (2003). Psychological and educational assessment of gifted children. In N. Colangelo, & G. A. Davis (Eds.), *Handbook of Gifted Education* (pp. 124-145). Boston: Pearson Education.

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2008–2013). *NAPLAN National Reports.* Sydney, AU: Australian Curriculum, Assessment and Reporting Authority. Retrieved from http://www.nap.edu.au/results-and-reports/national-reports.html

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2014). *National Assessment Program Literacy and Numeracy: Technical Report.* Sydney, AU: ACARA. Retrieved from https://www.nap.edu.au/_resources/NAPLAN_2013_technical_report.pdf

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2015). *My School.* Retrieved September 9th, 2015, from Guide to understanding ICSEA (Index of Community Socioeducational Advantage) values: http://www.acara.edu.au/verve/_resources/Guide_to_understanding_icsea_values.pdf

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2015). *Student Diversity.* Retrieved September 7, 2015, from Gifted and talented students: http://www.australiancurriculum.edu.au/studentdiversity/gifted-and-talented-students

Australian Curriculum, Assessment and Reporting Authority (ACARA). (2016). *NAPLAN - General*. Retrieved from http://www.nap.edu.au/information/faqs/naplan--general

Australian Government. (2014). *Budget 2014-15.* Retrieved April 8, 2015, from http://www.budget.gov.au/2014-15/content/overview/html/overview_18.htm

Australian Psychological Society. (2013). *Why can't Jonny read? Bringing theory into cognitive assessment*. Retrieved January 26, 2016, from https://www.psychology.org.au/inpsych/2013/december/jacobs/

Bagnato, S. J. (2005). The authentic alternative for assessment in early intervention: An emerging evidence-based practice. *Journal of Early Intervention, 28*(1), 17-22. doi:10.1177/105381510502800102

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100*(3), 603-617.

Bakeman, R., & Gottman, J. M. (1997). *Observing Interaction: An Introduction to Sequential Analysis* (2nd ed.). Cambridge, GB: Cambridge University Press.

Baker, J. A., Bridger, R., & Evans, K. (1998). Models of underachievement among gifted preadolescents: The role of personal, family, and school factors. *Gifted Child Quarterly, 42*, 5-14.

Bakermans-Kranenburg, M. J., & van Ijzendoorn, M. H. (2015). The hidden efficacy of interventions: Gene×environment experiments from a differential susceptibility perspective. *Annual Review of Psychology, 66*, 381-409. doi:10.1146/annurev-psych-010814-015407

Balduf, M. (2009). Underachievement among college students. *Journal of Advanced Academics, 20*(2), 274-294.

Bangel, N. J., Moon, S. M., & Capobianco, B. M. (2010). Preservice teachers' perceptions and experiences in a gifted education training model. *Gifted Child Quarterly, 54*(3), 209-221. doi:10.1177/0016986210369257

Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology, 34*(1), 215-220. doi:10.1093/ije/dyh299

Baslanti, U. (2008). Investigating the underachievement of university students in Turkey: Exploring subscales. *International Journal of Progressive Education, 4*(2), 1-29. Retrieved from http://files.eric.ed.gov/fulltext/ED501580.pdf

Baslanti, U., & McCoach, D. B. (2006). Factors related to the underachievement of university students in Turkey. *Roeper Review, 28*(4), 210-215.

Baudson, T. G., & Preckel, F. (2013). Teachers' implicit personality theories about the gifted: An experimental approach. *School Psychology Quarterly, 28*(1), 37-46. doi:10.1037/spq0000011

Baum, S. M., Renzulli, J. S., & Hébert, T. P. (1995). *The Prism Metaphor: A New Paradigm for Reversing Underachievement.* Storrs, CT: University of Connecticut, The National Research Center on the Gifted and Talented.

Begeny, J. C., Eckert, T. L., Montarello, S. A., & Storie, M. S. (2008). Teachers' perceptions of students' reading abilities: An examination of the relationship between teachers' judgements and students' performance across a continuum of rating methods. *School Psychology Quarterly*, *23*(1), 43-55. doi:10.1037/1045-3830.23.1.43

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice, 31*(3), 2-9.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment, 17*(2-3), 62-87.

Bennett, R. E., Gottesman, R. L., Rock, D. A., & Cerullo, F. (1993). Influence of behaviour perceptions and gender on teachers' judgements of students' academic skill. *Journal of Educational Psychology*, *85*(2), 347-356. doi:10.1037/0022-0663.85.2.347

Benson, N. F. (in press). Review of the Wechsler Intelligence Scale for Children-Fifth Edition. In J. F. Carlson, K. F. Geisinger, & J. L. Jonson (Eds.), *The Twentieth Mental Measurements Yearbook.* Lincoln, NE: Buros Center for Testing. Retrieved from Mental measurements Yearbook with Tests in Print database (Ovid)

Bergner, S., & Neubauer, A. C. (2011). Sex and training differences in mental rotation: A behavioural and neurophysiological comparison of gifted achievers, gifted underachievers and average intelligent achievers. *High Ability Studies, 22*(2), 155-177.

Berkowitz, E., & Cicchelli, T. (2004). Metacognitive strategy use in reading of gifted high achieving and gifted underachieving middle school students in New York City. *Education and Urban Society, 37*(1), 37-57.

Berndt, D. J., Kaiser, C. F., & van Aalst, F. (1982). Depression and self-actualization in gifted adolescents. *Journal of Clinical Psychology, 38*(1), 142-150. doi:10.1002/1097-4679(198201)38:1<142::AID-JCLP2270380123>3.0.CO;2-D

Beswick, J. F., Willms, J. D., & Sloat, E. A. (2005). A comparative study of teacher ratings of emergent literacy skills and student performance on a standardized measure. *Education Journal, 136*(1), 116-137.

Binet, A. (1903). *Etude expérimentale de l'intelligence.* Paris, FR: Schleicher frères & cie.

Blaas, S. (2014). The relationship between social-emotional difficulties and underachievement of gifted students. *Australian Journal of Guidance and Counselling, 24*(2), 243-255.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet, 327*, 307-310. doi:10.1016/S0140-6736(86)90837-8

Bland, J. M., & Altman, D. G. (1995). Comparing methods of measurement: Why plotting difference against standard method is misleading. *Lancet, 346*, 1085-1087.

Bland, J. M., & Altman, D. G. (1999). Measuring agrement in method comparison studies. *Statistical Methods in Medical Research, 8*(2), 135-160. doi:10.1191/096228099673819272

Bloom, B. S. (Ed.). (1985). *Developing Talent in Young People.* New York, NY: Ballantine Books.

Blyth, C. R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association, 67*(338), 364-366. doi:10.2307/2284382

Board of Studies Teaching & Educational Standards NSW (BOSTES). (2001–2011). *Results Analysis Package*. Retrieved from https://bosho.boardofstudies.nsw.edu.au/links/schoolsonline.html

Board of Studies Teaching & Educational Standards NSW (BOSTES). (2010). *HSC Marking*. Retrieved December 30, 2013, from http://www.boardofstudies.nsw.edu.au/hsc_exams/marking.html

Board of Studies Teaching & Educational Standards NSW (BOSTES). (2011). *Explanation of aligning and moderating procedures for the Higher School Certificate*. Retrieved December 30, 2013, from http://www.boardofstudies.nsw.edu.au/hsc-results/moderation.html

Board of Studies Teaching & Educational Standards NSW (BOSTES). (2013). *2013 HSC results released to 74,000 students.* Retrieved December 30, 2013, from http://www.boardofstudies.nsw.edu.au/news-media/pdf_doc/131218-2013-HSC-results-released-to-74000-students.pdf

Board of Studies Teaching & Educational Standards NSW (BOSTES). (2013). *HSC Exam people, papers, and processes*. Retrieved December 30, 2013, from http://www.boardofstudies.nsw.edu.au/hsc_exams/hsc-people-papers-processes.html

Board of Studies Teaching & Educational Standards NSW (BOSTES). (2016, April 16). *Common Grade Scale for Preliminary Courses*. Retrieved from https://www.boardofstudies.nsw.edu.au/rosa/principals-teachers/common-grade-scale-pre-courses.html

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis.* Chichester, GB: Wiley.

Borland, J. H. (1989). *Planning and Implementing Programs for the Gifted.* New York, NY: Teachers College Press.

Borsboom, D. (2012). Whose consensus is it anyway? Scientific versus legalistic conceptions of validity. *Measurement: Interdisciplinary Research and Perspectives, 10*, 38-41.

Bouffard, T., Roy, M., & Vezeau, C. (2005). Self-perceptions, temperament, socioemotional adjustment and the perceptions of parental support of chronically underachieving children. *International Journal of Educational Research, 43*(4), 215-235. doi:10.1016/j.ijer.2006.06.003

Brennan, R. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement, 36*, 295-317.

Brennan, R. L. (2013). Commentary on “Validating the interpretations and uses of test scores". *Journal of Educational Measurement, 50*, 74-83. doi:10.1111/jedm.12001

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misusses, and alternatives. *Educational and Psychological Measurement, 41,* 687-699.

British Psychological Society. (2011). *Response to the American Psychiatric Association: DSM-5 Development.* Leicester, GB: The British Psychological Society. Retrieved from http://apps.bps.org.uk/_publicationfiles/consultation-responses/DSM-5%202011%20-%20BPS%20response.pdf

Bryman, A. (2004). *Social Research Methods* (2nd ed.). Oxford, GB: Oxford University Press.

Bush, A. (2005). Paying close attention at school: Some observations and psychoanalytic perspectives on the educational underachievement of teenage boys. *Infant Observation, 8*(1), 69-79.

Cabrera, J. A., Sanchez-Quintana, D., Farre, J., Navarro, F., Rubio, J. M., Cabestrero, F., . . . Ho, S. Y. (2002). Ultrasonic characterization of the pulmonary venous wall: Echographic and histological correlation. *Circulation, 106*(8), 968-973.

Calaguas, G. M. (2012). Academic achievement and school ability: Implications to guidance and counseling programs. *Researchers World, 3*(2), 49-55.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Card, D., & Giuliano, L. (2015). Can universal screening increase the representation of low income and minority students in gifted education? *National Bureau of Economic Research Working Paper*. doi:10.3386/w21519

Carlson, J. F., Geisinger, K. F., & Jonson, J. L. (Eds.). (2014). *The Nineteenth Mental Measurements Yearbook.* Lincoln, NE: University of Nebraska Press.

Carnevale, A. P., Rose, S. J., & Cheah, B. (2011). *The College Payoff: Education, Occupations, Lifetime Earnings.* Washington, WA: Georgetown University Center on Education and the Workforce.

Carr, M., Borkowski, J. G., & Maxwell, S. E. (1991). Motivational components of underachievement. *Developmental Psychology, 27*(1), 108-118.

Carroll, J. (1993). *Human Cognitive Abilities.* Cambridge, GB: Cambridge University Press.

Carvacho, H., Zick, A., Haye, A., González, R., Manzi, J., Kocik, C., & Bertl, M. (2013). On the relation between social class and prejudice: The roles of education, income, and ideological attitudes. *European Journal of Social Psychology, 45*(4), 272-285. doi:10.1002/ejsp.1961

Catholic Education Office (CEO). (2014). *Gifted Education Policy.* Retrieved December 12, 2014, from http://www.ceosyd.catholic.edu.au/About/Documents/policy-gifted-education.pdf

Cavilla, D. (2015). Observation and analysis of three gifted underachievers in an underserved, urban high school setting. *Gifted Education International,* 1-14. doi:10.1177/0261429414568181

Cernovsky, Z. Z. (2002). A frequent misunderstanding associated with point biserial and phi coefficients. *Psychological Reports, 90,* 65-66. doi:10.2466/pr0.2002.90.1.65

Chaffey, G. W., Bailey, S. B., & Vine, K. W. (2003). Identifying high academic potential in Australian Aboriginal children using dynamic testing. *Australasian Journal of Gifted Education, 12*(1), 42-55.

Chaffey, G. W., Bailey, S. B., & Vine, K. W. (2015). Identifying high academic potential in Australian Aboriginal children using dynamic testing. *Australasian Journal of Gifted Education, 24*(2), 24-37.

Chaffey, G. W., McCluskey, K., & Halliwell, G. (2005). Using Coolabah Dynamic assessment to identify Canadian Aboriginal children with high academic potential: A cross-cultural study. *Gifted and Talented International, 20*(2), 50-59.

Chang, M., Paulson, S. E., Finch, W. H., Mcintosh, D. E., & Rothlisberg, B. A. (2013). Joint confirmatory factor analysis of the woodcock-johnson tests of cognitive abilities, third edition, and the stanford-binet intelligence scales, fifth edition, with a preschool population. *Psychology in the Schools, 51*(1), 32-57. doi:10.1002/pits.21734

Chapelle, C. A. (2012). Validity argument for language assessment: the framework is simple... *Language Testing, 29*(1), 19-27.

Chapelle, C. A., Enright, M. E., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3-13.

Christensen, A. H., Gjorup, T., Hilden, J., Fenger, C., Henriksen, B., Vyberg, M., . . . Hansen, B. F. (1992). Observer homogeneity in the histologic diagnosis of Helicobacter pylori. Latent class analysis, kappa coefficient, and repeat frequency. *Scandinavian Journal of Gastroenterology, 27*(11), 933-939.

Cizek, G. J., Bowen, D., & Church, K. (2010). Sources of validity evidence for educational and psychological tests: A follow-up study. *Educational and Psychological Measurement, 70*(5), 732-743.

Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. *Educational and Psychological Measurement, 68*(3), 397-412.

Clark, B. (2002). *Growing Up Gifted: Developing the Potential of Children at Home and at School* (6th ed.). New York, NY: Prentice Hall.

Clauser, B. E., Harik, P., & Clyman, S. G. (2000). The generalizability of scores for a performance assessment scored with a computer-automated scoring system. *Journal of Educational Measurement, 37*, 245-262.

Cochran, W. G. (1950). The comparison of percentages in matched samples. *Biometrika, 37*(3-4), 256-266. doi:10.1093/biomet/37.3-4.256

Coe, R. (2008). Comparability of GCSE examinations in different subjects: An application of the Rasch model. *Oxford Review of Education, 34*(5), 609-636.

Coe, R. (2010). Understanding comparability of examination standards. *Research Papers in Education, 25*(3), 271-284. doi:10.1080/02671522.2010.498143

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 213-220.

Cohen, J. F., Chalumeau, M., Cohen, R., Korevaar, D. A., Khoshnood, B., & Bossuyt, P. M. (2015). Cochran's Q test was useful to assess heterogeneity in likelihood ratios in

studies of diagnostic accuracy. *Journal of Clinical Epidemiology, 68*(3), 299-306. doi:10.1016/j.jclinepi.2014.09.005

Cohen, L. (2012). Considerations of the Actiotope model of giftedness. *High Ability Studies, 23*(1), 43-45. doi:10.1080/13598139.2012.679089

Colangelo, N., Assouline, S. G., & Gross, M. U. M. (2004). *A Nation Deceived: how schools hold back America's brightest students.* National Association for Gifted Children. Iowa City: The Connie Belin & Jacqueline N. Blank International Center for Gifted Education and Talent Development.

Colangelo, N., Assouline, S. G., Marron, M. A., Castellano, J. A., Clinkenbeard, P. R., Rogers, K., . . . Smoth, D. (2010). Guidelines for developing an academic acceleration policy. *Journal of Advanced Academics, 21*, 180-203.

Colangelo, N., Kerr, B., Christensen, P., & Maxey, J. (1993). A comparison of gifted underachievers and gifted high achievers. *Gifted child quarterly, 37,* 155-160.

Coleman, L. J., & Cross, T. L. (2014). Is Being Gifted a Social Handicap? *Journal for the Education of the Gifted, 37*(1), 5-17.

Collins, J., & Huynh, M. (2014). Estimation of diagnostic test accuracy without full verification: A review of latent class methods. *Statistics in Medicine, 33*(24), 4141-4169. doi:10.1002/sim.6218

Commonwealth of Australia. (2000). *Senate employment, workplace relations, small business and education references committee terms of reference: The education of gifted children.* Canberra, AU: Parliament of Australia.

Cone, T. E., & Wilson, L. R. (1981). Quantifying a severe discrepancy: A critical analysis. *Learning Disability Quarterly, 4*, 359-371.

Connelly, L. M. (2008). Bland-Altman plots. *Medsurg Nursing, 17*(3), 175-176.

Correia, M. A., Mello, M. J., Petribu, N. C., Silva, E. J., Bezerra, P. G., Duarte, M. C., & Correia, J. B. (2011). Agreement on radiological diagnosis of acute lower respiratory tract infection in children. *Journal of Tropical Pediatrics, 57*(3), 204-207. doi:10.1093/tropej/fmq071

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 443-507). Washington, WA: American Council on Education.

Cross, J. R., & Cross, T. L. (2015). Clinical and mental health issues in counseling the gifted individual. *Journal of Counseling & Development, 93*(2), 163-172. doi:10.1002/j.1556-6676.2015.00192.x

Cross, T. L. (2013). *Suicide Among Gifted Children and Adolescents: Understanding the Suicidal Mind.* Waco, TX: Prufrock Press.

Csikszentmihalyi, M. (1996). *Creativity: Flow and the Psychology of Discovery and Invention.* New York, NY: Harper.

Cummings, J. A. (1995). Woodcock–Johnson Tests of Cognitive Ability. In J. C. Conoley, & J. C. Impara (Eds.), *The Twelfth Mental Measurements Yearbook.* Lincoln, NE: Buros Center for Testing. Retrieved from Mental Measurements Yearbook with Tests in Print database (Ovid)

Cutler, D. M., & Lleras-Muney, A. (2010). Understanding differences in health behaviour by education. *Journal of Health Economics, 29*, 1-28.

Czodrowski, P. (2014). Count on kappa. *Journal of Computer-Aided Molecular Design, 28*(11), 1049-1055. doi:10.1007/s10822-014-9759-6

Dai, D. Y. (2012). Giftedness in the making: A response to Ziegler and Phillipson. *High Ability Studies, 23*(1), 47-50. doi:10.1080/13598139.2012.679090

Dai, D. Y., & Renzulli, J. S. (2008). Snowflakes, living systems, and the mystery of giftedness. *Gifted Child Quarterly, 52*(2), 114-130. doi:10.1177/0016986208315732

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*(2), 126-138.

Dare, L., & Nowicki, E. (2015). Conceptualizing concurrent enrollment: Why high-achieving students go for it. *Gifted Child Quarterly, 59*(4), 249-264. doi:10.1177/0016986215597749

Davies, A. (2012). Kane, validity and soundness. *Language Testing, 29*(1), 37-42.

Davis, G. A., & Rimm, S. B. (1998). *Education of the Gifted and Talented.* Boston, MA: Allyn & Bacon.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117-135.

Deary, I. J. (2006). Follow-up studies of the Scottish mental surveys of 1932 and 1947. In R. A. Peel, & M. Zeki (Eds.), *Human Ability: Genetic and Environmental Influences* (pp. 91-105). London, GB: Galton Institute.

DeHaan, R. F., & Havighurst, R. J. (1957). *Educating Gfted Children.* Chicago, IL: Chicago University Press.

Delaney, B. C., Holder, R. L., Allan, T. F., Kenkre, J. E., & Hobbs, F. D. (2003). A comparison of Bayesian and maximum likelihood methods to determine the performance of a point of care test for Helicobacter pylori in the office setting. *Medical Decision Making, 23*, 21-30.

Demazeux, S., & Singy, P. (Eds.). (2015). *The DSM-5 in Perspective: Philosophical Reflections on the Psychiatric Babel.* Dordrecht, NL: Springer.

Department of Education and Training (DET). (2004). *Policy and implementation strategies for the education of gifted and talented students.* Retrieved December 12, 2015, from http://www.curriculumsupport.education.nsw.gov.au/policies/gats/assets/pdf/polimp.pdf

Department of Education, Employment and Workplace Relations (DEEWR). (2011, December). *Review of Funding for Schooling - Final Report.* Retrieved December 30th, 2015, from https://docs.education.gov.au/system/files/doc/other/review-of-funding-for-schooling-final-report-dec-2011.pdf

Dewitte, K., Fierens, C., Stockl, D., & Thienpont, L. M. (2002). Application of the Bland-Altman plot for interpretation of method-comparison studies: A critical investigation of its practice. *Clinical Chemistry, 48*(5), 799-801.

Diaz, E. I. (1998). Perceived factors influencing the academic underachievement of talented students of Puerto Rican descent. *Gifted Child Quarterly, 42*, 105-122.

Dittrich, E. (2014). Underachievement leading to downgrading at the highest level of secondary education in the Netherlands: A longitudinal case study. *Roeper Review, 36*(2), 104-113. doi:10.1080/02783193.2014.884201

Donnelly, J. E. (2010). Use of a web-based academic alert system for identification of underachieving students at an urban research institution. *College and University, 85*(4), 39-42.

Dowdall, C. B., & Colangelo, N. (1982). Underachieving gifted students: Review and implications. *Gifted Child Quarterly, 26*, 179-184.

Duncan, A. L. (2009). *Examining the relationship between the WISC-IV, the OLSAT-7, and the EQAO achievement test*. (Master's thesis). Retrieved from ProQuest Dissertations and Theses database. (MR48862)

Dunn, D. S., & Andrews, E. E. (2015). Person-first and identity-first language: Developing psychologists' cultural competence using disability language. *American Psychologist, 70*(3), 255-264. doi:10.1037/a0038636

Dunne, M., & Gazeley, L. (2008). Teachers, social class and underachievement. *British Journal of Sociology of Education, 29*(5), 451-463.

Dyer, C. O. (1985). Review of Otis–Lennon School Ability Test. In J. V. Mitchell (Ed.), *The Ninth Mental Measurements Yearbook* (pp. 1107-1111). Lincoln, NE: Buros Institute of Mental Measurements.

Eddles-Hirsch, K. A., Vialle, W., Rogers, K. B., & McCormick, J. (2010). Just challenge those high-ability learners and they'll be all right! *Journal of Advanced Academics, 22*, 106-128.

Elijah, K. (2011). Meeting the guidance and counseling needs of gifted students in school settings. *Journal of School Counseling, 9*(14), 1-19.

Elwood, M. (2007). *Critical Appraisal of Epidemiological Studies and Clinical Trials.* Oxford, GB: Oxford University Press.

Emerick, L. J. (2004). Academic underachievement among the gifted: Students' perceptions of factors that reverse the pattern. In S. M. Moon (Ed.), *Social/Emotional Issues, Underachievement, and Counseling of Gifted and Talented Students* (pp. 105-118). Thousand Oaks, CA, California: Corwin Press.

Espeland, M. A., & Handelman, S. L. (1989). Using latent class models to characterize and assess relative error in discrete measurements. *Biometrics, 45*, 587-599.

Ewe, S. H., Delgado, V., van der Geest, R., Westenberg, J. J., Haeck, M. L., Witkowski, T. G., . . . Siebelink, H. M. (2013). Accuracy of three-dimensional versus two-dimensional echocardiography for quantification of aortic regurgitation and validation by three-dimensional three-directional velocity-encoded magnetic resonance imaging. *The American Journal of Cardiology, 112*(4), 560-566. doi:10.1016/j.amjcard.2013.04.025

Farquhar, W. W., & Payne, D. A. (1964). A classification and comparison of techniques used in selecting under- and over-achievers. *Personnel and Guidance Journal, 42*, 874-884. doi:10.1002/j.2164-4918.1964.tb04746.x

Feinstein, L., Duckworth, K., & Sabates, R. (2008). *Education and the Family: Passing Success Across the Generations.* London, GB: Routledge.

Feinstein, L., Sabates, R., Anderson, T. M., Sorhaindo, A., & Hammond, C. (2006). What are the effects of education on health? *Measuring the Effects of Education on Health and Civic Engagement: Proceedings of the Copenhagen Symposium* (pp. 171-354). Paris: OECD.

Feldhusen, J. F., & Moon, S. M. (1992). Grouping gifted students: Issues and concerns. *Gifted Child Quarterly, 36*(2), 63-67. doi:10.1177/001698629203600202

Field, A. (2009). *Discovering Statistics Using SPSS* (3rd ed.). Los Angeles, CA: SAGE Publications.

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (4th ed.). Los Angeles, CA: SAGE Publications.

Figg, S. D., Rogers, K. B., McCormick, J., & Low, R. (2012). Differentiating low performance of the gifted learner: Achieving, underachieving, and selective consuming students. *Journal of Advanced Academics, 23*(1), 53-71. doi:10.1177/1932202X11430000

Fisher, E. J. (2005). Black student achievement and the oppositional culture model. *Journal of Negro Education, 74*(3), 201-209.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical Methods for Rates and Proportions* (3rd ed.). New York, NY: John Wiley & Sons.

Flint, L. J. (2002). Retrieved July 12, 2012, from Self-Interventions of Gifted Underachievers: Stories of Success [PhD: Dissertation]: https://getd.libs.uga.edu/pdfs/flint_lori_j_200208_phd.pdf

Flint, L. J. (2009). Using life-story research in gifted education. *Gifted Children, 3*(2), 6-13.

Fong, C. J., & Krause, J. M. (2014). Lost confidence and potential: A mixed methods study of underachieving college students' sources of self-efficacy. *Social Psychology of Education, 17*(2), 249-268. doi:10.1007/s11218-013-9239-1

Ford, D. Y. (2003). Equity and excellence: Culturally diverse students in gifted education. In N. Colangelo, & G. A. Davis (Eds.), *Handbook of Gifted Education* (pp. 506-520). Boston, MA: Pearson Education.

Foreman, J. L., & Gubbins, E. J. (2015). Teachers see what ability scores cannot. *Journal of Advanced Academics, 26*(1), 5-23. doi:10.1177/1932202X14552279

Forget-Dubois, N., Lemelin, J. -P., Boivin, M., Dionne, G., Séguin, J. R., Vitaro, F., & Tremblay, R. E. (2007). Predicting early school achievement with the EDI: A longitudingal population-based study. *Early Education and Development*, 405-426. doi:10.1080/10409280701610796

Fraenkel, J. R., & Wallen, N. E. (2006). *How to Design and Evaluate Research in Education* (6th ed.). New York, NY: McGraw-Hill.

Franke, G. R. (2001). Applications of meta-analysis for marketing and public policy: A review. *Journal of Public Policy & Marketing, 20*(2), 186-200.

Frisbie, D. A. (2005). Measurement 101: Some fundamentals revisited. *Educational Measurement: Issues and Practice, 24*(3), 21-28.

Funk-Werblo, D. (2003). The invisible gifted child. In F. J. Smutny (Ed.), *Underserved Gifted Populations* (pp. 27-51). Cresskill, NJ: Hampton Press.

Gagné, F. (1985). Giftedness and talent: Reexamining a reexamination of the definitions. *Gifted Child Quarterly, 29*(3), 103-112.

Gagné, F. (1995). From giftedness to talent: A developmental model and its impact on the language of the field. *Roeper Review, 18*(2), 103-111.

Gagné, F. (1998). A proposal for subcategories within the gifted or talented populations. *Gifted Child Quarterly, 42*, 87-95.

Gagné, F. (2003). Transforming gifts into talents: The DMGT as a developmental theory. In N. Colangelo, & G. Davis (Eds.), *Handbook of Gifted Education* (3rd ed., pp. 60-74). New York, NY: Pearson Education.

Gagné, F. (2007). Ten commandments for academic talent development. *Gifted child quarterly, 51*(2), 93-118.

Gagné, F. (2009). Building gifts into talents: Detailed overview of the DMGT 2.0. In *Leading Change in Gifted Education: The Festschrift of Dr. Joyce Vantassel-Baska* (pp. 61-80). Waco, TX: Prufrock Press Inc.

Gagné, F. (2009). Debating giftedness: Pronat vs antinat. In L. V. Shavinina (Ed.), *International Handbook on Giftedness* (pp. 155-198). Dordrecht, NL: Springer.

Gagné, F. (2011). Academic talent development and the equity issue in gifted education. *Talent Development and Excellence, 3*, 3-22.

Gagné, F. (2013). The DMGT: Changes within, beneath, and beyond. *Talent Development and Excellence, 5*, 5-19.

Gallagher, J. J. (2015). Education of gifted students: A civil rights issue? *Journal for the Education of the Gifted, 38*(1), 64-69.

Gallagher, S., Smith, S. R., & Merrotsy, P. (2011). Teachers' perceptions of the socioemotional development of intellectually gifted primary aged students and their attitudes towards ability grouping and acceleration. *Gifted and Talented International, 26*(2), 11-24.

Gallant, D. J. (2009). Predictive validity evidence for an assessment program based on the work sampling system in mathematics and language and literacy. *Early Childhood Research Quarterly, 24*(2), 133-141. doi:10.1016/j.ecresq.2009.03.003

Galton, F. (1869). *Hereditary Genius: An Inquiry into its Laws and Consequences.* London, GB: Macmillan.

Galton, F. (1883). *Inquiries into Human Faculty and its Development.* London, GB: JM Dent.

Gardner, H. (1983). *Frames of Mind: The Theory of Multiple Intelligences.* New York, NY: Basic Books.

Gardner, H. (2004). *Changing Minds: The art and science of changing our own and other people's minds.* Watertown, MA: Harvard Business School Press.

Gardner, H. (2011). *Frames of Mind: The Theory of Multiple Intelligences* (3rd ed.). New York, NY: Basic Books.

Garson, G. D. (2012). *Measures of Association.* Asheboro, NC: Statistical Associates Publishers.

Geake, J. G., & Gross, M. U. M. (2008). Teachers' negative affect toward academically gifted students: An evolutionary psychological study. *Gifted Child Quarterly, 52*(3), 217-231. doi:10.1177/0016986208319704

Gelman, A., & Loken, E. (2014). The satistical crisis in science. *American Scientist, 102*, 460-465. Retrieved from http://www.americanscientist.org/issues/feature/2014/6/the-statistical-crisis-in-science

Ghanizadeh, A. (2013). Agreement between Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, and the proposed DSM-V attention deficit hyperactivity

disorder diagnostic criteria: An exploratory study. *Comprehensive Psychiatry, 54*(1), 7-10. doi:10.1016/j.comppsych.2012.06.001

Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochemia Medica, 25*(2), 141-151. doi:10.11613/BM.2015.015

Glantz, S. A. (2011). *Primer of Biostatistics* (7th ed.). San Francisco, CA: McGraw-Hill Education.

Glock, S., Krolak-Schwerdt, S., Klapproth, F., & Böhmer, M. (2013). Beyond judgement bias: How students' ethnicity and academic profile consistency influence teachers' tracking judgements. *Social Psychology of Education, 16*(4), 555-573. doi:10.1007/s11218-013-9227-5

Goldberg, J. D., & Wittes, J. T. (1978). The estimation of false negatives in medical screening. *Biometrics, 34*, 77-86.

Goldschmidt, P., Martinez, J. F., Niemi, D., & Baker, E. L. (2007). Relationships among measures as empirical evidence of validity: Incorporating multiple indicators of achievement and school context. *Educational Assessment, 12*(3&4), 239-266. doi:10.1080/10627190701578297

Goodman, S. (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology, 45*(3), 135-140. doi:10.1053/j.seminhematol.2008.04.003

Gorard, S., & Smith, E. (2004). What is 'underachievement' at school? *School Leadership & Management, 24*(2), 205-225.

Gottfredson, L. S. (2000). Skill gaps, not tests, make racial proportionality impossible. *Psychology, Public Policy, and Law, 6*(1), 129-143.

Gottfredson, L. S. (2003). The science and politics of intelligence in gifted education. In N. Colangelo, & G. A. Davis (Eds.), *Handbook of Gifted Education* (pp. 24-40). Boston, MA: Allyn & Bacon.

Gottfredson, L. S. (2006). Social consequences of group differences in cognitive ability (Consequencias sociais das diferencas de grupo em habilidade cognitiva). In C. E. Flores-Mendoza, & R. Colom (Eds.), *Introducau a psicologia das diferencas individuais* (pp. 433-456). Porto Allegre, BR: ArtMed Publishers.

Gottfredson, L. S. (2011). Intelligence: Instant expert 13. *New Scientist, 211*(2819), i-viii.

Gottfredson, L. S. (2013). Resolute ignorance on race and Rushton. *Personality and Individual Differences, 55*(3), 218-223. doi:10.1016/j.paid.2012.10.021

Gowan, J. C. (1955). The underachieving gifted child: A problem for everyone. *Exceptional Children, 21*, 247-249, 270-271.

Grantham, T. C., & Ford, D. Y. (1998). A case study of the social needs of Danisha: An underachieving gifted African-American female. *Roeper Review, 21*(2), 96-101.

Grobman, J. (2006). Underachievement in exceptionally gifted adolscents and young adults: A psychiatrist's view. *Journal of Secondary Gifted Education, 17*(4), 199-210.

Groot, W., & van den Brink, H. M. (2010). The effects of education on crime. *Applied Economics, 42*(3), 279-289. doi:10.1080/00036840701604412

Gross, M. U. M. (1995). Current research on the school acceleration of gifted and talented students: Address to the Meeting of Fellows of the NSW Chapter of the Australian College of Education, 29 November, 1994. *Australian College of Education NSW Chapter Monograph*, 2-8.

Gross, M. U. M. (2006). To group or not to group. In C. M. Smith (Ed.), *Including the Gifted and Talented* (pp. 119-137). London: Routledge.

Gross, M. U. M. (2010). *Miraca Gross, In Her Own Write: A Lifetime in Gifted Education.* Sydney, AU: GERRIC.

Grossman, M. (2000). The human capital model. In A. J. Culyer, & J. P. Newhouse (Eds.), *Handbook of Health Economics* (Vol. 1a, pp. 347-408). Amsterdam, NL: Elsevier.

Guilmette, T. J., Kennedy, M. L., & Queally, P. T. (2001). A comparison of the WISC-III and the Otis–Lennon School Ability Test with students referred for learning disabilities. *Journal of Psychoeducational Assessment, 19*(3), 239-244.

Guldemond, H., Bosker, R., Kuyper, H., & van der Werf, G. (2007). Do highly gifted students really have problems? *Educational Research and Evaluation, 13*(6), 555-568.

Hadgu, A., Dendukuri, N., & Hilden, J. (2005). Evaluation of nucleic acid amplification tests in the absence of a perfect gold-standard test. *Epidemiology, 16*(5), 604-612.

Haertel, E. (2013). Getting the help we need. *Journal of Educational Measurement, 50*(1), 84-90.

Hair, J., Anderson, R., Black, B., Babin, B., & Black, W. C. (2010). *Multivariate Data Analysis* (7th ed.). London, GB: Pearson Education.

Haladyna, T. (2006). Perils of standardized achievement testing. *Educational Horizons, 85*(1), 30-43.

Hammes, P. S., Bigras, M., & Crepaldi, M. A. (2014). Validity and bias of academic achievement measures in the first year of elementary school. *International Journal of Research & Method in Education*, 1-16. doi:10.1080/1743727X.2014.933473

Hanneman, S. K. (2008). Design, analysis, and interpretation of method-comparisoon studies. *AACN Advanced Critical Care, 19*(2), 223-234. doi:10.1097/01.AACN.0000318125.41512.a3

Hanses, P., & Rost, D. H. (1998). Das "drama" der hochbegabten underachiever. "Gewöhnliche" oder "außergewöhnliche" underachiever? [The "drama" of gifted underachievers. "Common" or "exceptional" underachievers?]. *Zeitschrift für Pädagogische Psychologie, 12*, 53-71.

Harcourt Educational Measurement. (2003). *Otis–Lennon School Ability Test Eighth Edition: Technical Manual.* New York, NY: Harcourt Educational Measurement.

Harder, B., Vialle, W., & Ziegler, A. (2014). Conceptions of giftedness and expertise put to the empirical test. *High Ability Studies, 25*(2), 83-120. doi:10.1080/13598139.2014.968462

Harlen, W. (2005). Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education, 20*(3), 245-270. doi:10.1080/02671520500193744

Harrison, G. E., & Van Haneghan, J. P. (2011). The gifted and the shadow of the night: Dabrowski's overexcitabilities and their correlation to insomnia, death anxiety, and fear of the unknown. *Journal for the Education of the Gifted, 34*(4), 669-697, 699-701.

Harter, S. (2006). The development of self-representations in childhood and adolescence. In W. Damon, & R. Lerner (Eds.), *Handbook of Child Psychology* (6th ed.). New York, NY: Wiley.

Hayes, M. L., & Sloat, R. S. (1989). Gifted students at risk for suicide. *Roeper Review, 12*(2), 102-107.

Hébert, T. P. (2001). "If I had a new notebook, I know things would change": Bright underachieving young men in urban classrooms. *Gifted Child Quarterly, 45*(3), 174-194. doi: 10.1177/001698620104500303

Hébert, T. P., & Olenchak, F. R. (2000). Mentors for gifted underachieving males: Developing potential and realizing promise. *Gifted Child Quarterly, 44*(3), 196-207.

Hecht, S. A., & Greenfield, D. B. (2001). Comparing the predictive validity of first grade teacher ratings and reading-related tests on third grade levels of reading skills in young children exposed to poverty. *School Psychology Review, 30*(1), 50-69.

Heller, K. A. (2004). Identification of gifted and talented students. *Psychology Science, 46*(3), 302-323.

Herrnstein, R. J., & Murray, C. (1994). *The Bell Curve: Intelligence and Class Structure in American Life.* New York, NY: Free Press.

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences, 24*(3), 383-403.

Hertzog, N. B., & Chung, R. U. (2015). Outcomes for students on a fast track to college: Early college entrance programs at the University of Washington. *Roeper Review, 37*(1), 39-49. doi:10.1080/02783193.2014.976324

Hill, H. C., Charalambous, C. Y., Blazar, D., McGinn, D., Kraft, M. A., Beisiegel, M., . . . Lynch, K. (2012). Validating arguments for observational instruments: Attending to multiple sources of variation. *Educational Assessment, 17*(2-3), 88-106.

Hinnant, J. B., O'Brien, M., & Ghazarian, S. R. (2009). The longitudinal relations of teacher expectations to achievement in the early school year. *Journal of Educational Psychology, 101*(3), 662-670. doi:10.1037/a0014306

Ho, P.-L., Leung, S. M.-H., Tse, H., Chow, K.-H., Cheng, V. C.-C., & Que, T.-L. (2014). Novel selective medium for isolation of Staphylococcus Lugdunensis from wound specimens. *Journal of Clinical Microbiology, 52*(7), 2633-2636. doi:10.1128/JCM.00706-14

Hoffmann, N. G., & Ninonuevo, F. G. (1994). Concurrent validation of substance abusers self-reports against collateral information: Percentage agreement vs. k vs. Yule's Y. *Alchoholism: Clinical and Experimental Research, 18*(2), 231-237.

Hoover-Schultz, B. (2005). Gifted underachievement: Oxymoron or educational enigma. *Gifted Child Today, 28*(2), 46-49.

Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta Psychologica, 26*, 107-129.

Hsieh, P., Sullivan, J. R., & Guerra, N. S. (2007). A closer look at college students: Self-efficacy and goal orientation. *Journal of Advanced Academics, 18*(3), 454-476.

Hubley, A. M., & Zumbo, B. D. (2011). Validity and the consequences of test interpretation and use. *Social Indicators Research, 103*(2), 219-230.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings.* Thousand Oaks, CA: SAGE.

Hurwitz, J. T., Elliott, S. N., & Braden, J. P. (2007). The influence of test familiarity and student disability status upon teachers' judgements of students' test performance. *School Psychology Quarterly, 22*(2), 115-144. doi:10.1037/1045-3830.22.2.115

Hwang, M. H., Lee, D., Lim, H. J., Seon, H. Y., Hutchison, B., & Pope, M. (2014). Academic underachievement and recovery: Student perspectives on effective career interventions. *The Career Development Quarterly, 62*(1), 81-94.

Jackson, M. A., Perolini, C. M., Fietzer, A. W., Altschuler, E., Woerner, S., & Hashimoto, N. (2011). Career-related success-learning experiences of academically underachieving urban middle school students. *The Counselling Psychologist, 39*(7), 1024-1060.

Jagger, C., Matthews, R., Melzer, D., Matthews, F., Brayne, C., & Study, M. C. (2007). Educational differences in the dynamics of disability incidence, recovery and mortality: Findings from the MRC Cognitive Function and Ageing Study (MRC CFAS). *International Journal of Epidemiology, 36*, 358-365.

Jarjoura, C., Tayeh, P. A., & Zgheib, N. K. (2015). Using team-based learning to teach grade 7 biology: Student satisfaction and improved performance. *Journal of Biological Education, 49*(4), 401-419. doi:10.1080/00219266.2014.967277

Jencks, C., & Phillips, M. (1998). *The Black-White Test Score Gap.* Washington, WA: Brookings Institute.

Jessurun, J. H., Shearer, C. B., & Weggeman, M. C. (2015). A universal model of giftedness - an adaptation of the Munich Model. *High Ability Studies, 27*(2), 113-128. doi:10.1080/13598139.2015.1108184

Johnson, J. A., & D'Amato, R. C. (2005). Review of the Stanford-Binet Intelligence Scales: Fifth Edition. In R. S. Spies, & B. S. Plake (Eds.), *The sixteenth mental measurements yearbook* (pp. 976-979). Lincoln, NE: Buros Institute of Mental Measurements.

Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis* (6th ed.). New York, NY: Pearson.

Johnson, S. (2013). On the reliability of high-stakes teacher assessment. *Research Papers in Education, 28*(1), 91-105.

Johnson, W., Chumlea, W. C., Czerwinski, S. A., & Demerath, E. W. (2012). Concordance of the recently published body adiposity index with measured body fat percent in European-American adults. *Obesity, 20*(4), 900-903. doi:10.1038/oby.2011.346

Johny, L., Lukose, L., & Magno, C. (2012). The assessment of academic self-regulation and learning strategies: Can they predict school ability? *Educational Measurement and Evaluation Review, 3*, 77-89.

Jones, P. H. (2010). *Introducing Neuroeducational Research.* Abingdon, GB: Taylor & Francis.

Jones, S. (2005). The invisibility of the underachieving girl. *International Journal of Inclusive Education, 9*(3), 269-286.

Jones, S., & Myhill, D. (2004). Seeing things differently: Teachers' constructions of underachievement. *Gender and Education, 16*(4), 531-546.

Jovanović, V., Teovanović, P., Mentus, T., & Petrović, M. (2010). The gifted underachiever in school: A student who has a problem or a 'rebel' making problems? *Psihologija, 43*(3), 263-279. doi:10.2298/PSI1003263J

Jung, J. Y. (2014). Predictors of attitudes to gifted programs/provisions: Evidence from preservice educators. *Gifted Child Quarterly, 58*(4), 247-258. doi:10.1177/0016986214547636

Jung, J. Y., McCormick, J., & Gross, M. U. M. (2012). The forced choice dilemma: A model incorporating Idiocentric/Allocentric cultural orientation. *Gifted Child Quarterly, 56*(1), 15-24. doi:10.1177/0016986211429169

Kaiser, J., Retelsdorf, J., Südkamp, A., & Möller, J. (2013). Achievement and engagement: How student characteristics influence teacher judgements. *Learning and Instruction, 28*, 73-84. doi:10.1016/j.learninstruc.2013.06.001

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17-64). Westport, CT: Praeger.

Kane, M. T. (2013). Validating the interpretations and ues of test scores. *Journal of Educational Measurement, 50*(1), 1-73. doi:10.1111/jedm.12000

Kanevsky, L., & Keighley, T. (2003). To produce or not to produce? Understanding boredom and the honor in underachievement. *Roeper Review, 26*, 20-28.

Kaplan, R. M., & Saccuzzo, D. P. (2010). *Psychological Testing: Principles, Applications, & Issues* (8th ed.). Belmont, CA: Wadsworth, Cengage learning.

Kaufman, A. S. (2013). Intelligent testing with Wechsler's fourth editions: Perspectives on the Weiss et al. studies and the eight commentaries. *Journal of Psychoeducational Assessment, 31*(2), 224-234. doi:10.1177/0734282913478049

Keith, T. (in press). Review of the Wechsler Intelligence Scale for Children--Fifth Edition. In J. F. Carlson, K. F. Geisinger, & J. L. Jonson (Eds.), *The Twentieth Mental*

Measurements Yearbook. Lincoln, NE: Buros Center for Testing. Retrieved from Mental Measurements Yearbook with Tests in Print database (Ovid)

Ki-Soon, H., & Marvin, C. (2000). A five year follow-up study of the Nebraska project: Still a long way to go... *Roeper Review, 23*(1), 25-33.

Kohn, A. (2000). Burnt at the high stakes. *Journal of Teacher Education, 51*(4), 315-327. doi:10.1177/0022487100051004007

Kroesbergen, E. H., van Hooijdonk, M., Van Viersen, S., Middel-Lalleman, M. M., & Reijnders, J. J. (2016). The psychological well-being of early identified gifted children. *Gifted Child Quarterly, 60*(1), 16-30. doi:10.1177/0016986215609113

Kulik, J. A., & Kulik, C. C. (1992). Meta-analytic findings on grouping programs. *Gifted Child Quarterly, 36*(2), 73-77. doi:10.1177/001698629203600204

Kumazawa, T. (2013). Evaluating validity for in-house placement test score interpretations and uses. *Japan Association for Language Teaching Journal, 35*(1), 73-100.

Kundel, H. L., & Polansky, M. (2003). Measurement of observer agreement. *Radiology, 228*(2), 303-308. doi:http://dx.doi.org/10.1148/radiol.2282011860

Kush, J. C. (2005). Review of the Stanford-Binet Intelligence Scales: Fifth Edition. In R. S. Spies, & B. S. Plake (Eds.), *The Sixteenth Mental Measurements Yearbook* (pp. 979-984). Lincoln, NE: Buros Institute of Mental Measurements.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. doi:10.2307/2529310

Landis, R. N., & Reschly, A. L. (2013). Reexamining gifted underachievement and dropout through the lens of student engagement. *Journal for the Education of the Gifted, 36*, 220-249.

Lane, S. (2012). Consequences of assessment and accountability systems are integral to the argument-based approach to validity. *Measurement: Interdisciplinary Research and Perspectives, 10*(1), 71-74.

Lane, S. (2013). The need for a principled approach for examining indirect effects of test use. *Measurement: Interdisciplinary Research and Perspectives, 11*(1), 44-46.

Lane, S. P., & Sher, K. J. (2015). Limits of current approaches to diagnosis severity based on criterion counts: An example with DSM-5 alcohol use disorder. *Clinical Psychological Science, 3*(6), 819-835. doi:10.1177/2167702614553026

Lau, K., & Chan, D. W. (2001). Identification of underachivers in Hong Kong: Do different methods select different underachievers? *Educational Studies, 27*(2), 187-200.

Lau, K., & Chan, D. W. (2001). Motivational characteristics of under-achievers in Hong Kong. *Educational Psychology: An International Journal of Experimental Educational Psychology, 21*(4), 417-430.

Lee, S. W., & Stefany, E. F. (1995). Review of the Woodcock-Johnson Psycho-Educational Battery---Revised. In J. C. Conoley, & J. C. Impara (Eds.), *The Twelfth Mental Measurements Yearbook.* Lincoln, NE: Buros Center for Testing. Retrieved from Mental Measurements Yearbook with Tests in Print database (Ovid)

Lee-Corbin, H., & Evans, R. (1996). Factors influencing success or underachievement of the able child. *Early Child Development and Care, 117*(1), 133-144.

Lew, M. J. (2012). Bad statistical practice in pharmacology (and other basic biomedical disciplines): You probably don't know P. *British Journal of Pharmacology, 166*(5), 1559-1567.

Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics, 45*, 255-268.

Lin, L. I.-K. (2000). A note on the concordance correlation coefficient. *Biometrics, 56*, 324-325.

Lin, L. I.-K., Hedayat, A. S., & Wu, W. (2007). A unified approach for assessing agreement for continuous and categorical data. *Journal of Biopharmaceutical Statistics, 17*, 629-652. doi:10.1080/10543400701376498

Lindsley, M. D., Mekha, N., Baggett, H. C., Surinthong, Y., Autthateinchai, R., Sawatwong, P., . . . Poonwan, N. (2011). Evaluation of a newly developed lateral flow immunoassay for the diagnosis of cryptococcosis. *Clinical Infectious Diseases, 53*(4), 321-325. doi:10.1093/cid/cir379

Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software, 42*(10), 1-29. Retrieved from http://www.jstatsoft.org/v42/i10/

Lohman, D. F. (2005). An aptitude perspective on talent identification: Implications for identification of academically gifted minority students. *Journal for the Education of the Gifted, 28*, 333-360.

Lovett, B. J., & Sparks, R. L. (2013). The identification and performance of gifted students with learning disability diagnoses: a quantitative synthesis. *Journal of Learning Disability, 46*(4), 304-316. doi:10.1177/0022219411421810

Lupart, J., & Pyryt, M. (1996). "Hidden gifted" students: Underachiever prevalence and profile. *Journal for the Education of the Gifted, 20,* 36-53.

MacCann, R. G., & Stanley, G. (2010). Classification consistency when scores are converted to grades: examination marks versus moderated school assessments. *Assessment in Education: Principles, Policy & Practice, 17*(3), 255-272.

Machin, S., Marie, O., & Vujić, S. (2011). The crime reducing effect of education. *The Economic Journal, 121*(552), 463-484. doi:10.1111/j.1468-0297.2011.02430.x

Maddux, C. D. (2010). Review of the Otis-Lennon School Ability Test(r), Eighth Edition. In R. A. Spies, J. F. Carlson, & K. F. Geisinger (Eds.), *The Eighteenth Mental Measurements Yearbook.* Lincoln, NE: Buros Center for Testing. Retrieved from Mental Measurements Yearbook with Tests in Print database (Ovid)

Magno, C. (2009). Investigating the effect of school ability on self-efficacy, learning approaches, and metacognition. *Asia-Pacific Education Researcher, 18*(2), 233-244.

Maki, K. E., Floyd, R. G., & Roberson, T. (2015). State learning disability eligibility criteria: A comprehensive review. *School Psychology Quarterly, 30*(4), 457-469. doi:10.1037/spq0000109

Mammadov, S., Ward, T. J., & Riedl, J. (2016). Use of latent profile analysis in studies of gifted students. *Roeper Review, 38*(3), 175-184. doi:10.1080/02783193.2016.1183739

Marland, S. P. (1972). *Education of the gifted and talented: Report to the Congress of the United States by the U.S. Commissioner of Education and background papers submitted to the U.S. Office of Education, 2 vols.* Washington, WA: U.S. Government Printing Office.

Martin, S. D., & Shapiro, E. S. (2011). Examining the accuracy of teachers' judgements of DIBELS performance. *Psychology in the Schools, 48*(4), 343-356. doi:10.1002/pits.20558

Masden, C. A., Leung, O. N., Shore, B. M., Schneider, B. H., & Udvari, S. J. (2015). Social-perspective coordination and gifted adolescents' friendship quality. *High Ability Studies, 26*(1), 3-38. doi:10.1080/13598139.2015.1028613

Mashburn, A. J., & Henry, G. T. (2004). Assessing school readiness: Validity and bias in preschool and kindergarten teachers' ratings. *Educational Measurement: Issues and Practice, 23*(4), 16-30. doi:10.1111/j.1745-3992.2004.tb00165.x

Matthews, M. S., & McBee, M. T. (2007). School factors and the underachievement of gifted students in a talent search summer program. *Gifted Child Quarterly, 51*(2), 167-181.

Maynard, T., Waters, J., & Clement, J. (2013). Child-initiated learning, the outdoor environment and the 'underachieving' child. *Early Years, 33*(3), 212-225. doi:10.1080/09575146.2013.771152

McBride, G. B. (2005). *A proposal for strength-of-agreement criteria for Lin's Concordance Correlation Coefficient.* HAM2005-062: NIWA Client Report. Retrieved from http:www.niwa.co.nz

McCall, R. B. (1994). Academic underachievers. *Current Directions in Psychological Science, 3*(1), 15-19.

McCall, R. B., Beach, S. R., & Lau, S. (2000). The nature and correlates of underachievement among elementary schoolchildren in Hong Kong. *Child Development, 71*(3), 785-801.

McCall, R. B., Evahn, C., & Kratzer, L. (1992). *High school underachievers: What do they achieve as adults?* New York, NY: SAGE.

McClain, M. -C., & Pfeiffer, S. I. (2012). Education for the gifted in the United States today: A look at state definitions, policies, and practices. *Journal of Applied School Psychology, 28*(1), 59-88.

McCoach, D. B., & Siegle, D. (2003). Factors that differentiate underachieving gifted students from high-achieving gifted students. *Gifted Child Quarterly, 47*(2), 144-154.

McCoach, D. B., & Siegle, D. (2003). The school attitude assessment survey-revised: A new instrument to identify academically able students who underachieve. *Educational and Psychological Measurement, 63*(3), 414-429.

McCoach, D. B., & Siegle, D. (2008). Underachievers. In J. A. Plucker, & C. M. Callahan (Eds.), *Critical issues and practices in gifted education: What the research says* (pp. 721-734). Waco, TX: Prufrock Press.

McGaghie, W. C., Cohen, E. R., & Wayne, D. B. (2011). Are United States medical licensing exam step 1 and 2 scores valid measures for postgraduate medical residency selection decisions? *Academic Medicine, 86*(1), 48-52.

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*(1), 1-10. doi:10.1016/j.intell.2008.08.004

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276-282. doi:10.11613/BM.2012.031

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*(2), 153-157. doi:10.1007/BF02295996

Medallon, M. C., & Cataquis, R. E. (2011). Predictive validity of the Otis-Lennon School Ability Test (OLSAT) to the frst semester performance of incoming students at Lyceum of the Philippines - Laguna. *Lyceum of the Philippines-Laguna Research Journal, 1*(1), 71-77.

Meisels, S. J., Bickel, D. D., Nicholson, J., Xue, Y., & Atkins-Burnett, S. (2001). Trusting teachers' judgements: A validity study of curriculum-embedded performance assessment in kindergarten to grade 3. *American Educational Research Journal, 38*(1), 73-95. doi:10.3102/00028312038001073

Merrotsy, P. (2013). Invisible gifted students. *Talent Development & Excellence, 5*(2), 31-42.

Merrotsy, P. (2016). Teaching gifted Aboriginal and Torres Strait Islander children. In N. Harrison, & J. Sellwood (Eds.), *Learning and Teaching in Aboriginal and Torres Strait Islander Education* (pp. 100-117). South Melbourne, AU: Oxford University Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.

Messick, S. (1996). Validity of performance assessment. In G. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1-18). Washington, WA: National Center for Education Statistics.

Miller, M. D. (2010). Review of the Wechsler Individual Achievement Test-Third Edition. In R. A. Spies, J. F. Carlson, & K. F. Geisinger (Eds.), *The Eighteenth Mental Measurements Yearbook.* Lincoln, NE: Buros Center for Testing. Retrieved from Mental Measurements Yearbook with Tests in Print database (Ovid)

Miller, W. H., Kerr, B., & Ritter, G. (2008). School performance measurement. *The American Review of Public Administration, 38*(1), 100-117. doi:10.1177/0275074007304387

Mishel, L., Bivens, J., Gould, E., & Shierholz, H. (2012). *The State of Working America* (12th ed.). Ithaca, NY: Cornell University Press.

Missett, T. C. (2013). Exploring the relationship between mood disorders and gifted individuals. *Roeper Review, 35*(1), 47-57. doi:10.1080/02783193.2013.740602

Moon, S. M. (2009). Myth 15: high-ability students don't face problems and challenges. *Gifted child quarterly, 53*(4), 274-276.

Morgan, H. (1996). An analysis of Gardner's theory of multiple intelligence. *Roeper Review, 18*, 263-270.

Morisano, D., & Shore, B. M. (2010). Can personal goal setting tap the potential of the gifted underachiever? *Roeper Review, 32*(4), 249-258.

Morse, D. (2010). Review of Otis-Lennon School Ability Test(r), Eighth Edition. In R. A. Spies, J. F. Carlson, & K. F. Geisinger (Eds.), *The Eighteenth Mental Measurements Yearbook.* Lincoln, NE: Buros Center for Testing. Retrieved from Mental Measurements Yearbook with Tests in Print database (Ovid)

Moss, P. (1994). Can there be validity without reliability? *Educational Researcher, 23*, 5-12.

Moss, P. A. (2013). Validity in action: Lessons from studies of data use. *Journal of Educational Measurement, 50,* 91-98. doi:10.1111/jedm.12003

Muir-Broaddus, J. E. (1995). Gifted underachievers: Insights from the characteristics of strategic functioning associated with giftedness and achievement. *Learning and Individual Differences, 7*(3), 189-206.

Murphy, R. (2007). Response to commentary on chapter 8. In P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards* (pp. 329-330). London, GB: QCA.

Naglieri, J. A., & Ford, D. Y. (2015). Misconceptions about the Naglieri Nonverbal Ability Test: A Commentary of Concerns and Disagreements. *Roeper Review, 37*(4), 234-240.

Neihart, M., Reis, S. M., Robinson, N. M., & Moon, S. M. (2002). *The social and emotional development of gifted children: What do we know?* Waco, TX: Prufrock Press.

Neubauer, A. C., & Opriessnig, S. (2014). The development of talent and excellence - do not dismiss psychometric intelligence, the (potentially) most powerful predictor. *Talent Development & Excellence, 6*(2), 1-15.

Neumeister, S. K., & Hébert, T. P. (2003). Underachievement versus selective acheivement: Delving deeper and discovering the difference. *Journal for the Education of the Gifted, 26*(3), 221-238. doi:10.1177/016235320302600305

New South Wales. Department of Training and Education Co-ordination (DTEC). (1997). *Securing their future.* Sydney: NSW Government. Retrieved December 30, 2013, from https://www.det.nsw.edu.au/media/downloads/reviews/hscwhite.pdf

Newton, P. E. (2005). Examination standards and the limits of linking. *Assessment in Education, 12*(2), 105-123.

Newton, P. E. (2013). Two kinds of argument? *Journal of Educational Measurement, 50,* 105-109. doi:10.1111/jedm.12004

Nichols, P. D., & Williams, N. (2009). Consequences of test score use as validity evidence: roles and responsibilities. *Educational Measurement: Issues and Practice, 28*(1), 3-9.

Nisbett, R. E., Aronson, J., Blair, C., Dickens, W., Flynn, J., Halpern, D. F., & Turkheimer, E. (2012). Intelligence: New findings and theoretical developments. *American Psychologist, 67*(2), 130-159. doi:10.1037/a0026699

Noble, K. D., Vaughan, R. C., Chan, C., Childers, S., Chow, B., Federow, A., & Hughes, S. (2007). Love and work: The legacy of early university entrance. *Gifted Child Quarterly, 51*, 152-166.

Nurmi, J., Onatsu, T., & Haavisto, T. (1995). Underachievers' cognitive and behavioural strategies - self-handicapping at school. *Contemporary Educational Psychology, 20,* 188-200. doi:10.1006/ceps.1995.1012

Nussbaum, E. M. (2015). *Categorical and Nonparametric Data Analysis.* New York, NY: Taylor & Francis.

Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature, 506,* 150-152. Retrieved from http://www.nature.com/news/scientific-method-statistical-errors-1.14700

Oakland, T. (1985). Review of Otis–Lennon School Ability Test. In J. V. Mitchell (Ed.), *The Ninth Mental Measurements Yearbook* (pp. 1111-1112). Lincoln, NE: Buros Institute of Mental Measurements.

Obergriesser, S., & Stoeger, H. (2015). The role of emotions, motivation, and learning behavior in underachievement and results of an intervention. *High Ability Studies, 26*(1), 167-190. doi:10.1080/13598139.2015.1043003

Olszewski-Kubilius, P. (1995). A summary of research regarding early entrance to college. *Roeper Review, 18*, 121-126.

Panter, J., Costa, S., Dalton, A., Jones, A., & Ogilvie, D. (2014). Development of methods to objectively identify time spent using active and motorised modes of travel to work: How do self-reported measures compare? *International Journal of Behavioral Nutrition and Physical Activity, 11*, 116-131. doi:10.1186/s12966-014-0116-x

Park, G., Lubinski, D., & Benbow, C. P. (2013). When less is more: Effects of grade skipping on adult STEM productivity among mathematically precocious adolescents. *Journal of Educational Psychology, 105*(1), 176-198. doi:10.1037/a0029481

Park, Y. S., Riddle, J., & Tekian, A. (2014). Validity evidence of resident competency ratings and the identification of problem residents. *Medical Education, 48*(6), 614-622. doi:10.1111/medu.12408

Pearson. (2015). *Otis-Lennon School Ability Test, Eighth Edition*. Retrieved December 17, 2015, from http://www.pearsonassessments.com/learningassessments/products/100000003/otis-lennon-school-ability-test-eighth-edition-olsat-8-olsat-8.html?Pid=OLSAT

Pellegrini, A. D. (2001). Practioner Review: The role of direct observation in the assessment of young children. *Journal of Child Psychology and Psychiatry, 42*(7), 861-869. doi:10.1111/1469-7610.00783

Peng, R. (2015). The reproducibility crisis in science: A statistical counterattack. *Significance, 12*(3), 30-32.

Perleth, C., & Heller, K. A. (1994). The Munich longitudinal study of giftedness. In R. F. Subotnik, & K. D. Arnold (Eds.), *Beyond Terman: Contemporary longitudinal studies of giftedness and talent* (pp. 77-114). Norwood: Ablex.

Peters, S. J., & Engerrand, K. G. (2016). Equity and excellence: Proactive efforts in the identification of underrepresented students for gifted and talented services. *Gifted Child Quarterly, 60*(3), 159-171. doi:10.1177/0016986216643165

Peters, W. A., & van Boxtel, H. W. (1999). Irregular error patterns in Raven's Standard Progressive Matrices: A sign of underachievement in testing situations? *High Ability Studies, 10*(2), 213-232.

Peterson, J. S. (2000). A follow-up study of one group of achievers and underachievers four years after high school graduation. *Roeper Review, 22*(4), 217-224.

Peterson, J. S. (2001). Successful adults who were once adolescent underachievers. *Gifted Child Quarterly, 45*(4), 236-250.

Peterson, J. S., & Colangelo, N. (1996). Gifted achievers and underachievers: a comparison of patterns found in school files. *Journal of Counseling and Development, 74*, 399-407.

Pett, M. A. (1997). *Nonparametric Statistics for Health Care Research: Statistics for small samples and unusual distributions* (6th ed.). Thousand Oaks, CA: SAGE.

Pfeiffer, S. I. (2012). Current perspectives on the identification and assessment of gifted students. *Journal of Psychoeducational Assessment, 30*(1), 3-9. doi:10.1177/0734282911428192

Phillipson, S. N., & Tse, A. K. (2007). Discovering patterns of achievement in Hong Kong students: An application of the Rasch measurement model. *High Ability Studies, 18*(2), 173-190. doi:10.1080/13598130701709640

Pietschnig, J., & Voracek, M. (2015). One century of global IQ gains. *Perspectives on Psychological Science, 10*(3), 282-306. doi:10.1177/1745691615577701

Pipkin, C. S., Winters, S. M., & James, W. (2007). Effects of instruction, goals and reinforcement on academic behavior: Assessing skill versus reinforcement deficits. *Journal of Early and Intensive Behavior Intervention, 4*(4), 648-657.

Plewis, I. (1991). Underachievement: A case for conceptual confusion. *British Educational Research Journal, 17*(4), 377-385.

Plucker, J. A., & Callahan, C. M. (2014). Research on giftedness and gifted education: Status of the field and considerations for the future. *Exceptional Children, 80*(4), 390-406.

Plucker, J. A., Callahan, C. C., & Tomchin, E. M. (1996). Wherefore art thou, multiple intelligences? Alternative assessments for identifying talent in ethnically diverse and low income students. *Gifted Child Quarterly, 40*, 81-92.

Preckel, F., Holling, H., & Vock, M. (2006). Academic underachievement: Relationship with cognitive motivation, achievement motivation, and conscientiousness. *Psychology in the Schools, 43*(3), 401-411.

Preiss, D., & Fisher, J. (2008). A measure of confidence in Bland-Altman analysis for the interchangeability of two methods of measurement. *Journal of Clinical Monitoring and Computing, 22*, 257-259. doi:10.1007/s10877-008-9127-y

Punch, K. F. (2005). *Introduction to Social Research Quantitative and Qualitative Approaches* (2nd ed.). London, GB: SAGE.

Pyryt, M. C. (2000). Finding "g": Easy viewing through higher order factor analysis. *Gifted Child Quarterly, 44*, 190-192.

Rafidi, M. (2008). Gifted achievement and underachievement in the classroom. *English in Australia, 43*(2), 63-65.

Rayneri, L. J., Gerber, B. L., & Wiley, L. P. (2003). Gifted achievers and gifted underachievers: The impact of learning style preferences in the classroom. *Journal of Secondary Gifted Education, 14*(4), 197-204. doi:10.4219/jsge-2003-434

Rayneri, L. J., Gerber, B. L., & Wiley, L. P. (2006). The relationship between classroom environment and the learning style preferences of gifted middle school students and the impact on levels of performance. *Gifted child quarterly, 50*(2), 104-118.

Reckase, M. D. (1998). Consequential validity from the test developer's perspective. *Educational Measurement Issues and Practice, 17*(2), 13-16.

Redding, R. E. (1990). Learning preferences and skill patterns among underachieving gifted adolescents. *Gifted Child Quarterly, 34*, 72-75.

Reis, S. M., & McCoach, D. B. (2000). The underachievement of gifted students: What do we know and where do we go? *Gifted Child Quarterly, 44*(3), 152-170.

Reis, S. M., & Renzulli, J. S. (2010). Is there still a need for gifted education? An examination of current research. *Learning and Individual Differences, 20*, 308-317.

Reis, S. M., Baum, S. M., & Burke, E. (2014). An operational definition of twice-exceptional learners: Implications and applications. *Gifted Child Quarterly, 58*(3), 217-230. doi:10.1177/0016986214534976

Reis, S. M., Burns, D. E., & Renzulli, J. S. (1992). *Curriculum compacting: The complete guide to modifying the regular curriculum for high ability students.* Mansfield Center, CT: Creative Learning Press.

Reis, S. M., Colbert, R. D., & Hébert, T. P. (2005). Understanding resilience in diverse, talented students in an urban high school. *Roeper Review*, *27*(2), 110-120.

Renzulli, J. S. (1978). What makes giftedness? Reexamining a definition. *Phi Delta Kappan, 60*(3), 180-184.

Renzulli, J. S. (1986). The three-ring conception of giftedness: A developmental model for creative productivity. In R. J. Sternberg, & J. Davidson (Eds.), *Conceptions of Giftedness* (pp. 53-92). New York: Cambridge University Press.

Reschly, D. J., & Hosp, J. L. (2004). State SLD identification and policies and practices. *Learning Disabilities Quarterly, 27*(4), 197-213.

Rettew, D. C., Lynch, A. D., Achenbach, T. M., Dumenci, L., & Ivanova, M. Y. (2009). Meta-analyses of agreement between diagnoses made from clinical evaluations and standardized diagnostic interviews. *International Journal of Methods in Psychiatric Research, 18*(3), 169-184.

Richert, E. S. (2003). Excellence with justice in identification and programming. In N. Colangelo, & G. A. Davis (Eds.), *Handbook of Gifted Education* (3rd ed., pp. 146-158). Boston, MA, MA: Pearson Education.

Rimm, S. B. (1991). *Underachievement Syndrome: Causes and cures.* Watertown, MA: Apple Publishing.

Rimm, S. B. (1997). An underachievement epidemic. *Educational Leadership, 54*(7), 18-22.

Rimm, S. B. (2003). Underachievement: A national epidemic. In N. Colangelo, & G. A. Davis (Eds.), *Handbook of Gifted Education* (3rd ed., pp. 424-443). Boston, MA: Allyn & Bacon.

Ritchotte, J. A., Matthews, M. S., & Flowers, C. P. (2014). The validity of the achievement-orientation model for gifted middle school students: An exploration study. *Gifted Child Quarterly, 58*(3), 183-198. doi:10.1177/0016986214534890

Ritchotte, J., Rubenstein, L., & Murry, F. (2015). Reversing the underachievement of gifted middle school students. *Gifted Child Today, 38*(2), 103-113. doi:10.1177/1076217514568559

Ritts, V. M., Patterson, M. L., & Tubbs, M. E. (1992). Expectations, impressions, and judgements of physically attractive students: A review. *Review of Educational Research, 62*(4), 413-426. doi:10.2307/1170486

Robbins, B; Society for Humanistic Psychology. (2011). *Open Letter to the DSM-5*. Retrieved January 4th, 2016, from http://www.ipetitions.com/petition/dsm5/

Roberts, S. W., Sheffield, J. S., McIntire, D. D., & Alexander, J. M. (2011). Urine screening for chlamydia trachomatis during pregnancy. *Obstetrics & Gynecology, 117*(4), 883-885. doi:10.1097/AOG.0b013e3182107d47

Robertson, S., & Pfeiffer, S. (2016). Development of a procedural guide to implement response to intervention (RtI) with high-ability learners. *Roeper Review, 38*(1), 9-23. doi:10.1080/02783193.2015.1112863

Robieson, W. Z. (1999). *On Weighted Kappa and Concordance Correlation Coefficient, Ph.D. thesis.* Chicago, MI: University of Illinois.

Robinson, A., Shore, B. M., & Enersen, D. L. (2007). *Best Practices in Gifted Education: An Evidence-Based Guide.* Waco, TX: Prufrock Press.

Robinson, N. (2006). A report card on the state of research in the field of gifted education. *Gifted Child Quarterly, 50*(4), 342-345.

Robinson, N. M., Reis, S. M., Neihart, M., & Moon, S. M. (2002). Social and emotional issues: What have we learned and what should we do now? In M. Neihart, S. M. Reis, N. M. Robinson, & S. M. Moon (Eds.), *The Social and Emotional Development of Gifted Children: What Do We Know?* (pp. 267-288). Waco, TX: Prufrock Press.

Roche, G. (1979). Much ado about mentors. *Harvard Business Review, 57,* 14-28.

Rock, D. A., & Stenner, J. A. (2005). Assessment issues in the testing of children at school entry. *The Future of Children, 15*(1), 15-34. doi:10.1353/foc.2005.0009

Rosado, J. I., Pfeiffer, S., & Petscher, Y. (2015). Identifying gifted students in Puerto Rico. *Gifted Education International, 31*(2), 162-175. doi:10.1177/0261429413507178

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52*(1), 59-82.

Rothman, K. J., Lash, T. L., & Greenland, S. (2008). *Modern Epidemiology* (3rd ed.). Baltimore, MD: Lippincott Williams & Wilkins.

Rushton, J. P., & Jensen, A. R. (2005). Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law, 11*(2), 235-294.

Ryan, J. (2007). *Internal Grant Final Reports.* Warrensburg, MO: University of Central Missouri. Retrieved February 20, 2016, from https://www.ucmo.edu/osp/reports.cfm?print=yes

Ryan, T. G., & Coneybeare, S. (2013). The underachievement of gifted students: A synopsis. *Journal of the International Association of Special Education, 14*(1), 58-66.

Salinas-Jimenez, M., Artes, J., & Salinas-Jimenez, J. (2011). Education as a positional good: A life satisfaction approach. *Social Indicators Research, 103*(3), 409-426.

Salmela, M., & Määttä, K. (2015). Even the best have difficulties: A study of Finnish straight-A graduates' resource-oriented students. *Gifted Child Quarterly, 59*(2), 124-135. doi:10.1177/0016986214568720

Santelices, M. V., & Taut, S. (2011). Convergent validity evidence regarding the validity of the Chilean standards-based teacher evaluation system. *Assessment in Education: Principles, Policy & Practice, 18*(1), 73-93.

Schober, B., Reimann, R., & Wagner, P. (2004). Is research on gender-specific underachievement in gifted girls an obsolete topic? New findings on an often discussed issue. *High Ability Studies, 15*(1), 43-62.

Schultz, R. A. (2002). Understanding giftedness and underachievement: At the edge of possiblity. *Gifted Child Quarterly, 46*(3), 193-208.

Sharp, C., Kendall, L., & Schagen, I. (2003). Different for girls? An exploration of the impact of "playing for success". *Educational Research, 45*(3), 309-324.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405-450.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and practice, 16*(2), 5-8, 13, 24.

Siegler, R. S. (1992). The other Alfred Binet. *Developmental Psychology, 28*(2), 179-190. doi:10.1037/0012-1649.28.2.179

Sikora, J., & Saha, L. J. (2011). *Lost talent? The occupational ambitions and attainments of young Australians.* Commonwealth of Australia, Department of Education, Employment and Workplace Relations. Canberra: National Centre for Vocational Education Research (NCVER).

Silverman, L. K. (1997). The construct of asynchronous development. *Peabody Journal of Education, 72*(3/4), 36-58.

Silverman, W., Miezejeski, C., Ryan, R., Zigman, W., Krinsky-McHale, S., & Urv, T. (2010). Stanford-Binet and WAIS IQ differences and their implications for adults

with intellectual disability (aka mental retardation). *Intelligence, 38*(2), 242-248. doi:10.1016/j.intell.2009.12.005

Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy, 85*, 257-268.

Singhal, N., & Siddhu, A. (2011). Durnin and Womersley revisited: Need for Bland-Altman plots. *Medicine & Science in Sports & Exercise, 43*(8), 1598-1599. doi:10.1249/MSS.0b013e318220a122

Sink, C. A., & Eppler, C. (2007). Review of the Stanford-Binet Intelligence Scales for Early Childhood, Fifth Edition. In K. F. Geisinger, R. A. Spies, J. F. Carlson, & B. S. Plake (Eds.), *The Seventeenth Mental Measurements Yearbook.* Lincoln, NE: Buros Center for Testing. Retrieved from Mental Measurements Yearbook with Tests in Print database (Ovid)

Sireci, S. G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement, 50*, 99-104. doi:10.1111/jedm.12005

Smith, E. (2003). Failing boys and moral panics: Perspectives on the underachievement debate. *British journal of educational studies, 51*(3), 282-295.

Smith, E. (2007). *Analysing Underachievement in Schools.* London, GB: Continuum International Publishing Group.

Smith, E. (2010). Underachievement, failing youth and moral panics. *Evaluation and Research in Education, 23*(1), 37-49. doi:10.1080/09500791003605102

Sosniak, L. A. (2006). Retrospective interviews in the study of expertise and expert performance. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.),

The Cambridge Handbook of Expertise and Expert Performance (pp. 287-301). Cambridge, GB: Cambridge University Press.

Spearman, C. (1904). General intelligence, "objectively determined and measured". *The American Journal of Psychology, 15*(2), 201-292.

Stamatis, D. H. (2002). *Six Sigma and Beyond: Statistics and Probability* (Vol. III). New York, NY: St Lucie Press.

Staudt, B., & Neubauer, A. C. (2006). Achievement, underachievement and cortical activation: A comparative EEG study of achieving and underachieving adolescents of average and above-average intelligence. *High Ability Studies, 17*, 3-16.

Steenbergen-Hu, S., & Olszewski-Kubilius, P. (2016). How to conduct a good meta-analysis in gifted education. *Gifted Child Quarterly, 60*(2), 134-154. doi:10.1177/0016986216629545

Stepnowsky, C., Zamora, T., Barker, R., Liu, L., & Sarmiento, K. (2013). Accuracy of positive airway pressure device—measured apneas and hypopneas: Role in treatment followup. *Sleep Disorders*, 6. doi:10.1155/2013/314589

Sternberg, R. J., & Davidson, J. E. (Eds.). (2005). *Conceptions of Giftedness* (2nd ed.). Cambridge, GB: Cambridge University Press.

Stoeger, H., & Ziegler, A. (2005). Evaluation of an elementary classroom self-regulated learning program for gifted mathematics underachievers. *International Education Journal, 6*(2), 261-271.

Stoeger, H., Suggate, S., & Ziegler, A. (2013). Identifying the causes of underachievement: A plea for the inclusion of fine motor skills. *Psychological Test and Assessment Modeling, 55*(3), 274-288.

Strand, S. (2012). The white British-Black caribbean achievement gap: Tests, tiers and teacher expectations. *British Educational Research Journal, 38*(1), 75-101. doi:10.1080/01411926.2010.526702

Strand, S. (2014). Ethnicity, gender, social class and achievement gaps at age 16: Intersectionality and 'Getting it' for the white working class. *Research Papers in Education, 29*(2), 131-171. doi:10.1080/02671522.2013.767370

Subotnik, R. F., Olszewski-Kubilius, P., & Worrell, F. C. (2011). Rethinking giftedness and gifted education: A proposed direction forward based on psychological science. *Psychological Science in the Public Interest, 12*(1), 3-54. doi:10.1177/1529100611418056

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgements of students' academic achievement: A Meta-analysis. *Journal of Educational Psychology, 104*(3), 743-762. doi:10.1037/a0027627

Swan, B., Coulombe-Quach, X.-L., Huang, A., Godek, J., Becker, D., & Zhou, Y. (2015). Meeting the needs of gifted and talented students: Case study of a virtual learning lab in a rural middle school. *Journal of Advanced Academics, 26*(4), 294-319.

Swann. (1985). *Education for All: The Report of the Committee of Inquiry into the Education of Children from Ethnic Minority Groups.* London, GB: Her Majesty's Stationery Office.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240,* 1285-1293.

Tang, W., He, H., & Tu, X. M. (2012). *Applied Categorical and Count Data Analysis.* Boca Raton, FL: Taylor & Francis Group.

Tannenbaum, A. J. (1983). *Gifted children: Psychological and educational perspectives.* New York: Macmillan.

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

Terman, L. M. (1925). *Genetic Studies of Genius* (Vol. I). Stanford, CA: Stanford University Press.

Terman, L. M., & Oden, M. H. (1947). *Genetic Studies of Genius ...: The gifted child grows up; twenty-five years' follow-up of a superior group.* Stanford, CA: Stanford University Press.

Thompson, D. D., & McDonald, D. (2007). Examining the influence of teacher-constructed and student-constructed assignments on the achievement patterns of gifted and advanced sixth-grade students. *Journal for the Education of the Gifted, 31*(2), 198-226.

Thurstone, L. L. (1926). The mental age concept. *Psychological Review, 33*(4), 268-278.

Thurstone, L. L. (1938). *Primary mental abilities.* Chicago: University of Chicago Press.

Tiedemann, J. (2002). Teachers' gender stereotypes as determinants of teacher perceptions in elementary school mathematics. *Educational Studies in Mathematics, 50*(1), 49-62. doi:10.1037/1045-3830.22.2.115

Timmermans, M., vanLier, P. A., & Koot, H. M. (2009). Pathways of behavior problems from childhood to late adolescence leading to delinquency and academic

underachievement. *Journal of Clinical Child & Adolescent Psychology, 38*(5), 630-638.

Trafimow, D., & Marks, M. (2015). Editorial. *Basical and Applied Social Psychology, 37*, 1-2.

Trochim, W., & Donnelly, J. P. (2007). *The Research Methods Knowledge Base* (3rd ed.). Cincinnati, OH: Atomic Dog Publishing.

Tuss, P., Zimmer, J., & Ho, H.-Z. (1995). Causal attributions of underachieving fourth-grade students in china, Japan and the United States. *Journal of Cross-Cultural Psychology, 26*(4), 408-425.

Tyler-Wood, T., & Carri, L. (1991). Identification of gifted children: The effectiveness of various measures of cognitive ability. *Roeper Review, 14*(2), 63-64.

Tze, V. M., Daniels, L. M., & Klassen, R. M. (2016). Evaluating the relationship between boredom and academic outcomes: A meta-analysis. *Educational Psychology Review, 28*(1), 119-144. doi:10.1007/s10648-015-9301-y

United Nations. (2006). *Social Justice in an Open World: The role of the United Nations.* New York: United Nations. Retrieved January 18, 2016, from http://www.un.org/esa/socdev/documents/ifsd/SocialJustice.pdf

Universities Admissions Centre (UAC). (2001–2014). *Report on the Scaling of the 2001–2013 NSW Higher School Certificate.* Sydney, AU: Universities Admissions Centre. Retrieved from http://www.uac.edu.au/publications/atar.shtml

Uttl, B. (2005). Measurement of individual differences: lessons from memory assessment in research and clinical practice. *Psychological Science, 16*(6), 460-467.

Vacca, J. J. (2007). Review of the Stanford-Binet Intelligence Scales for early childhood, Fifth Edition. In K. F. Geisinger, R. A. Spies, J. F. Carlson, & B. S. Plake (Eds.), *The Seventeenth Mental Measurements Yearbook.* Lincoln, NE: Buros Center for Testing. Retrieved from Mental Measurements Yearbook with Tests in Print database (Ovid)

Vaillant, G. (1977). *Adaption to life.* Boston, MA: Little-Brown.

Van Boxtel, H. W., & Mönks, F. J. (1992). General, social, and academic self concepts of gifted adolescents. *Journal of Youth and Adolescence, 21*(2), 169-186.

Van den Broeck, W. (2002a). The misconception of the tegression-based fiscrepancy operationalization in the definition and research of learning disabilities. *Journal of Learning Disabilities, 35*(3), 194-204.

Van den Broeck, W. (2002b). Will the real discrepant learning disability please stand up? *Journal of Learning Disabilities, 35*(3), 209-213.

Van Nijlen, D., & Janssen, R. (2015). Examinee non-effort on contextualized and non-contextualized mathematics items in large-scale assessments. *Applied Measurement in Education, 28*(1), 68-84. doi:10.1080/08957347.2014.973559

Veas, A., Gilar, R., Miñano, P., & Castejón, J. (2016). Estimation of the proportion of underachieving students in compulsory secondary education in Spain: An application of the rasch model. *Frontiers in Psychology, 7*, 1-9. doi:10.3389/fpsyg.2016.00303

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine, 37*(5), 360-364.

Visser, B. A., Ashton, M. C., & Vernon, P. A. (2006). g and the measurement of multiple intelligences: A response to Gardner. *Intelligence, 34*(5), 507-510. doi:10.1016/j.intell.2006.04.006

Vladut, A., Vialle, W., & Ziegler, A. (2015). Learning resources within the Actiotope: A validation study of the QELC (Questionnaire of Educational Learning Capital). *Psychological Test and Assessment Modeling, 57*(1), 40-56.

Vlahovic-Stetic, V., Vidovic, V., & Arambasic, L. (1999). Motivational characteristics in mathematical achievement: a study of gifted high-achieving, gifted underachieving and non-gifted pupils. *High Ability Studies, 10*(1), 37-49.

Vogl, K., & Preckel, F. (2014). Full-time ability grouping of gifted students: Impacts on social self-concept and school-related attitudes. *Gifted Child Quarterly, 58*(1), 51-68. doi:10.1177/0016986213513795

Walter, S. D., Frommer, D. J., & Cook, R. J. (1991). The estimation of sensitivity and specificity in colorectal cancer screening methods. *Cancer Detection and Prevention, 15*, 465-469.

Walts, A. E., Bose, S., Fan, X., Frishberg, D., Scharre, K., de Peralta-Venturina, M., . . . Marchevsky, A. M. (2011). A simplified Bethesda system for reporting thyroid cytopathology using only four categories improves intra- and inter-observer diagnostic agreement and provides non-overlapping estimates of malignancy risks. *Diagnostic Cytopathology, 40*(S1), E62-E68. doi:10.1002/dc.21697

Wang, H., Choi, I., Schmidgall, J., & Bachman, L. F. (2012). Review of Pearson Test of English Academic: Building an assessment use argument. *Language Testing, 29*(4), 603-619.

Wang, L., & Hui, S. S. (2015). Validity of four commercial bioelectrical impedance scales in measuring body fat among Chinese children and adolescents. *BioMed Research International, 2015,* 1-8. doi:10.1155/2015/614858

Ward, A. W., & Murray-Ward, M. (1999). *Assessment in the Classroom.* Belmont, CA: Wadsworth.

Warne, R. T. (2016). Five reasons to put the g back into giftedness: An argument for applying the Cattell–Horn–Carroll theory of intelligence to gifted education research and practice. *Gifted Child Quarterly, 60*(1), 3-15. doi:10.1177/0016986215605360

Warrens, M. J. (2011). Cohen's kappa is a weighted average. *Statistical Methodology, 8*(6), 473-484. doi:10.1016/j.stamet.2011.06.002

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistican*. doi:10.1080/00031305.2016.1154108

Waterhouse, L. (2006). Multiple intelligences, the Mozart effect, and emotional intelligence: A critical review. *Educational Psychologist, 41*(4), 207-225. doi:10.1207/s15326985ep4104_1

Weinert, F. E., & Schneider, W. (Eds.). (1999). *Individual Development from 3 to 12: Findings from the Munich Longitudinal Study.* New York, NY: Cambridge University Press.

Weiss, L. G., Keith, T. Z., Zhu, J., & Chen, H. (2013). WAIS-IV and clinical validation of the four- and five-factor interpretative approaches. *Journal of Psychoeducational Assessment, 31*(2), 94-113. doi:10.1177/0734282913478030

Weiss, L. G., Keith, T. Z., Zhu, J., & Chen, H. (2013). WISC-IV and clinical validation of the four- and five-factor interpretative approaches. *Journal of Psychoeducational Assessment, 31*(2), 114-131. doi:10.1177/0734282913478032

Wellisch, M. (2016). Gagné's DMGT and underachievers: The need for an alternative inclusive gifted model. *Australasian Journal of Gifted Education, 25*(1), 18-30. doi:10.21505/ajge.2016.0003

Wellisch, M., & Brown, J. (2011). Where are the underachievers in the DMTG's academic talent development? *Talent Development and Excellence, 3*, 115-117.

Wellisch, M., & Brown, J. (2012). An integrated identification and intervention model for intellectually gifted children. *Journal of Advanced Academics, 23*(2), 145-167.

Weschler, D. (1991). *Manual for the Weschler intelligence scale for children - Revised.* San Antonio, TX: Psychological Corporation.

William, D. (2001). Reliability, validity and all that jazz. *Education 3-13, 29*(3), 17-21.

William, D. (2003). National curriculum assessment: How to make it better. *Research Papers in Education, 18*(2), 129-36.

Willie, C. V. (2001). The contextual effects of socioeconomic status on student achievement test scores by race. *Urban Education, 36*(4), 461-478. doi:10.1177/0042085901364002

Willse, J. T. (2010). Review of the Wechsler Individual Achievement Test-Third Edition. In R. A. Spies, J. F. Carlson, & K. F. Geisinger (Eds.), *The Eighteenth Mental Measurements Yearbook.* Lincoln, NE: Buros Center for Testing. Retrieved from Mental Measurements Yearbook with Tests in Print database (Ovid)

Wilson, V. L., & Reynolds, C. R. (1985). Another look at evalutating aptitude-achievement discrepancies in the diagnosis of learning disabilities. *Journal of Special Education, 18*, 477-497.

Wolming, S., & Wikström, C. (2010). The concept of validity in theory and practice. *Assessment in Education: Principles, Policy and Practice, 17*(2), 117-132.

Wood, E. (2003). The power of pupil perspectives in evidence-based practice: The case of gender and underachievement. *Research Papers in Education, 18*(4), 365-383.

Wools, S., Eggen, T. J., & Béguin, A. A. (2016). Constructing validity arguments for test combinations. *Studies in Educational Evaluation, 48*, 10-18. doi:10.1016/j.stueduc.2015.11.001

Wools, S., Eggen, T., & Sanders, P. (2010). Evaluation of validity and validation by means of the argument-based approach. *CADMO, 8*, 63-82.

Worrell, F. C., & Erwin, J. O. (2011). Best practices in identifying students for gifted and talented education programs. *Journal of Applied School Psychology, 27*(4), 319-340. doi:10.1080/15377903.2011.615817

Yilmaz, D. (2015). A qualitative study to understand the social and emotional needs of the gifted adolescents, who attend the science and arts centers in Turkey. *Educational Research and Reviews, 10*(8), 1109-1120.

Zaki, R., Bulgiba, A., Ismail, R., & Ismail, N. A. (2012). Statistical methods used to test for agreement of medical instruments measuring continuous variables in method comparison studies: A systematic review. *PLOS One, 7*(5), e37908. doi: 10.1371/journal.pone.0037908

Zeidner, M., & Schleyer, E. J. (1999). Test anxiety in intellectually gifted school students. *Anxiety, Stress & Coping, 12*(2), 163-189. doi:10.1080/10615809908248328

Ziegler, A. (2005). The actiotope model of giftedness. In R. J. Sternberg, & J. E. Davidson (Eds.), *Conceptions of Giftedness* (2nd ed., pp. 411-436). Cambridge, GB: Cambridge University Press.

Ziegler, A. (2008). *Hochbegabung [Giftedness]*. Munich, Germany: Reinhardt.

Ziegler, A., & Phillipson, S. N. (2012). Towards a systemic theory of gifted education. *High Ability Studies, 23*(1), 3-30. doi:10.1080/13598139.2012.679085

Ziegler, A., & Stoeger, H. (2010). How fine motor skills influence the assessment of high abilities and underachievement in math. *Journal for the Education of the Gifted, 34*(2), 195-219.

Ziegler, A., Ziegler, A., & Stoeger, H. (2012). Shortcomings of the IQ-based construct of underachievement. *Roeper Review, 34*(2), 123-132.

Ziliak, S. T. (2010). The validus medicus and a new gold standard. *The Lancet, 376*, 324-325.

Zirkel, P. A., & Thomas, L. B. (2010). State laws for RTI: An updated snapshot. *Teaching Exceptional Children, 42*(3), 56-63.

Appendix 1 Summary of Articles that Identify Underachievement

Publication	Method used to identify underachievement	Threshold/criteria for identification	Expected Achievement instrument	Actual Achievement Instrument	% GUA identified
Abelman, 2007	Mixed: Nomination + miscellaneous	profile	WISC	SAT	26
Albaili, 2003	Absolute split I	below average	Test of Nonverbal Intelligence	school grades	32
Baker, Bridger, & Evans, 1998	Absolute split (alt)	Academic failure/removal from gifted programs	Standardised ability tests	School grades	n/a
Baslanti & McCoach, 2006	Absolute split II	GPA below 2.0	Student Selection and Placement Examination	university grades	17
Baslanti, 2008	Absolute split II	GPA below 2.0	Student Selection and Placement Examination	university grades	17
Bergner & Neubauer, 2011	Simple Difference	1.1 standard deviations	IQ	school grades	45
Berkowitz & Cicchelli, 2004	Absolute split II	Grades below 85%	English Language Arts test	report grades	26
Cavilla, 2015	Nomination	profile			n/a
Colangelo, Kerr, Christensen & Maxey, 1993	Absolute split II	GPA below 2.25	American College Testing Program	GPA	0
Diaz, 1998	Nomination	fit qualitative profile; low grades	previous enrolment in gifted program, or any previous evidence of superior achievement	school grades	n/a

Emerick, 2004	Absolute split I	below average actual achievement	high performance on any indicator	school grades	n/a
Figg, Rogers, McCormicj, & Low, 2012	Absolute split I	below 85th percentile (rank)	OLAST	General Achievement Test (standardised achievement test); academic ranking in the grade	23
Flint, 2002, 2009	Nomination	self-identified			n/a
Grantham & Ford, 1998	Absolute split I & II	GPA at average to low grades	Iowa Test of Basic Skills	GPA	n/a
Grobman, 2006	Nomination	referrals to private psychiatrist			n/a
Guldemond, Bosker, Kuyper & van der Werf, 2007	Absolute split I	below average	Groningen Intelligence test	school grades	15
Hanses & Rost, 1998	Absolute split I	achievement < average	IQ	school grades	n/a
Hébert & Olenchak, 2000	Absolute split II	Grades of C or lower	enrolled in gifted programs previously	school grades	n/a
Hébert, 2001	Combination of nomination, absolute split II, other factors	GPA below 2.0, drop out, subject choice	past school acheivement (non-standardised) or intelligence tests		n/a
Jovanović, Teovanović, Mentus, & Petrović, 2010	Absolute split I	bottom 10% of gifted students	ability tests	school grades	24
Ki-Soon & Marvin, 2000	Nomination	profile			50
Lau & Chan, 2001a	Absolute split I, simple difference, regression, nomination	below average; one standard deviation; one standard error	Ravens	school grades	6; 21; 22; 12

Lau & Chan, 2001b	Simple difference	One standard deviation	Ravens, verbal ability	school grades	21
Lee-Corbin & Evans, 1996	Nomination	Teacher ratings	Ravens		48
Lupart & Pyryt, 1996	Regression	one standard error	IQ	GPA	n/a
Matthews & McBee, 2007	Absolute split II	GPA below 3.49	SAT or ACT (standardised school achievement)	GPA	9
McCoach & Siegle, 2003a	Absolute split I & II	GPA below 2.5, or bottom half	IQ	GPA	32
McCoach & Siegle, 2003b	Absolute splitI & II	GPA below 2.5, or bottom half	IQ	GPA	32
Muir-Broaddus, 1995	Combination of absolute split II and nomination	GPA below 3.0	IQ	GPA	n/a
Neumeister & Hébert, 2003	Nomination	fit qualitative profile	n/a	n/a	n/a
Obergriesser & Stoeger, 2015	Simple Difference	one standard deviation	IQ	GPA	28
Perleth & Heller, 1994	Simple Difference	1.4 standard deviations	IQ	school grades	
Peters & van Boxtel, 1999	Regression	1.96 standard errors	Ravens	school grades	6
Peterson & Colangelo, 1996	absolute split II	GPA below 3.35	WISC, OLSAT or SAT	GPA	32
Peterson, 2000	absolute split II	GPA below 3.35	WISC, OLSAT or SAT	GPA	
Peterson, 2001	Nomination	self-identified			
Rafidi, 2008	Nomination	profile			n/a
Rayneri, Gerber & Wiley, 2003	Absolute split II	GPA below 85%	enrolled in gifted programs	GPA	20
Rayneri, Gerber & Wiley, 2006	Absolute split II	GPA below 80%	enrolled in gifted programs	GPA	20

Redding, 1990	Regression	One standard error	WISC	GPA	n/a
Reis, Colbert & Hébert, 2005	Absolute split II	GPA below 2.20; educational pathway	high performance on any indicator	GPA	49
Schultz, 2002	Absolute split II + nomination	GPA below 2.75	IQ	GPA	n/a
Sikora & Saha, 2011	Absolute split (alt)	fail to achieve educational plans	academic achievement	educational pattern	15
Staudt & Neubauer, 2006	Absolute split I	below median score	IQ	school grades	47
Stoeger & Ziegler, 2005	Simple Difference	one standard deviation	Ravens	school grades	n/a
Stoeger, Suggate, & Ziegler, 2013	Simple Difference	one standard deviation	Culture Fair Intelligence Test	school grades	49
Thompson & McDonald, 2007	Nomination	profile			32
Timmermans, van Lier, & Koot, 2009	Absolute split (alt)	educational pathway			
Van Boxtel & Mönks, 1992	Regression	statistically significant difference (5% level)	Intelligence Structure Test	GPA	28
Vlahovic-Stetic, Vlasta & Arambasic, 1999	Absolute split I	below average actual achievement	Ravens, PRONAD, PROFNAD, B-Serija	mathematics test	50
Wood, 2003	Nomination	profile			
Ziegler & Stoeger, 2010	Combination of absolute split II and nomination	GPA below 3.0	IQ	GPA	n/a

Appendix 2 Human Research Ethics Approval HC 13060

HUMAN RESEARCH ETHICS COMMITTEE (HREC)

24-Apr-2013

Dr Jae Jung

Sydney NSW 2052

Dear Dr Jung,

HREC Ref: **# HC13060**

Predicting the Underachievement of Gifted Students

The Human Research Ethics Committee considered the above protocol at its meeting held on 23-Apr-2013 and is pleased to advise it is satisfied that this protocol meets the requirements as set out in the National Statement on Ethical Conduct in Human Research*. Having taken into account the advice of the Committee, the Deputy Vice-Chancellor (Research) has approved the project to proceed.

Would you please note:-

- approval is valid from 23-Apr-2013 to 22-Apr-2018;
- you will be required to provide annual reports on the studys progress to the HREC, as recommended by the National Statement;
- you are required to immediately report to the Ethics Secretariat anything which might warrant review of ethical approval of the protocol (National Statement 3.3.22, 5.5.7: http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/e72.pdf) including:
    - serious or unexpected outcomes experienced by research participants (using the Serious Adverse Event proforma on the University website at http://research.unsw.edu.au/human-ethics-forms-and-proformas ;
    - proposed changes in the protocol; and

    - unforeseen events or new information (eg. from other studies) that might affect continued ethical acceptability of the project or may indicate the need for amendments to the protocol;
- any modifications to the project must have prior written approval and be ratified by any other relevant Human Research Ethics Committee, as appropriate;
- if there are implantable devices, the researcher must establish a system for tracking the participants with implantable devices for the lifetime of the device (with consent) and report any device incidents to the TGA;
- if the research project is discontinued before the expected date of completion, the researcher is required to inform the HREC and other relevant institutions (and where possible, research participants), giving reasons. For multi-site research, or where there has been multiple ethical review, the researcher must advise how this will be communicated before the research begins (National Statement 3.3.22, 5.5.7: http://www.nhmrc.gov.au/_files_nhmrc/publications/attachments/e72.pdf);
- consent forms are to be retained within the archives of the EDUC - School of Education and made available to the Committee upon request.

Sincerely,

Michael Grimm
Presiding Member
Human Research Ethics Committee

* http://www.nhmrc.gov.au

Appendix 3 Principal's Letter Granting Approval For Research

[redacted]

To whom it may concern,

Rahmi Jackson is currently employed as a full-time high school teacher at [redacted]
We have given Rahmi permission to access and analyse our school records including (but not limited to): student reports, academic testing results, health related records, behavioural records, student backgrounds etc.

We request the following conditions for the use of this data.

- All data is to remain the property of [redacted]
- Rahmi has permission to access and make copies of this data for his research at UNSW. Copies of data should only be kept for the minimum time period required by the research.
- All identifiable data is to remain on the physical site of [redacted]
- Identifiable data is not to be made public.

Kind regards,

[redacted]
School Principal

Appendix 4 Human Research Ethics Approval HC 15176

Human Research Ethics Advisory (HREA) Panel B: Arts, Humanities and Law
The University of New South Wales
UNSW Sydney, NSW, Australia, 2052
E: HREApanelB@unsw.edu.au

31-May-2015

Dear Jae Yup Jung,

Project Title	Validation of Methods to Identify and Measure the Underachievement of Gifted Students
HC No	HC15176
Re	Notification of Ethics Approval
Approval Period	20-May-2015 - 19-May-2020

Thank you for submitting the above research project to the **HREAP B: Arts, Humanities & Law** for ethical review. This project was considered by the **HREAP B: Arts, Humanities & Law** at its meeting on 20-May-2015.

I am pleased to advise you that the **HREAP B: Arts, Humanities & Law** has granted ethical approval of this research project, subject to the following conditions being met:

Conditions of Approval Specific to Project:
N/A

Conditions of Approval – All Projects:

- The Chief Investigator will immediately report anything that might warrant review of ethical approval of the project.
- The Chief Investigator will notify the **HREAP B: Arts, Humanities & Law** of any event that requires a modification to the protocol or other project documents and submit any required amendments in accordance with the instructions provided by the **HREAP B: Arts, Humanities & Law**. These instructions can be found at https://research.unsw.edu.au/research-ethics-and-compliance-support-recs.
- The Chief Investigator will submit any necessary reports related to the safety of research participants in accordance with **HREAP B: Arts, Humanities & Law** policy and procedures. These instructions can be found at https://research.unsw.edu.au/research-ethics-and-compliance-support-recs.
- The Chief Investigator will report to the **HREAP B: Arts, Humanities & Law** annually in the specified format and notify the HREC when the project is completed at all sites.
- The Chief Investigator will notify the **HREAP B: Arts, Humanities & Law** if the project is discontinued at a participating site before the expected completion date, with reasons provided.
- The Chief Investigator will notify the **HREAP B: Arts, Humanities & Law** of any plan to extend the duration of the project past the approval period listed above and will submit any associated required documentation. Iinstructions for obtaining an extension of approval can be found at

https://research.unsw.edu.au/research-ethics-and-compliance-support-recs.

- The Chief Investigator will notify the **HREAP B: Arts, Humanities & Law** of his or her inability to continue as Coordinating Chief Investigator including the name of and contact information for a replacement.

A copy of this ethical approval letter must be submitted to all Investigators and sites prior to commencing the project.

The **HREAP B: Arts, Humanities & Law** Terms of Reference, Standard Operating Procedures, membership and standard forms are available from https://research.unsw.edu.au/research-ethics-and-compliance-support-recs.

Should you require any further information, please contact the Ethics Administrator at:

E: HREApanelB@unsw.edu.au
W:https://research.unsw.edu.au/human-research-ethics-home

The HREAP B: Arts, Humanities & Law wishes you every continued success in your research.

Kind Regards

Professor Colin Evers
Convenor HREA Panel B: Arts, Humanities and Law

Appendix 5 Principal's Letter Granting Approval for Nomination Survey

School Logo & contact details redacted

16th March 2015

To whom it may concern

Rahmi Jackson is currently employed as a full-time high school teacher. We have given Rahmi permission to present to staff on his research topic and to invite staff to voluntarily participate in his research by completing an electronic survey.

Yours sincerely

Principal

Appendix 6 Participant Information Statement

ONLINE PARTICIPANT INFORMATION STATEMENT
Teachers
Validation of Methods to Identify and Measure the Underachievement of Gifted Students
Dr Jae Jung

The research study is being carried out by the following researchers:		
Role	**Name**	**Organisation**
Chief Investigator	Dr Jae Jung	UNSW
Co-Investigator/s		
Student Investigator/s	Mr Rahmi Jackson is conducting this study as part of the requirements to complete a PhD at The University of New South Wales. This will take place under the supervision of Dr Jae Jung.	UNSW [redacted]
Research Funder	This research is not funded.	

What is the research study about?
You are invited to take part in an online research study. You have been invited because you will be teaching at least one class a[redacted] over the period of January to July 2015.

The research study is aiming to compare the methods of identifying gifted underachievement to determine whether the available methods are equivalent, and if not, which method may be considered the 'best' available method.

Do I have to take part in this research study?
This Participant Information Statement tells you about the research study and the research tasks involved. Knowing what is involved will help you decide if you wish to take part.

Please read this information carefully. Before deciding whether or not to take part, you might wish to discuss it with a relative or friend.

Participation in this research is voluntary. If you do not wish to take part, you do not have to. Your decision will not affect your relationship with UNSW or [redacted]

What does participation in this research require, and are there any risks involved?
If you decide to take part in the research study, you will be asked to complete an online questionnaire. The questionnaire will ask you questions about the achievement of students in your classes and your perception of whether they could perform at a significantly higher level. We expect this activity to take up to 30 minutes.

Will I be paid to participate in this project?
There are no costs associated with participating in this research study, nor will you be paid.

What are the possible benefits to participation?
We hope to use information we obtain from this research study to benefit underachieving gifted students, researchers of gifted underachievement, teachers and school administrators.

What will happen to information about me?
By consenting to the research, you agree to allow the research team to collect and use information from the questionnaire that you fill out. We will keep your data in a safe and secure location within the School of Education at UNSW for a period of 7 years. Thereafter, the information will be destroyed according to the standard procedures for destroying confidential documents.

ONLINE PARTICIPANT INFORMATION STATEMENT
Teachers
Validation of Methods to Identify and Measure the Underachievement of Gifted Students
Dr Jae Jung

It is anticipated that the results of this research study will be published and/or presented in a variety of forums. In any publication and/or presentation, information will be presented in such a way that you will not be individually identifiable.

How and when will I find out what the results of the research study are?
The findings of the research will be accessible to participants in a future school newsletter article, after the completion of the study.

What if I want to withdraw from the research study?
The submission of your completed questionnaire will be treated as an indication of your consent to participate in the study. You may elect to withdraw your responses at any time prior to submission of the questionnaire. Furthermore, you may withdraw your responses after the submission of the questionnaire by advising Rahmi Jackson (telephone [redacted] or email: jacksonr@[redacted]sw.edu.au).

What should I do if I have further questions about my involvement in the research study?
If you require any further information concerning this project, or if you have any problems which may be related to your involvement in the project, you may contact Rahmi Jackson (telephone [redacted] or email: jacksonr@[redacted]sw.edu.au).

What if I have a complaint or any concerns about the research study?
If you have any complaints about any aspect of the project, or the way it is being conducted, then you may contact:

Complaints Contact

Position	Human Research Ethics Coordinator
Telephone	+ 61 2 9385 6222
Email	humanethics@unsw.edu.au
HC Reference Number	HC15176

UNSW AUSTRALIA

ONLINE PARTICIPANT INFORMATION STATEMENT
Teachers
Validation of Methods to Identify and Measure the Underachievement of Gifted Students
Dr Jae Jung

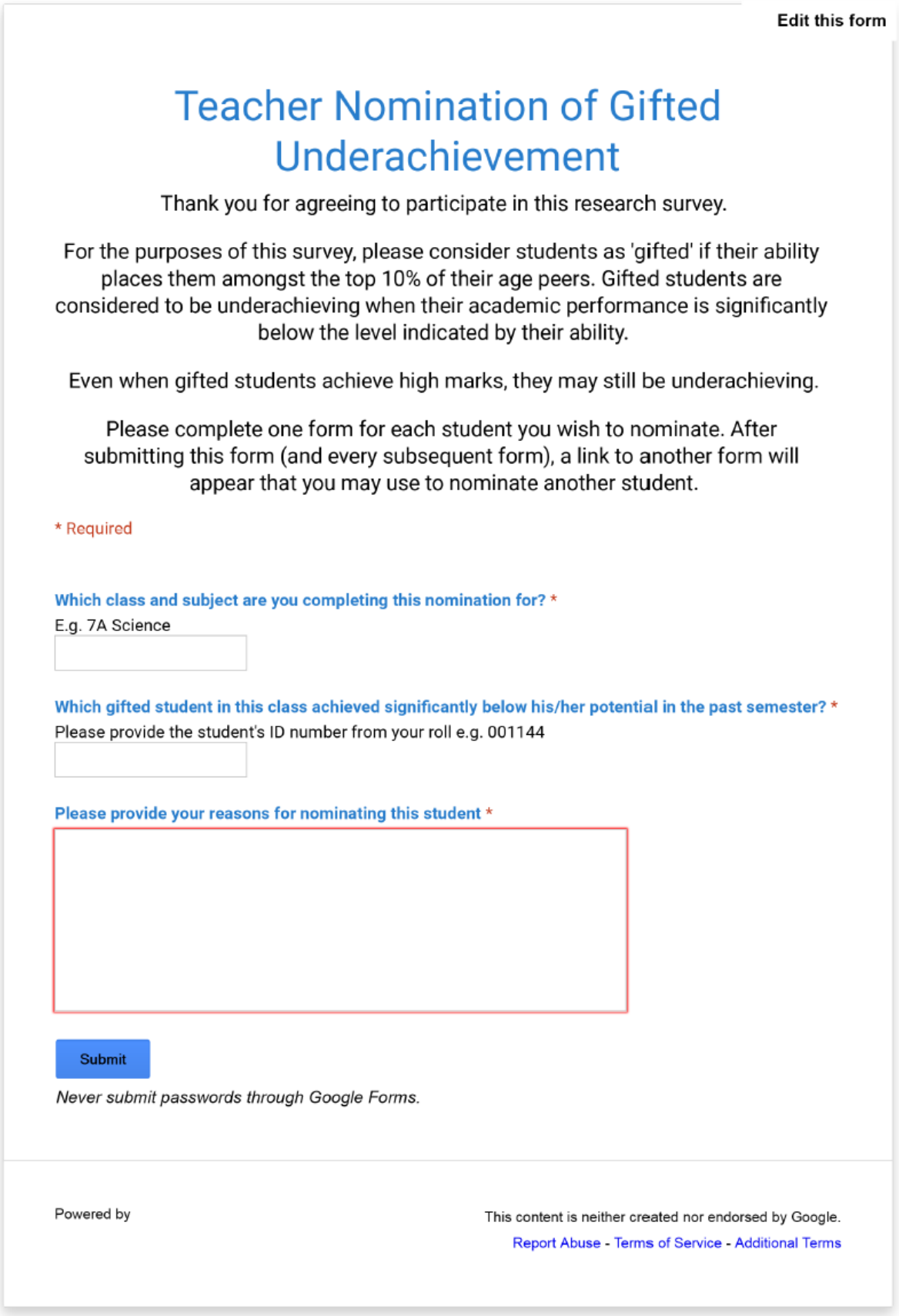
Consent Form – Participant providing own consent

Declaration by the participant

- ☐ I have read the Participant Information Sheet;
- ☐ I understand the purposes, study tasks and risks of the research described in the project;
- ☐ I have had an opportunity to ask questions and I am satisfied with the answers I have received;
- ☐ I freely agree to participate in this research study as described and understand that I am free to withdraw at any time during the project and withdrawal will not affect my relationship with any of the named organisations and/or research team members;
- ☐ I understand that I can download a copy of this consent form from here.

I agree, start questionnaire

Appendix 7 Teacher Nomination Survey



Edit this form

Teacher Nomination of Gifted Underachievement

Thank you for agreeing to participate in this research survey.

For the purposes of this survey, please consider students as 'gifted' if their ability places them amongst the top 10% of their age peers. Gifted students are considered to be underachieving when their academic performance is significantly below the level indicated by their ability.

Even when gifted students achieve high marks, they may still be underachieving.

Please complete one form for each student you wish to nominate. After submitting this form (and every subsequent form), a link to another form will appear that you may use to nominate another student.

* Required

Which class and subject are you completing this nomination for? *

E.g. 7A Science

Which gifted student in this class achieved significantly below his/her potential in the past semester? *

Please provide the student's ID number from your roll e.g. 001144

Please provide your reasons for nominating this student *

Submit

Never submit passwords through Google Forms.

Powered by

This content is neither created nor endorsed by Google.

Report Abuse - Terms of Service - Additional Terms

Appendix 8 Email Invitation

Dear staff member

"Research project: The underachievement of gifted students"

My name is Jae Yup Jared Jung, and I am a senior lecturer in the School of Education at The University of New South Wales. I am writing to invite you to participate in research on the identification of gifted underachievers, which will require you to complete an online nomination form for each underachieving gifted student that you teach at[redacted].

On each nomination form, you will be asked to provide details of the class and identification number of each nominated student. This information will enable the linking of each nomination form to the nominated student's school/class assessment results.

The following is a link to an information statement about the research (the first two pages) and to an online nomination form (the third page; please use as many online nomination forms as necessary):

https://www.dropbox.com/s/1cd8vdon2r2yrmr/Participant%20Information%20Statement.docx?dl=0

(please copy and paste the URL into your internet browser). **Please do not forward this invitation to anyone else, as only the recipients of this email are eligible for participation.** Please complete all nomination forms by 30 July 2015.

If you have any questions, please do not hesitate to contact me (email: jae.jung@unsw.edu.au; phone: 9385 8629; mobile phone:). This research has received ethics approval (approval no. HC15176) from HREA Panel B at The University of New South Wales.

Best regards,

Jae Yup Jared Jung, PhD

Appendix 9 Contingency Tables

OLSAT - NAPLAN

Table 37

Contingency table for OLSAT verbal and NAPLAN Literacy data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	44	0	44		
method	GA	21	263	284		
Total		65	263	328	206	0.79
Absolute	GUA	51	31	82		
Split method	GA	14	232	246		
Total		65	263	328	124	0.61
		Regression method				
Absolute	GUA	42	40	82		
Split method	GA	2	244	246		
Total		44	284	328	135	0.64

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Table 38

Contingency table for OLSAT Non-verbal and NAPLAN Numeracy data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	62	1	63		
method	GA	30	403	433		
Total		92	404	496	305	0.78
Absolute	GUA	49	21	70		
Split method	GA	43	383	426		
Total		92	404	496	143	0.54
		Regression method				
Absolute	GUA	49	21	70		
Split method	GA	14	412	426		
Total		63	433	496	241	0.70

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Table 39

Contingency table for OLSAT School Ability Index and NAPLAN Literacy data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	63	0	63		
method	GA	46	281	327		
Total		109	281	390	194	0.70
Absolute	GUA	80	27	107		
Split method	GA	29	254	283		
Total		109	281	390	161	0.64
		Regression method				
Absolute	GUA	62	45	107		
Split method	GA	1	282	283		
Total		63	327	390	190	0.70

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Table 40

Contingency table for OLSAT School Ability Index and NAPLAN Numeracy data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	43	7	50		
method	GA	0	337	337		
Total		43	344	387	326	0.92
Absolute	GUA	27	9	36		
Split method	GA	16	335	351		
Total		43	344	387	164	0.65
		Regression method				
Absolute	GUA	32	4	36		
Split method	GA	18	333	351		
Total		50	337	387	204	0.73

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

OLSAT and SC

Table 41

Contingency table for OLSAT Verbal and School Certificate English data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	1	1	2		
method	GA	2	19	21		
Total		3	20	23	2.64	0.34
Absolute Split	GUA	1	1	2		
method 1	GA	2	19	21		
Total		3	20	23	2.64	0.34
Absolute Split	GUA	1	0	1		
method 2	GA	2	20	22		
Total		3	20	23	7	0.55
		Regression method				
Absolute Split	GUA	2	0	2		
method 1	GA	0	21	21		
Total		2	21	23	23	1.00
Absolute Split	GUA	1	0	1		
method 2	GA	1	21	22		
Total		2	21	23	11	0.69
		Absolute split method 1				
Absolute Split	GUA	1	0	1		
method 2	GA	1	21	22		
Total		2	21	23	11	0.69

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Table 42

Contingency table for OLSAT Non-verbal and School Certificate Mathematics data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	4	1	5		
method	GA	4	37	41		
Total		8	38	46	15.3	0.58
Absolute Split	GUA	4	3	7		
method 1	GA	4	35	39		
Total		8	38	46	9.08	0.44
Absolute Split	GUA	4	4	8		
method 2	GA	4	34	38		
Total		8	38	46	7.2	0.39
		Regression method				
Absolute Split	GUA	5	2	7		
method 1	GA	0	39	39		
Total		5	41	46	31	0.82
Absolute Split	GUA	5	3	8		
method 2	GA	0	38	38		
Total		5	41	46	27	0.76
		Absolute split method 1				
Absolute Split	GUA	7	1	8		
method 2	GA	0	38	38		
Total		7	39	46	39	0.92

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Table 43

Contingency table for OLSAT School Ability Index and School Certificate English data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	5	0	5		
method	GA	1	33	34		
Total		6	33	39	31.5	0.90
Absolute Split	GUA	5	4	9		
method 1	GA	1	29	30		
Total		6	33	39	14.5	0.61
Absolute Split	GUA	4	0	4		
method 2	GA	2	33	35		
Total		6	33	39	24.5	0.79
		Regression method				
Absolute Split	GUA	5	4	9		
method 1	GA	0	30	30		
Total		5	34	39	19.1	0.70
Absolute Split	GUA	4	0	4		
method 2	GA	1	34	35		
Total		5	34	39	30	0.88
		Absolute split method 1				
Absolute Split	GUA	4	0	4		
method 2	GA	5	30	35		
Total		9	30	39	14.9	0.62

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Table 44

Contingency table for OLSAT School Ability Index and School Certificate Mathematics data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	4	1	5		
method	GA	0	34	34		
Total		4	35	39	30.3	0.88
Absolute Split	GUA	4	1	5		
method 1	GA	0	34	34		
Total		4	35	39	30.3	0.88
Absolute Split	GUA	4	3	7		
method 2	GA	0	32	32		
Total		4	35	39	20.4	0.72
		Regression method				
Absolute Split	GUA	5	0	5		
method 1	GA	0	34	34		
Total		5	34	39	39	1.00
Absolute Split	GUA	5	2	7		
method 2	GA	0	32	32		
Total		5	34	39	26.2	0.82
		Absolute split method 1				
Absolute Split	GUA	5	2	7		
method 2	GA	0	32	32		
Total		5	34	39	26.2	0.82

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

OLSAT - HSC

Table 45

Contingency table for OLSAT School Ability Index and Higher School Certificate English data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	9	0	9		
method	GA	36	10	46		
Total		45	10	55	2	0.21
Absolute Split	GUA	44	6	50		
method 1	GA	1	4	5		
Total		45	10	55	14	0.51
Absolute Split	GUA	33	0	33		
method 2	GA	12	10	22		
Total		45	10	55	18	0.58
		Regression method				
Absolute Split	GUA	9	41	50		
method 1	GA	0	5	5		
Total		9	46	55	1	0.14
Absolute Split	GUA	9	24	33		
method 2	GA	0	22	22		
Total		9	46	55	7	0.36
		Absolute split method 1				
Absolute Split	GUA	33	0	33		
method 2	GA	17	5	22		
Total		50	5	55	8	0.39

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Table 46

Contingency table for OLSAT School Ability Index and Higher School Certificate Mathematics data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	3	0	3		
method	GA	18	3	21		
Total		21	3	24	0.49	0.14
Absolute Split	GUA	21	0	21		
method 1	GA	0	3	3		
Total		21	3	24	24	1.00
Absolute Split	GUA	14	0	14		
method 2	GA	7	3	10		
Total		21	3	24	5	0.45
		Regression method				
Absolute Split	GUA	3	18	21		
method 1	GA	0	3	3		
Total		3	21	24	0.49	0.14
Absolute Split	GUA	3	11	14		
method 2	GA	0	10	10		
Total		3	21	24	2	0.32
		Absolute split method 1				
Absolute Split	GUA	14	0	14		
method 2	GA	7	3	10		
Total		21	3	24	5	0.45

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Table 47

Contingency for OLSAT Verbal score and Higher School Certificate English data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	7	0	7		
method	GA	22	7	29		
Total		29	7	36	2	0.24
Absolute Split	GUA	28	3	31		
method 1	GA	1	4	5		
Total		29	7	36	14	0.61
Absolute Split	GUA	20	0	20		
method 2	GA	9	7	16		
Total		29	7	36	11	0.55
		Regression method				
Absolute Split	GUA	7	24	31		
method 1	GA	0	5	5		
Total		7	29	36	1	0.20
Absolute Split	GUA	7	13	20		
method 2	GA	0	16	16		
Total		7	29	36	7	0.44
		Absolute split method 1				
Absolute Split	GUA	20	0	20		
method 2	GA	11	5	16		
Total		31	5	36	7	0.45

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Table 48

Contingency table for OLSAT Non-Verbal score and Higher School Certificate Mathematics data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	2	0	2		
method	GA	26	1	27		
Total		28	1	29	0.08	0.05
Absolute Split	GUA	26	0	26		
method 1	GA	2	1	3		
Total		28	1	29	9	0.56
Absolute Split	GUA	16	0	16		
method 2	GA	12	1	13		
Total		28	1	29	1	0.21
		Regression method				
Absolute Split	GUA	2	24	26		
method 1	GA	0	3	3		
Total		2	27	29	0.25	0.09
Absolute Split	GUA	2	14	16		
method 2	GA	0	13	13		
Total		2	27	29	2	0.25
		Absolute split method 1				
Absolute Split	GUA	16	0	16		
method 2	GA	10	3	13		
Total		26	3	29	4	0.38

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

NAPLAN - SC

Table 49

Contingency table for NAPLAN Literacy score and School Certificate English data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	5	2	7		
method	GA	0	32	32		
Total		5	34	39	26	0.82*
Absolute Split	GUA	4	0	4		
method 1	GA	1	34	35		
Total		5	34	39	30	0.88*
Absolute Split	GUA	2	0	2		
method 2	GA	3	34	37		
Total		5	34	39	14	0.61#
		Regression method				
Absolute Split	GUA	4	0	4		
method 1	GA	3	32	35		
Total		4	32	39	20	0.72*
Absolute Split	GUA	2	0	2		
method 2	GA	5	32	37		
Total		7	32	39	9.6	0.50^
		Absolute split method 1				
Absolute Split	GUA	2	0	2		
method 2	GA	2	35	37		
Total		4	35	39	18	0.69*

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Table 50

Contingency table for NAPLAN Numeracy score and School Certificate Mathematics data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	7	0	7		
method	GA	1	68	69		
Total		8	68	76	66	0.93*
Absolute Split	GUA	3	1	4		
method 1	GA	5	67	72		
Total		8	68	76	19	0.50*
Absolute Split	GUA	3	1	4		
method 2	GA	5	67	72		
Total		8	68	76	19	0.50*
		Regression method				
Absolute Split	GUA	3	1	4		
method 1	GA	4	68	72		
Total		7	69	76	22	0.54*
Absolute Split	GUA	3	1	4		
method 2	GA	4	68	72		
Total		7	69	76	22	0.54*
		Absolute split method 1				
Absolute Split	GUA	4	0	4		
method 2	GA	0	72	72		
Total		4	72	76	76	1.00*

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

NAPLAN - HSC

Table 51

Contingency table for NAPLAN Literacy score and Higher School Certificate English data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	4	0	4		
method	GA	31	6	37		
Total		35	6	41	0.76	0.14
Absolute Split	GUA	33	0	33		
method 1	GA	2	6	8		
Total		35	6	41	29	0.84
Absolute Split	GUA	18	0	18		
method 2	GA	17	6	23		
Total		35	6	41	6	0.37
		Regression method				
Absolute Split	GUA	4	29	33		
method 1	GA	0	8	8		
Total		4	37	41	1	0.16
Absolute Split	GUA	4	14	18		
method 2	GA	0	23	23		
Total		4	37	41	6	0.37
		Absolute split method 1				
Absolute Split	GUA	18	0	18		
method 2	GA	15	8	23		
Total		33	8	41	8	0.44

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Table 52

Contingency table for NAPLAN Numeracy score and Higher School Certificate Mathematics data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	6	0	6		
method	GA	36	3	39		
Total		42	3	45	0.49	0.10
Absolute Split	GUA	39	2	41		
method 1	GA	3	1	4		
Total		42	3	45	2	0.23
Absolute Split	GUA	28	0	28		
method 2	GA	14	3	17		
Total		42	3	45	5	0.34
		Regression method				
Absolute Split	GUA	6	35	41		
method 1	GA	0	4	4		
Total		6	39	45	0.68	0.12
Absolute Split	GUA	6	22	28		
method 2	GA	0	17	17		
Total		6	39	45	4	0.31
		Absolute split method 1				
Absolute Split	GUA	28	0	28		
method 2	GA	13	4	17		
Total		41	4	45	7	0.40

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

SC - HSC

Table 53

Contingency table for School Certificate English and Higher School Certificate English data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	12	0	12		
method	GA	35	22	47		
Total		47	22	69	7	0.31
Absolute Split	GUA	46	4	50		
method 1	GA	1	18	19		
Total		47	22	69	48	0.83
Absolute Split	GUA	30	0	30		
method 2	GA	17	22	39		
Total		47	22	69	25	0.60
		Regression method				
Absolute Split	GUA	12	38	50		
method 1	GA	0	19	19		
Total		12	47	69	6	0.28
Absolute Split	GUA	12	18	30		
method 2	GA	0	39	39		
Total		12	47	69	19	0.52
		Absolute split method 1				
Absolute Split	GUA	30	0	30		
method 2	GA	20	19	39		
Total		50	19	69	20	0.54

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Table 54

Contingency table for School Certificate Mathematics and Higher School Certificate Mathematics data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	14	0	14		
method	GA	82	10	92		
Total		96	10	106	1.68	0.13
Absolute Split	GUA	90	2	92		
method 1	GA	6	8	14		
Total		96	10	106	43	0.64
Absolute Split	GUA	65	0	65		
method 2	GA	31	10	41		
Total		96	10	106	18	0.41
		Regression method				
Absolute Split	GUA	14	78	92		
method 1	GA	0	14	14		
Total		14	92	106	2.45	0.15
Absolute Split	GUA	14	51	65		
method 2	GA	0	41	41		
Total		14	92	106	10	0.31
		Absolute split method 1				
Absolute Split	GUA	65	0	65		
method 2	GA	27	14	41		
Total		92	14	106	26	0.49

Note. GUA = Gifted underachievement; GA = Gifted achievement.

$^{*}p<0.05$

OLSAT - Junior SA

Table 55

Contingency table for OLSAT School Ability Index and Average Junior English achievement data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	21	0	21		
method	GA	38	99	137		
Total		59	99	158	41	0.51*
Absolute Split	GUA	39	11	50		
method 1	GA	20	88	108		
Total		59	99	158	52	0.57*
Absolute Split	GUA	59	61	120		
method 2	GA	0	38	38		
Total		59	99	158	30	0.43*
		Regression method				
Absolute Split	GUA	21	29	50		
method 1	GA	0	108	108		
Total		21	147	158	52	0.58*
Absolute Split	GUA	21	99	120		
method 2	GA	0	38	38		
Total		21	147	158	8	0.22
		Absolute split method 1				
Absolute Split	GUA	50	70	120		
method 2	GA	0	38	38		
Total		50	108	158	23	0.38*

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Table 56

Contingency table for OLSAT School Ability Index and Average Junior Mathematics achievement data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	24	0	24		
method	GA	34	100	134		
Total		58	100	158	49	0.56*
Absolute Split	GUA	47	11	58		
method 1	GA	11	89	100		
Total		58	100	158	78	0.70*
Absolute Split	GUA	44	12	56		
method 2	GA	14	88	102		
Total		58	100	158	65	0.64*
		Regression method				
Absolute Split	GUA	24	34	58		
method 1	GA	0	100	100		
Total		24	134	158	49	0.56*
Absolute Split	GUA	23	33	56		
method 2	GA	1	101	102		
Total		24	134	158	45	0.53*
		Absolute split method 1				
Absolute Split	GUA	54	2	56		
method 2	GA	4	98	102		
Total		58	100	158	133	0.92*

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Table 57

Contingency table for OLSAT Verbal Score and Average Junior English achievement data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	14	0	14		
method	GA	21	88	109		
Total		35	88	123	40	0.57*
Absolute Split	GUA	23	8	31		
method 1	GA	12	80	92		
Total		35	88	123	43	0.59*
Absolute Split	GUA	34	52	86		
method 2	GA	1	36	37		
Total		35	88	123	17	0.37*
		Regression method				
Absolute Split	GUA	14	17	31		
method 1	GA	0	92	92		
Total		14	109	123	47	0.62*
Absolute Split	GUA	14	72	86		
method 2	GA	0	37	37		
Total		14	109	123	7	0.24
		Absolute split method 1				
Absolute Split	GUA	31	55	86		
method 2	GA	0	37	37		
Total		31	92	123	18	0.38*

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Table 58

Contingency table for OLSAT Non-Verbal Score and Average Junior mathematics achievement data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	29	0	29		
method	GA	65	120	185		
Total		94	120	214	43	0.45*
Absolute Split	GUA	66	20	86		
method 1	GA	28	100	128		
Total		94	120	214	63	0.54*
Absolute Split	GUA	65	22	87		
method 2	GA	29	98	127		
Total		94	120	214	56	0.51*
		Regression method				
Absolute Split	GUA	29	57	86		
method 1	GA	0	128	128		
Total		29	185	214	50	0.48*
Absolute Split	GUA	29	58	87		
method 2	GA	0	127	127		
Total		29	185	214	49	0.48*
		Absolute split method 1				
Absolute Split	GUA	83	4	87		
method 2	GA	3	124	127		
Total		86	128	214	186	0.93*

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

OLSAT - Senior SA

Table 59

Contingency table for OLSAT School Ability Index and Average Senior English achievement data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	12	0	12		
method	GA	34	29	63		
Total		46	29	75	9	0.35*
Absolute Split	GUA	41	4	45		
method 1	GA	5	25	30		
Total		46	29	75	42	0.75*
Absolute Split	GUA	44	18	62		
method 2	GA	2	11	13		
Total		46	29	75	14	0.43*
		Regression method				
Absolute Split	GUA	12	33	45		
method 1	GA	0	30	30		
Total		12	63	75	10	0.36*
Absolute Split	GUA	12	50	62		
method 2	GA	0	13	13		
Total		12	63	75	3	0.20
		Absolute split method 1				
Absolute Split	GUA	45	17	62		
method 2	GA	0	13	13		
Total		45	30	75	24	0.56*

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Table 60

Contingency table for OLSAT School Ability index and Average Senior mathematics achievement data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	12	0	12		
method	GA	31	32	63		
Total		43	32	75	11	0.38*
Absolute Split	GUA	35	3	38		
method 1	GA	8	29	37		
Total		43	32	75	38	0.71*
Absolute Split	GUA	37	10	47		
method 2	GA	6	22	28		
Total		43	32	75	24	056*
		Regression method				
Absolute Split	GUA	12	26	38		
method 1	GA	0	37	37		
Total		12	63	75	14	0.43*
Absolute Split	GUA	12	35	47		
method 2	GA	0	28	28		
Total		12	63	75	9	0.34*
		Absolute split method 1				
Absolute Split	GUA	38	9	47		
method 2	GA	0	28	28		
Total		38	37	75	46	0.78*

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Table 61

Contingency table for OLSAT Verbal Score and Average Senior English achievement data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	7	0	7		
method	GA	22	26	48		
Total		29	26	55	7	0.36
Absolute Split	GUA	27	4	31		
method 1	GA	2	22	24		
Total		29	26	55	34	0.78*
Absolute Split	GUA	28	16	44		
method 2	GA	1	10	11		
Total		29	26	55	11	0.44*
		Regression method				
Absolute Split	GUA	7	24	31		
method 1	GA	0	24	24		
Total		7	48	55	6	0.34
Absolute Split	GUA	7	37	44		
method 2	GA	0	11	11		
Total		7	48	55	2	0.19
		Absolute split method 1				
Absolute Split	GUA	31	13	44		
method 2	GA	0	11	11		
Total		31	24	55	18	0.57*

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Table 62

Contingency table for OLSAT Non-Verbal Score and Average Senior Mathematics achievement data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	16	0	16		
method	GA	39	37	76		
Total		55	37	92	13	0.38*
Absolute Split	GUA	44	5	49		
method 1	GA	11	32	43		
Total		55	37	92	39	0.65*
Absolute Split	GUA	46	17	63		
method 2	GA	9	20	29		
Total		55	37	92	15	0.40*
		Regression method				
Absolute Split	GUA	16	33	49		
method 1	GA	0	43	43		
Total		16	76	92	17	0.43*
Absolute Split	GUA	16	47	63		
method 2	GA	0	29	29		
Total		16	76	92	9	0.31*
		Absolute split method 1				
Absolute Split	GUA	49	14	63		
method 2	GA	0	29	29		
Total		49	43	92	48	0.72*

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

NAPLAN - Junior SA

Table 63

Contingency table for NAPLAN Literacy and Average Junior English achievement data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	20	1	21		
method	GA	2	141	143		
Total		22	142	164	139	0.92*
Absolute Split	GUA	14	10	24		
method 1	GA	8	132	140		
Total		22	142	164	49	0.55*
Absolute Split	GUA	22	71	93		
method 2	GA	0	71	71		
Total		22	142	164	19	0.34*
		Regression method				
Absolute Split	GUA	15	9	24		
method 1	GA	6	134	140		
Total		21	143	164	62	0.62*
Absolute Split	GUA	21	72	93		
method 2	GA	0	71	71		
Total		21	143	164	18	0.33*
		Absolute split method 1				
Absolute Split	GUA	24	69	93		
method 2	GA	0	71	71		
Total		24	140	164	22	0.36*

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Table 64

Contingency table for NAPLAN Numeracy and Average Junior Mathematics achievement data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	51	0	51		
method	GA	96	171	267		
Total		147	171	318	71	0.47*
Absolute Split	GUA	87	17	104		
method 1	GA	60	154	214		
Total		147	171	318	87	0.52*
Absolute Split	GUA	86	21	107		
method 2	GA	61	150	211		
Total		147	171	318	76	0.49*
		Regression method				
Absolute Split	GUA	44	60	104		
method 1	GA	7	207	214		
Total		51	267	318	79	0.50*
Absolute Split	GUA	44	63	107		
method 2	GA	7	204	211		
Total		51	267	318	75	0.49*
		Absolute split method 1				
Absolute Split	GUA	101	6	107		
method 2	GA	3	208	211		
Total		104	214	318	279	0.94*

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

NAPLAN - Senior SA

Table 65

Contingency table for NAPLAN Literacy and Average Senior English achievement data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	6	0	6		
method	GA	24	45	69		
Total		30	45	75	10	0.36*
Absolute Split	GUA	26	9	35		
method 1	GA	4	36	40		
Total		30	45	75	32	0.65*
Absolute Split	GUA	26	25	51		
method 2	GA	4	20	24		
Total		30	45	75	8	0.33*
		Regression method				
Absolute Split	GUA	6	29	35		
method 1	GA	0	40	40		
Total		6	69	75	7	0.32
Absolute Split	GUA	6	45	51		
method 2	GA	0	24	24		
Total		6	69	75	3	0.20
		Absolute split method 1				
Absolute Split	GUA	34	17	51		
method 2	GA	1	23	24		
Total		35	40	75	26	0.58*

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Table 66

Contingency table for NAPLAN Numeracy and Average Senior Mathematics achievement data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	21	0	21		
method	GA	61	49	110		
Total		82	49	131	15	0.34*
Absolute Split	GUA	62	5	67		
method 1	GA	20	44	64		
Total		82	49	131	53	0.63*
Absolute Split	GUA	64	16	80		
method 2	GA	18	33	51		
Total		82	49	131	27	0.45*
		Regression method				
Absolute Split	GUA	21	46	67		
method 1	GA	0	64	64		
Total		21	110	131	24	0.43*
Absolute Split	GUA	21	59	80		
method 2	GA	0	51	51		
Total		21	110	131	16	0.35*
		Absolute split method 1				
Absolute Split	GUA	67	13	80		
method 2	GA	0	51	51		
Total		67	64	131	87	0.82*

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Junior SA - SC

Table 67

Contingency table for Average junior assessment English and School Certificate English achievement data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	0	8	8		
method	GA	0	67	67		
Total		0	75	75	NA	NA
Absolute Split	GUA	0	4	4		
method 1	GA	0	71	71		
Total		0	75	75	NA	NA
Absolute Split	GUA	0	1	1		
method 2	GA	0	74	74		
Total		0	75	75	NA	NA
		Regression method				
Absolute Split	GUA	4	0	4		
method 1	GA	4	67	71		
Total		8	67	75	35	0.69*
Absolute Split	GUA	1	0	1		
method 2	GA	7	67	74		
Total		8	67	75	8	0.34*
		Absolute split method 1				
Absolute Split	GUA	1	0	1		
method 2	GA	3	71	74		
Total		4	71	75	18	0.49*

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Table 68

Contingency table for Average junior assessment Mathematics and School Certificate Mathematics achievement data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	1	3	4		
method	GA	0	67	67		
Total		1	70	71	17	0.49*
Absolute Split	GUA	1	1	2		
method 1	GA	0	69	69		
Total		1	70	71	35	0.70*
Absolute Split	GUA	1	3	4		
method 2	GA	0	67	67		
Total		1	70	71	17	0.49*
		Regression method				
Absolute Split	GUA	2	0	2		
method 1	GA	2	67	69		
Total		4	67	71	35	0.70*
Absolute Split	GUA	2	2	4		
method 2	GA	2	65	67		
Total		4	67	71	16	0.47*
		Absolute split method 1				
Absolute Split	GUA	2	2	4		
method 2	GA	0	67	67		
Total		2	69	71	35	0.70*

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

SC - Senior SA

Table 69

Contingency table for School Certificate English and Average Senior English assessment

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	12	0	12		
method	GA	17	47	64		
Total		29	47	76	23	0.55*
Absolute Split	GUA	29	8	37		
method 1	GA	0	39	39		
Total		29	47	76	49	0.81*
Absolute Split	GUA	27	11	38		
method 2	GA	2	36	38		
Total		29	47	76	35	0.68*
		Regression method				
Absolute Split	GUA	12	25	37		
method 1	GA	0	39	39		
Total		12	64	76	15	0.44*
Absolute Split	GUA	12	26	38		
method 2	GA	0	38	38		
Total		12	64	76	14	0.43*
		Absolute split method 1				
Absolute Split	GUA	32	6	38		
method 2	GA	5	33	38		
Total		37	39	76	38	0.71*

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Table 70

Contingency table for School Certificate Mathematics and Average Senior Mathematics assessment

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	36	0	36		
method	GA	51	104	155		
Total		87	104	191	53	0.53*
Absolute Split	GUA	69	20	89		
method 1	GA	18	84	102		
Total		87	104	191	69	0.60*
Absolute Split	GUA	72	51	123		
method 2	GA	15	53	68		
Total		87	104	191	24	0.35*
		Regression method				
Absolute Split	GUA	33	56	89		
method 1	GA	3	99	102		
Total		36	145	191	36	0.44*
Absolute Split	GUA	33	90	123		
method 2	GA	3	65	68		
Total		36	155	191	14	0.27*
		Absolute split method 1				
Absolute Split	GUA	87	36	123		
method 2	GA	2	66	68		
Total		89	102	191	81	0.65*

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Junior SA - HSC

Table 71

Contingency table for Average Junior English assessment and HSC English data

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	6	0	6		
method	GA	43	30	73		
Total		49	30	79	4	0.22
Absolute Split	GUA	44	13	57		
method 1	GA	5	17	22		
Total		49	30	79	20	0.50
Absolute Split	GUA	23	0	23		
method 2	GA	26	30	56		
Total		49	30	79	20	0.50
		Regression method				
Absolute Split	GUA	6	51	57		
method 1	GA	0	22	22		
Total		6	73	76	0.68	0.10
Absolute Split	GUA	6	17	23		
method 2	GA	0	56	56		
Total		6	73	76	16	0.45
		Absolute split method 1				
Absolute Split	GUA	23	0	23		
method 2	GA	34	22	56		
Total		57	22	79	13	0.40

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Table 72

Contingency table for Average Junior Mathematics assessment and HSC Mathematics

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	4	0	4		
method	GA	35	10	45		
Total		39	10	49	1	0.15
Absolute Split	GUA	36	3	39		
method 1	GA	3	7	10		
Total		39	10	49	19	0.62
Absolute Split	GUA	25	0	25		
method 2	GA	14	10	24		
Total		39	10	49	13	0.52
		Regression method				
Absolute Split	GUA	4	35	39		
method 1	GA	0	10	10		
Total		4	45	49	1	0.15
Absolute Split	GUA	4	21	25		
method 2	GA	0	24	24		
Total		4	45	49	4	0.29
		Absolute split method 1				
Absolute Split	GUA	25	0	25		
method 2	GA	14	10	24		
Total		39	10	49	13	0.52

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Senior SA - HSC

Table 73

Contingency table for Average Senior English assessment and HSC English

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	8	0	8		
method	GA	9	29	38		
Total		17	29	46	17	0.60
Absolute Split	GUA	15	2	17		
method 1	GA	2	27	29		
Total		17	29	46	30	0.81
Absolute Split	GUA	5	0	5		
method 2	GA	12	29	41		
Total		17	29	46	10	0.46
		Regression method				
Absolute Split	GUA	8	9	17		
method 1	GA	0	29	29		
Total		8	38	46	17	0.60
Absolute Split	GUA	4	1	5		
method 2	GA	4	37	41		
Total		8	38	46	15	0.58
		Absolute split method 1				
Absolute Split	GUA	5	0	5		
method 2	GA	12	29	41		
Total		17	29	46	10	0.46

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Table 74

Contingency table for Average Senior Mathematics assessment and HSC Mathematics

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	4	0	4		
method	GA	17	18	35		
Total		21	18	39	4	0.31
Absolute Split	GUA	21	3	24		
method 1	GA	0	15	15		
Total		21	18	39	28	0.85
Absolute Split	GUA	12	0	12		
method 2	GA	9	18	27		
Total		21	18	39	15	0.62
		Regression method				
Absolute Split	GUA	4	20	24		
method 1	GA	0	15	15		
Total		4	35	39	3	0.27
Absolute Split	GUA	4	8	12		
method 2	GA	0	27	27		
Total		4	35	39	10	0.51
		Absolute split method 1				
Absolute Split	GUA	12	0	12		
method 2	GA	12	15	27		
Total		24	15	39	11	0.53

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Junior SA - Senior SA

Table 75

Contingency table for Average Junior English assessment and Average Senior English assessment

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	4	0	4		
method	GA	8	77	85		
Total		12	77	89	27	0.55*
Absolute Split	GUA	12	11	23		
method 1	GA	0	66	66		
Total		12	77	89	40	0.67*
Absolute Split	GUA	12	32	44		
method 2	GA	0	45	45		
Total		12	77	89	14	0.40#
		Regression method				
Absolute Split	GUA	4	19	23		
method 1	GA	0	66	66		
Total		4	85	89	12	0.37#
Absolute Split	GUA	4	40	44		
method 2	GA	0	45	45		
Total		4	85	89	4	0.22
		Absolute split method 1				
Absolute Split	GUA	21	23	44		
method 2	GA	2	43	45		
Total		23	66	89	22	0.49*

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

Table 76

Contingency table for Average Junior Mathematics assessment and Average Senior Mathematics assessment

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	4	0	4		
method	GA	7	66	73		
Total		11	66	77	25	0.57*
Absolute Split	GUA	11	6	17		
method 1	GA	0	60	60		
Total		11	66	77	45	0.77*
Absolute Split	GUA	11	25	36		
method 2	GA	0	41	41		
Total		11	66	77	15	0.44#
		Regression method				
Absolute Split	GUA	4	13	17		
method 1	GA	0	60	60		
Total		4	73	77	15	0.44#
Absolute Split	GUA	4	32	36		
method 2	GA	0	41	41		
Total		4	73	77	5	0.25
		Absolute split method 1				
Absolute Split	GUA	17	19	36		
method 2	GA	0	41	41		
Total		17	60	77	25	0.57*

Note. GUA = Gifted underachievement; GA = Gifted achievement.

*$p<0.05$

OLSAT-SA + Nomination

Table 77

Contingency table for OLSAT School Ability Index and Average School Assessment for Semester 1 2015

		GUA	GA	Total	Chi-square	Phi
		Simple Difference method				
Regression	GUA	46	0	46		
method	GA	70	88	158		
Total		116	88	204	45	0.47
Absolute Split	GUA	99	7	106		
method 1	GA	17	81	98		
Total		116	88	204	120	0.77
Absolute Split	GUA	99	17	116		
method 2	GA	17	71	88		
Total		116	88	204	89	0.66
Nomination	GUA	26	96	122		
Method	GA	90	78	168		
Total		116	174	290	31	0.33
		Regression method				
Absolute Split	GUA	46	60	106		
method 1	GA	0	98	98		
Total		46	158	204	55	0.52
Absolute Split	GUA	46	70	116		
method 2	GA	0	88	88		
Total		46	158	204	45	0.47
Nomination	GUA	9	113	122		
Method	GA	37	131	168		
Total		46	244	290	11	0.20
		Absolute split method 1				
Absolute Split	GUA	98	18	116		
method 2	GA	8	80	88		
Total		106	98	204	114	0.75
Nomination	GUA	26	96	122		
Method	GA	80	88	168		
Total		106	184	290	21	0.27
		Absolute split method 2				
Nomination	GUA	26	96	122		
Method	GA	90	78	168		
Total		116	174	290	31	0.33

Note. GUA = Gifted underachievement; GA = Gifted achievement.
*$p<0.05$

Appendix 10 Bland–Altman Plots

OLSAT-NAPLAN:

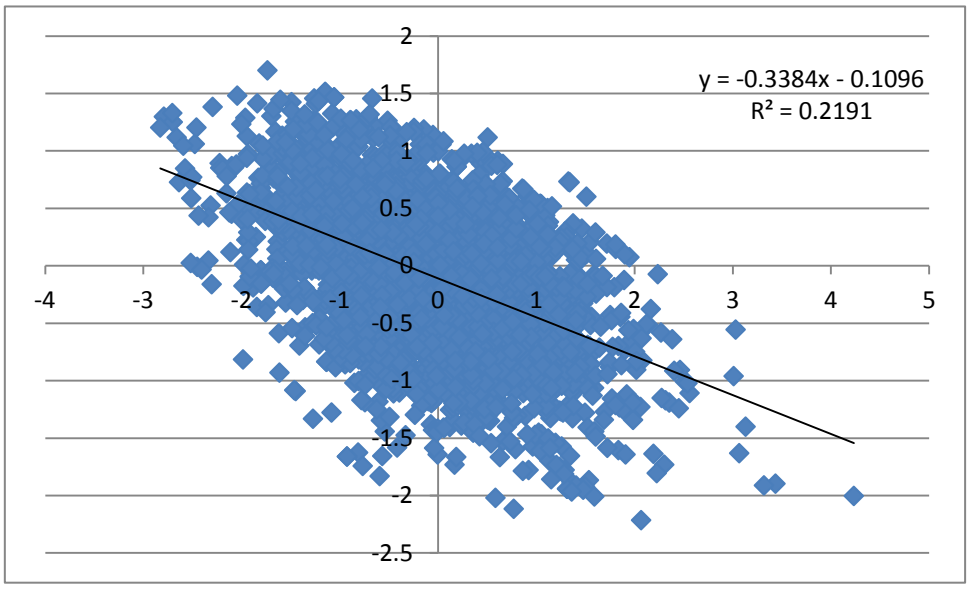


Figure 37. SAI-LIT Bland-Altman plot

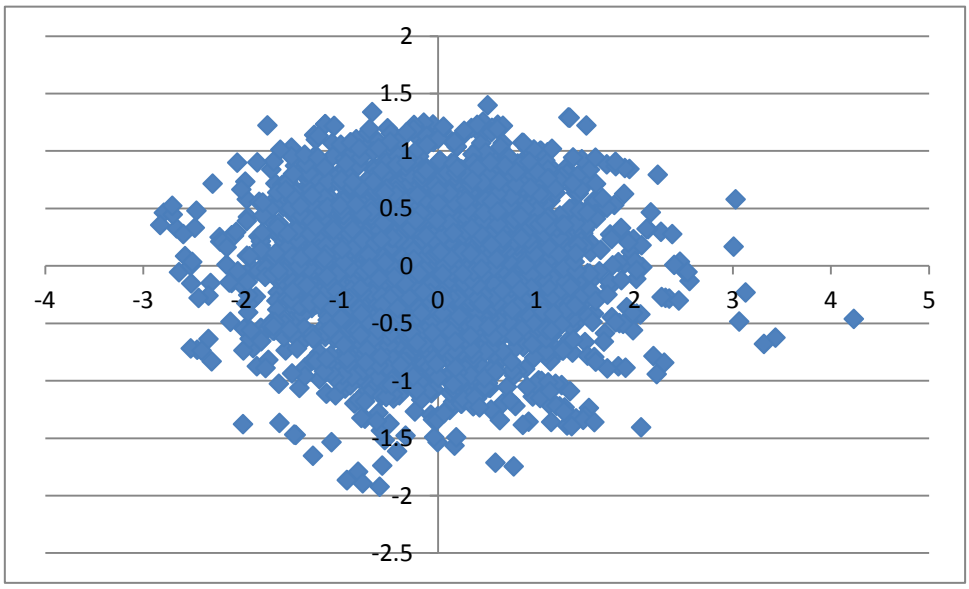


Figure 38. SAI-LIT Bland-Altman plot with systematic bias removed

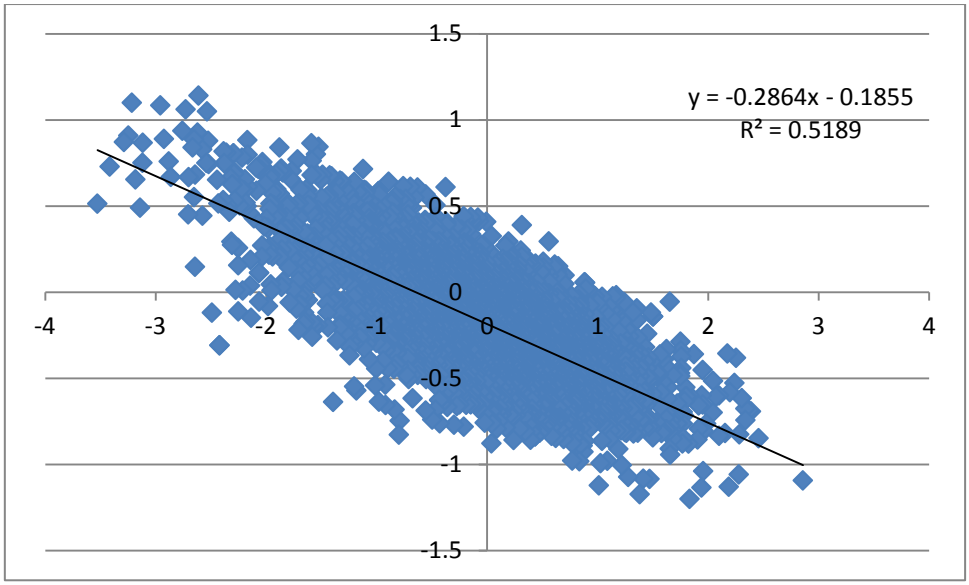


Figure 39. SAI-NUM Bland-Altman plot

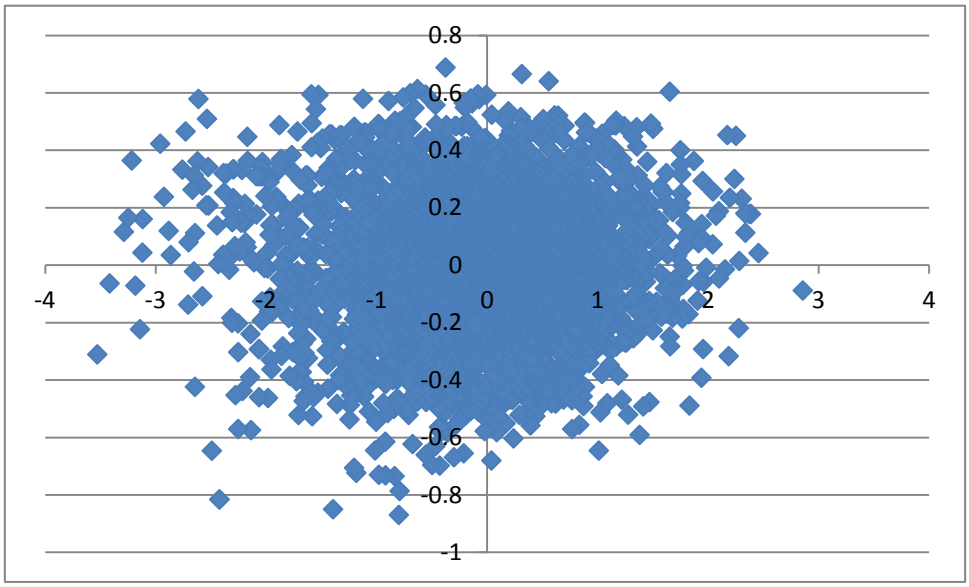


Figure 40. SAI-NUM Bland-Altman plot with systematic bias removed

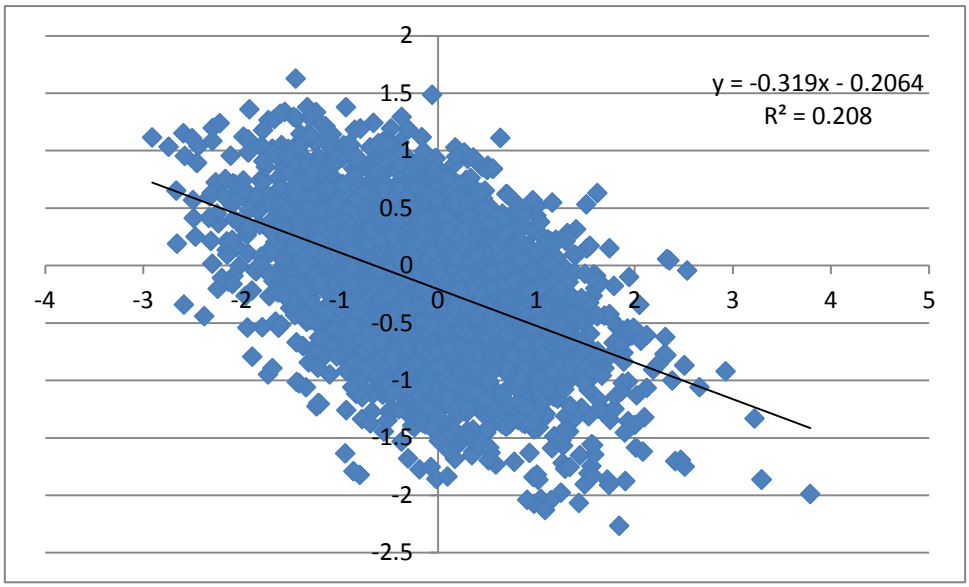


Figure 41. VS-LIT Bland-Altman plot

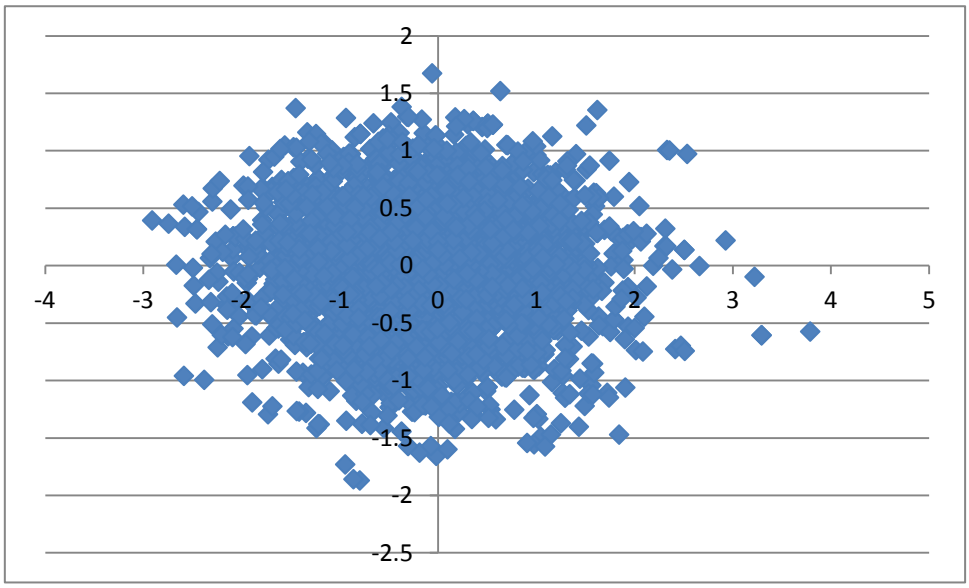


Figure 42. VS-LIT Bland-Altman plot

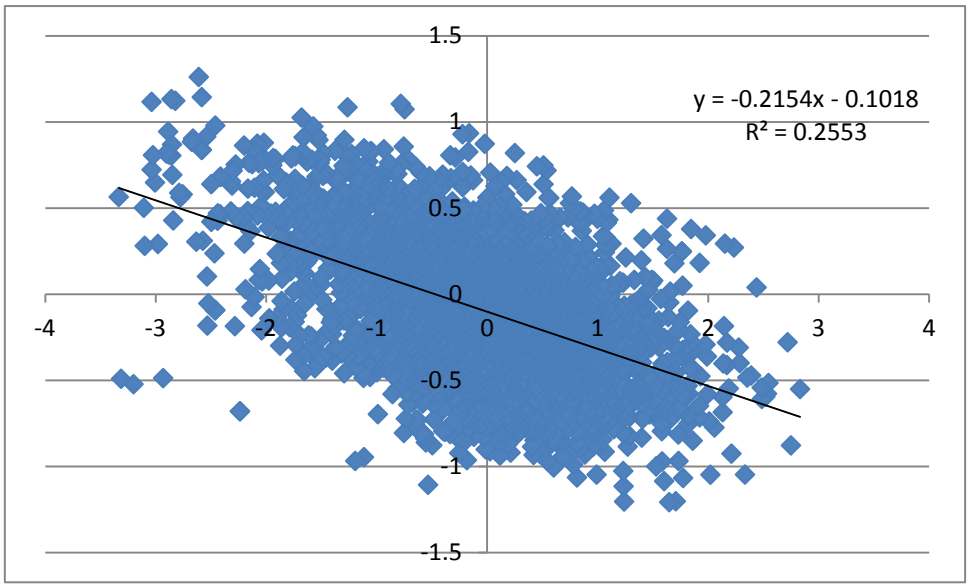


Figure 43. NV-NUM Bland-Altman plot

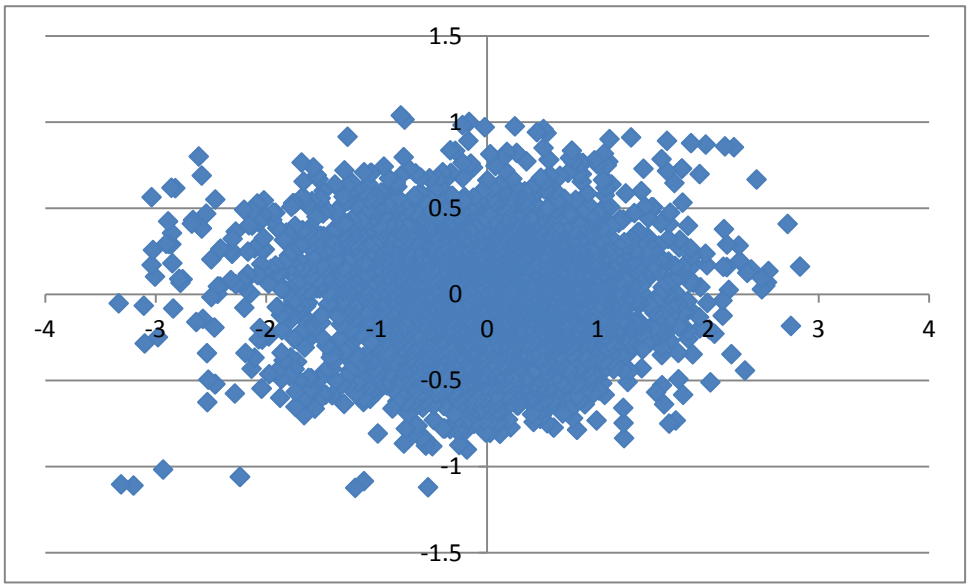


Figure 44. NV-NUM Bland-Altman plot with systematic bias removed

OLSAT- SC

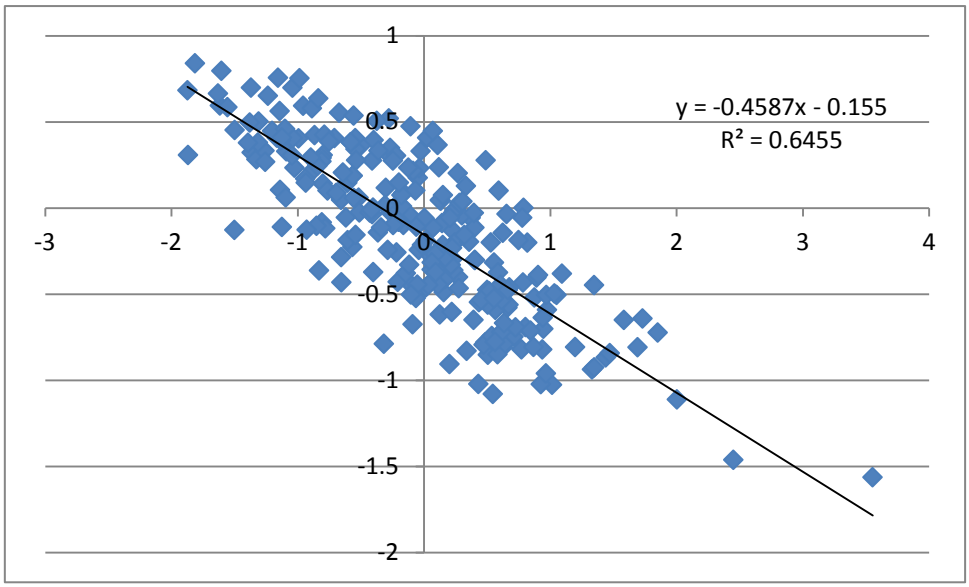


Figure 45. SAI-M Bland-Altman plot

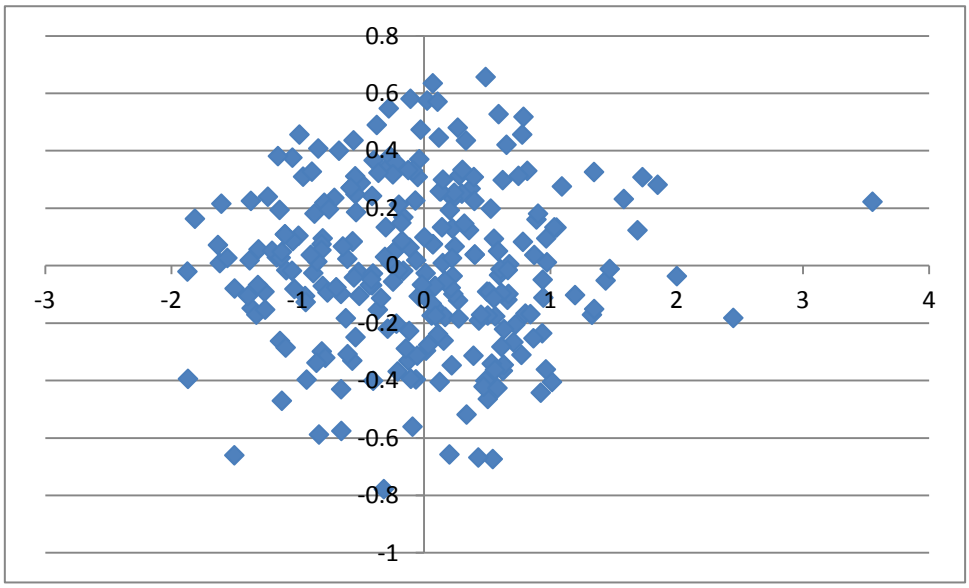


Figure 46. SAI-M Bland-Altman plot with systematic bias removed

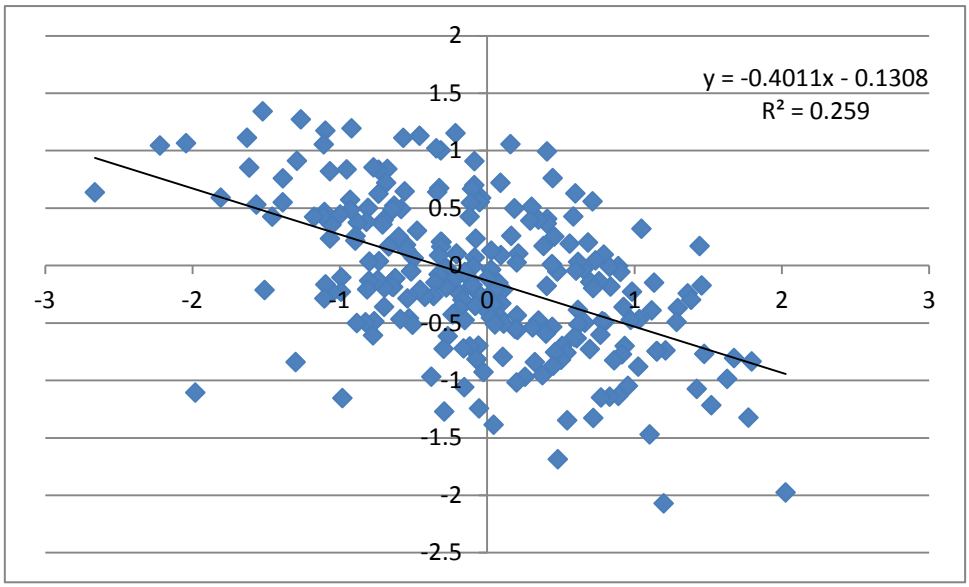


Figure 47. SAI-E Bland-Altman plot

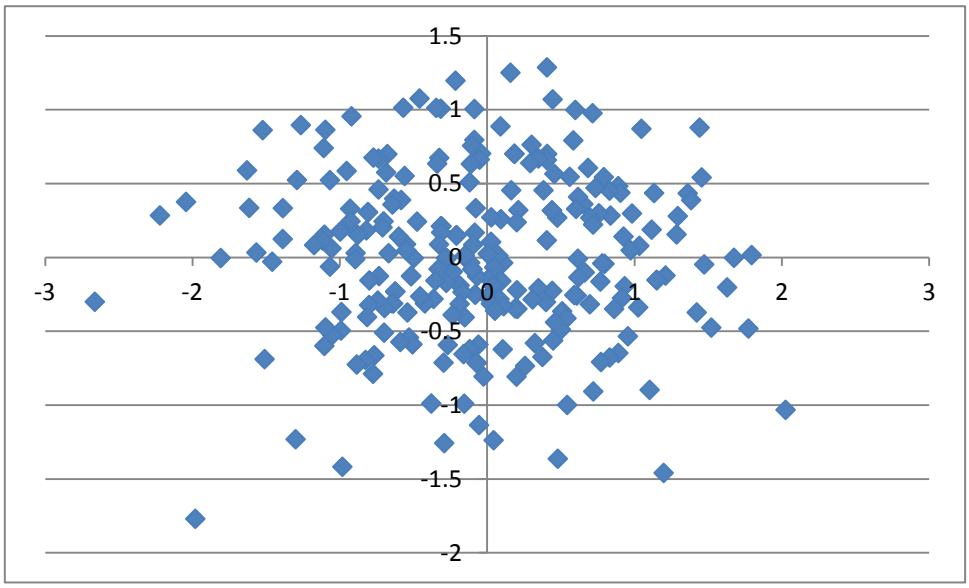


Figure 48. SAI-E Bland-Altman plot with systematic bias removed

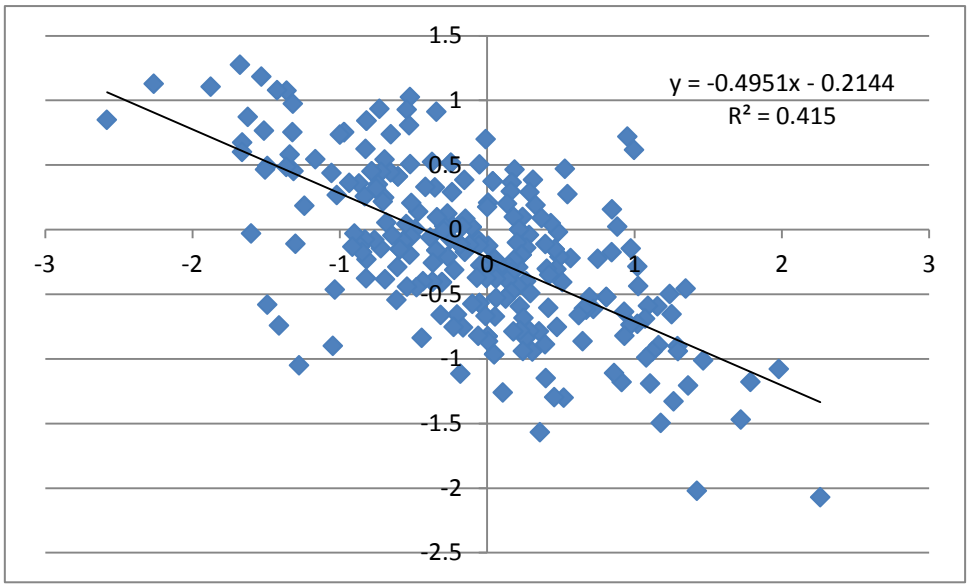


Figure 49. VS-E Bland-Altman plot

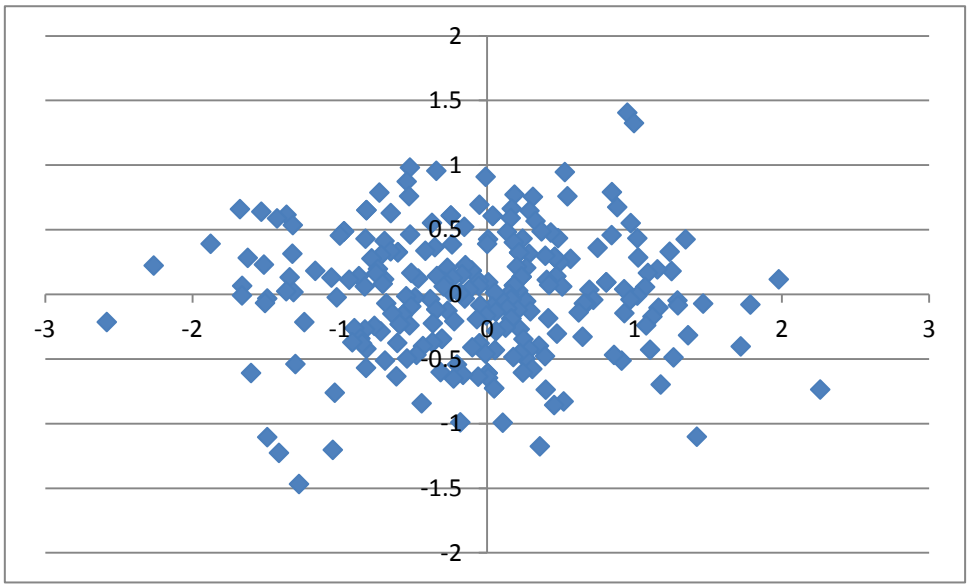


Figure 50. VS-E Bland-Altman plot with systematic bias removed

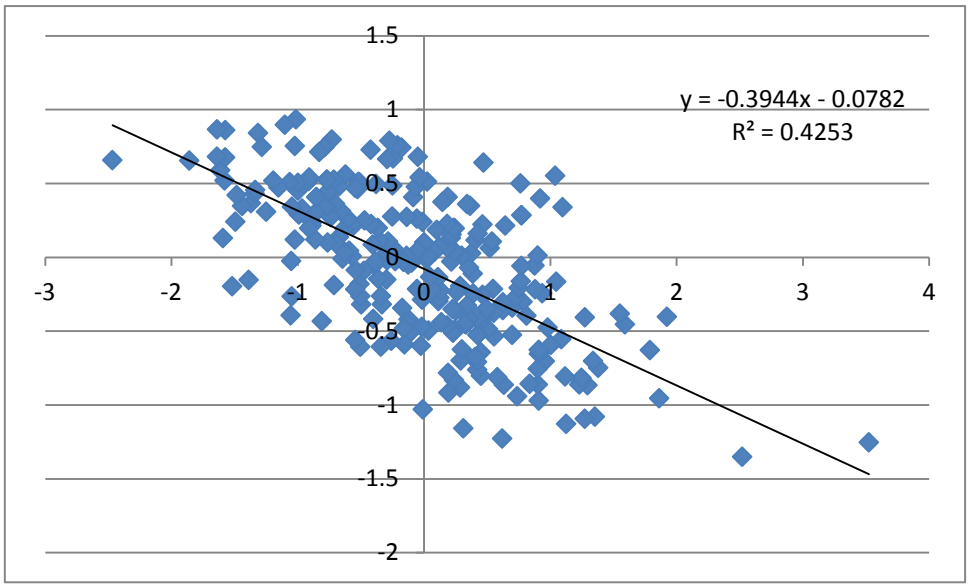


Figure 51. NV-M Bland-Altman plot

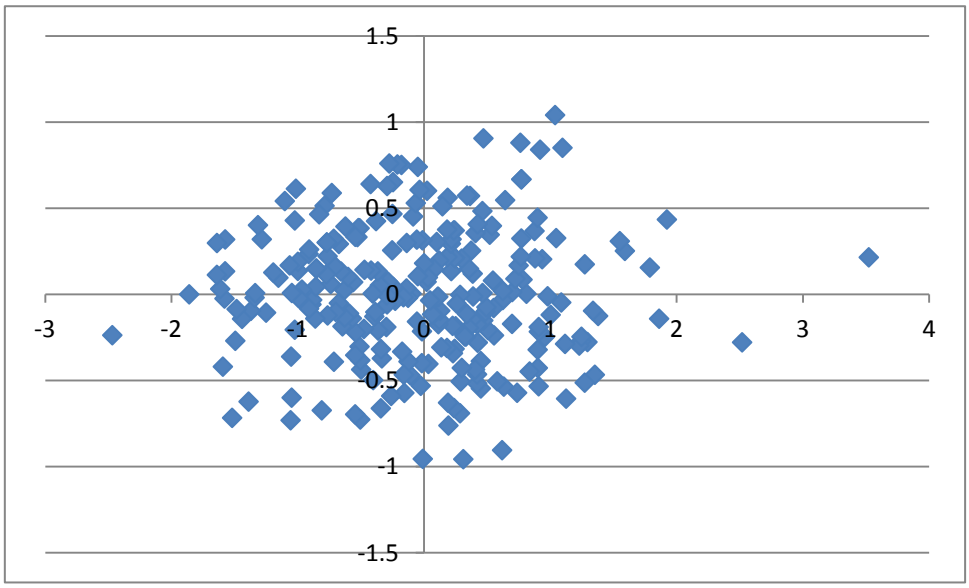


Figure 52. NV-M Bland-Altman plot with systematic bias removed

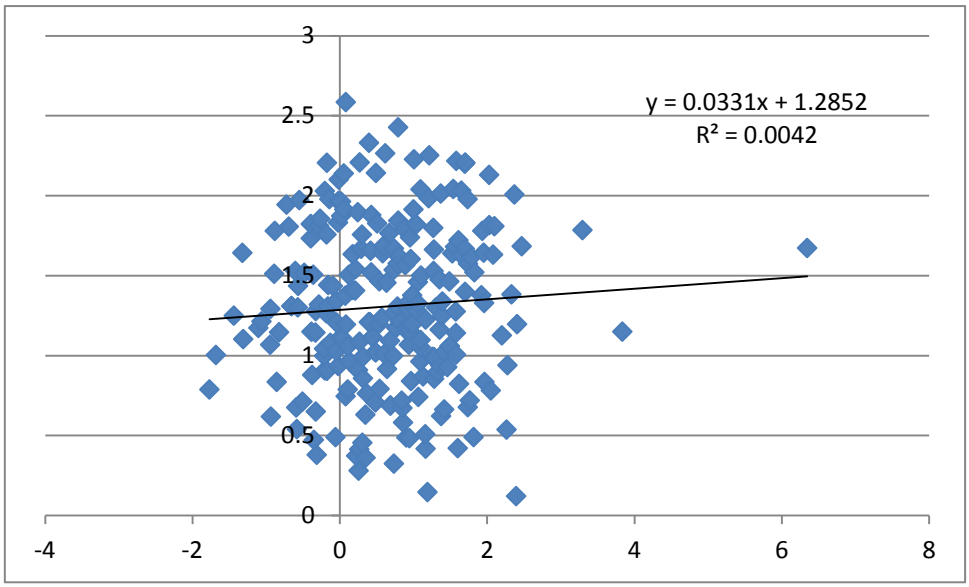


Figure 53. SAI-E Bland-Altman plot

No adjustments were necessary to the original Bland-Altman plot for the SAI-E data combination, as no linear pattern that may be indicative of systematic bias was identified.

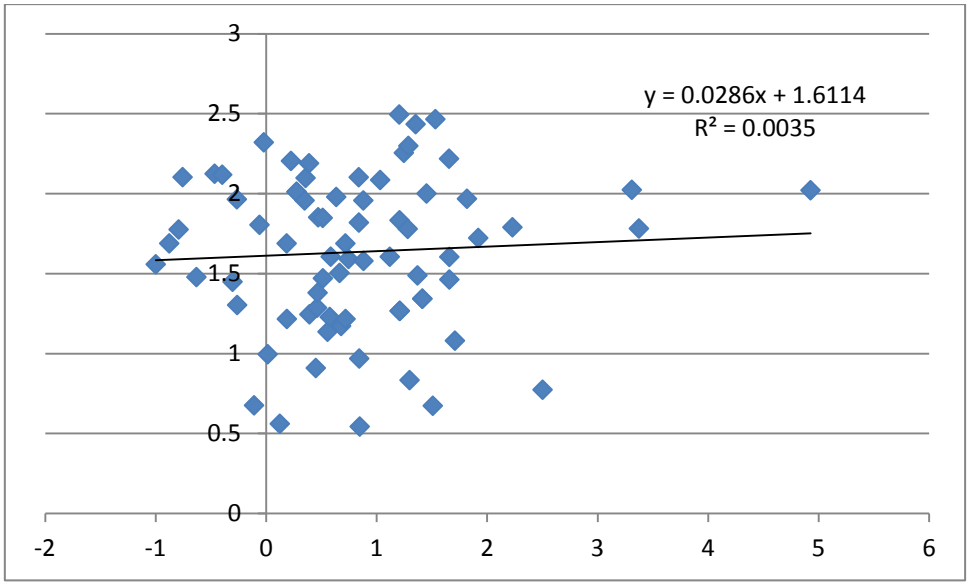


Figure 54. SAI-M Bland-Altman plot

No adjustments were necessary to the original Bland-Altman plot for the SAI-M data combination, as no linear pattern that may be indicative of systematic bias was identified.

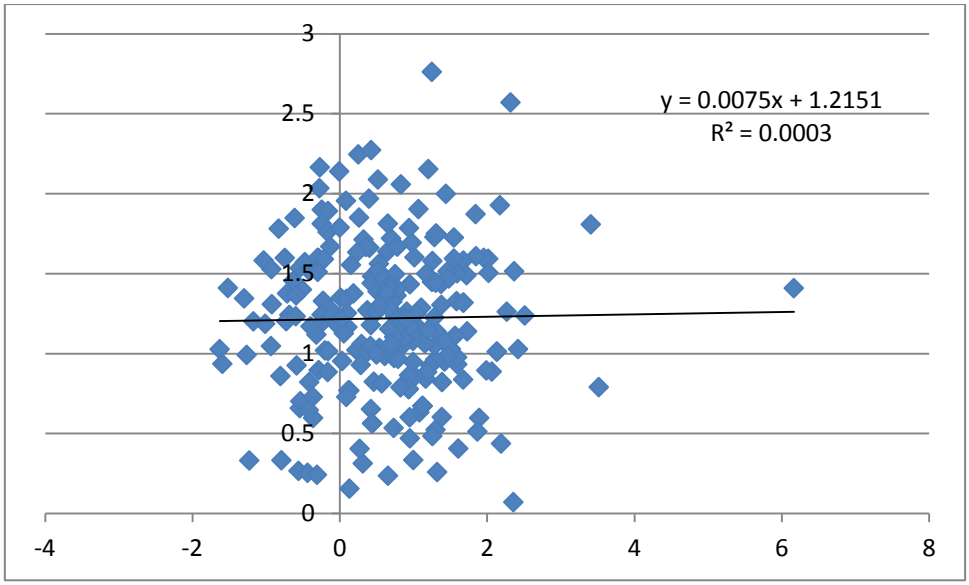


*Figure 55.*VS-E Bland-Altman plot

No adjustments were necessary to the original Bland-Altman plot for the VS-E data combination, as no linear pattern that may be indicative of systematic bias was identified.

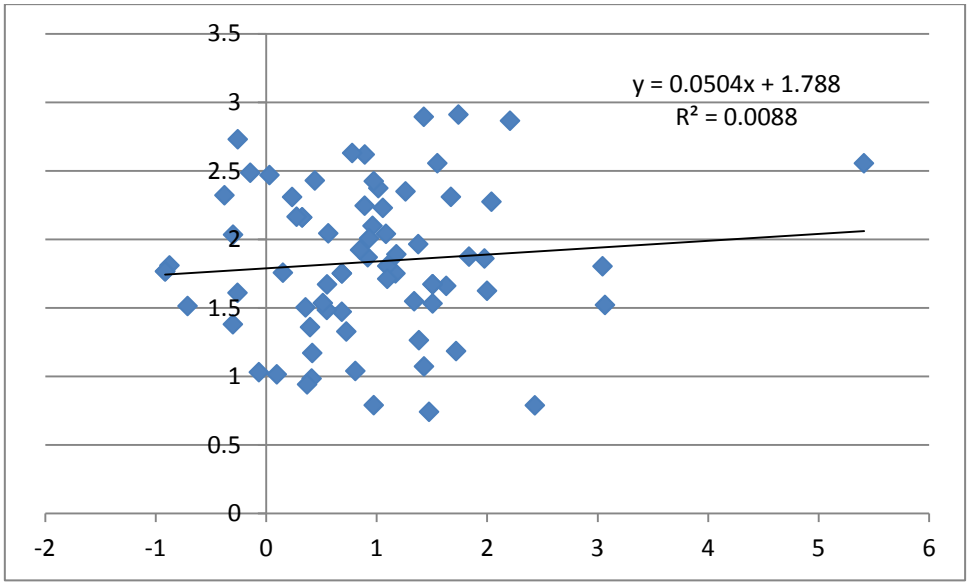


Figure 56. NV-M Bland-Altman plot

No adjustments were necessary to the original Bland-Altman plot for the NV-M data combination, as no linear pattern that may be indicative of systematic bias was identified.

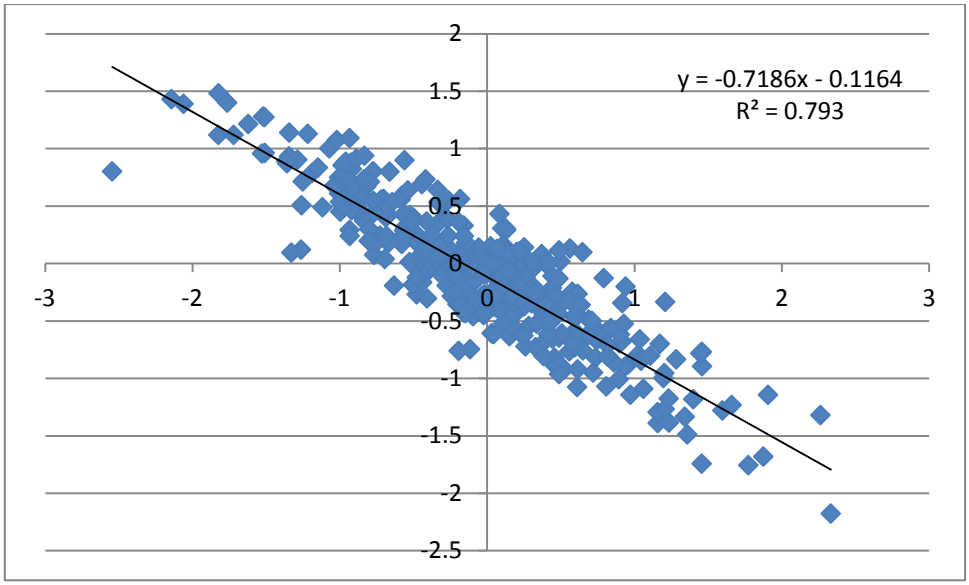


Figure 57. LIT-E Bland-Altman plot

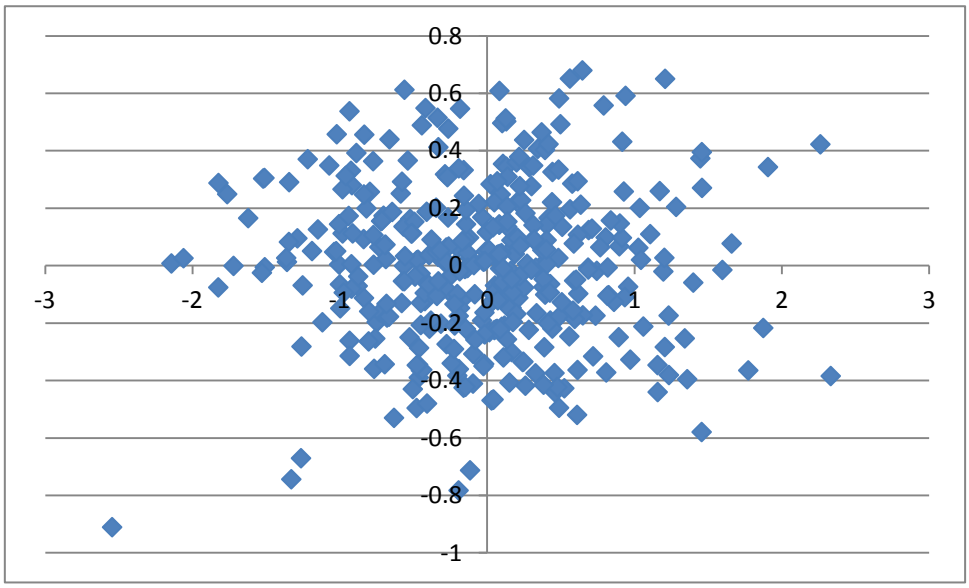


Figure 58. LIT-E Bland-Altman plot with systematic bias removed

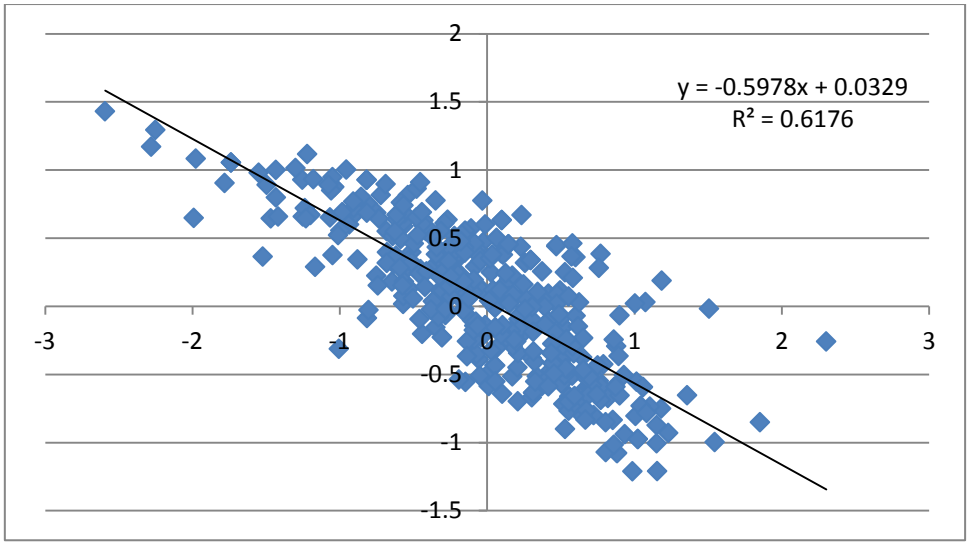


Figure 59. NUM-M Bland-Altman plot

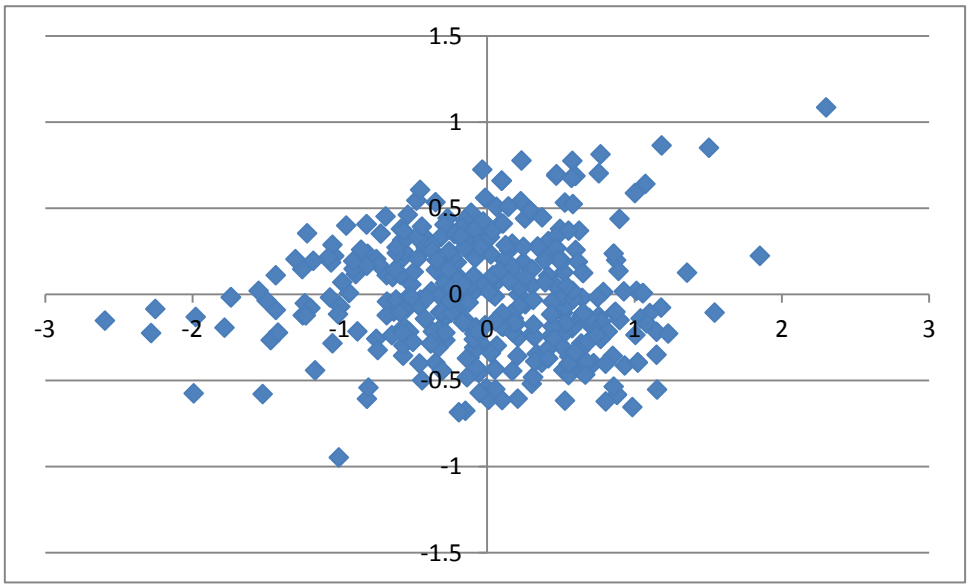


Figure 60. NUM-M Bland-Altman plot with systematic bias removed

NAPLAN-HSC

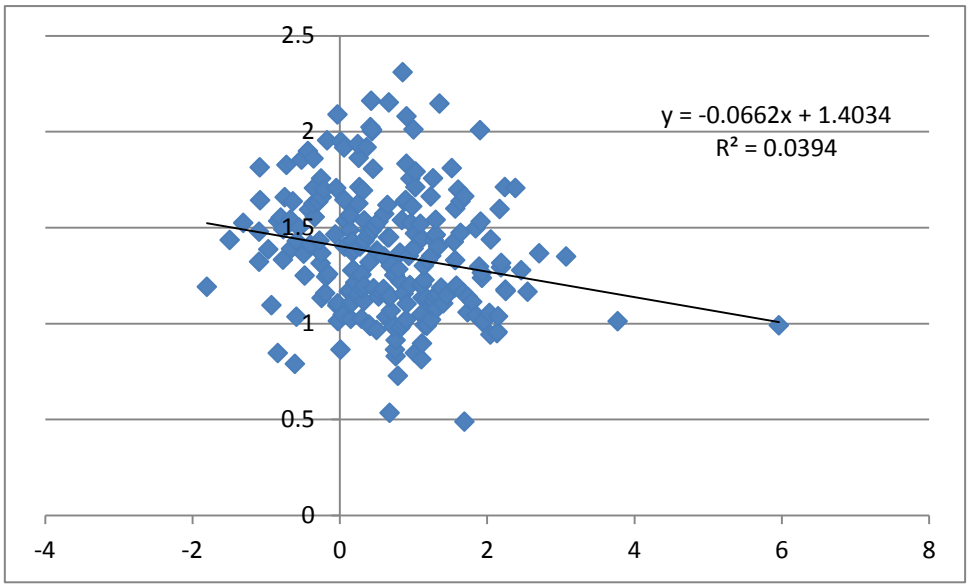


Figure 61. LIT-E Bland-Altman plot

No adjustments were necessary to the original Bland-Altman plot for the LIT-E data combination, as no linear pattern that may be indicative of systematic bias was identified.

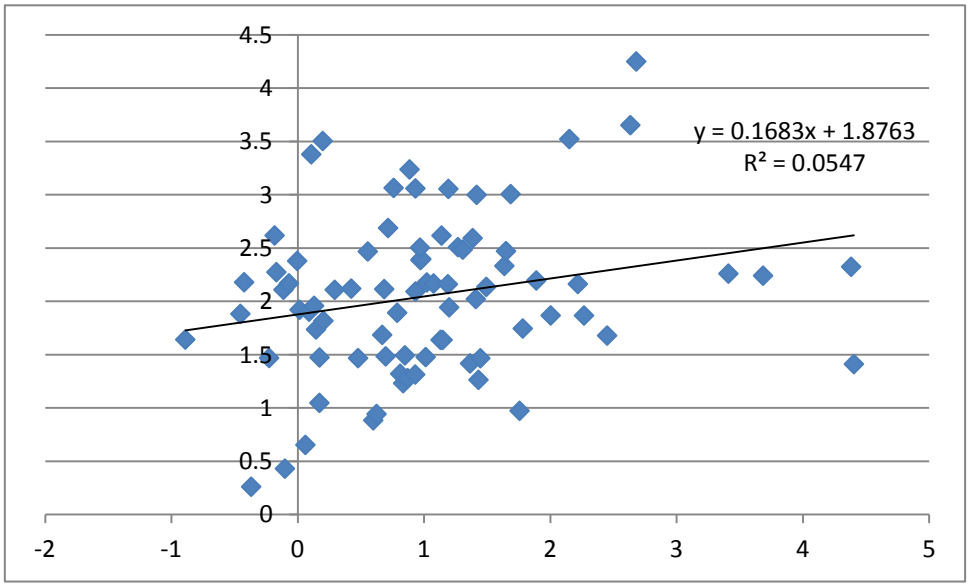


Figure 62. NUM-M Bland-Altman plot

No adjustments were necessary to the original Bland-Altman plot for the NUM-M data combination, as no linear pattern that may be indicative of systematic bias was identified.

SC-HSC

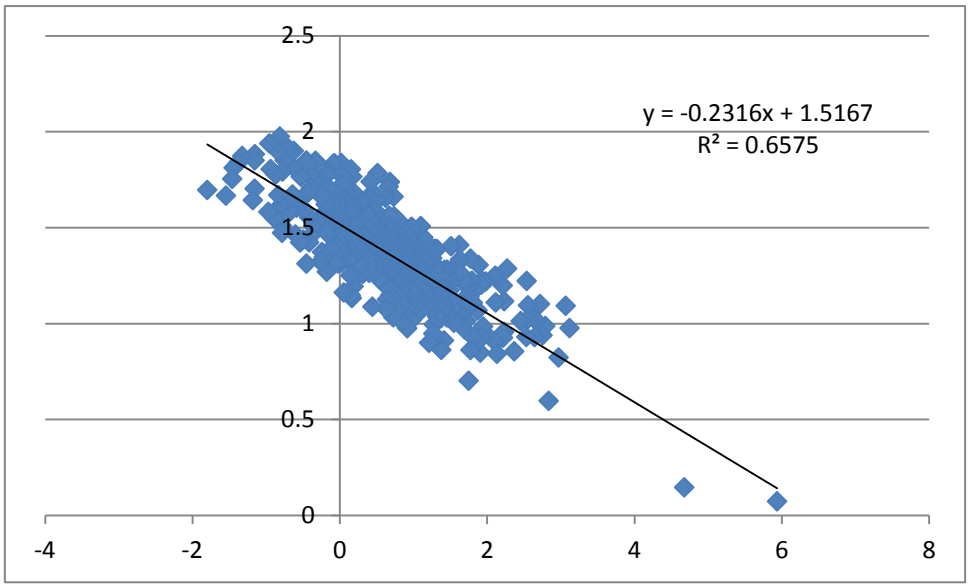


Figure 63. E-E Bland-Altman plot

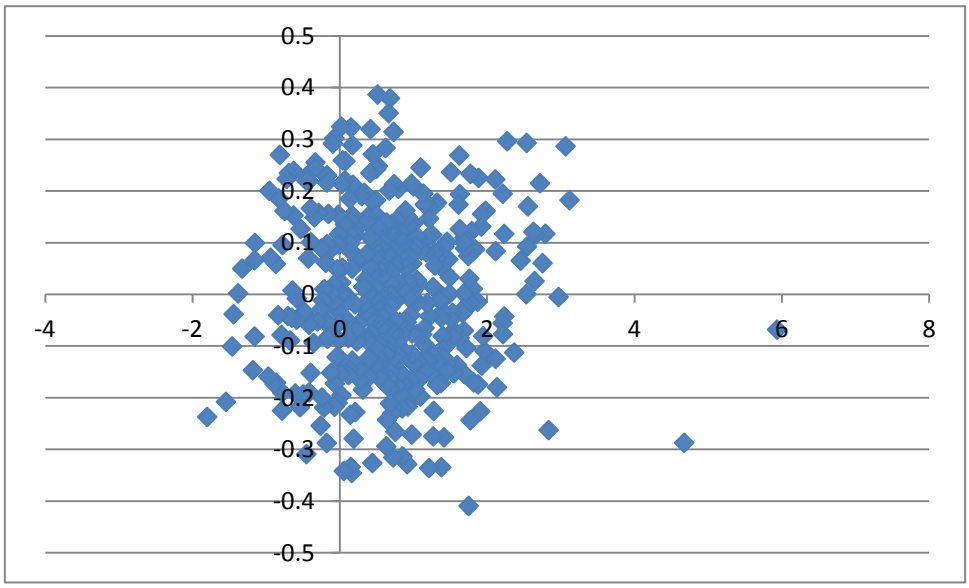


Figure 64. E-E Bland-Altman plot with systematic bias removed

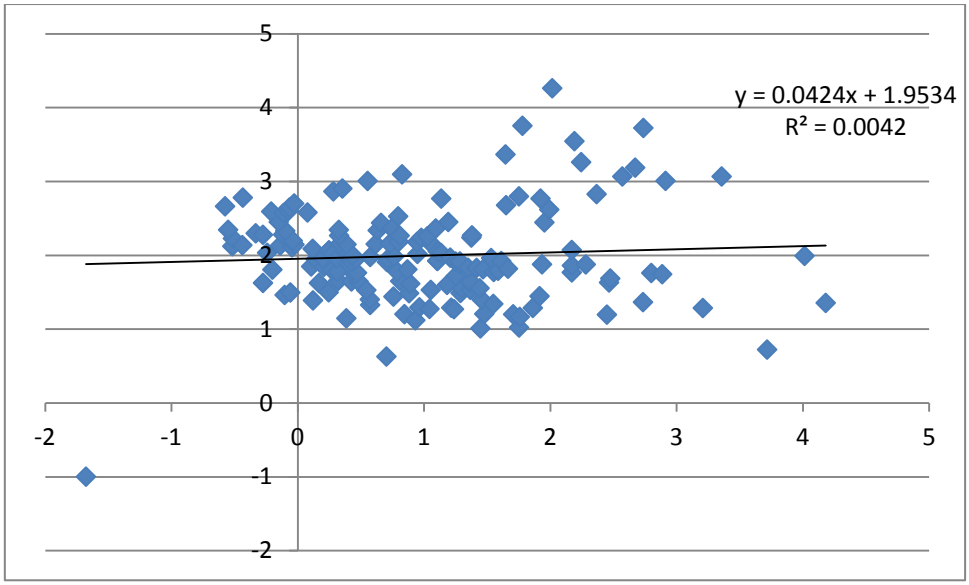


Figure 65. M-M Bland-Altman plot

No adjustments were necessary to the original Bland-Altman plot for the M-M data combination, as no linear pattern that may be indicative of systematic bias was identified.

OLSAT-Junior SA

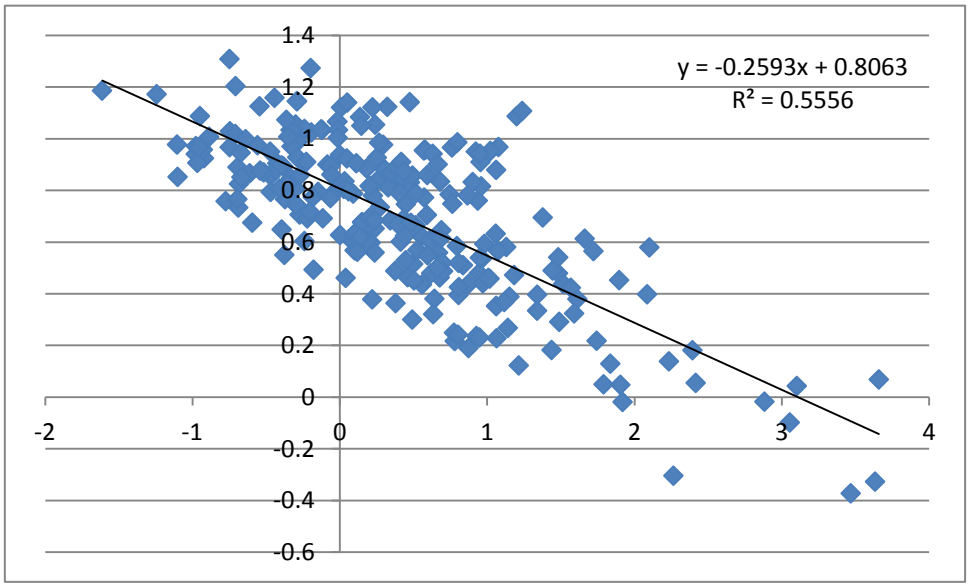


Figure 66. SAI-E Bland-Altman plot

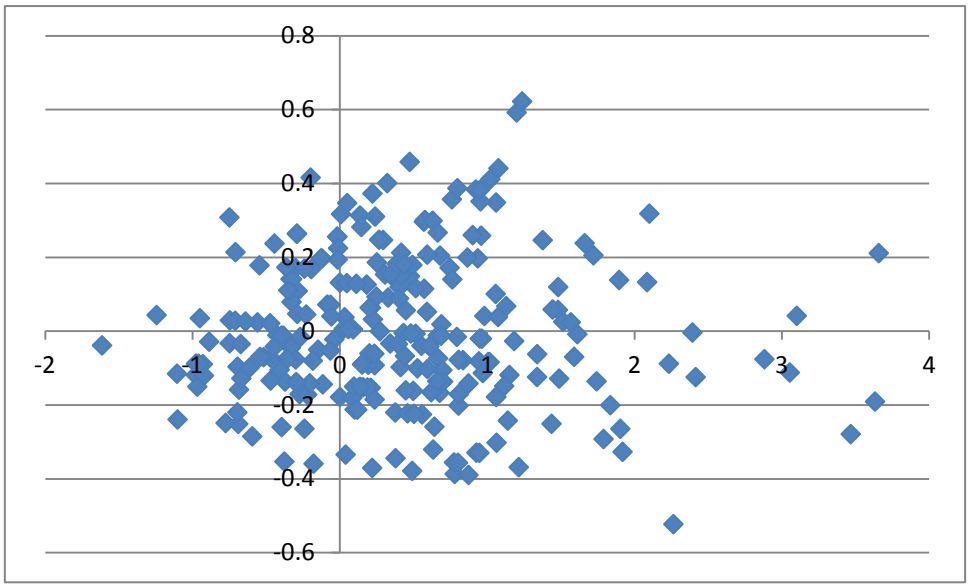


Figure 67. SAI-E Bland-Altman plot with systematic bias removed

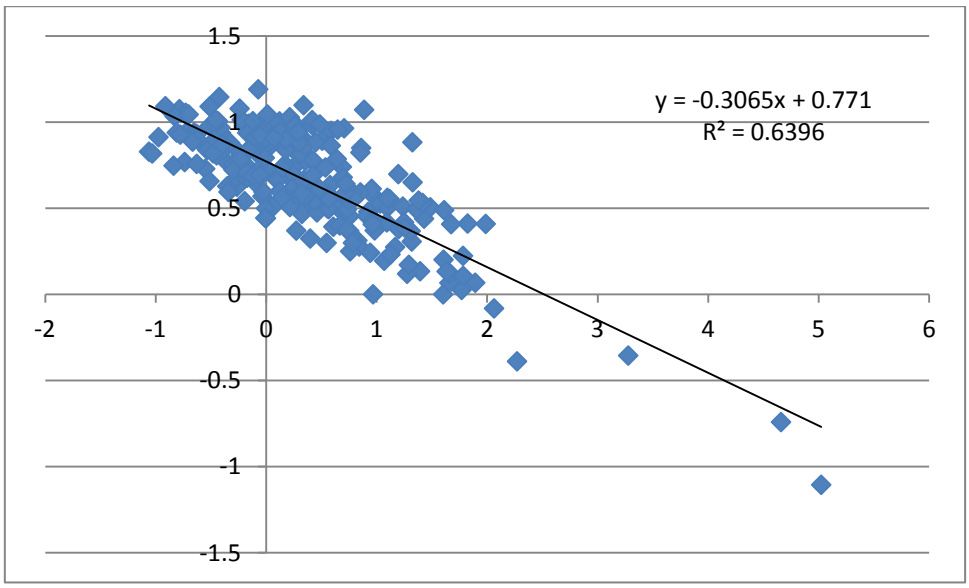


Figure 68. SAI-M Bland-Altman plot

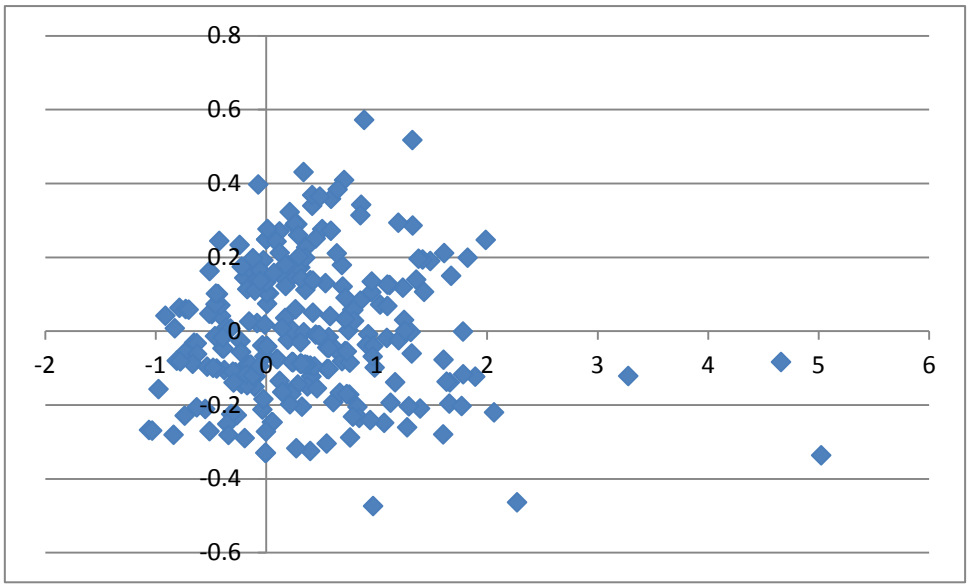


Figure 69. SAI-M Bland-Altman plot with systematic bias removed

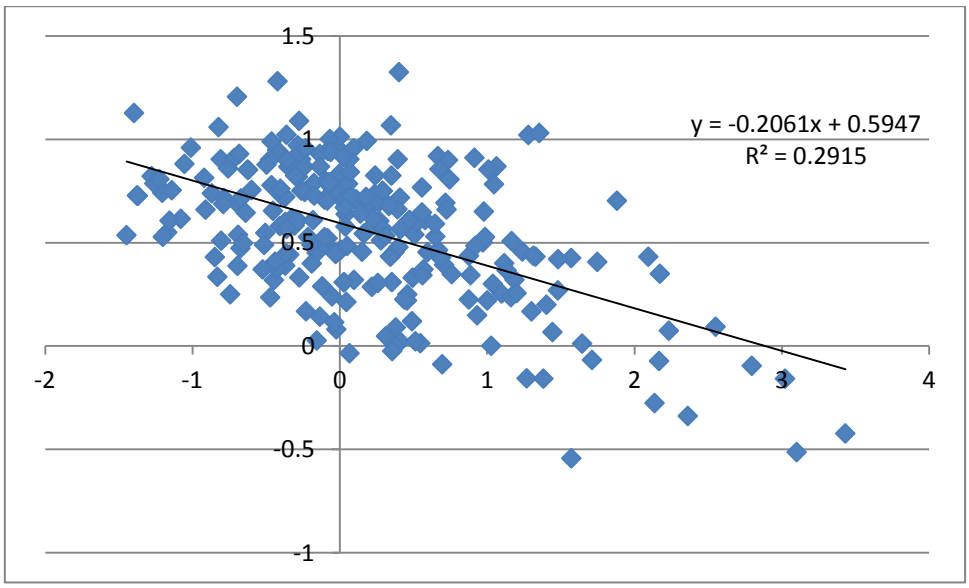


Figure 70. VS-E Bland-Altman plot

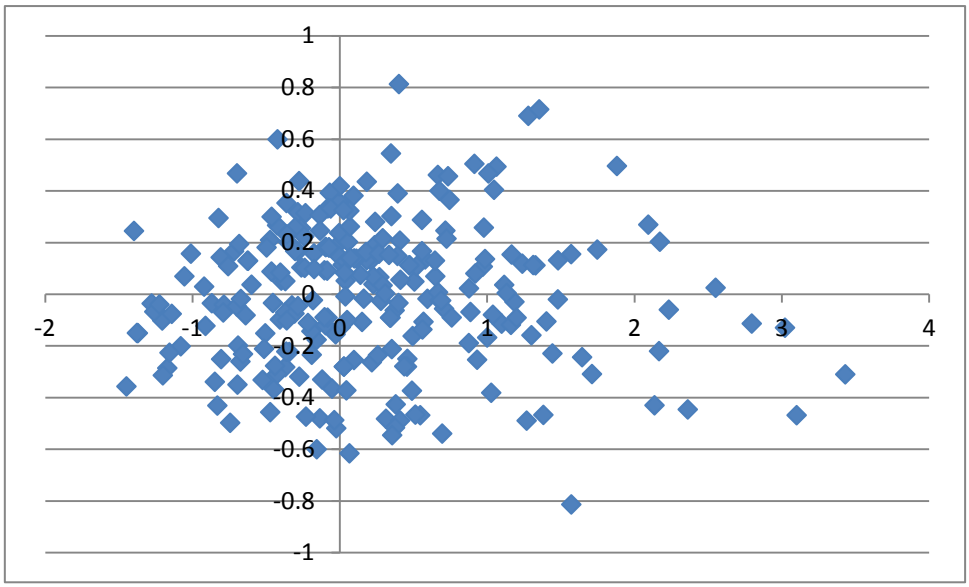


Figure 71. VS-E Bland-Altman plot with systematic bias removed

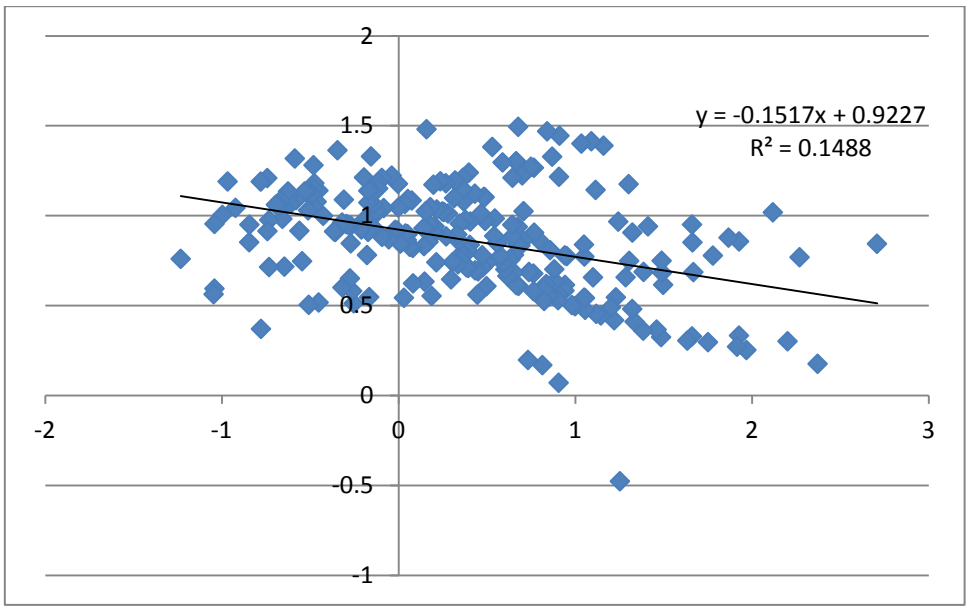


Figure 72. NV-M Bland-Altman plot

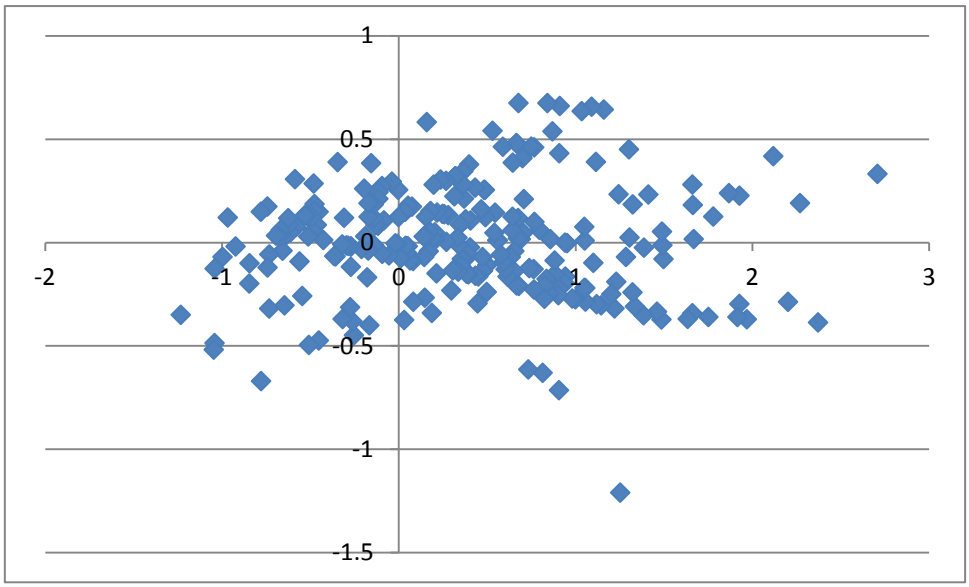


Figure 73. NV-M Bland-Altman plot with systematic bias removed

OLSAT-JSA

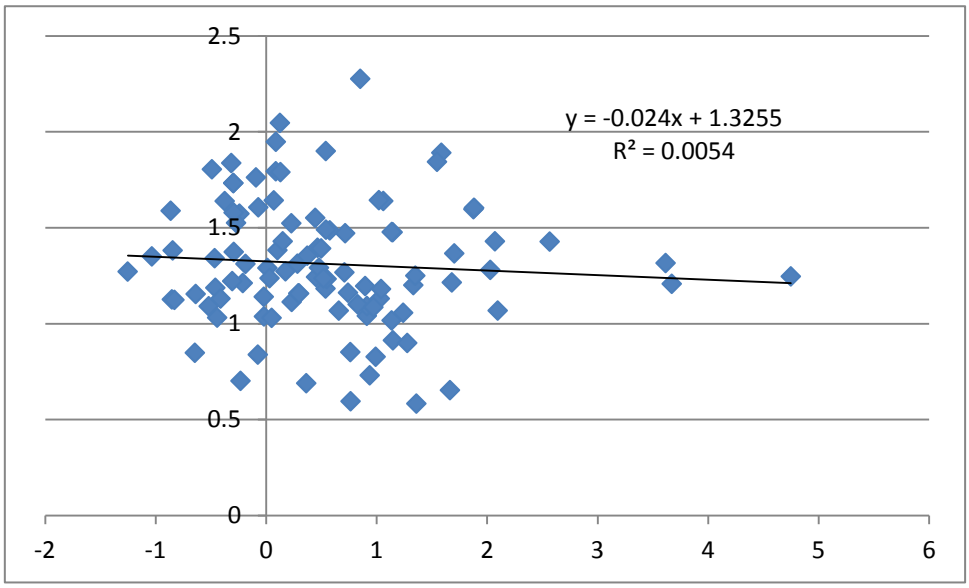


Figure 74. SAI-E Bland-Altman plot

No adjustments were necessary to the original Bland-Altman plot for the SAI-E data combination, as no linear pattern that may be indicative of systematic bias was identified.

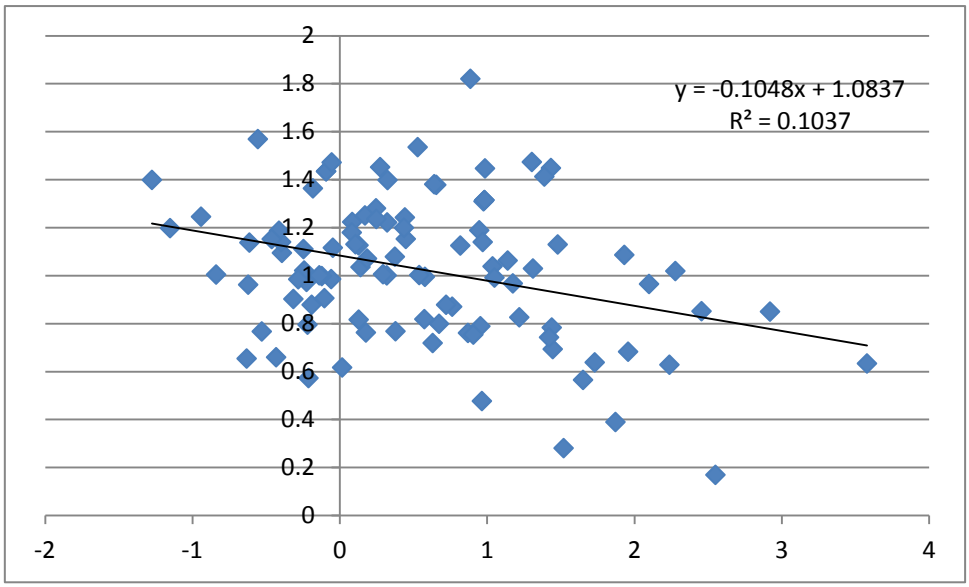


Figure 75. SAI-M Bland-Altman plot

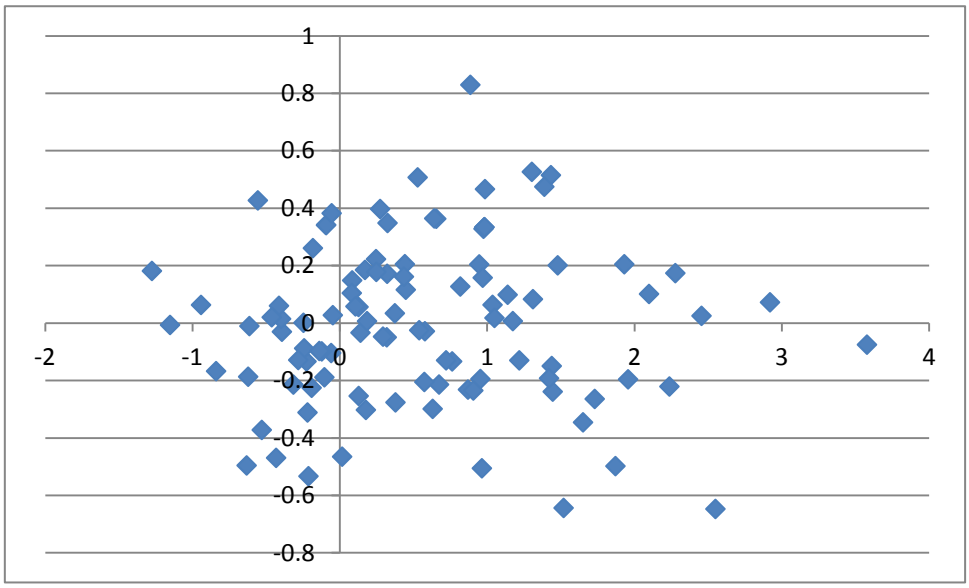


Figure 76. SAI-M Bland-Altman plot with systematic bias removed

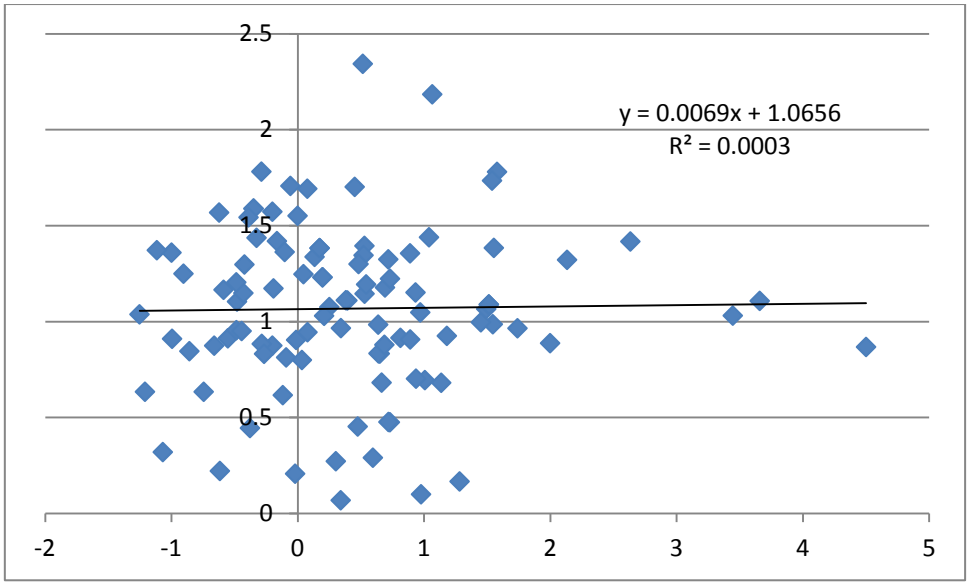


Figure 77. VS-E Bland-Altman plot

No adjustments were necessary to the original Bland-Altman plot for the VS-E data combination, as no linear pattern that may be indicative of systematic bias was identified.

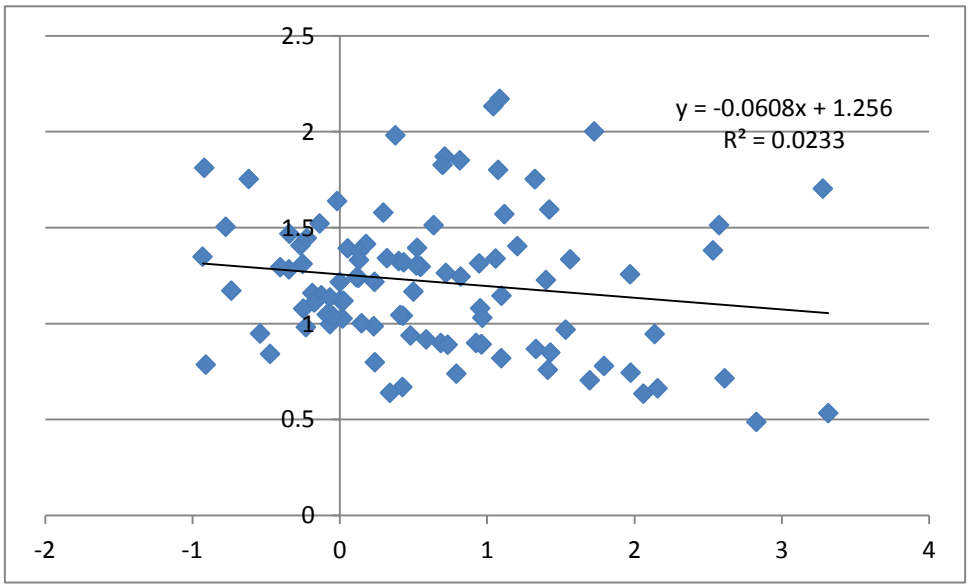


Figure 78. NV-M Bland-Altman plot

No adjustments were necessary to the original Bland-Altman plot for the NV-M data combination, as no linear pattern that may be indicative of systematic bias was identified.

NAPLAN JSA

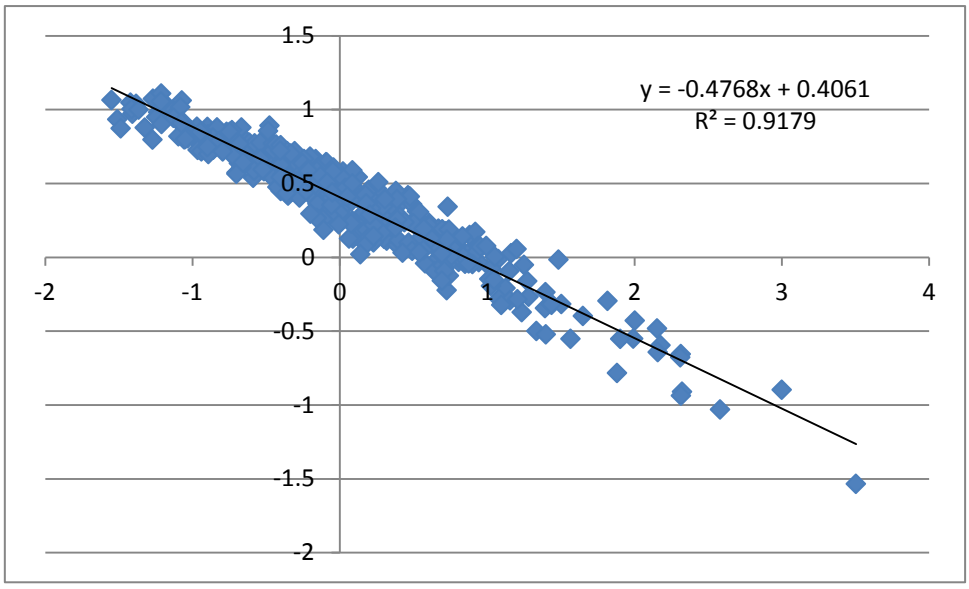


Figure 79. LIT-E Bland-Altman plot

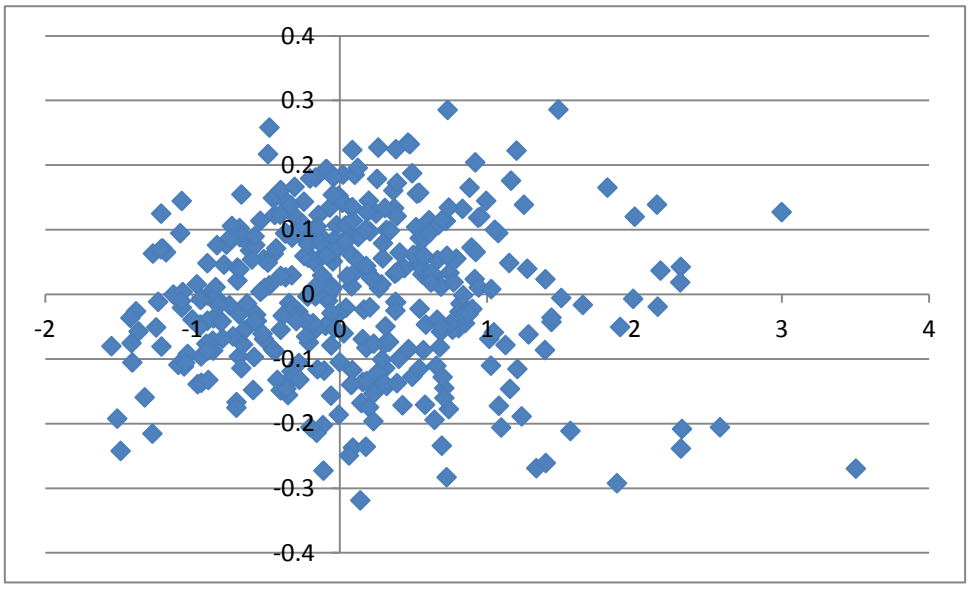


Figure 80. LIT-E Bland-Altman plot with systematic bias removed

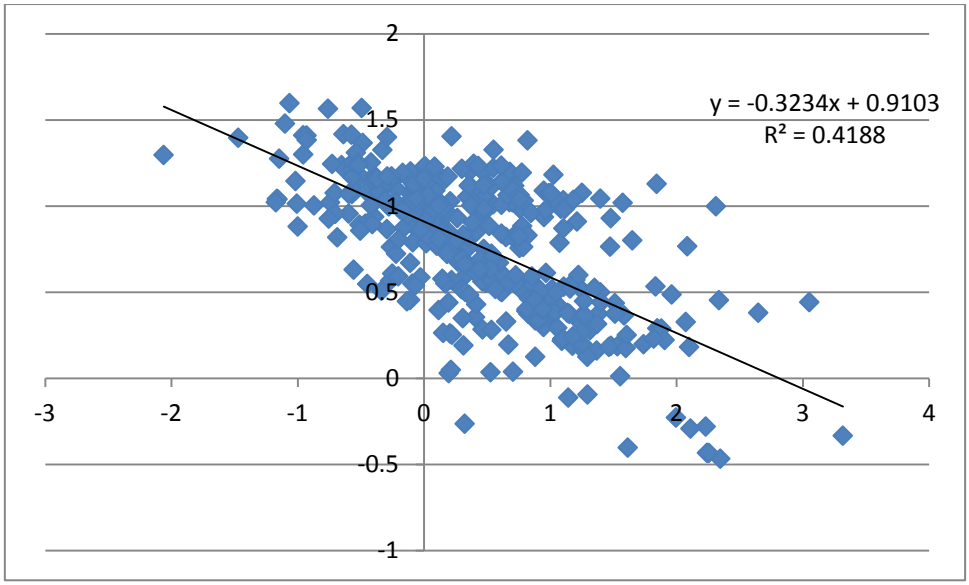


Figure 81. NUM-M Bland-Altman plot

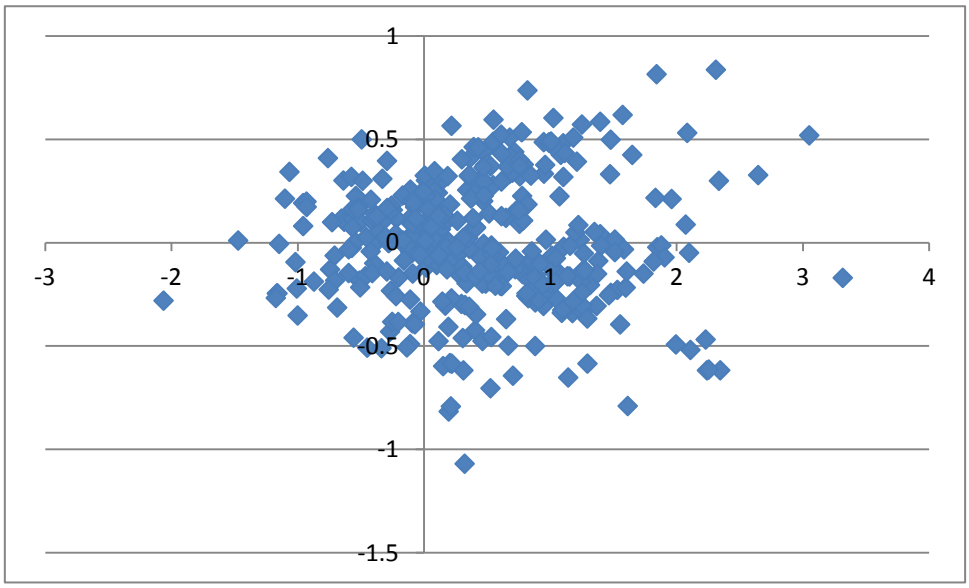


Figure 82. NUM-M Bland-Altman plot with systematic bias removed

NAPLAN-SSA

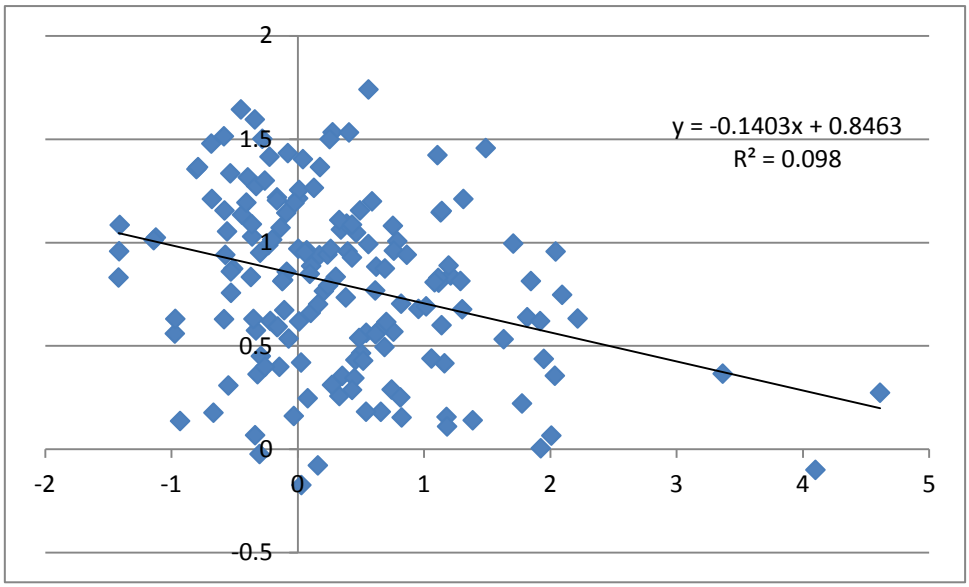


Figure 83. LIT-E Bland-Altman plot

No adjustments were necessary to the original Bland-Altman plot for the LIT-E data combination, as no linear pattern that may be indicative of systematic bias was identified.

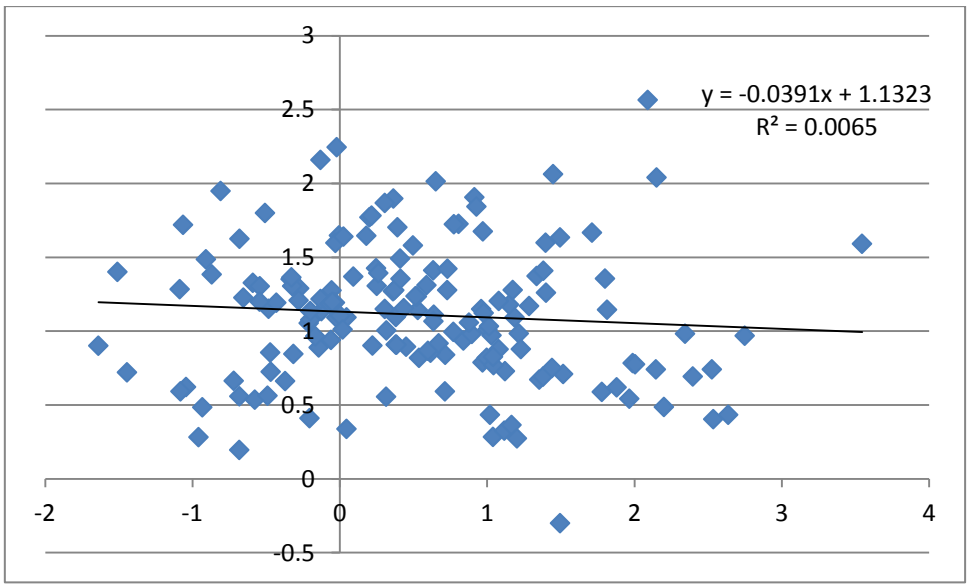


Figure 84. NUM-M Bland-Altman plot

No adjustments were necessary to the original Bland-Altman plot for the NUM-M data combination, as no linear pattern that may be indicative of systematic bias was identified.

JSA-SC

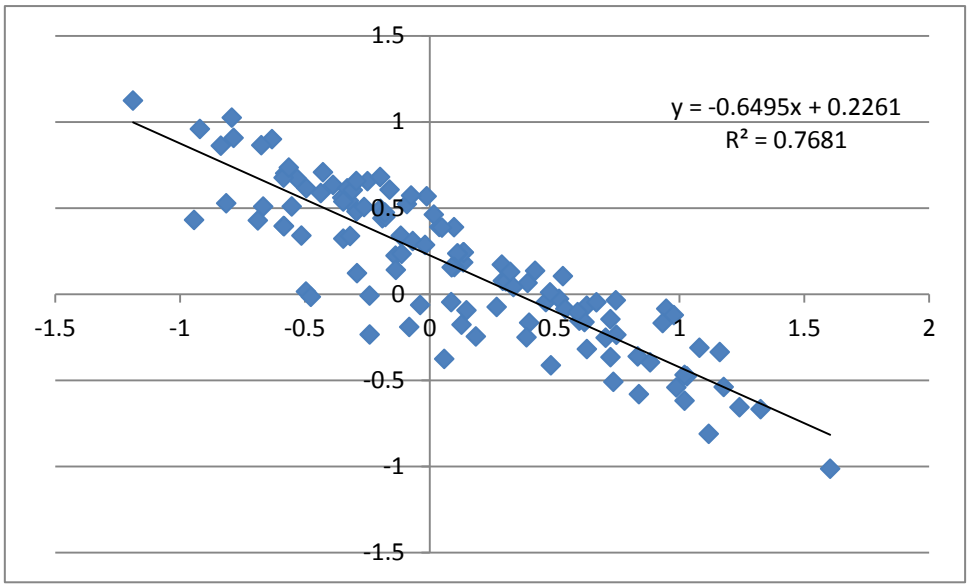


Figure 85. E-E Bland-Altman plot

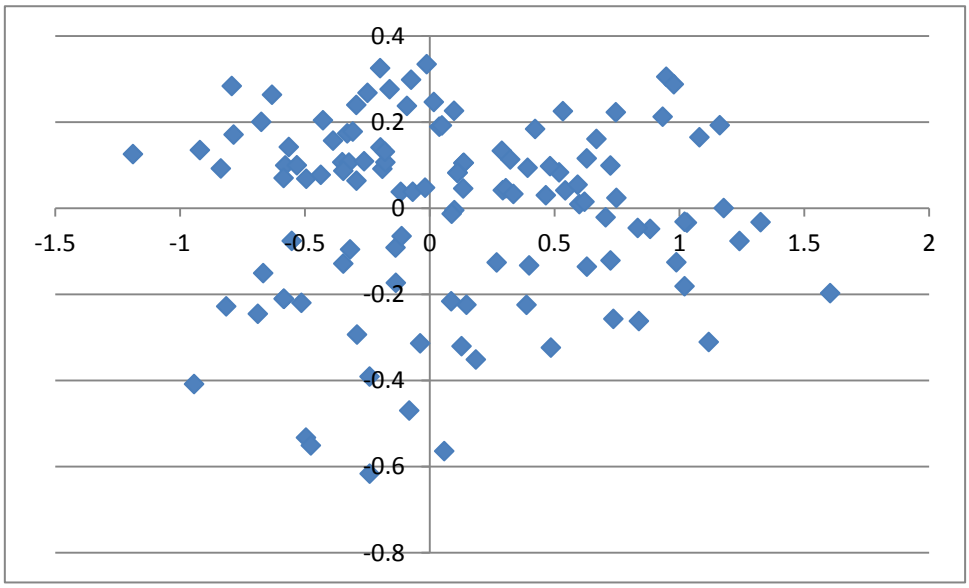


Figure 86. E-E Bland-Altman plot with systematic bias removed

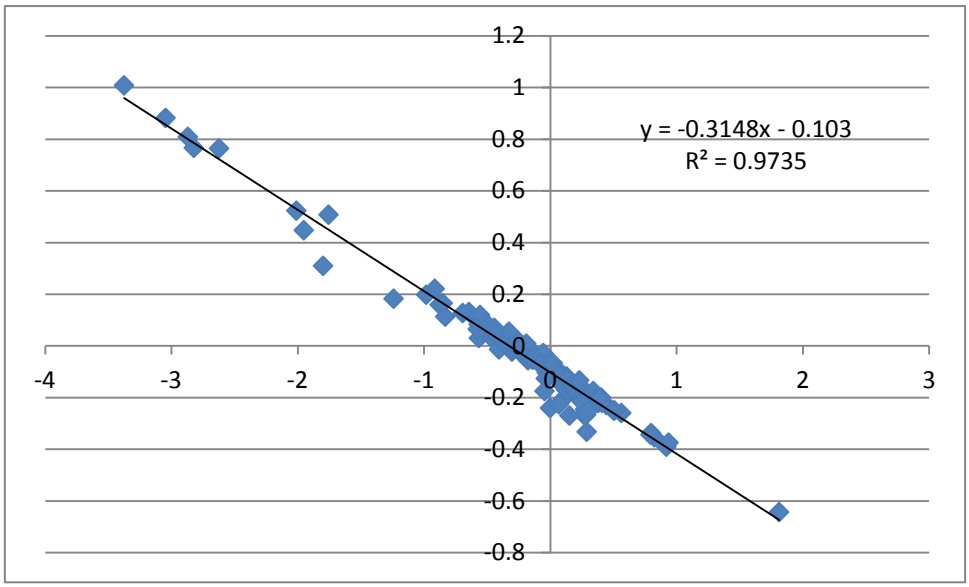


Figure 87. M-M Bland-Altman plot

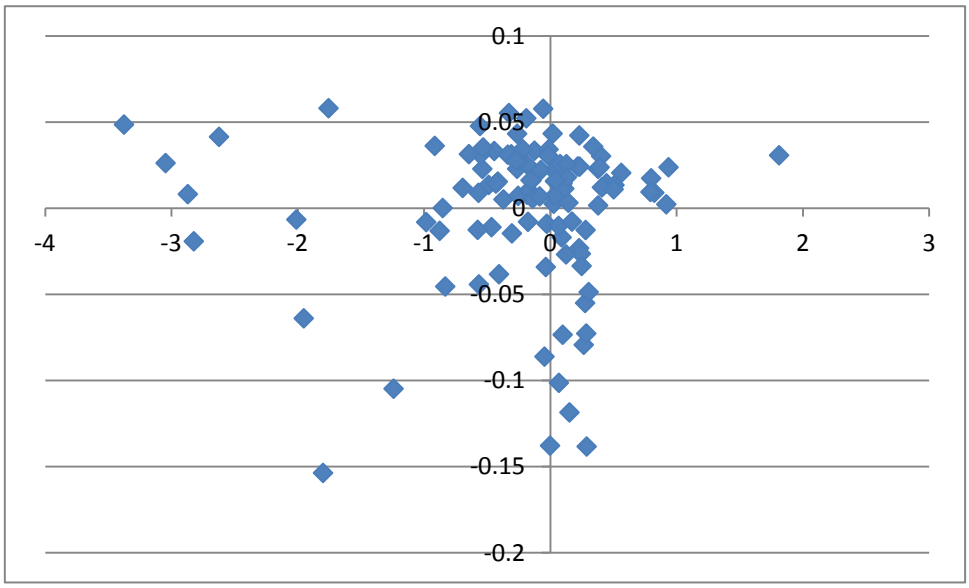


Figure 88. M-M Bland-Altman plot with systematic bias removed

SC-SSA

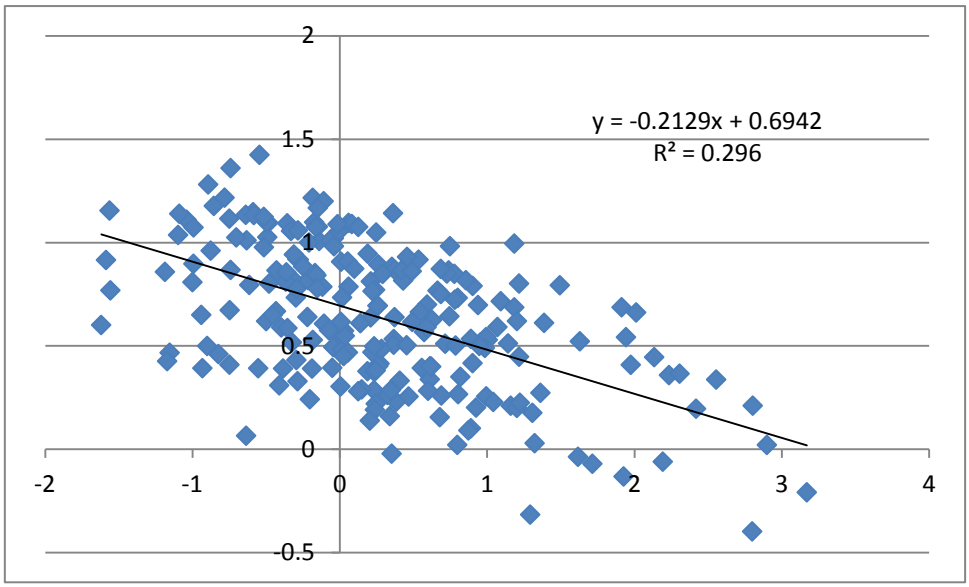


Figure 89. E-E Bland-Altman plot

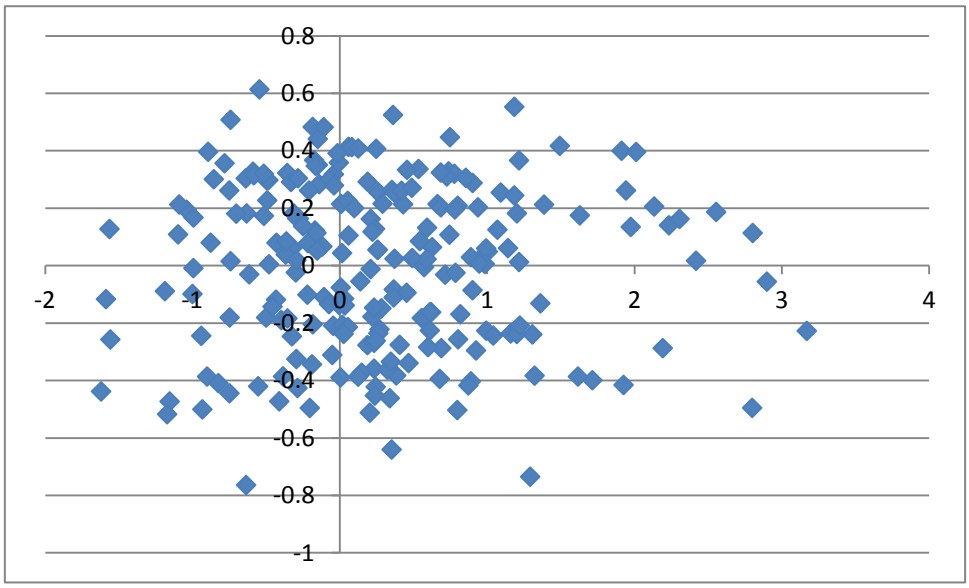


Figure 90. E-E Bland-Altman plot with systematic bias removed

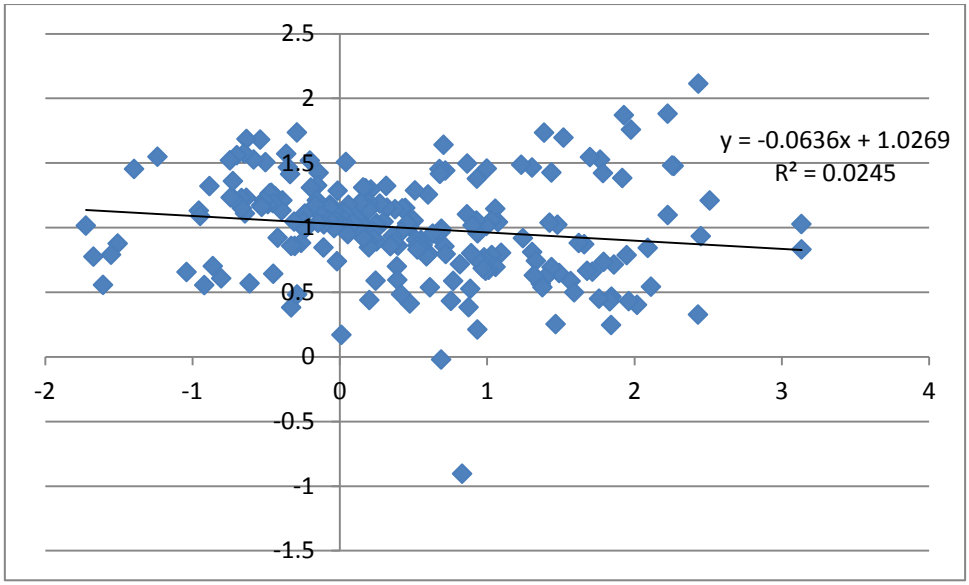


Figure 91. M-M Bland-Altman plot

No adjustments were necessary to the original Bland-Altman plot for the M-M data combination, as no linear pattern that may be indicative of systematic bias was identified.

JSA-HSC

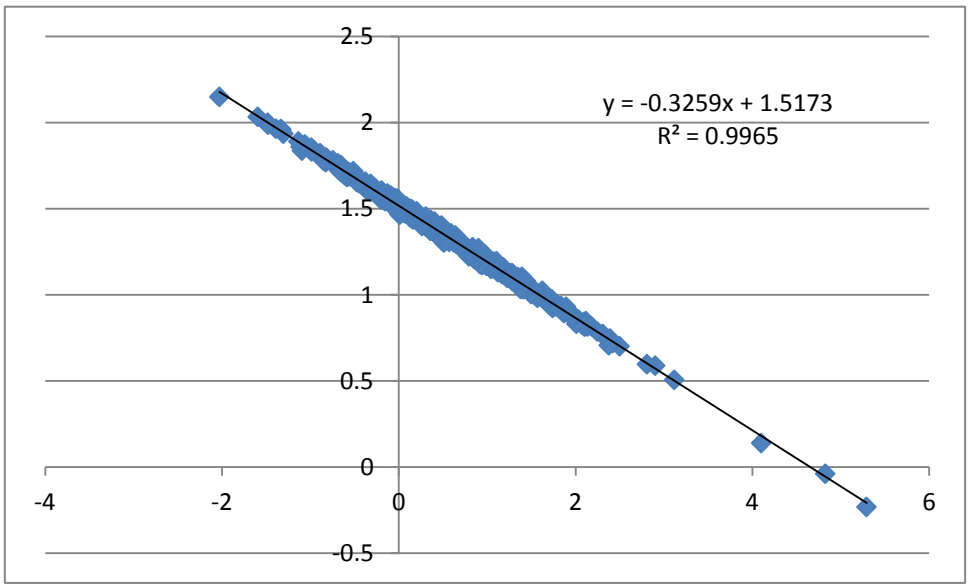


Figure 92. E-E Bland-Altman plot

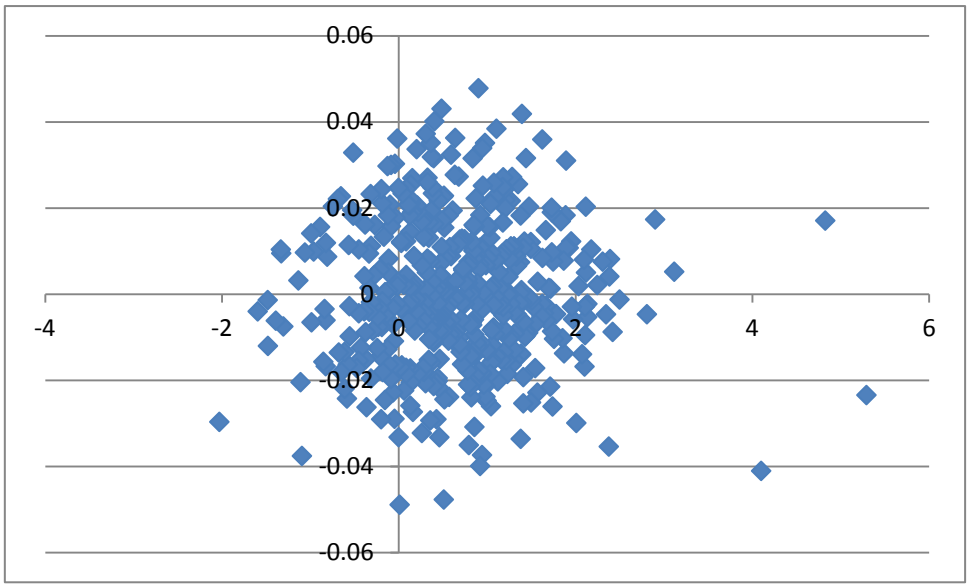


Figure 93. E-E Bland-Altman plot with systematic bias removed

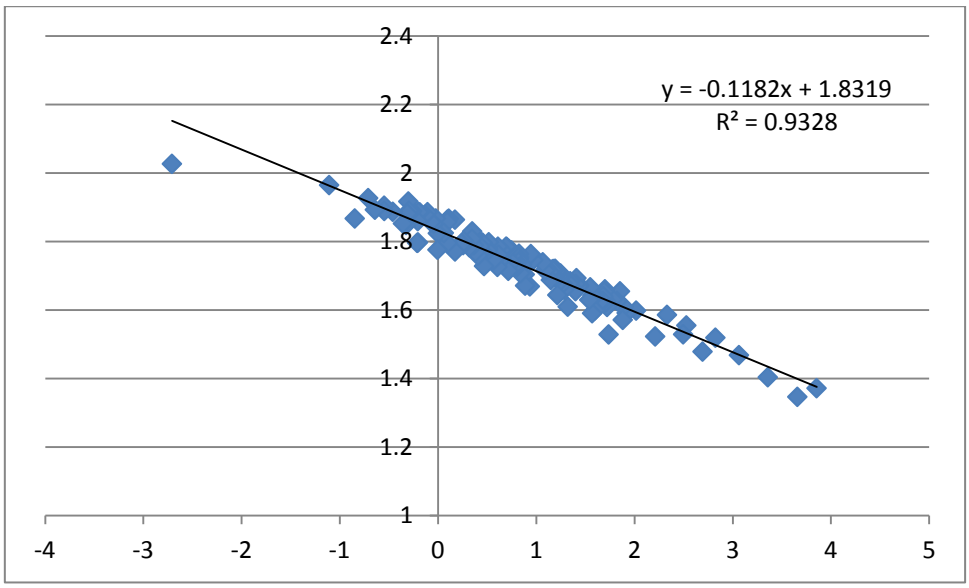


Figure 94. M-M Bland-Altman plot

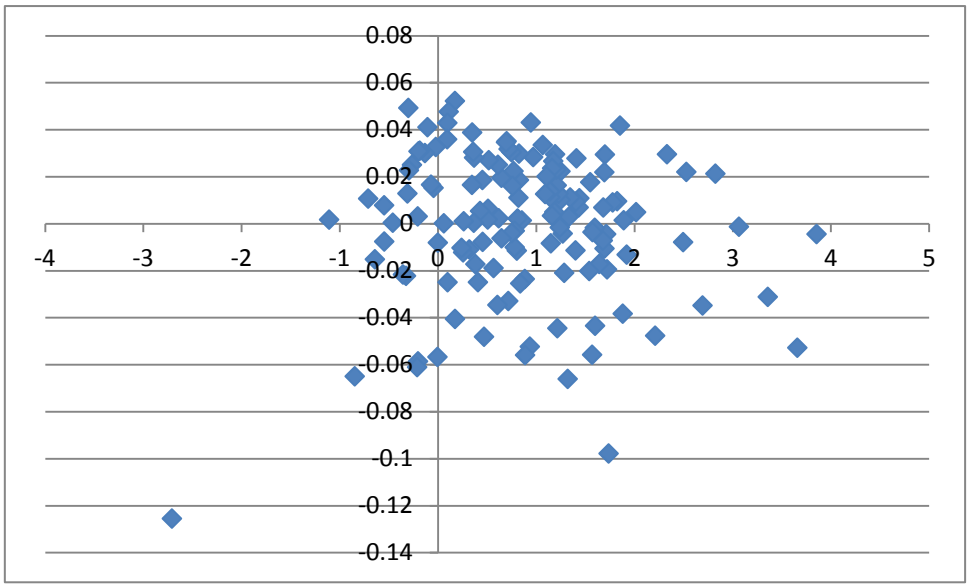


Figure 95. M-M Bland-Altman plot with systematic bias removed

SSA-HSC

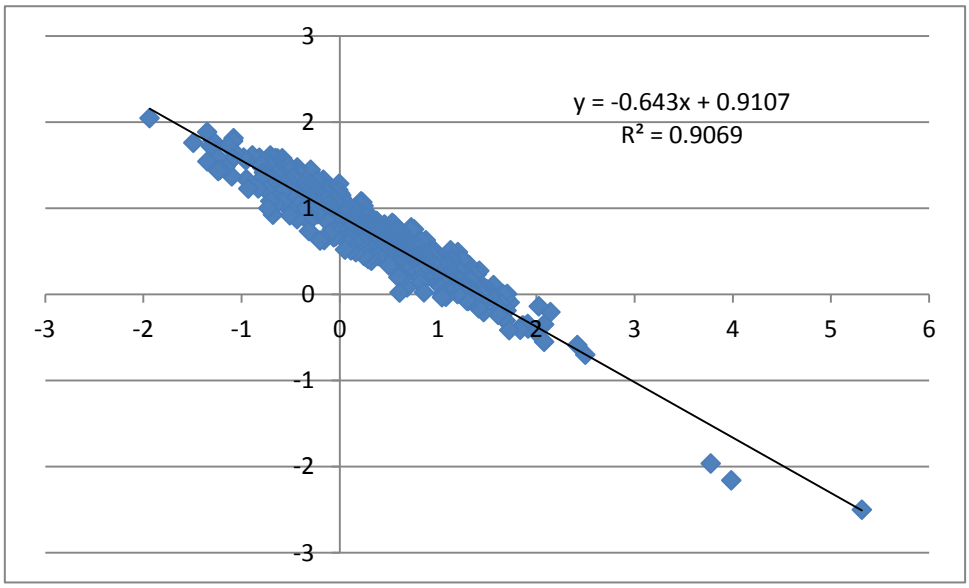


Figure 96. E-E Bland-Altman plot

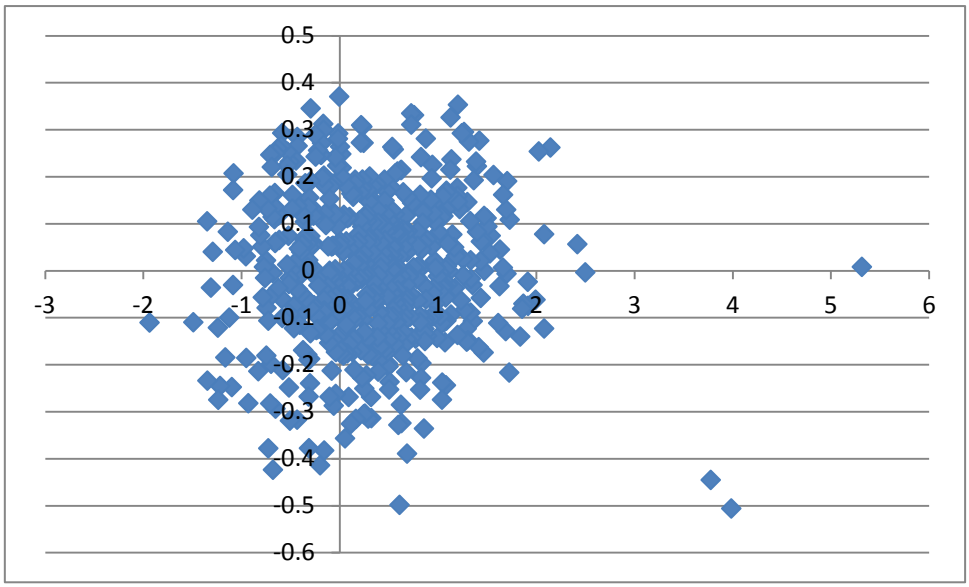


Figure 97. E-E Bland-Altman plot with systematic bias removed

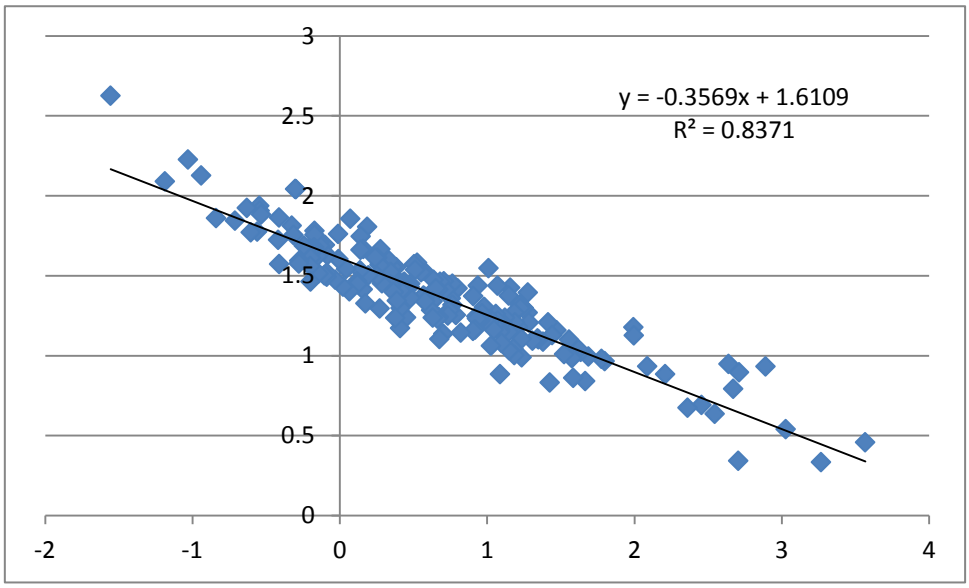


Figure 98. M-M Bland-Altman plot

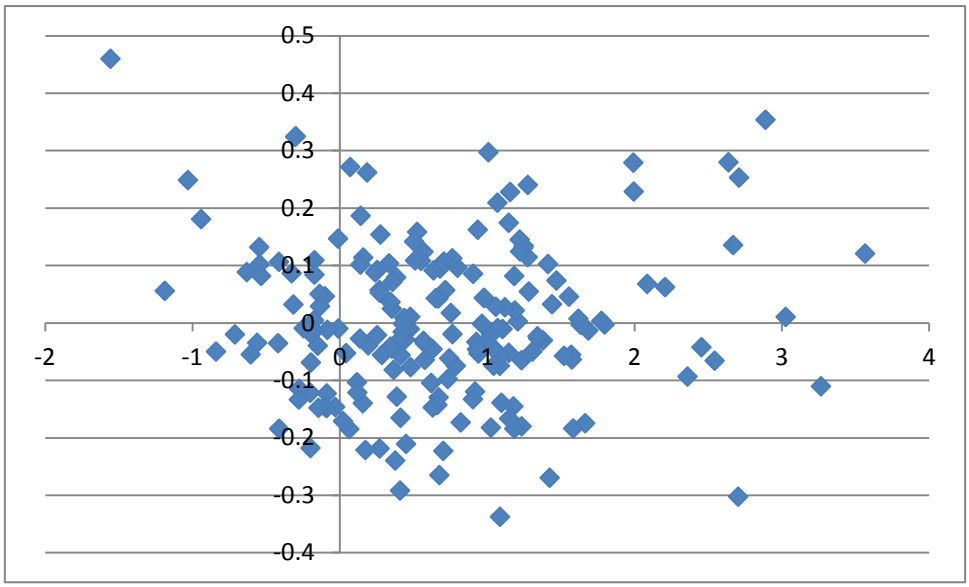


Figure 99. M-M Bland-Altman plot with systematic bias removed

JSA-SSA

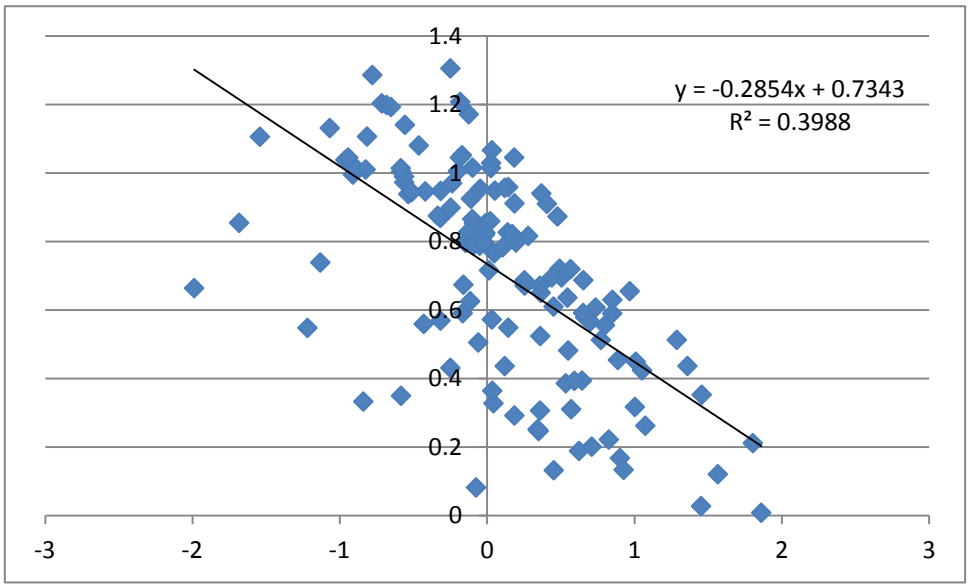


Figure 100. E-E Bland-Altman plot

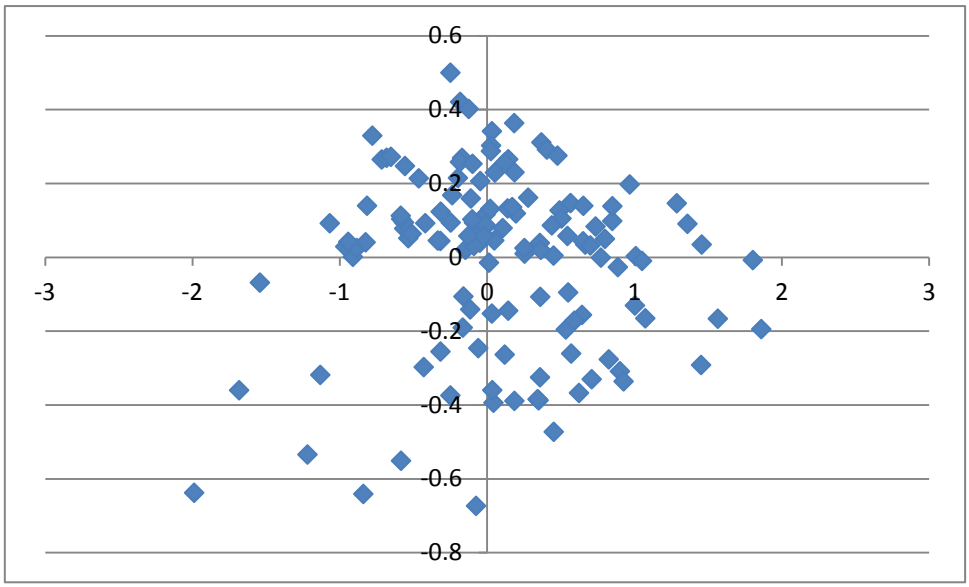


Figure 101. E-E Bland-Altman plot with systematic bias removed

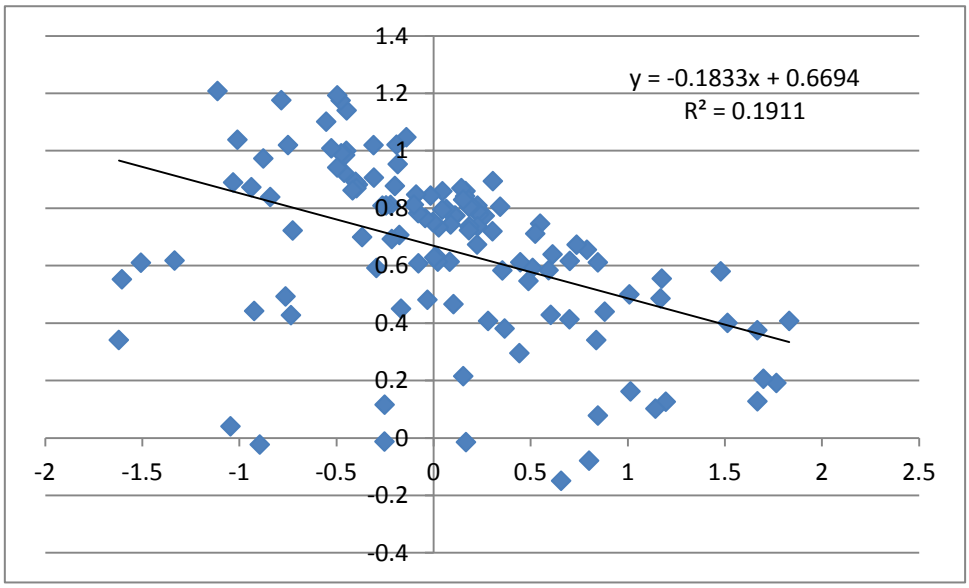


Figure 102. M-M Bland-Altman plot

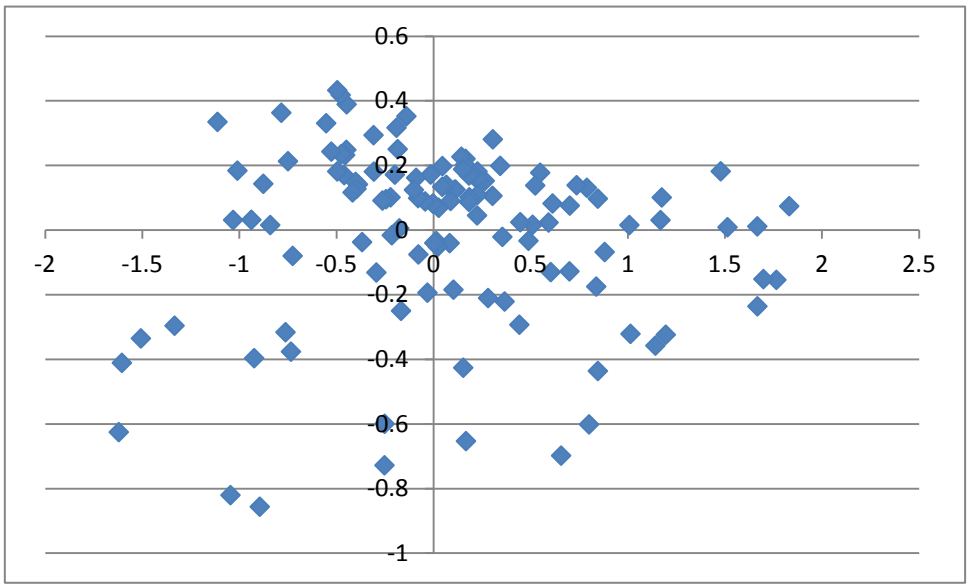


Figure 103. M-M Bland-Altman plot with systematic bias removed

OLSAT-SEM 1 2015

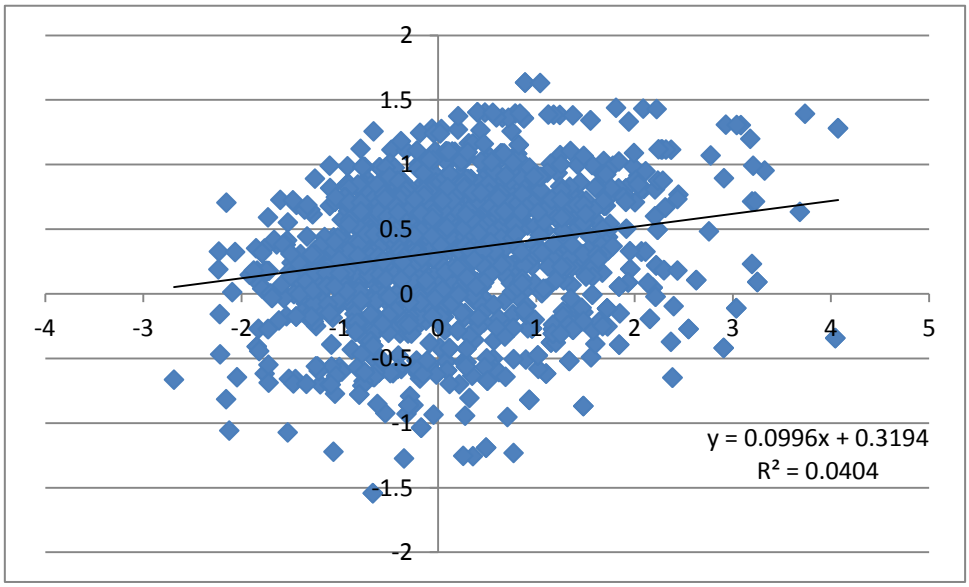


Figure 104. SAI-SA Bland-Altman plot

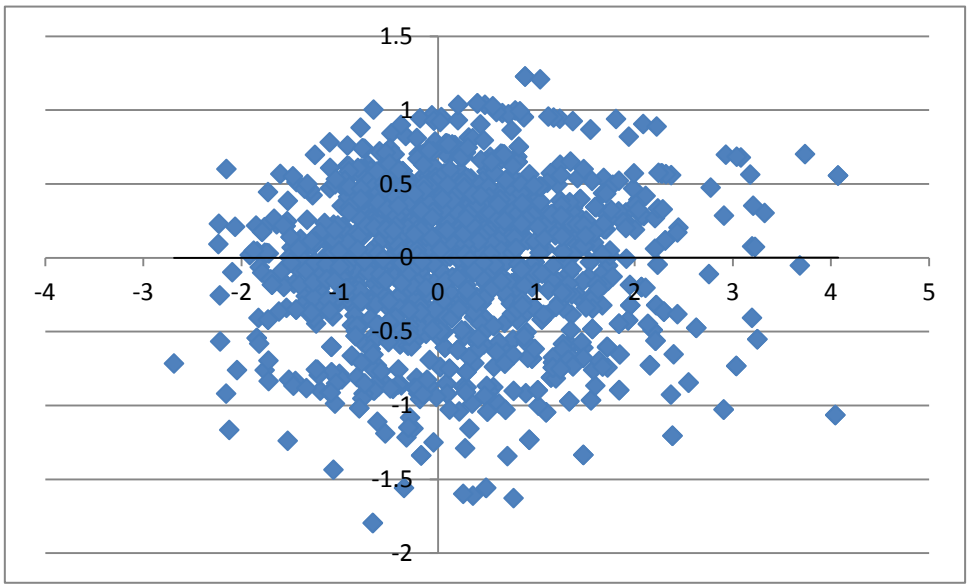


Figure 105. SAI-SA Bland-Altman plot with systematic bias removed

Appendix 11 Multiple Regression Assumptions Test Results

% GUA: ABSI

Correlations

		ABSIprop	gradient	intercept	correlation
Pearson Correlation	ABSIprop	1.000	-.397	-.663	-.735
	gradient	-.397	1.000	-.126	.753
	intercept	-.663	-.126	1.000	.425
	correlation	-.735	.753	.425	1.000
Sig. (1-tailed)	ABSIprop	.	.005	.000	.000
	gradient	.005	.	.216	.000
	intercept	.000	.216	.	.003
	correlation	.000	.000	.003	.
N	ABSIprop	41	41	41	41
	gradient	41	41	41	41
	intercept	41	41	41	41
	correlation	41	41	41	41

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.835[a]	.696	.672	.13061	.696	28.295	3	37	.000	2.902

a. Predictors: (Constant), correlation, intercept, gradient

b. Dependent Variable: ABSIprop

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	.625	.069		9.000	.000	.485	.766					
	gradient	-.175	.201	-.181	-.872	.389	-.581	.231	-.397	-.142	-.079	.190	5.261
	intercept	-.220	.063	-.526	-3.486	.001	-.348	-.092	-.663	-.497	-.316	.360	2.780
	correlation	-.423	.257	-.375	-1.647	.108	-.944	.098	-.735	-.261	-.149	.158	6.318

a. Dependent Variable: ABSIprop

Casewise Diagnostics[a]

Case Number	Std. Residual	ABSIprop	Predicted Value	Residual
10	-2.574	.38	.7162	-.33621
12	-2.319	.41	.7129	-.30294

a. Dependent Variable: ABSIprop

Case Summaries[a]

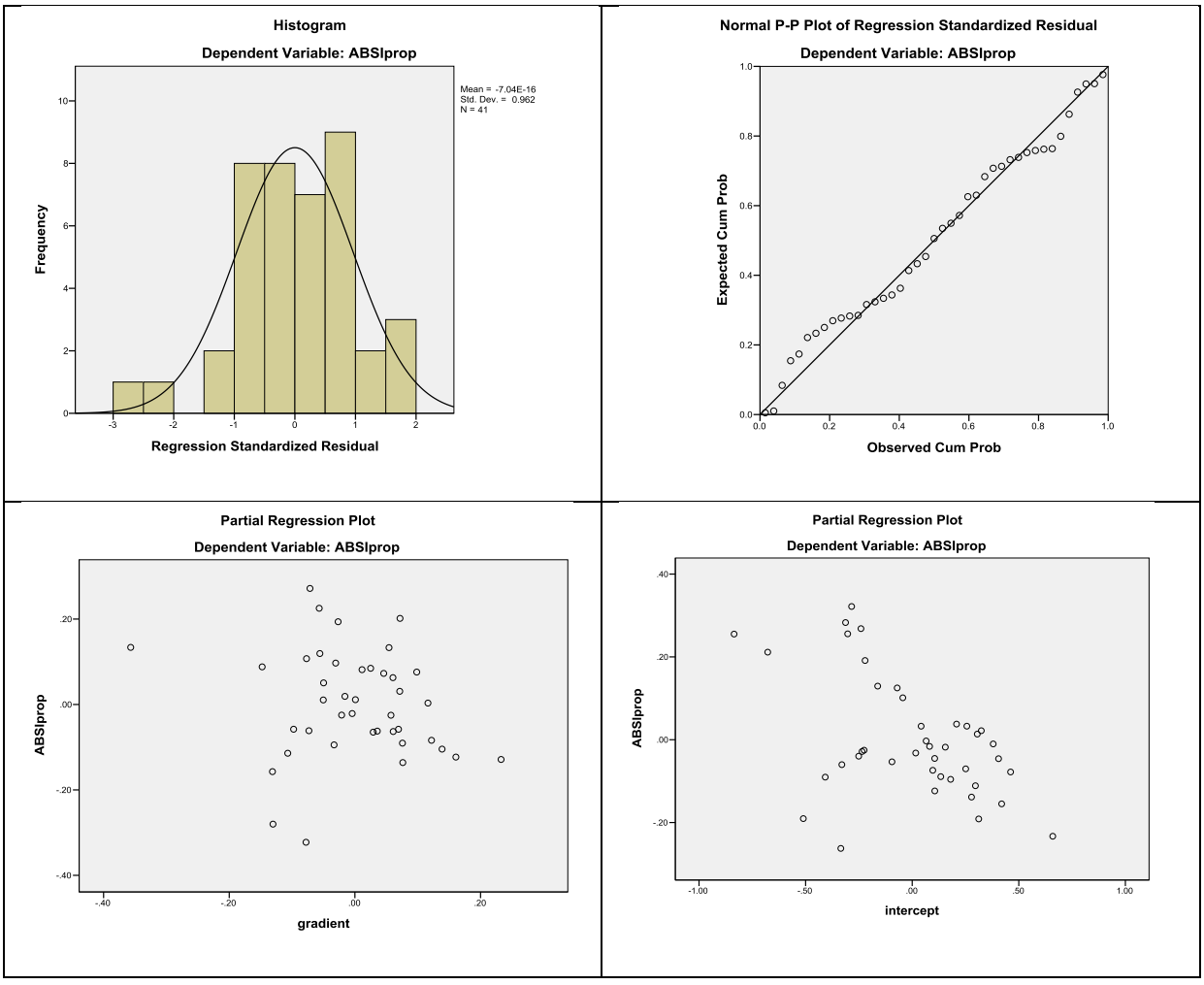
		Standardized Residual	Mahalanobis Distance	Cook's Distance	Centered Leverage Value	COVRATIO	Standardized DFFIT	Standardized DFBETA Intercept	Standardized DFBETA gradient	Standardized DFBETA intercept	Standardized DFBETA correlation
1		.32067	1.09419	.00148	.02735	1.16298	.07599	.00682	-.02557	.00676	.02739
2		-.47902	2.57837	.00614	.06446	1.19163	-.15510	-.02147	-.09743	-.10558	.06709
3		.12434	.98591	.00021	.02465	1.17131	.02857	.00840	-.00315	.00969	.00275
4		-.35110	1.39436	.00206	.03486	1.16939	-.08977	-.01365	-.03930	-.05289	.02437
5		.01400	1.31224	.00000	.03281	1.18349	.00350	.00070	-.00113	.00058	.00104
6		-.11572	1.69911	.00026	.04248	1.19393	-.03164	.00120	-.01078	-.01485	.00349
7		-.76844	1.33083	.00959	.03327	1.10590	-.19481	-.01577	.04131	-.04156	-.05273
8		-.16824	1.19771	.00043	.02994	1.17612	-.04092	.00422	.00116	-.00896	-.00881
9		1.98452	2.38320	.09853	.05958	.74320	.65869	.27650	-.24964	-.31234	.07737
10		-2.57421	3.51754	.23616	.08794	.51036	-1.07305	-.44915	.38259	.51827	-.09250
11		1.64879	2.16407	.06282	.05410	.86847	.51543	.24161	-.16047	-.21385	.02307
12		-2.31947	4.34582	.23806	.10865	.61756	-1.05508	-.21632	.57912	.71427	-.34936
13		-.21840	1.96019	.00102	.04900	1.19752	-.06303	.02254	.00754	-.00185	-.02442
14		-.57402	3.45152	.01153	.08629	1.20519	-.21288	.11652	.09584	.06980	-.14985
15		1.44770	1.89141	.04358	.04729	.93439	.42500	.05436	-.06515	-.23226	.03244
16		-1.37937	4.60925	.08972	.11523	1.01355	-.60940	-.24443	.32861	.32161	-.14508
17		1.09209	3.00162	.03655	.07504	1.07093	.38412	-.03222	.10156	-.13018	-.05864
18		-1.01721	3.51558	.03686	.08789	1.10590	-.38484	-.17282	.18850	.18324	-.07123
19		.08731	.15373	.00006	.00384	1.14727	.01489	.00353	.00011	.00280	.00059
20		.70127	.41729	.00460	.01043	1.09371	.13467	-.01591	-.03390	-.01535	.05295
21		-.42920	.33999	.00162	.00850	1.13020	-.07959	-.03457	-.02403	-.03824	.02282
22		.71840	.77957	.00619	.01949	1.10041	.15640	-.02474	-.08828	-.05844	.09869
23		.56214	3.84728	.01232	.09618	1.22024	.22001	.21688	.05891	.11639	-.13394
24		.61878	1.70454	.00737	.04261	1.14378	.17029	.15713	.04636	.09710	-.09351
25		.33192	3.82750	.00427	.09569	1.25102	.12915	.12704	.04086	.07300	-.08317
26		.68248	1.65500	.00877	.04137	1.13129	.18605	.17123	.02790	.08992	-.08459
27		-.40295	1.61701	.00301	.04043	1.17092	-.10846	.03275	-.03990	-.01988	.00347
28		.54587	13.12408	.06263	.32810	1.63912	.49681	-.18930	-.45854	-.33798	.46577
29		.18115	2.95829	.00099	.07396	1.23267	.06218	.05596	.03540	.04424	-.04898
30		.63943	1.17259	.00613	.02931	1.12503	.15537	.13456	.01169	.06743	-.05572
31		-.56837	2.59854	.00870	.06496	1.17899	-.18492	.07438	.06953	.02838	-.11025
32		-.67403	6.45370	.03182	.16134	1.28954	-.35457	-.05100	-.29396	-.26224	.22139
33		.71315	1.27321	.00803	.03183	1.11491	.17803	.13045	.11348	.11767	-.13156
34		.83873	.37588	.00636	.00940	1.06660	.15896	.03835	-.07404	-.02944	.05989
35		1.63978	5.67564	.16081	.14189	.92927	.82802	-.33260	.22390	-.30691	-.00872
36		-.61432	8.73327	.03993	.21833	1.39568	-.39691	.05889	-.17164	.09161	.10901
37		.47703	6.72343	.01679	.16809	1.34017	.25662	-.19067	-.13258	-.19165	.20209
38		-.72753	11.71665	.09009	.29292	1.50170	-.59843	.24458	-.26225	.12893	.08671
39		-.45721	.19544	.00162	.00489	1.12295	-.07972	-.02989	-.02085	-.03019	.01809
40		-.93929	.94600	.01169	.02365	1.05904	-.21605	-.14851	-.11539	-.14870	.13266
41		-.59145	1.27846	.00553	.03196	1.13578	-.14750	-.12567	-.07223	-.08363	.10172
Total	N	41	41	41	41	41	41	41	41	41	41

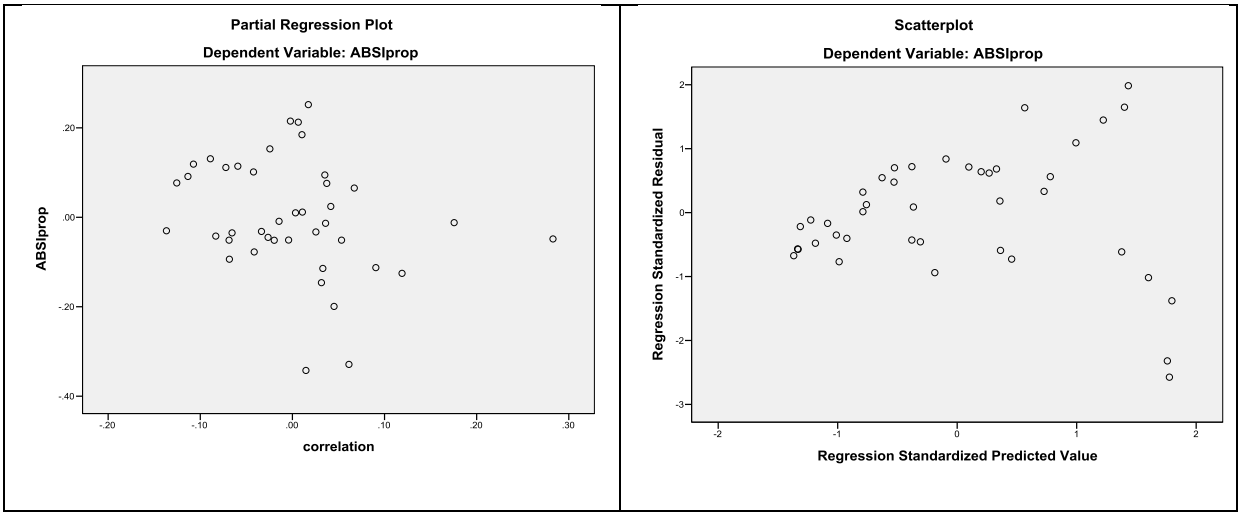
a. Limited to first 100 cases.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	.1180	.7202	.3783	.19026	41
Std. Predicted Value	-1.368	1.797	.000	1.000	41
Standard Error of Predicted Value	.022	.078	.039	.013	41
Adjusted Predicted Value	.1308	.7594	.3796	.19411	41
Residual	-.33621	.25919	.00000	.12561	41
Std. Residual	-2.574	1.985	.000	.962	41
Stud. Residual	-2.732	2.073	-.005	1.021	41
Deleted Residual	-.37876	.28295	-.00134	.14186	41
Stud. Deleted Residual	-3.016	2.176	-.011	1.063	41
Mahal. Distance	.154	13.124	2.927	2.897	41
Cook's Distance	.000	.238	.034	.058	41
Centered Leverage Value	.004	.328	.073	.072	41

a. Dependent Variable: ABSlprop





%GUA ABSII

Correlations

		ABSIIprop	gradient	intercept	correlation
Pearson Correlation	ABSIIprop	1.000	-.503	-.019	-.611
	gradient	-.503	1.000	-.164	.755
	intercept	-.019	-.164	1.000	.386
	correlation	-.611	.755	.386	1.000
Sig. (1-tailed)	ABSIIprop	.	.001	.456	.000
	gradient	.001	.	.166	.000
	intercept	.456	.166	.	.009
	correlation	.000	.000	.009	.
N	ABSIIprop	37	37	37	37
	gradient	37	37	37	37
	intercept	37	37	37	37
	correlation	37	37	37	37

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.677[a]	.458	.409	.18286	.458	9.310	3	33	.000	1.385

a. Predictors: (Constant), correlation, intercept, gradient

b. Dependent Variable: ABSIIprop

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	.875	.098		8.950	.000	.676	1.074					
	gradient	.390	.289	.401	1.349	.187	-.198	.978	-.503	.229	.173	.185	5.399
	intercept	.205	.092	.471	2.224	.033	.017	.392	-.019	.361	.285	.367	2.727
	correlation	-1.257	.365	-1.096	-3.442	.002	-2.001	-.514	-.611	-.514	-.441	.162	6.176

a. Dependent Variable: ABSIIprop

Casewise Diagnostics[a]

Case Number	Std. Residual	ABSIIprop	Predicted Value	Residual
15	2.108	.76	.3745	.38552

a. Dependent Variable: ABSIIprop

Case Summaries[a]

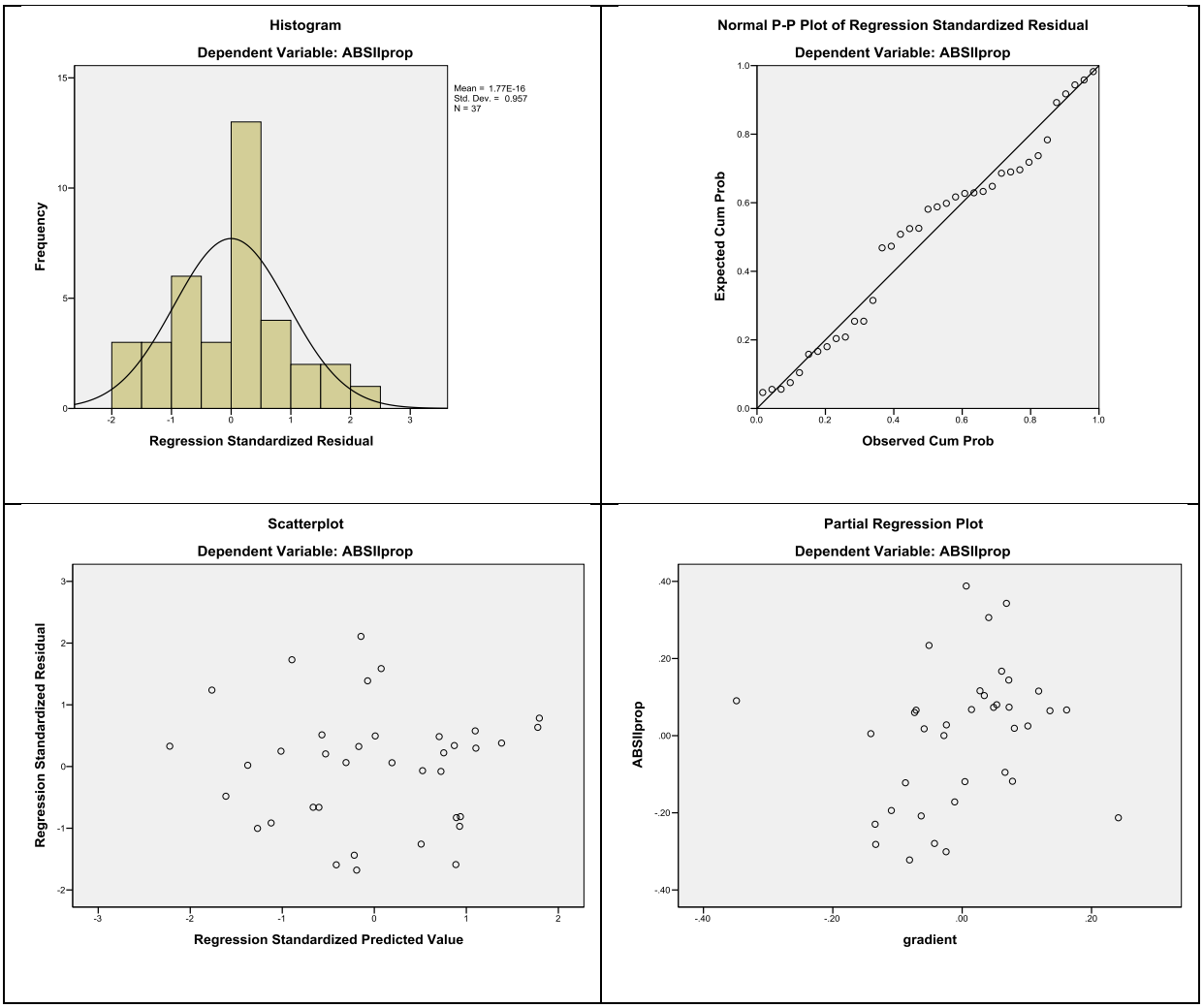
		Standardized Residual	Mahalanobis Distance	Cook's Distance	Centered Leverage Value	COVRATIO	Standardized DFFIT	Standardized DFBETA Intercept	Standardized DFBETA gradient	Standardized DFBETA intercept	Standardized DFBETA correlation
1		-1.43700	1.45282	.03999	.04036	.91854	-.40778	-.08873	.10619	-.09367	-.10531
2		-.66126	2.01566	.01079	.05599	1.16362	-.20610	-.00117	-.07491	-.10733	.02793
3		-1.59284	1.53063	.05095	.04252	.86087	-.46414	-.05147	.06915	-.13070	-.10466
4		-.66061	1.42148	.00833	.03949	1.14436	-.18101	.00976	-.00469	-.05348	-.02996
5		.48465	2.11150	.00602	.05865	1.19890	.15338	.06221	-.06092	-.07470	.02101
6		-1.58889	3.18919	.09330	.08859	.89052	-.62940	-.25098	.23790	.31667	-.06789
7		.22221	1.89723	.00116	.05270	1.22099	.06719	.03064	-.02210	-.02869	.00427
8		-1.25681	4.01925	.07381	.11165	1.04463	-.55059	-.10243	.31226	.38066	-.19247
9		-.91482	2.13020	.02160	.05917	1.10589	-.29354	.08624	.01830	-.03530	-.09669
10		-.48147	3.36944	.00904	.09360	1.24551	-.18800	.09342	.07500	.04536	-.12362
11		.06134	1.62957	.00008	.04527	1.21852	.01751	.00208	-.00290	-.00960	.00150
12		-.96897	4.24700	.04656	.11797	1.15531	-.43221	-.16439	.24134	.23654	-.11257
13		.32455	2.63109	.00326	.07309	1.23907	.11257	-.00979	.03002	-.03652	-.01791
14		-.82806	3.19056	.02535	.08863	1.16287	-.31732	-.13685	.16142	.15656	-.06575
15		2.10824	.25109	.04049	.00697	.64214	.42719	.10789	.02293	.11317	.00498
16		.20465	.48514	.00046	.01348	1.17250	.04230	-.00319	-.00810	-.00066	.01453
17		1.58664	.47076	.02739	.01308	.84601	.33974	.14616	.10997	.17906	-.09783
18		.51317	.76210	.00350	.02117	1.14890	.11705	-.01463	-.06002	-.03371	.06950
19		.78369	3.46855	.02465	.09635	1.18403	.31255	.30767	.08581	.16854	-.18821
20		.29734	1.60646	.00184	.04462	1.20427	.08455	.07711	.02424	.05015	-.04562
21		.63459	3.46558	.01615	.09627	1.21970	.25204	.24739	.08172	.14536	-.16074
22		.57712	1.53689	.00671	.04269	1.16383	.16219	.14792	.02673	.08216	-.07236
23		1.72998	1.73522	.06582	.04820	.80931	.53202	-.13138	.20977	.14032	-.03542
24		1.23827	12.18060	.34776	.33835	1.31479	1.20642	-.44417	-1.10044	-.78776	1.12589
25		.37974	2.78238	.00469	.07729	1.23824	.13517	.12046	.07822	.09855	-.10531
26		.33958	1.11247	.00188	.03090	1.18281	.08559	.07319	.00811	.03964	-.03005
27		-1.00300	2.67129	.03152	.07420	1.09613	-.35573	.12298	.11192	.02228	-.19218
28		-1.67871	6.81164	.24800	.18921	.90964	-1.03903	-.16953	-.85252	-.79211	.64129
29		-.07920	1.26460	.00011	.03513	1.20496	-.02074	-.01495	-.01334	-.01428	.01505
30		1.38841	.37277	.01944	.01035	.91464	.28334	.07223	-.11870	-.03174	.10003
31		.24865	5.06986	.00375	.14083	1.34692	.12069	-.04820	.03377	-.04161	-.00296
32		.06381	7.92504	.00044	.22014	1.50132	.04150	-.00636	.01790	-.00925	-.01169
33		.32904	6.01456	.00809	.16707	1.38066	.17749	-.12993	-.08991	-.12586	.13843
34		.02020	10.65499	.00007	.29597	1.67045	.01670	-.00674	.00750	-.00315	-.00271
35		.49524	.30030	.00233	.00834	1.13673	.09545	.03601	.02744	.04174	-.02237
36		-.06727	1.03164	.00007	.02866	1.19698	-.01655	-.01107	-.00894	-.01180	.00987
37		-.81223	1.19045	.01122	.03307	1.10414	-.21086	-.17794	-.10571	-.12540	.14384
Total	N	37	37	37	37	37	37	37	37	37	37

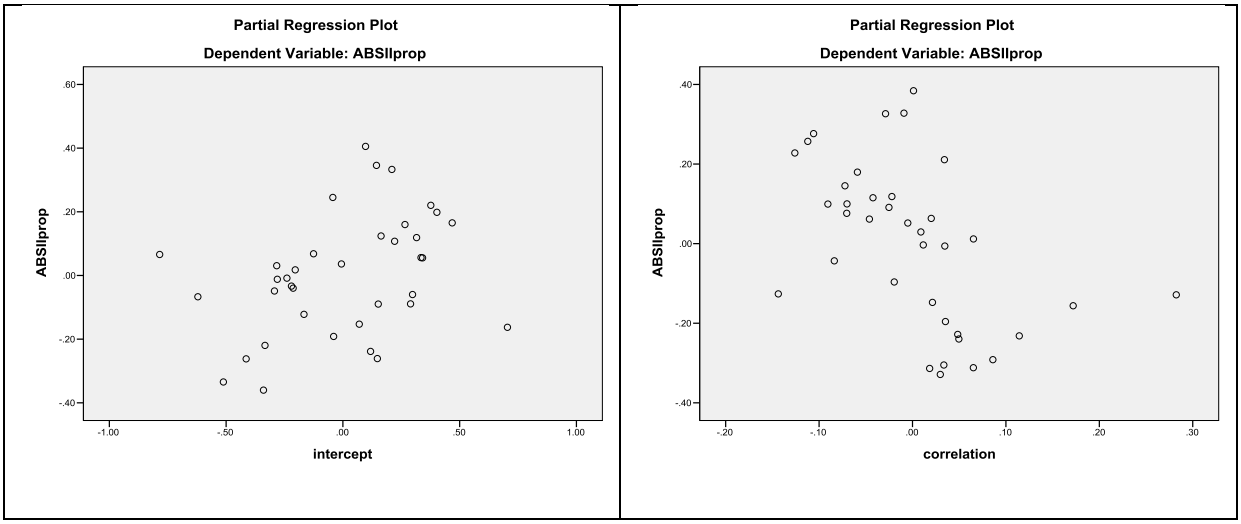
a. Limited to first 100 cases.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	.0398	.6867	.3978	.16107	37
Std. Predicted Value	-2.223	1.793	.000	1.000	37
Standard Error of Predicted Value	.034	.111	.057	.018	37
Adjusted Predicted Value	-.0168	.6676	.3984	.16939	37
Residual	-.30697	.38552	.00000	.17508	37
Std. Residual	-1.679	2.108	.000	.957	37
Stud. Residual	-1.896	2.145	-.002	1.014	37
Deleted Residual	-.39166	.39909	-.00052	.19769	37
Stud. Deleted Residual	-1.978	2.277	-.001	1.039	37
Mahal. Distance	.251	12.181	2.919	2.738	37
Cook's Distance	.000	.348	.034	.069	37
Centered Leverage Value	.007	.338	.081	.076	37

a. Dependent Variable: ABSIIprop





%GUA: SD

Correlations

		propSD	gradient	intercept	correlation
Pearson Correlation	propSD	1.000	-.452	-.618	-.724
	gradient	-.452	1.000	-.126	.753
	intercept	-.618	-.126	1.000	.425
	correlation	-.724	.753	.425	1.000
Sig. (1-tailed)	propSD	.	.002	.000	.000
	gradient	.002	.	.216	.000
	intercept	.000	.216	.	.003
	correlation	.000	.000	.003	.
N	propSD	41	41	41	41
	gradient	41	41	41	41
	intercept	41	41	41	41
	correlation	41	41	41	41

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.820[a]	.672	.645	.13560	.672	25.274	3	37	.000	2.603

a. Predictors: (Constant), correlation, intercept, gradient

b. Dependent Variable: propSD

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	.613	.072		8.496	.000	.467	.759					
	gradient	-.385	.208	-.399	-1.850	.072	-.807	.037	-.452	-.291	-.174	.190	5.261
	intercept	-.249	.066	-.596	-3.799	.001	-.382	-.116	-.618	-.530	-.358	.360	2.780
	correlation	-.191	.267	-.170	-.717	.478	-.732	.349	-.724	-.117	-.068	.158	6.318

a. Dependent Variable: propSD

Casewise Diagnostics[a]

Case Number	Std. Residual	propSD	Predicted Value	Residual
10	-2.492	.38	.7179	-.33786
12	-2.093	.44	.7238	-.28379

a. Dependent Variable: propSD

Case Summaries[a]

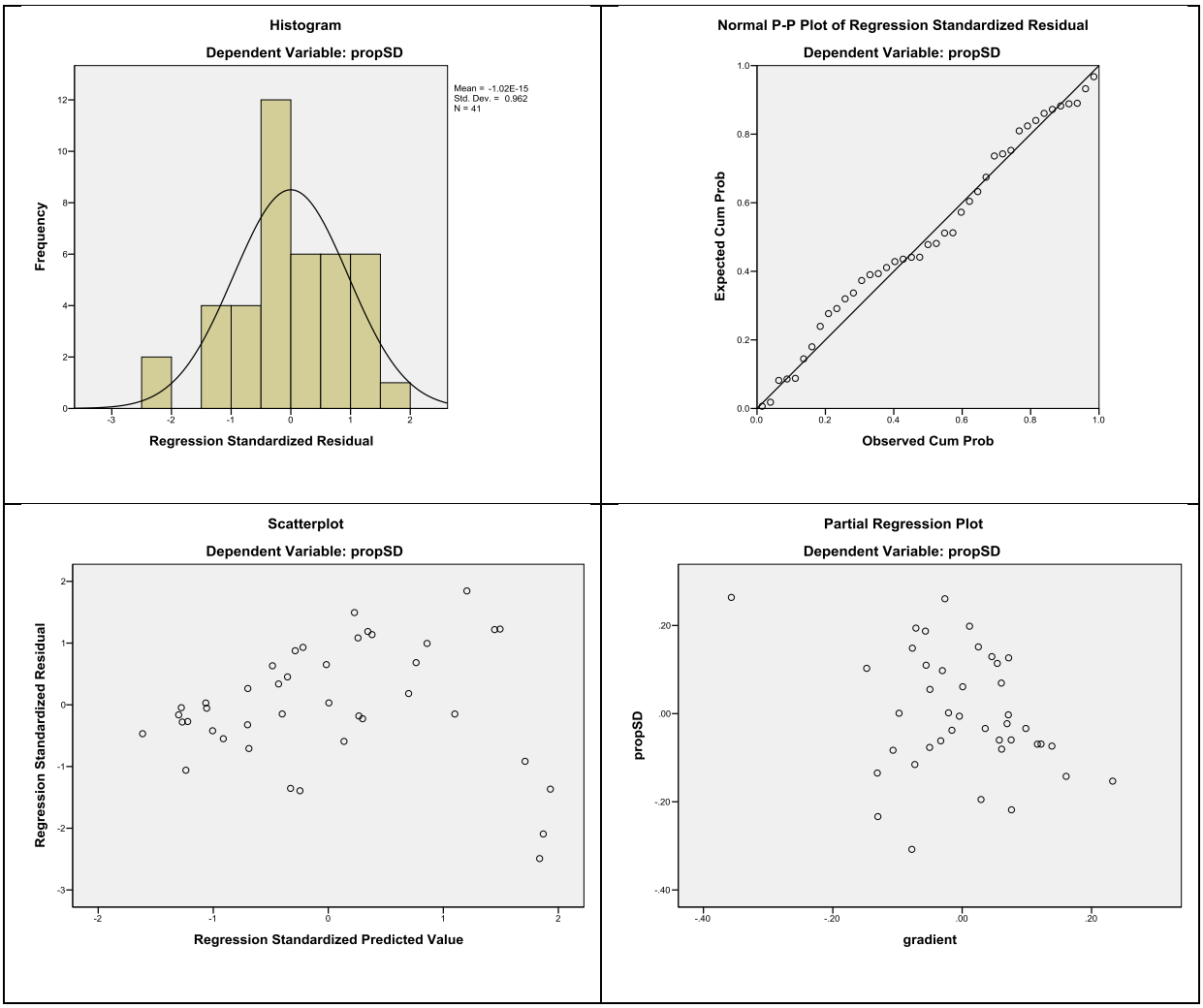
		Standardized Residual	Mahalanobis Distance	Cook's Distance	Centered Leverage Value	COVRATIO	Standardized DFFIT	Standardized DFBETA Intercept	Standardized DFBETA gradient	Standardized DFBETA intercept	Standardized DFBETA correlation
1		.26376	1.09419	.00100	.02735	1.16741	.06247	.00561	-.02102	.00556	.02252
2		-.16267	2.57837	.00071	.06446	1.22079	-.05251	-.00727	-.03298	-.03575	.02271
3		-.32350	.98591	.00142	.02465	1.15947	-.07442	-.02188	.00819	-.02524	-.00717
4		.02888	1.39436	.00001	.03486	1.18599	.00737	.00112	.00323	.00434	-.00200
5		-.70823	1.31224	.00807	.03281	1.11690	-.17850	-.03555	.05767	-.02943	-.05288
6		-.27928	1.69911	.00150	.04248	1.18502	-.07643	.00289	-.02605	-.03587	.00843
7		-.54958	1.33083	.00490	.03327	1.14360	-.13874	-.01123	.02942	-.02960	-.03755
8		-.05525	1.19771	.00005	.02994	1.17952	-.01343	.00138	.00038	-.00294	-.00289
9		1.22708	2.38320	.03767	.05958	1.01565	.39169	.16442	-.14845	-.18573	.04601
10		-2.49166	3.51754	.22126	.08794	.54371	-1.03046	-.43132	.36740	.49770	-.08883
11		1.21847	2.16407	.03431	.05410	1.01335	.37363	.17514	-.11632	-.15502	.01672
12		-2.09288	4.34582	.19382	.10865	.71540	-.93467	-.19163	.51303	.63275	-.30949
13		-.04681	1.96019	.00005	.04900	1.20390	-.01350	.00483	.00162	-.00040	-.00523
14		-.27067	3.45152	.00256	.08629	1.24356	-.09999	.05473	.04502	.03279	-.07038
15		1.84531	1.89141	.07080	.04729	.79165	.55306	.07075	-.08478	-.30225	.04221
16		-1.36654	4.60925	.08806	.11523	1.01833	-.60337	-.24202	.32536	.31844	-.14364
17		.99427	3.00162	.03030	.07504	1.09840	.34861	-.02924	.09218	-.11814	-.05322
18		-.91613	3.51558	.02990	.08789	1.13332	-.34554	-.15518	.16925	.16453	-.06395
19		.45213	.15373	.00153	.00384	1.12235	.07733	.01833	.00059	.01455	.00305
20		.63208	.41729	.00373	.01043	1.10521	.12122	-.01432	-.03052	-.01382	.04767
21		-.14799	.33999	.00019	.00850	1.15095	-.02738	-.01189	-.00827	-.01316	.00785
22		.87618	.77957	.00921	.01949	1.06898	.19144	-.03028	-.10806	-.07153	.12080
23		.68300	3.84728	.01818	.09618	1.19760	.26794	.26413	.07174	.14175	-.16312
24		1.08347	1.70454	.02259	.04261	1.04139	.30168	.27838	.08213	.17203	-.16567
25		.18231	3.82750	.00129	.09569	1.26293	.07086	.06970	.02242	.04005	-.04563
26		1.18702	1.65500	.02654	.04137	1.01122	.32816	.30202	.04921	.15860	-.14920
27		-.42113	1.61701	.00329	.04043	1.16889	-.11338	.03424	-.04171	-.02078	.00362
28		.93062	13.12408	.18203	.32810	1.48727	.85733	-.32667	-.79129	-.58325	.80377
29		-.18078	2.95829	.00099	.07396	1.23269	-.06205	-.05585	-.03533	-.04415	.04888
30		1.49456	1.17259	.03349	.02931	.90584	.37313	.32315	.02807	.16194	-.13381
31		-1.05996	2.59854	.03026	.06496	1.06988	-.34907	.14041	.13125	.05357	-.20811
32		-.46847	6.45370	.01537	.16134	1.33085	-.24547	-.03531	-.20351	-.18155	.15327
33		.03095	1.27321	.00002	.03183	1.18217	.00767	.00562	.00489	.00507	-.00567
34		.65103	.37588	.00383	.00940	1.10105	.12290	.02965	-.05724	-.02276	.04630
35		1.13677	5.67564	.07728	.14189	1.12781	.56030	-.22506	.15151	-.20768	-.00590
36		-.14892	8.73327	.00235	.21833	1.46881	-.09560	.01419	-.04134	.02207	.02626
37		.33791	6.72343	.00843	.16809	1.36079	.18143	-.13481	-.09374	-.13549	.14288
38		-.59274	11.71665	.05980	.29292	1.54539	-.48581	.19855	-.21290	.10467	.07039
39		-1.35438	.19544	.01425	.00489	.93204	-.24173	-.09063	-.06323	-.09154	.05486
40		-1.39289	.94600	.02571	.02365	.93445	-.32543	-.22370	-.17381	-.22399	.19982
41		-.22533	1.27846	.00080	.03196	1.17560	-.05595	-.04767	-.02740	-.03172	.03859
Total	N	41	41	41	41	41	41	41	41	41	41

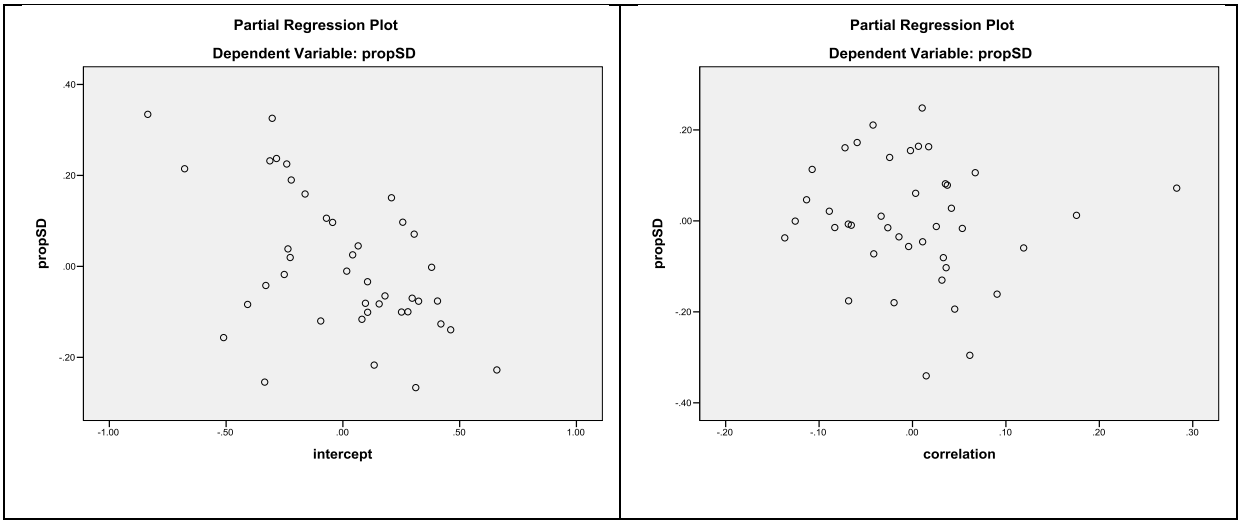
a. Limited to first 100 cases.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	.0735	.7353	.3749	.18669	41
Std. Predicted Value	-1.614	1.931	.000	1.000	41
Standard Error of Predicted Value	.023	.081	.040	.013	41
Adjusted Predicted Value	.0880	.7673	.3752	.19011	41
Residual	-.33786	.25022	.00000	.13041	41
Std. Residual	-2.492	1.845	.000	.962	41
Stud. Residual	-2.645	1.915	-.001	1.016	41
Deleted Residual	-.38062	.26954	-.00031	.14598	41
Stud. Deleted Residual	-2.897	1.990	-.008	1.048	41
Mahal. Distance	.154	13.124	2.927	2.897	41
Cook's Distance	.000	.221	.031	.053	41
Centered Leverage Value	.004	.328	.073	.072	41

a. Dependent Variable: propSD





Mean SD:

Correlations

		meanSD	gradient	intercept	correlation
Pearson Correlation	meanSD	1.000	-.548	-.743	-.830
	gradient	-.548	1.000	-.126	.753
	intercept	-.743	-.126	1.000	.425
	correlation	-.830	.753	.425	1.000
Sig. (1-tailed)	meanSD	.	.000	.000	.000
	gradient	.000	.	.216	.000
	intercept	.000	.216	.	.003
	correlation	.000	.000	.003	.
N	meanSD	41	41	41	41
	gradient	41	41	41	41
	intercept	41	41	41	41
	correlation	41	41	41	41

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.985[a]	.970	.968	.11656	.970	405.294	3	37	.000	2.108

a. Predictors: (Constant), correlation, intercept, gradient

b. Dependent Variable: meanSD

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	1.705	.062		27.482	.000	1.579	1.830					
	gradient	-1.965	.179	-.711	-10.976	.000	-2.327	-1.602	-.548	-.875	-.310	.190	5.261
	intercept	-1.033	.056	-.863	-18.329	.000	-1.147	-.919	-.743	-.949	-.518	.360	2.780
	correlation	.235	.229	.073	1.023	.313	-.230	.699	-.830	.166	.029	.158	6.318

a. Dependent Variable: meanSD

Casewise Diagnostics[a]

Case Number	Std. Residual	meanSD	Predicted Value	Residual
39	-2.534	.51	.8053	-.29534
40	-2.701	.50	.8148	-.31485

a. Dependent Variable: meanSD

Case Summaries[a]

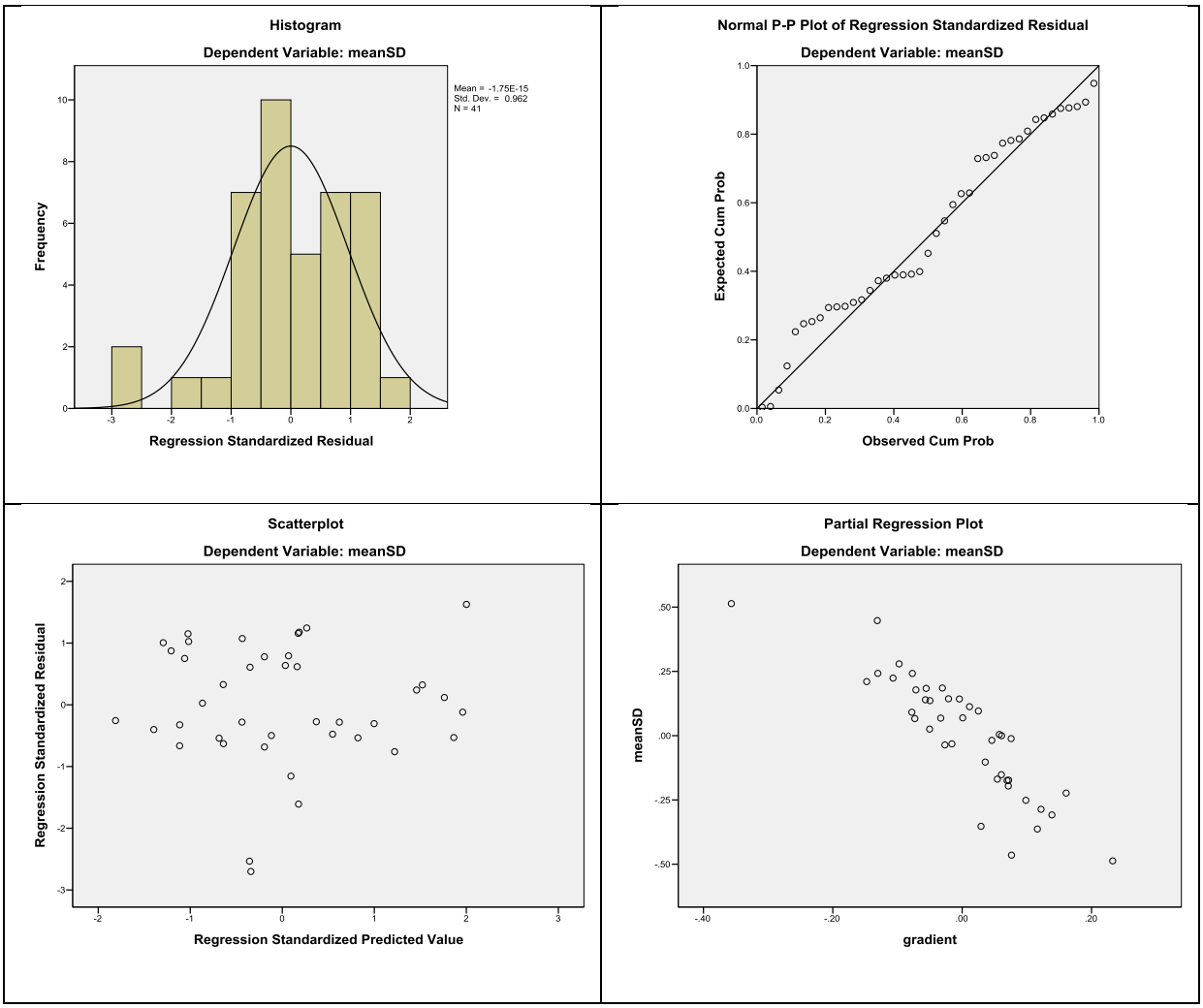
		Standardized Residual	Mahalanobis Distance	Cook's Distance	Centered Leverage Value	COVRATIO	Standardized DFFIT	Standardized DFBETA Intercept	Standardized DFBETA gradient	Standardized DFBETA intercept	Standardized DFBETA correlation
1		.32780	1.09419	.00155	.02735	1.16237	.07769	.00697	-.02614	.00691	.02800
2		-.40100	2.57837	.00430	.06446	1.20144	-.12971	-.01795	-.08147	-.08829	.05611
3		-.54121	.98591	.00397	.02465	1.13478	-.12483	-.03670	.01375	-.04234	-.01203
4		-.32476	1.39436	.00177	.03486	1.17179	-.08301	-.01262	-.03634	-.04891	.02253
5		-.62955	1.31224	.00638	.03281	1.13064	-.15842	-.03155	.05118	-.02612	-.04693
6		1.00620	1.69911	.01944	.04248	1.06158	.27917	-.01054	.09515	.13103	-.03078
7		.02631	1.33083	.00001	.03327	1.18401	.00661	.00054	-.00140	.00141	.00179
8		1.15092	1.19771	.02012	.02994	1.01115	.28528	-.02940	-.00808	.06246	.06141
9		.32241	2.38320	.00260	.05958	1.20324	.10076	.04230	-.03819	-.04778	.01183
10		-.53067	3.51754	.01004	.08794	1.21447	-.19849	-.08308	.07077	.09587	-.01711
11		.23922	2.16407	.00132	.05410	1.20276	.07180	.03366	-.02235	-.02979	.00321
12		-.11920	4.34582	.00063	.10865	1.28477	-.04948	-.01014	.02716	.03349	-.01638
13		.87418	1.96019	.01633	.04900	1.10038	.25497	-.09119	-.03051	.00750	.09879
14		.75101	3.45152	.01973	.08629	1.17085	.27953	-.15301	-.12585	-.09166	.19676
15		-.75967	1.89141	.01200	.04729	1.12321	-.21794	-.02788	.03341	.11911	-.01663
16		1.62715	4.60925	.12484	.11523	.91635	.72798	.29200	-.39255	-.38420	.17331
17		-.53521	3.00162	.00878	.07504	1.19696	-.18565	.01557	-.04909	.06292	.02834
18		.11873	3.51558	.00050	.08789	1.25480	.04421	.01986	-.02166	-.02105	.00818
19		.60896	.15373	.00277	.00384	1.10160	.10440	.02474	.00080	.01964	.00411
20		1.07377	.41729	.01077	.01043	1.01386	.20817	-.02460	-.05241	-.02373	.08185
21		-.28040	.33999	.00069	.00850	1.14367	-.05192	-.02255	-.01567	-.02495	.01489
22		.77782	.77957	.00726	.01949	1.08923	.16955	-.02682	-.09571	-.06335	.10699
23		-.28164	3.84728	.00309	.09618	1.25648	-.10983	-.10826	-.02941	-.05810	.06686
24		.61791	1.70454	.00735	.04261	1.14392	.17004	.15691	.04629	.09697	-.09338
25		-.47661	3.82750	.00881	.09569	1.23307	-.18579	-.18275	-.05878	-.10502	.11965
26		1.24487	1.65500	.02919	.04137	.99417	.34489	.31741	.05172	.16669	-.15680
27		1.02577	1.61701	.01949	.04043	1.05452	.27973	-.08448	.10290	.05128	-.00894
28		-1.60871	13.12408	.54395	.32810	1.09086	-1.54057	.58700	1.42189	1.04807	-1.44433
29		-1.15545	2.95829	.04038	.07396	1.05102	-.40459	-.36412	-.23035	-.28788	.31872
30		1.15808	1.17259	.02011	.02931	1.00860	.28527	.24705	.02146	.12381	-.10230
31		-.66370	2.59854	.01187	.06496	1.16248	-.21631	.08701	.08133	.03320	-.12897
32		-.25524	6.45370	.00456	.16134	1.35853	-.13340	-.01919	-.11059	-.09866	.08329
33		-.49800	1.27321	.00391	.03183	1.14907	-.12385	-.09075	-.07895	-.08186	.09153
34		.63727	.37588	.00367	.00940	1.10326	.12027	.02902	-.05602	-.02227	.04532
35		-.27462	5.67564	.00451	.14189	1.32533	-.13265	.05328	-.03587	.04917	.00140
36		-.30521	8.73327	.00986	.21833	1.45397	-.19619	.02911	-.08484	.04528	.05388
37		-.68350	6.72343	.03447	.16809	1.29737	-.36919	.27431	.19074	.27571	-.29074
38		.79360	11.71665	.10719	.29292	1.47744	.65411	-.26733	.28665	-.14093	-.09478
39		-2.53374	.19544	.04986	.00489	.52290	-.48611	-.18224	-.12715	-.18409	.11033
40		-2.70111	.94600	.09669	.02365	.46320	-.68894	-.47357	-.36797	-.47419	.42303
41		1.17721	1.27846	.02192	.03196	1.00561	.29809	.25397	.14597	.16900	-.20556
Total	N	41	41	41	41	41	41	41	41	41	41

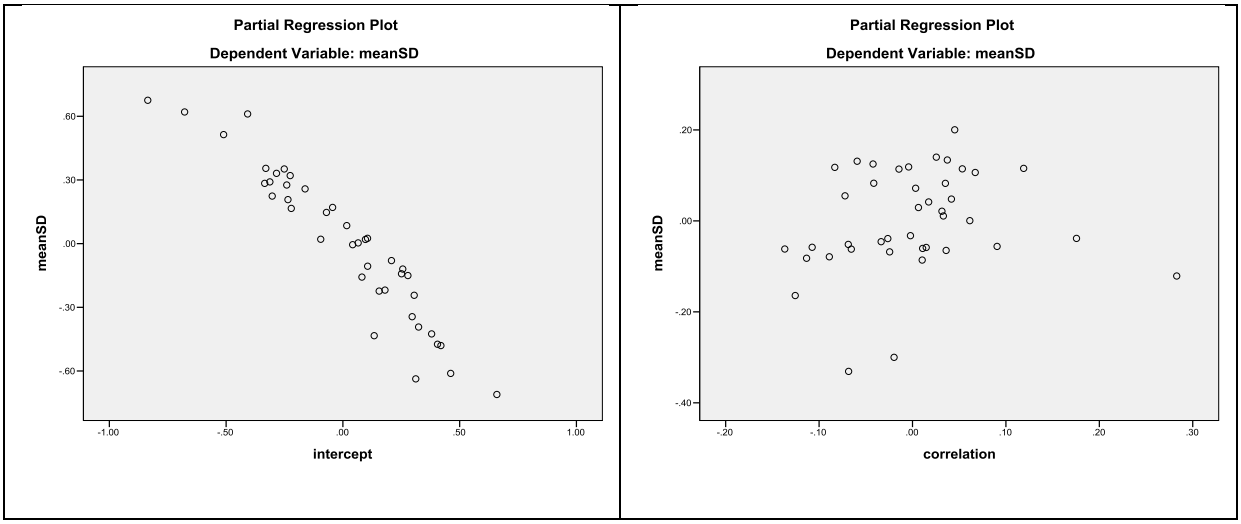
a. Limited to first 100 cases.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-.1302	2.3203	1.0341	.64265	41
Std. Predicted Value	-1.812	2.001	.000	1.000	41
Standard Error of Predicted Value	.020	.069	.035	.012	41
Adjusted Predicted Value	-.1235	2.2960	1.0360	.64283	41
Residual	-.31485	.18966	.00000	.11211	41
Std. Residual	-2.701	1.627	.000	.962	41
Stud. Residual	-2.768	1.754	-.007	1.013	41
Deleted Residual	-.33074	.22044	-.00186	.12531	41
Stud. Deleted Residual	-3.067	1.807	-.019	1.055	41
Mahal. Distance	.154	13.124	2.927	2.897	41
Cook's Distance	.000	.544	.032	.087	41
Centered Leverage Value	.004	.328	.073	.072	41

a. Dependent Variable: meanSD





Mean: REG

Correlations

		MeanREG	gradient	intercept	correlation
Pearson Correlation	MeanREG	1.000	-.138	.271	.140
	gradient	-.138	1.000	-.126	.753
	intercept	.271	-.126	1.000	.425
	correlation	.140	.753	.425	1.000
Sig. (1-tailed)	MeanREG	.	.194	.043	.192
	gradient	.194	.	.216	.000
	intercept	.043	.216	.	.003
	correlation	.192	.000	.003	.
N	MeanREG	41	41	41	41
	gradient	41	41	41	41
	intercept	41	41	41	41
	correlation	41	41	41	41

Model Summary[b]

					Change Statistics					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	R Square Change	F Change	df1	df2	Sig. F Change	Durbin-Watson
1	.401[a]	.160	.092	.12508	.160	2.357	3	37	.087	1.856

a. Predictors: (Constant), correlation, intercept, gradient

b. Dependent Variable: MeanREG

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	-.031	.067		-.470	.641	-.166	.104					
	gradient	-.374	.192	-.673	-1.948	.059	-.763	.015	-.138	-.305	-.293	.190	5.261
	intercept	-.026	.060	-.108	-.429	.670	-.148	.097	.271	-.070	-.065	.360	2.780
	correlation	.450	.246	.692	1.828	.076	-.049	.949	.140	.288	.275	.158	6.318

a. Dependent Variable: MeanREG

Casewise Diagnostics[a]

Case Number	Std. Residual	MeanREG	Predicted Value	Residual
39	-2.619	-.33	-.0024	-.32758
40	-2.342	-.31	-.0170	-.29297
41	2.250	.25	-.0314	.28140

a. Dependent Variable: MeanREG

Case Summaries[a]

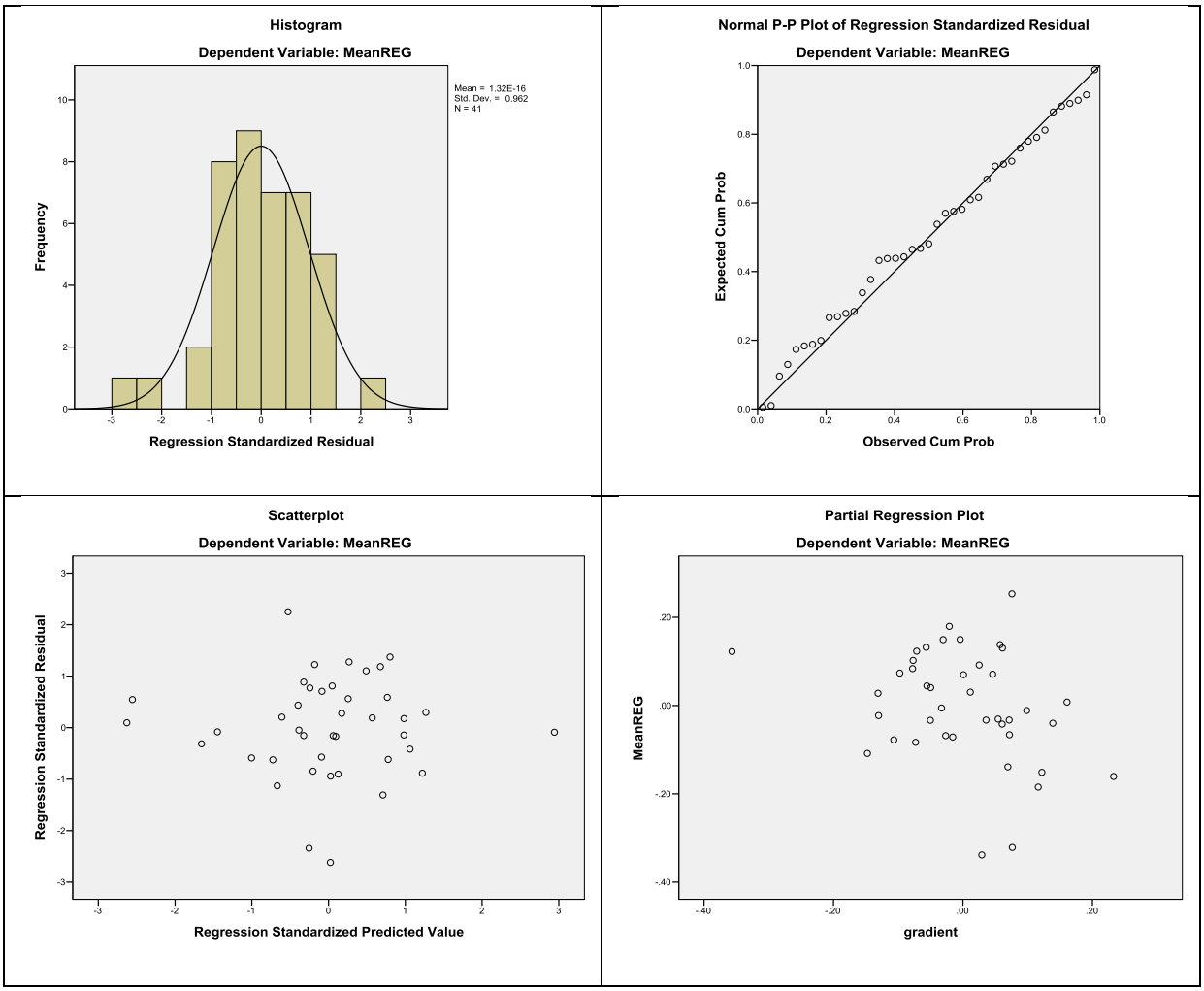
		Standardized Residual	Mahalanobis Distance	Cook's Distance	Centered Leverage Value	COVRATIO	Standardized DFFIT	Standardized DFBETA Intercept	Standardized DFBETA gradient	Standardized DFBETA intercept	Standardized DFBETA correlation
1		.17592	1.09419	.00045	.02735	1.17257	.04165	.00374	-.01401	.00371	.01501
2		-.84489	2.57837	.01910	.06446	1.12416	-.27557	-.03814	-.17310	-.18758	.11920
3		-.61716	.98591	.00516	.02465	1.12338	-.14253	-.04190	.01569	-.04834	-.01374
4		-.90257	1.39436	.01363	.03486	1.07890	-.23310	-.03544	-.10204	-.13733	.06328
5		-.41689	1.31224	.00280	.03281	1.16011	-.10457	-.02083	.03378	-.01724	-.03098
6		1.27586	1.69911	.03125	.04248	.98573	.35728	-.01350	.12177	.16769	-.03940
7		-.14337	1.33083	.00033	.03327	1.18131	-.03605	-.00292	.00764	-.00769	-.00976
8		1.18427	1.19771	.02130	.02994	1.00183	.29389	-.03029	-.00833	.06434	.06326
9		.77093	2.38320	.01487	.05958	1.13489	.24269	.10188	-.09198	-.11508	.02851
10		.43628	3.51754	.00678	.08794	1.22814	.16296	.06821	-.05810	-.07871	.01405
11		.88494	2.16407	.01810	.05410	1.10340	.26849	.12586	-.08359	-.11139	.01202
12		-.57165	4.34582	.01446	.10865	1.23540	-.23844	-.04889	.13088	.16142	-.07896
13		1.37147	1.96019	.04020	.04900	.96091	.40684	-.14551	-.04869	.01197	.15764
14		.29556	3.45152	.00306	.08629	1.24142	.10920	-.05978	-.04917	-.03581	.07687
15		-.62471	1.89141	.00811	.04729	1.14828	-.17873	-.02286	.02740	.09767	-.01364
16		-.16992	4.60925	.00136	.11523	1.29220	-.07283	-.02921	.03927	.03843	-.01734
17		-.08161	3.00162	.00020	.07504	1.23803	-.02819	.00236	-.00745	.00955	.00430
18		-.94105	3.51558	.03154	.08789	1.12678	-.35520	-.15951	.17398	.16913	-.06574
19		.56118	.15373	.00235	.00384	1.10854	.09613	.02278	.00074	.01809	.00379
20		1.10246	.41729	.01136	.01043	1.00655	.21392	-.02528	-.05385	-.02438	.08412
21		-.15595	.33999	.00021	.00850	1.15064	-.02886	-.01253	-.00871	-.01386	.00827
22		.58688	.77957	.00413	.01949	1.12225	.12745	-.02016	-.07194	-.04762	.08042
23		-.15418	3.84728	.00093	.09618	1.26511	-.06007	-.05922	-.01608	-.03178	.03657
24		.70491	1.70454	.00956	.04261	1.12857	.19431	.17931	.05290	.11081	-.10671
25		-.04830	3.82750	.00009	.09569	1.26773	-.01876	-.01846	-.00594	-.01061	.01208
26		.80896	1.65500	.01233	.04137	1.10646	.22114	.20352	.03316	.10688	-.10054
27		1.22453	1.61701	.02778	.04043	.99940	.33619	-.10152	.12367	.06163	-.01074
28		-.08875	13.12408	.00166	.32810	1.72100	-.08028	.03059	.07410	.05462	-.07527
29		-1.12861	2.95829	.03852	.07396	1.05909	-.39481	-.35533	-.22479	-.28092	.31102
30		.27719	1.17259	.00115	.02931	1.16884	.06703	.05805	.00504	.02909	-.02404
31		-.88441	2.59854	.02107	.06496	1.11543	-.28974	.11654	.10894	.04447	-.17274
32		-.58815	6.45370	.02423	.16134	1.30848	-.30883	-.04442	-.25604	-.22841	.19283
33		.20496	1.27321	.00066	.03183	1.17662	.05082	.03724	.03240	.03359	-.03756
34		.18961	.37588	.00033	.00940	1.15021	.03560	.00859	-.01658	-.00659	.01341
35		-.31386	5.67564	.00589	.14189	1.32136	-.15166	.06092	-.04101	.05621	.00160
36		.09531	8.73327	.00096	.21833	1.47156	.06117	-.00908	.02645	-.01412	-.01680
37		-1.30796	6.72343	.12624	.16809	1.09147	-.72191	.53639	.37297	.53912	-.56851
38		.54415	11.71665	.05040	.29292	1.55915	.44549	-.18207	.19523	-.09598	-.06455
39		-2.61889	.19544	.05327	.00489	.49248	-.50623	-.18978	-.13241	-.19171	.11490
40		-2.34221	.94600	.07270	.02365	.59547	-.57893	-.39795	-.30921	-.39847	.35548
41		2.24973	1.27846	.08007	.03196	.63202	.60371	.51436	.29564	.34228	-.41632
Total	N	41	41	41	41	41	41	41	41	41	41

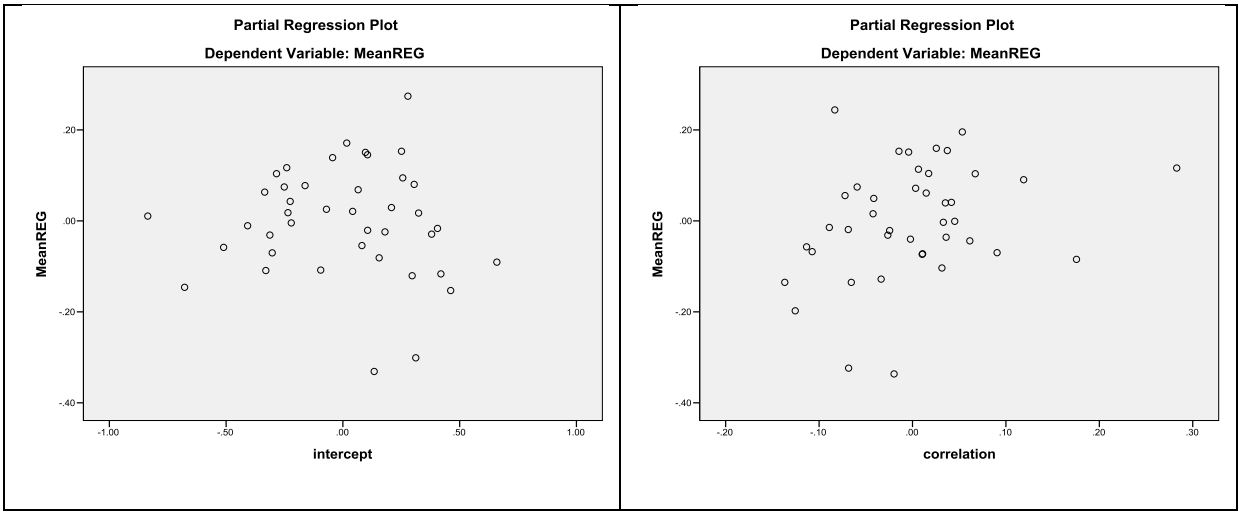
a. Limited to first 100 cases.

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-.1419	.1511	-.0037	.05259	41
Std. Predicted Value	-2.629	2.942	.000	1.000	41
Standard Error of Predicted Value	.021	.074	.037	.012	41
Adjusted Predicted Value	-.1697	.1571	-.0029	.05603	41
Residual	-.32758	.28140	.00000	.12030	41
Std. Residual	-2.619	2.250	.000	.962	41
Stud. Residual	-2.658	2.316	-.003	.997	41
Deleted Residual	-.33746	.29821	-.00076	.12952	41
Stud. Deleted Residual	-2.915	2.470	-.009	1.036	41
Mahal. Distance	.154	13.124	2.927	2.897	41
Cook's Distance	.000	.126	.019	.026	41
Centered Leverage Value	.004	.328	.073	.072	41

a. Dependent Variable: MeanREG




Partial Regression Plot
Dependent Variable: MeanREG
MeanREG
.20
.00
-.20
-.40
-1.00
-.50
.00
.50
1.00
intercept
Partial Regression Plot
Dependent Variable: MeanREG
MeanREG
.20
.00
-.20
-.40
-.20
-.10
.00
.10
.20
.30
correlation

Appendix 12 A Guide to Calculations

This thesis contains a large number of different statistical calculations, many of which were calculated manually using *Excel*. Therefore, this appendix serves as a reference for all calculations that were conducted.

Measures of gifted underachievement

Simple difference

Equation 1. The simple difference measurement of gifted underachievement (adapted from Phillipson & Tse, 2007, p. 175).

$$SD = z_{expected\ achievement} - z_{actual\ achievement}$$

The measurement of gifted underachievement from the simple difference method is defined by Equation 1. The *z* terms indicate that both the expected achievement and actual achievement scores must be standardised and expressed in standard deviation units.

Regression

Equation 2. Standard Error (Field, 2009, p. 42).

$$SE = \sqrt{\frac{\sum(Y_i - Y`_i)^2}{N}}$$

When a model makes a prediction (Y`) there is always some degree of error (the difference between the real value and the predicted value). The standard error is the standard deviation of the all errors measured. For a simple regression model, the $Y_i$ values are the values of the dependent variable, $Y`_i$ are the values predicted by the regression model, and N is the sample size.

Equation 3. The regression measurement of gifted underachievement (Lau & Chan, 2001a).

$$REG = \frac{Y` - Y}{SE}$$

The measurement of gifted underachievement from the regression model is defined by Equation 3. The units of this measurement are standard errors.

Convergence Statistics and Statistical Tests

Cochran's *Q* test

Cochran's Q test is used to determine whether the proportion of cases identified (by any method) are the same across multiple samples or (as in this investigation) multiple methods. The test requires a calculation of the degree of variation, *Q*, of the proportion of cases across the samples, which is then tested for significance using a chi-square test. The calculation of *Q* is complex and is fully explained in Fleiss, Levin, and Paik (2003, p. 389), however, is summarised in Equation 4, where p_i is the overall proportion of cases identified as gifted underachievement by each of identification methods, m is the number of methods examined (5), N is the total number of studies (41), $\bar{p}$ is the average overall proportion of identifications across the different methods (in this study, 34%), and P_n is the average proportion of identifications across the different methods for each study. This chi-square test has $m - 1$ degrees of freedom.

Equation 4. Cochran's Q test (Fleiss, Levin, & Paik, 2003, p. 389).

$$Q = \frac{N^2(m-1)}{m} \times \frac{\sum(p_i - \bar{p})^2}{N\bar{p}(1-\bar{p}) - \sum(P_n - \bar{p})^2}$$

Contingency Table

Many of the following analyses rely on a contingency table. A contingency table is used to compare two classification methods for the same sample. A generic contingency table is reproduced below.

		Method 1		Row total
		Yes	No	
Method 2	Yes	a	b	a + b
	No	c	d	c + d
Column total		a + c	b + d	TOTAL

McNemar's Test

McNemar's test is used with a contingency table to determine whether there is a statistically significant difference between two sets of dichotomous classifications made of the same group of people, and is therefore an appropriate test to determine whether any two methods have different or equal proportions of identification. The test is carried out by calculating a McNemar test statistic (refer Equation 5) which is compared to a chi-square distribution with one degree of freedom. The difference in proportion of classifications (δ) is often reported with the McNemar test results (refer Equation 6).

Equation 5. McNemar's test statistic (Nussbaum, 2015, p. 109)

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

Equation 6. Difference in proportion of identifications (Nussbaum, 2015, p. 109)

$$\delta = \frac{b-c}{a+b+c+d}$$

The Chi-Square test

The chi-square test is a very common statistic test used to determine the statistical significance of the difference between an observed frequency and the expected frequency. The test relies on the calculation of a chi-square test statistic, which is then compared to a chi-square distribution of a suitable degree of freedom. The calculation of the chi-square test statistic is shown in Equation 7.

Equation 7. Chi-square test statistic (Agresti, 2013, p.75).

$$\chi^2 = \sum \frac{(\text{observed frequency-expected frequency})^2}{\text{expected frequency}}$$

The expected frequency refers to the row total in the contingency table multiplied by the column total and divided by the total sample size for each cell. The chi-square statistic has $(r-1)(c-1)$ degrees of freedom, where *r* is the number of rows and *c* is the number of columns in the contingency table. In this investigation, $df = 1$.

Fisher's Exact Test

Fisher's exact test is used when one of the assumptions of the chi-square test (i.e., all of the expected values are greater than one, and less than 20% of the expected values are less than five) is not met. The calculation of the relevant statistic is shown in Equation 8.

Equation 8. Fisher's exact test (Agresti, 2013, p. 91).

$$p = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\,b!\,c!\,d!\,n!}$$

The Phi Coefficient

The phi coefficient (ϕ) is a measure of the degree to which two categorical variables are associated and is calculated using a contingency table (refer Equation 9). It is mathematically related to the chi-square statistic (refer Equation 10).

Equation 9. Phi coefficient (Nussbaum, 2015, p. 92).

$$\phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(b+d)(a+c)}}$$

Equation 10. Relationship between phi coefficient and the chi-square statistic (Nussbaum, 2015, p. 92).

$$\phi = \sqrt{\frac{\chi^2}{N}}$$

Pearson's correlation coefficient

Pearson's correlation coefficient (r) is a measure of the correlation between two continuous variables (refer Equation 11, where *N* is the sample size of measurements, x_i and y_i are the *i*th measurements, $\bar{x}$ and $\bar{y}$ are means, and s_x and s_y are the standard deviations of the respective variables).

Equation 11. Pearson's correlation coefficient (Field, 2009, p. 170).

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

Correlation *t*-test

A *t*-test statistic is able to determine whether a correlation coefficient is statistically significantly greater than zero. The statistic (refer Equation 12) depends on the size of the

correlation coefficient (r) and the sample size (N), and is compared to the t distribution with $N - 2$ degrees of freedom.

Equation 12. t-statistic for testing the significance of a correlation coefficient (Field, 2009, p.172).

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

Percentage agreement

The percentage agreement statistic is a measure of the percentage of cases for which classifications agree when different classification methods are used. The percentage agreement statistic may be calculated using data in a contingency table (refer Equation 13).

Equation 13. Percentage agreement (Agresti, 2013, p. 434).

$$\Pr(a) = \frac{a+d}{a+b+c+d}$$

Cohen's kappa statistic

The Cohen's kappa statistic (κ) is a measure of the level of agreement between two classification methods that takes into consideration the agreement that may be expected by chance. The relevant calculations may be undertaken in two steps using a contingency table: (a) the probability of chance agreement is estimated (refer Equation 14) and (b) the kappa value is calculated by comparing the observed percentage agreement to the expected agreement (refer Equation 15).

Equation 14. Probability of chance agreement (Agresti, 2013, p. 434).

$$\Pr(e) = \frac{(a+c)\times(a+b)+(b+d)\times(c+d)}{(a+b+c+d)^2}$$

Equation 15. Cohen's kappa (Fleiss, Levin, & Paik, 2003, p. 603).

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

Concordance correlation coefficient

The Concordance correlation coefficient (*CCC*) is a measure of agreement that is equivalent to Cohen's kappa, for two continuous variables. The coefficient is calculated using the standard deviation of each variable (σ), their mean values ($\bar{x}$, $\bar{y}$), and their Pearson correlation coefficient (r_{xy}), as shown in Equation 16.

Equation 16. Concordance correlation coefficient (Tang, He, & Tu, 2012, p. 298).

$$CCC = \frac{2r_{xy}\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\bar{x} - \bar{y})^2}$$

Paired *t*-test

A paired *t*-test is able to determine whether the difference between two means (*d*) is statistically significant. A *t* statistic is calculated from the difference between the two means, the standard deviation of the differences, and the sample size of the measurements (refer Equation 17).

Equation 17. The paired *t*-test statistic (Field, 2009, p.327)

$$t = \frac{d}{\sigma/\sqrt{N}}$$

Meta-Analysis

The methods used in this thesis to carry out meta-analysis closely follow the guide produced by Borenstein et al. (2009).

Calculating an weighted average effect size across multiple studies

Often the first goal of meta-analysis is to calculate an overall effect size by combining the reported effect sizes from multiple studies. This overall value is calculated as a weighted average of all the effect sizes (refer Equation 18), where the weighting of each study is equal to the inverse of the variance of the effect size for each study (refer Equation 19).

Equation 18. Overall effect size as the average of weighted effect sizes (Borenstein et al., 2009, p. 66).

$$\theta_{overall} = \frac{\sum_k W_i \theta_i}{\sum_k W_i}$$

Equation 19. The weighting of each study based on the inverse of its variance (Borenstein et al., 2009, p. 65).

$$W_i = \frac{1}{\sigma_i^2}$$

In this investigation, the effect sizes examined were (almost) all considered *r*-type effect sizes (i.e., correlation coefficients rather than differences in means). For *r*-type effect sizes, Borenstein et al. (2009) recommended that the effect sizes be transformed using Fisher's *z*-scale (Fisher, 1915; refer Equation 20). The benefit of using this transformation is that the variance of effect sizes after taking Fisher's *z* transformation is accurately estimated using only the sample size of the study (refer Equation 21). To calculate an overall effect size therefore requires converting all of the effect sizes to Fisher's *z* metric, estimating the variance based on the sample size (and therefore calculating the weighting for each study), and then calculating the overall effect size from these values. The resulting value is in Fisher's *z* metric, and needs to be converted back into the original correlation units. This is achieved using Equation 22.

Equation 20. Fisher's *z* transformation (Borenstein et al., 2009, p. 42).

$$z = 0.5 \times \ln\left(\frac{1+r}{1-r}\right)$$

Equation 21. Estimated variance of an effect size after Fisher's *z* transformation (Borenstein et al., 2009, p. 42).

$$\sigma_z^2 = \frac{1}{N-3}$$

Equation 22. Conversion of Fisher's *z* metric to correlation units (Borenstein et al., 2009, p. 42).

$$r = \frac{e^{2z}-1}{e^{2z}+1}$$

The one statistic used in this investigation that was not an *r* type effect size was the difference between means. This statistic may be converted into an *r* measurement by first calculating its *t*-statistic (refer Equation17) and then converting this using the number of degrees of freedom ($df = N - 1$) as shown in Equation 23 below.

Equation 23. Conversion of a *t*-statistic into an *r*-type effect size (Field, 2009, p.332)

$$r = \sqrt{\frac{t^2}{t^2 + df}}$$

Testing homogeneity

After calculating a weighted average effect size over a number of studies, it is possible to conduct statistical tests to determine whether the effect is homogeneous across the multiple studies. The first step for this test is to calculate the total amount of variation between studies (refer Equation 24). Secondly, the amount of variation is compared to the expected amount of variation using a chi-square test for significance with the degrees of freedom determined by the number of studies examined (*k*, refer Equation 25).

Equation 24. The weighted sum of squares, a measure of the total variation of effect sizes between studies (Borenstein et al., 2009, p. 109).

$$Q = \sum W_i(\theta_i - \theta_{overall})^2 = \sum \frac{(\theta_i - \theta_{overall})^2}{\sigma_i^2} = \sum W_i\theta_i^2 - \frac{(\sum W_i\,\theta_i)^2}{\sum W_i}$$

Equation 25. Degrees of freedom (Borenstein et al., 2009, p. 110).

$$df = k - 1$$

Measuring heterogeneity

If heterogeneity is detected, the degree of heterogeneity may be measured using two statistics. The first statistic, T^2, represents the degree of variance in the units of the effect size examined (refer Equations 26 and 27). Therefore, the T statistic may be used to determine the 95% confidence intervals for the weighted average effect size (refer Equation 28). The second statistic, I^2, represents the proportion of observed variance which reflects real differences in effect sizes (i.e., not due to the within-study variance; refer Equation 29). This statistic is useful in determining whether the degree of measured heterogeneity is explainable by the (lack of) precision of the studies that were analysed.

Equation 26. Variance of effect sizes (Borenstein et al., 2009, p. 114).

$$T^2 = \frac{(Q - df)}{C}$$

Equation 27. Conversion factor (Borenstein et al., 2009, p. 114).

$$C = \sum W_i - \frac{\sum W_i^2}{\sum W_i}$$

Equation 28. 95% confidence interval for the weighted average effect size (Borenstein et al., 2009, p. 116).

$$CI = \pm 1.96 \times T$$

Equation 29. The proportion of true variance to observed variance (Borenstein et al., 2009, p. 117).

$$I^2 = \frac{(Q - df)}{Q} \times 100\%$$