THE PSYCHOMETRIC PROPERTIES OF THE PAPER AND PENCIL GAMES LEVEL 2 FOR TIGRIGNA-SPEAKING CHILDREN IN ERITREA.

by

HAGOS GHEBREMICAEL TECLE

Thesis submitted in fulfillment of the requirements for the degree

MASTER OF ARTS

in

PSYCHOLOGY

in the

FACULTY OF ARTS

at

STELLENBOSCH UNIVERSITY

Supervisor: Prof. GP de Bruin
Co-Supervisor: Mr. H Steel
APRIL 2003

Declaration

I, the undersigned, hereby declare that the work contained in this assignment /thesis is my own original work and that I have not previously in its entirety or in part submitted it at any university for a degree.

Signature: ....

Date: ....

Abstract

The aim of this study was to investigate the psychometric properties of a screening test of cognitive ability, the Paper and Pencil Games Level 2 (PPG Level 2), for Tigrigna-speaking schoolchildren in Eritrea. This study represents one of the first attempts to measure cognitive ability in Eritrea. The PPG was developed in South Africa (Claassen, 1996) as a group test of general cognitive ability for children in Grade 2 and 3 (PPG Level 2), and Grade 4 and 5 (PPG Level 3). The PPG provides Total, Verbal and Non-verbal ability scores. The Verbal Scale consists of two subtests, namely (a) Verbal and Quantitative Reasoning and (b) Comprehension. The Non-verbal Scale consists of three subtests, namely (a) Figure Classification, (b) Figure Series, and (c) Pattern Completion. Although the PPG has the appearance of a standardized intelligence test, Claassen emphasized that it is best used as a screen for academic difficulties or failures.

Participants were 577 Tigrigna-speaking Grade 3 children. The children were selected from schools in the capital, Asmara, from small towns, and villages. Eleven schools participated. The participants can be considered representative of the Grade 3 Tigrigna-speaking population in Eritrea. The PPG Level 2 was completed under supervision of the researcher.

Classical and Rasch item analyses were conducted on the Verbal and Non-verbal Scales, respectively. The internal consistency of the Non-verbal scale can be considered satisfactory for a screening instrument (Cronbach's $\alpha = .85$). Furthermore, the non-verbal items showed satisfactory fit to the Rasch model (INFIT values and OUTFIT values < 1.3 for all items), suggesting that they measure a unidimensional construct. In addition, the item difficulty estimates corresponded well with the serial order of the items, with easy items being presented earlier than more difficult items. The internal consistency of the Verbal scale was lower (Cronbach's $\alpha = .72$), which can probably be attributed to the relative easiness of the items for the particular group of participants. The verbal items also showed satisfactory fit to the Rasch model. The Rasch analysis, which expresses person ability and item difficulty on the same scale, clearly showed that the Verbal items

were too easy for the majority of the children. However, it should be kept in mind that the PPG is intended to discriminate among children with low ability. Hence, the observed mismatch between the abilities and item difficulties was not unexpected. It should also be noted that the serial order of the items did not correspond well with item difficulty, with some difficult items being presented early and some easy items presented late in the scale.

The five subtests of the PPG were subjected to a confirmatory factor analysis. Two models were specified and compared. Model 1 specified a single general factor; this provided a more parsimonious description of the data and showed a satisfactory fit with the data, though a bit weaker than that of Model 2. Model 2 specified two correlated factors, namely a Verbal and a Non-verbal factor. Model 2 also fitted the data well, but a high correlation between the factors was observed (r = .77; r^2 = .59), suggesting the presence of a general factor. The results provide support for two levels of interpretation, namely on the Total score level and the Verbal and Non-verbal level.

The validity of the PPG Level 2 was further investigated by examining the correlations between the PPG scores and teacher ratings of academic achievement. Because different schools had different raters, the correlations within each of the schools were pooled to obtain an estimate of the correlations between the PPG scales and academic achievement for the total group. The pooled correlation for the PPG Total score with academic achievement was .56, for the Non-Verbal score .53, and for the Verbal score .41. The correlations for the Total and Non-verbal scores are similar to those typically reported in the educational psychology literature and provide support for the validity of these scales as a screen for academic difficulties.

The results show that the PPG, which was developed in South Africa, may be fruitfully exported to Eritrea. It is recommended, however, that before the PPG Level 2 is routinely used for screening purposes with Tigrigna-speaking children, the functioning of the Verbal Scale should be re-examined and possibly some of the Verbal items should be rewritten.

OPSOMMING

Die doel van die onderhawige studie was om die psigometriese eienskappe van 'n siftingstoets van kognitiewe vermoë, naamlik die Paper and Pencil Games Level 2 (PPG Level 2), vir Tigrigna-sprekende kinders in Ertirea te bestudeer. Hierdie studie verteenwoordig een van die eerste pogings om kognitiewe vermoë in Eritrea te meet. Die PPG is in Suid-Afrika ontwikkel (Claassen, 1996) as 'n groeptoets van kognitiewe vermoë vir kinders in Grade 2 en 3 (PPG Level 2), en Grade 4 en 5 (PPG Level 3). Die PPG lewer tellings van kognitiewe vermoë op drie vlakke, naamlik Totaal, Verbaal en Nie-Verbaal. Die Verbale skaal bestaan uit twee subskale: (a) Verbale en Kwantitatiewe redenering en (b) Begrip. Die Nie-Verbale skaal bestaan uit drie subskale, naamlik (a) Figuur Klassifisering, (b) Figuurreekse en (c) Patroonvoltooiing. Alhoewel die PPG op die oog af soos 'n konvensionele intelligensietoets lyk, beklemtoon Claassen dat dit te beste geskik is as 'n siftingsinstrument vir akademiese probleme of mislukkings.

Die deelnemers was 577 Tigrigna-sprekende kinders in Graad 3. Die kinders is uit skole van die hoofstad, Asmara, klein dorpe en geselekteer. Elf skole het aan die studie deelgeneem. Daar kan aanvaar word dat die deelnemers verteenwoordigend is van die Graad 3 Tigrigna-sprekende populasie in Eritrea. Die PPG Level 2 is onder supervisie van die navorser voltooi.

Klassieke en Rasch item-ontledings is op die Verbale en Nie-Verbale skale, onderskeidelik, uitgevoer. Die interne konsekwentheid van die Nie-Verbale skaal kan as bevredigend vir 'n siftingsinstrument beskou word (Cronbach se $\alpha = .85$). Hierbenewens het die Nie-Verbale items 'n bevredigende passing met die Rasch model getoon (INFIT gemiddelde kwadrate en OUTFIT gemiddelde kwadrate < 1.3 vir alle items), wat daarop dui dat die items 'n essensiëel eendimensionele konstruk meet. Die moeilikheidswaardes van die Nie-Verbale items het ook sterk ooreengestem met die volgorde waarin die items in die skaal aangebied word – die maklike items is eerste aangebied en daarna die moeilike items. Die interne konsekwentheid van die Verbale skaal was laer (Cronbach se $\alpha = .72$). Die laer koëffisiënt kan waarskynlik aan die relatiewe lae moeilikheidsgraad van

die Verbale items toegeskryf word. Die Verbale items het egter ook 'n bevredigende passing met die Rasch model getoon. Die Rasch ontleding, wat vermoë en item moeilikheid op dieselfde skaal uitdruk, het duidelik getoon dat die Verbale items te maklik vir die meeste kinders was. Daar dien egter op gelet te word dat die PPG ontwerp is om te diskrimineer tussen kinders met relatiewe lae vermoëns. In hierdie lig gesien is die swak passing tussen vermoë en item moeilikhede nie te onverwags nie. Die volgorde waarin die items in die skaal aangebied word het ook nie goed ooreengestem met die item moeilikheidswaardes nie – sommige moeilike items is vroeg in die skaal aangebied en sommige maklike items laat in die skaal.

Die vyf subtoetse van die PPG is aan 'n bevestigende faktorontleding onderwerp. Twee modelle is gespesifiseer en vergelyk. Model 1 het 'n enkele algemene faktor gespesifisieer.

Dedication

This work is dedicated to my beloved wife, **Nebiat**, and my daughter **Deborah**.

Acknowledgements

First and foremost, I would like to thank my Creator and Saviour (The Almighty God) for the amazing opportunity afforded to me in this postgraduate study and research experience. Without His guiding hand and quiet listening I would not have had the strength and perseverance to finish this study.

I would like to express my sincere appreciation and gratitude to those who contributed directly or indirectly to the completion of this dissertation.

- My supervisor, Professor Gideon P. de Bruin, Director of the Institute for Child and Adult Guidance at Rand Afrikaans University, for his inspiration, wise advice and guidance, for being available when I needed him, as well as his constant encouragement and support throughout my postgraduate studies. Without his guidance, this study would not have materialized.

- Mr. Steel, my co-supervisor, for his valuable comments and facilitating the last work of the thesis.

- Dr. Nicholas Claassen and Dr. Anil Kanjee of the Human Sciences Research Council for the permission they gave me to use the Paper and Pencil Games for this research. I would also like to express my sincere appreciation to Dr. Claassen for his cooperation and suggestions in the research work.

- My wife, Nebiat Gebremeskel, who has endured two years of strenuous separation - for her inspiring love and encouragement all the time.

- My family members who instilled in me the determination for success, and for supporting me throughout my school life. They have been my greatest inspiration.

- My brethren and friends who have continually prayed and encouraged me through the loneliest time in my life.

- The research assistants (Aman Musgun, Eden Andehaimanot, Letu Hailemariam, Sabrina Amanuel, Semhar Mehreteab, Senait Berhane and Senait Michael), principals and teachers of all sample schools who participated so generously and who went out of their way to facilitate the test administration.

- Abraham Fessehaye, Gebretatios Ogbai, Teame Atewebrhan, Frezghi Sealay, Bereket Gezae, Samuel Bizen, Daniel Rezene, Fanuel Sahle, Solomn Haile who participated so diligently in the translation and back-translation of the instruction material of the test.

- The following individuals: Zecarias Zemicael, Teclom Habtom, Tesfamariam Amare, Luel Zewde, Negassi Bisrat, Tesfay Bahta, Tesfalẹm Tekleab, Yosief Mobae, who helped me in feeding the data into a computer.

- Mrs. Lula Gebreyesus and AIPA staff for their constant encouragement, care and good facilitation throughout my postgraduate studies.

- Last, but not least, the subjects of the study who willingly participated in this study and who gave generously of their time answering the many items so graciously and diligently. Thank you for your interest and honesty. Without you this study would have not been possible.

Table of Contents

Note. The style of this dissertation is in accordance with the requirements of the Department of Psychology of Stellenbosch University. Also, the use of the male gender is used generically to include both males and females.

List of Tables

List of Figures

Chapter One

Introduction, Motivation, and Aims

1.1 Orientation and motivation

Eritrea, one of the poorest countries on the African continent, obtained its independence from Ethiopia in May 1991. The colonization by Italy in the first part of the twentieth century and then annexation by Ethiopia contributed to a high illiteracy rate (Ministry of Education Eritrea, 1999). Children, in particular, have been the victims of chronic malnutrition and the traumatizing effects of war, that are likely to adversely affect their social and cognitive development. In the light of this it is necessary to establish a research program into ways that the educational situation of children in Eritrea may be improved.

This study aims to contribute to the improvement of education in Eritrea by examining the psychometric properties of the Paper and Pencil Games (PPG) Level 2 (Claassen, 1996) for Grade 3 Tigrigna-speaking children. The PPG Level 2 is a South African group intelligence test for children in Grades 2 and 3 that is used as a screen for children who are at risk of developing educational difficulties. It is hoped that instruments such as the PPG Level 2 may help in the identification of children who may experience educational difficulties in the Eritrean context so that appropriate early educational interventions may be designed and implemented.

1.2 Education in Eritrea

The successive colonial powers in Eritrea designed and implemented curricula that enabled them to retain power at the expense of the education of the Eritrean people (Ministry of Education Eritrea, 1999; Ghenet, 2000). Carnine (1994) stated that curriculum weaknesses are among the leading causes of the failure of diverse learners even in developed countries. Furthermore, access to formal educational opportunities in Eritrea until 1991, regardless of the quality, were restricted to a few elite groups and people around the urban centres (Ministry of Education Eritrea, 1999).

Since 1991 the Provisional Government of Eritrea has tried to provide equal educational opportunities to school-aged children at all sectors and levels of its society. This was not an easy task because of limited human and material resources (Ministry of Education Eritrea, 1999). Learners with special educational needs have not received much attention in Eritrea because of two major factors: (a) even in mainstream schools the number of learners per class is 70 and above, which makes it difficult for teachers to attend to special educational needs; and (b) there is a shortage of experts in the areas of educational and psychological measurement and educational remediation.

1.3 Assessment in education and early intervention

Assessment is a broad and generic concept that refers to the process of gathering information for the purpose of making a decision (McLean, in Blasco, 2001). Psychological assessment, in particular, helps individuals, groups and organizations to understand and make informed and appropriate choices about their overall functioning. Assessment can serve many functions. For instance, it provides comprehensive information about people's particular strengths and weaknesses and it can be used to delineate progress and development, to make systematic decisions about appropriateness for a particular field in both job market and education, and training and recruitment purposes, and to make an overall diagnosis (Foxcroft & Roodt, 2001).

In school and clinical settings assessment is used to make decisions regarding identifying target populations for intervention, continued evaluation of programmes for particular purposes, highlighting programmatic strengths and weakness, and pinpointing the needs and priorities of young children at risk (Meisels & Atkins-Burnett, 2001). In defining the "at risk" population, interventionists include well-known biological and environmental factors that can be identified and place children at risk for developmental delay (Blasco, 2001). This means that through a careful and thorough assessment process an interventionist would be able to establish a baseline or entry level on the desired outcome. In addition, the process enables the service provider to select appropriate and relevant intervention goals and objectives (Meisels & Atkins-Burnett, 2001). Keogh and Becker (in Schaefer & McDermott, 1999) argued that assessment activities are not intended only to document or confirm deficits or weaknesses, but also to provide intervention information that would prevent future problems.

One of the major advantages of early intervention is that it is less expensive than treatment at a later stage (Mcintosh, Gibney, Quinn, & Kundert, 2000). Among the target groups for early educational intervention programmes are slow learners or learners with learning difficulties. Slow learners are usually defined in terms of intelligence test scores, educational placement, and methods of instruction. However, slow learners are not only at risk for educational failure but also at risk for mental health problems (Carnine, 1994). The basic assumption is that slow learners are at risk for behaviour problems due to skill deficits. Hence, failure to provide early educational interventions may lead to increased frustration and future behavioural problems (Asuni, 1970; Carnine, 1994).

Jensen (1998) pointed out that one of the major problems in public education today is the lack of accommodation for the diversified individual learner's mental capabilities and deficits. In other words, there is wide individual variation in intellectual abilities in the school population and the associated problem teachers' face is how to deal with prominent differences in human ability. By and large it is the task of educational designers and policy makers to deal with the exceptional variation among learners' intellectual abilities. However, there is room for classroom teachers to adapt the nature of instruction to accommodate differences in ability, style, or preferences among individual learners to improve learning outcomes (Jonassen & Grabowski, 1993). This becomes difficult to practice in Eritrean elementary schools with large numbers of learners per class, as mentioned earlier, and the lack of trained teachers for students with special needs.

1.4 The statement of the problem

In the preceding paragraphs the importance of good assessment instruments in the educational context has been emphasised. It has also been stressed that Eritrea is in need of such instruments because no such instruments have been developed or adapted for the Eritrean context (Ghenet, 2000). The aims of this study flows from this problem and are described in the next section.

1.5 Aims of the study

The principal aim of this study is to investigate the psychometric properties of a group intelligence test, namely the Paper and Pencil Games Level 2 (Claassen, 1996) for Tigrigna-speaking Grade 3 children in Eritrea. According to Claassen (1996), the Paper and Pencil Games (PPG) measure "figural, quantitative and verbal skills closely related to scholastic achievement" (p. 1). The PPG can be described as a group intelligence test that can be used to identify children that may be in need of special education. It is important to note that the PPG is used as a screening instrument and that final decisions regarding the treatment of any child should be based on thorough and careful individual assessments.

The PPG was developed in South Africa and it cannot be assumed that the scores of Eritrean children will be similar in meaning and comparable to those of South African children. However, this does not necessarily imply that an imported instrument, such as the PPG Level 2, will be invalid for use in a "new" culture. What is necessary is to demonstrate the internal and external validity of the PPG in the Eritrean context (van de Vijver & Leung, 1997).

Two broad issues concerning the psychometric properties of the PPG Level 2 will be addressed in the study, namely (a) reliability and (b) validity. The reliability of test scores refers to the dependability and consistency thereof, whereas validity is concerned with whether test scores represent what they purport to represent. Reliability is a necessary but not sufficient condition for validity. Concerning reliability, one would expect the PPG to provide consistent scores over time across culture. Furthermore, one would expect consistency in the way that children respond to the individual items of the PPG. Concerning validity, one would expect the PPG (a) to adequately differentiate between children with different ability levels, and (b) to adequately predict whether children will be good or poor performers in terms of scholastic achievement. This assumption is based on the fact that human general mental ability is highly related to academic and job performances (Eysenck, 1998; Gage & Berliner, 1984; Gottfredson, 1998; Neisser et al., 1996.)

1.6 Research questions

The primary and secondary research questions are listed separately for a clear understanding of the research work.

1.6.1 Primary questions

There are three broad primary research questions:

(a) Do the items of the PPG Level 2 function appropriately for Tigrigna-speaking Grade 3 children?

(b) Does the PPG Level 2 provide reliable scores for Tigrigna-speaking Grade 3 children?

(c) Does the PPG Level 2 provide valid scores for Grade 3 children?

1.6.2 Secondary questions

There are two secondary research questions:

(a) Are there significant differences between boys and girls for the PPG Total, Verbal, Non-verbal, Figure Classification, Verbal and Quantitative Reasoning, Figure Series, Comprehension, and Pattern Completion scores?

(b) Are there significant differences between children from three regional groups (the capital, towns, and villages) for the PPG Total, Verbal, Non-verbal, Figure Classification, Verbal and Quantitative Reasoning, Figure Series, Comprehension, and Pattern Completion scores?

1.7 Rationale of the study

Valid and reliable data on pupils' cognitive strengths and weaknesses may enable educational policy makers to devise proper and appropriate instructional mechanisms (Gage & Berliner, 1988; Jensen, 1998). The responsible use of psychological tests may help in the early identification of children with special educational needs (Meisels & Atkins-Burnett, 2001). The most important advantage that screening instruments like the PPG can offer is that children with special educational needs may be identified very early and that appropriate interventions may be designed as preventive and ameliorative measures, which are relatively effective and less costly in the earlier years of a child's life. This could also help all children to fulfill their potential and lead to a reduction in the school dropout rate.

1.8 Chapter delineation

In Chapter Two theories of intelligence are discussed. Chapter Three presents the research methods employed in the study, while Chapter Four presents the results of the study. The study is concluded in Chapter Five with a discussion of the results.

Chapter Two

Intelligence

2.1 Introduction

This chapter outlines the relevant literature reviewed for this project. The literature review will begin by highlighting some of the definitions and theories of intelligence by prominent figures in the field. In addition, various uses of intelligence tests and the relationship between academic performance and intelligence will be discussed. Finally, issues such as bias in intelligence testing will be explored.

2.2 Definition of intelligence

Human intelligence is probably the most controversial and debated topic in the whole field of psychology (Gottfredson, 1998; Graham & Lilly, 1984). One of the problems surrounding the study of intelligence is that it is difficult to find a consistent definition and understanding thereof. However, it would appear worthwhile to mention some of the definitions given by prominent researchers in the field, which reflect the general theories and trends to which researchers adhere.

The issue of differences in broad abilities of human intelligence has been acknowledged since ancient times, though not always addressed and assessed properly (Deary, 2001). Alfred Binet defined intelligence at the turn of the 20th century as "the tendency to take and maintain a definite direction; the capacity to make adaptations for the purpose of attaining a desired end, and the power of autocriticism" (Kaplan & Saccuzzo, 2001, p. 254). Binet's successor, David Wechsler, defined intelligence as "the aggregate or global capacity of the individual to act purposefully, to think rationally and to deal effectively with the environment" (Kaplan & Saccuzzo, 2001, p. 282). His definition implies that the interrelationship among specific human abilities results in one aggregate that is general intelligence (Kaplan & Saccuzzo, 2001).

Terman (in Kline, 1991, p. 2) defined intelligence as "the power to think abstractly, to be self-critical and to be adaptive", while Gardner (1983) defined intelligence as "the ability to resolve genuine problems or difficulties as they are encountered" (Kaplan & Saccuzzo, 2001, p. 254). In general terms, Gottfredson (1998) described intelligence as "the ability to deal with cognitive complexity" (p. 25). She added that the definition overlaps greatly with lay insights of intelligence, such as 'smarter people' to mean 'intelligent people' and includes reasoning, problem solving, abstract thinking, and quick learning.

The views of 52 influential researchers on intelligence were recently summarized as follows:

> Intelligence is a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience. It is not merely book learning, a narrow academic skill, or test-taking smarts. Rather, it reflects a broader and deeper capability for comprehending our surroundings - 'catching on', 'making sense' of things, or 'figuring out' what to do (Deary, 2001, p. 17).

2.3 Theories of intelligence

Many different models or theories of intelligence and the structure of mental abilities have been proposed. There are two broad camps regarding theories of intelligence: (a) those who adhere to the unidimensional construct of general intelligence (*g*), and (b) those who support the idea of many intelligences (Sattler, 2001).

2.3.1 Spearman's two-factor theory

The statistical extraction of the general intelligence factor, *g,* was achieved by a technique called factor analysis, which was primarily invented by the British psychometrician, Charles Spearman, at the beginning of the 20th century (Gottfredson, 1998; Jensen, 1980). This method enables researchers in the field of cognitive testing to extract the minimum number of underlying latent variables or factors necessary to explain a pattern of correlations among observed variables (Gottfredson, 1998; Kaplan & Saccuzzo, 2001).

Spearman originally postulated a two-factor theory of intelligence, which contended that all mental tests measure a shared general factor, *g*, and a specific factor, *s*, unique to the particular test (Jensen, 1980). Spearman described the *g* factor as being used to a certain extent in all cognitive tasks and the specific factors, *s*, as unique to accomplish a particular task that is unrelated across tasks (Gardner, Kornhaber, & Wake, 1996).

Similarly, Gottfredson (1998) and Jensen (1998) pointed out that regardless of the specific content of a mental test, the test measures only the *g* factor and the *s* factor (specific to the particular test). The *g* factor itself explains mental capacity rather than stored knowledge, but a person's store of knowledge tends to relate strongly to his *g* level, possibly for the reason that the accumulation of information represents a skill in learning and in comprehending new information. Moreover, the *g* factor is the one trait that best discriminates among persons considered gifted, average or retarded (Gottfredson, 1998).

In contrast to the views stated above, some critics of intelligence research maintain that the notion of general intelligence is deceptive: that no such universal mental capacity exists and that apparent 'intelligence' is really just a consequence of one's opportunities to learn skills and information valued in a particular cultural context (Kline, 1991). Indeed, the concept of intelligence and the way in which individuals are ranked according to this criterion could be social products (Gardner et al., 1996). However, the fact that *g* is not specific to any particular domain of knowledge or mental skill suggests that *g* is free of cultural content (Gottfredson, 1998; Graham & Lilly, 1984).

2.3.2 Fluid and crystallized theory of intelligence

Cattell (1966) proposed a hierarchical theory of intelligence in which there are two important broad factors or two types of intelligence and three-second order factors (power of visualization, retrieval capacity or general fluency, and cognitive speed). The two broad factors are fluid intelligence and crystallized intelligence (Cattell, 1966). As general abilities, both play a considerable role in academic and occupational success (Jensen, 1980). Fluid intelligence represents basic reasoning capacity that is highly dependent on mental processes or the neurology of the brain. It is viewed as a culture-free mental competence. Crystallized intelligence, in contrast, refers to all skills and abilities acquired as a result of exposure to a particular cultural setting (Sattler, 2001). Fluid intelligence is measured by tests that are relatively free of school influences or low educational input, such as figure series, matrices, mazes and block designs. Tests of reading, vocabulary, and arithmetic reasoning can be used to measure crystallized intelligence, as it is related to the acquisition of knowledge from one's surroundings (Gardner et al., 1996).

Horn (1998, in Sattler, 2001) modified the theory on the basis of many factor analytic studies suggesting 55 primary abilities and incorporating nine second-order abilities, among them the fluid and crystallized abilities. The other seven second-order abilities are visual processing, auditory processing, short-term memory, long-term memory, processing speed, decision speed, and quantitative knowledge.

2.3.3 A three-stratum theory of intelligence

After reanalysing several hundred data sets, Carroll (1997, in Sattler, 2001) proposed a three-stratum factor analytic theory of intelligence. Like Cattell and Horn he considered fluid and crystallized intelligence to be among eight-second stratum broad factors. In contrast to Cattell and Horn, however, he also included a single general factor at the apex of the factorial hierarchy. Carroll's three-stratum theory is arranged in three levels in the following manner:

(i) Stratum I is at the bottom of the hierarchy and consists of 65 narrow mental abilities.

(ii) Stratum II comprises eight broad factors, namely fluid intelligence, crystallized intelligence, general memory and learning, broad visual perception, broad auditory perception, broad retrieval capacity, broad cognitive speediness, and processing speed (decision speed).

(iii) Stratum III is the general factor located at the apex of the hierarchy (Sattler, 2001).

2.3.4 Triarchic theory of intelligence

Sternberg (1985, in Gardner et al., 1996) developed a triachic theory of intelligence. He argued that classical and contemporary theories of intelligence are not incorrect, but rather incomplete. Sternberg's theory of human intelligence distinguishes between three parts of intelligence. These are:

(i) Componential intelligence, which refers to the internal, elementary information processes underlying intelligent thought. It, in turn, has three components, namely meta, performance, and knowledge acquisition components.

(ii) Experiential intelligence, which is concerned with the external and internal dimensions of intelligence. In other words, experiential intelligence is the ability to deal with novel situations and the ability to become efficient and automatic in thinking and problem solving.

(iii) Contextual intelligence, which is related to the individual's capacity to deal with the particular setting or culture and the environment in the external world (Gardner et al., 1996).

Sternberg (1985, in Neisser et al., 1996) pointed out that it is difficult to measure the experiential (sometimes called creative) and contextual (practical) aspects of intelligence using conventional intelligence tests. Often, existing tests are suitable to measure only componential (analytic) intelligence.

2.3.5 Primary mental abilities (Thurstone's multidimensional theory of intelligence)

Unlike Spearman, Thurstone (1938, in Graham & Lilly, 1984) focused on more specific group factors of ability, which he called primary mental abilities. Originally, he described seven primary mental abilities and no general factor at all (Graham & Lilly, 1984). He later added one primary factor. The full list of primary abilities is verbal, perceptual speed, inductive reasoning, number, rote memory, deductive reasoning, word fluency, and space or visualization. He further argued that each of the specific abilities carries the same weight. In his later work the existence of moderate correlations among the primary factors prompted him to postulate a second-order factor, which seems similar to the *g* factor (Sattler, 2001).

2.3.6 Multiple intelligences

In Gardner's view, intelligence is not something concrete but a latent ability that allows people to think properly about specific kinds of content (Gardner et al., 1996). Like Thurstone, Gardner (1983) rejected the idea of general intelligence and developed a theory of multiple intelligences

In contrast to Spearman's *g* factor, Gardner argued that there are many forms of mental abilities that are unrelated to each other. In his original work he proposed seven intelligences, but later noted that there may be more or fewer (Gardner et al., 1996). After revising his work, he identified eight major competencies and two tentative competencies; in this way he leaves room for more to be discovered (Sattler, 2001). The eight competencies are as follows:

(i) Linguistic intelligence involves the uses and implications of language for communicating in diversified social settings (e.g. meaning, speech sound, grammar).

(ii) Musical intelligence enables individuals to form, comprehend and communicate meaning of sound (e.g. rhythmic and pitching abilities).

(iii) Logical-mathematical intelligence concerns the use and valuing of abstract and logical relations (e.g. logical thinking and numerical ability).

(iv) Spatial intelligence focuses on people's ability to perceive visual and spatial stimuli accurately, to change and adjust the received information and to restructure visual images abstractly.

(v) Bodily-kinesthetic intelligence is strongly related to the ability to control over fine and gross motor actions and thereby to manipulate external objects (e.g. dancing, acting, and athletics).

(vi) Intrapersonal intelligence mainly involves the wisdom of a person about self - including the ability to identify one's feelings, intentions and motivations.

(vii) Interpersonal intelligence is an important component in various social encounters as it calls for an individual's capacity to recognize other individuals' emotions, feelings, beliefs as well as inner intentions and motives (Gardner et al., 1996; Sattler, 2001).

(viii) Naturalist intelligence concerns individuals' ability to detect patterns in nature (Sattler, 2001).

Gardner emphasized that his theory has vast educational implications, particularly for children with learning difficulties. For instance, if someone has difficulty in learning mathematics, there is no valid reason to assume that he is not intelligent. The person could be strong in one or many other fields. Also, he might compensate for his weakness in one area by using other areas of intelligence he is interested in or inclined to use, providing the learning opportunities allow him to do so (Gardner et al., 1996; Sattler, 2001).

In contrast to Gardner's (1983) claims, however, intelligence proposed by him such as verbal, mathematical and musical intelligences are positively correlated to each other as well as to general mental ability (Deary, 2001).

2.4 Uses of intelligence tests

Educators often employ intelligence test for guidance, selection, placement and admission purposes in various settings, in conjunction with other assessment procedures (Gage & Berliner, 1984). Jensen (1980) added that standardized mental tests are used for the prediction of future success and for selection purposes in schools, colleges and the work place. Neisser et al. (1996) highlighted the wide use of intelligence tests in Europe and American for evaluation, diagnosis and selection at all levels of the educational system.

Psychological tests have the advantage of obtaining estimates of current cognitive functioning that in turn may help in the design of appropriate intervention goals. Tests may also be used to estimate an individual's progress in an intervention programme (Schaefer & McDermott, 1999). On the basis of test results and other relevant information proper educational programmes can be designed for particular children with special needs. When used in this manner intelligence tests become useful instruments rather than means with which to investigate failure and segregate people (Jensen, 1980).

2.5 Academic achievement and intelligence

Several studies have indicated a moderately strong correlation between intelligence test scores and academic performance (Kline, 1991). However, there are also external variables in addition to intelligence that influence academic performance. Some of these variables include willingness to study, effort, persistence, interest in the subject material, quality of teaching, health, and home environment (Neisser et al., 1996).

A recent review of research on general intelligence by Gottfredson (1998) showed that intelligence is the single most effective known predictor of individual academic performance as well as performance in the work place. She further asserted that apart from the content of the particular test and the method of test administration, differences on diverse mental abilities in individual's performance are explained by a single *g* factor. Similarly, Eysenck (1998) emphasised that intelligence test scores predict a child's academic achievement or student's success at university.

Gage and Berliner (1984) and Neisser et al. (1996) reported that the correlations between general intelligence tests and academic performance are similar to those of tests of special abilities and academic performance. The correlation between academic performance and intelligence is often found to be about .50. Hence, intelligence explains approximately 25% of the variance in academic performance. The correlation of general intelligence and job success falls in the region of .30. Jensen (1980) summarized the relationship between intelligence and academic performance as follows: "Children with higher intelligence test scores generally acquire more scholastic knowledge more quickly and more easily, get better marks, like school and stay in school better" (p. 317).

Claassen (1996) reported a strong correlation ($r = .73$) between the Total PPG Level 2 score and a standardized Mathematics achievement test. However, the correlation between the Total PPG Level 2 score and teacher evaluation of Mathematics was lower (r = .40). These correlations suggest that scores on the PPG are related to academic achievement.

2.6 Test bias

Even though Binet did not invest time in studying test bias, he fully acknowledged that factors such as language, cultural background and differences in social background or experience could affect the fairness of intelligence tests (Kaplan & Saccuzzo, 2001). Vernon (1973) also identified environmental factors that may be problematic in testing intelligence cross-culturally. These factors, such as unfamiliarity with the test situation and lack of motivation, may mask the true ability of an individual (Jensen, 1980). There are three sources of test bias in cross-cultural assessment (van de Vijver & Poortinga, in van de Vijver & Tanzer 1997), which are real threats to validity in cross-cultural comparisons if their effects are not controlled or minimized properly. These threats are construct bias, method bias and item bias.

Construct bias arises when the construct under question is not the same for different cultural groups. This means that construct bias arises partly from differences in the socio-cultural experiences of different groups and failure to incorporate all the necessary features of the construct being measured in the test (van de Vijver & Tanzer, 1997). Eells (in Jensen, 1980) defined cultural bias in tests as "the differences in the extent to which the child being tested has had the opportunity to know and become familiar with the specific subject matter or specific process required by the test item" (p.369).

Method bias can be divided into three types, namely sample bias, administration bias and instrument bias. Sample bias refers to differences related to the selection of incompatible groups in relation to relevant attributes, for instance education and level of motivation. Administration bias is related to all the inconsistencies that could arise during test administration. This can be caused by a variety of factors, for instance differences in the environmental administration conditions (physical, technical, social), ambiguous instruction to participants, differential expertise of administrators, and language barriers between tester and testees. Finally, instrument bias is a common source of bias in mental tests that incorporates factors such as differential familiarity with stimulus material, response procedures and response styles (van de Vijver & Tanzer, 1997).

Item bias is fairly easy to identify. It describes a condition where the bias exists within the individual items making up the test (Rust & Golombok, 1989). Item bias is also known as differential item functioning (van de Vijver & Leung, 1997). An item is thought to be biased if individuals with equal ability, but from different groups, do not have the same probability of succeeding on the item. According to van de Vijver and Tanzer (1997), the major sources of item bias are poor item translation, inadequate item formulation, ambiguities in the original item, inappropriateness of item content in certain cultural groups, and specifics related to cultural idioms and idiosyncrasies.

2.7 Conclusion

Intelligence as a broad and general psychological construct refers to an individual's ability (a) to learn and deal with abstractions rather than with concrete things, and (b) to solve novel and complex problems when encountered (Gage & Berliner, 1984). Scholars in areas such as psychology, education, genetics and sociology agree that the following points should be incorporated to describe intelligence in addition to the above-mentioned aspects: (a) the capacity to acquire knowledge, (b) adaptation to one's environment, (c) creativity, and (d) general knowledge (Sattler, 2001). Intelligence refers mainly to the human beings general cognitive potential rather than specific abilities acquired from learning. Regardless of the diversified explanations of human intelligence, there is a general consensus among many psychologists that there is a general mental ability, the *g* factor, which tends to correlate with and explain the variation in specific abilities to some extent. They believe also in the existence of specific dimensions (specific abilities) besides the general mental ability (Gage & Berliner, 1984).

Finally, it is wrong to assume that all people do equally well in all the types of items included in intelligence tests. For instance, some people could be strong in the verbal domain rather than arithmetic and spatial domain. However, the different subtests of intelligence tests tend to correlate positively. That is, an individual who scored highly on the verbal items is generally expected to score highly in other subtests. Moreover, it is clear that mental tests do not measure all forms of intelligence or all important aspects of brain functioning including creativity, wisdom, practical sense, social sensitivity and others (Deary, 2001; Neisser et al., 1996).

The following chapter presents the research methods employed in the study.

Chapter Three

Methodology

3.1 Introduction

This study investigated the psychometric properties of the PPG Level 2 among Grade 3 Tigrignia-speaking Eritrean children. This chapter includes a description of the sampling procedure, the data collection methods, the research design, and the techniques used for data analysis.

3.2 Participants

There are nine ethnic or language groups in Eritrea. However, in this study the focus falls on Tigrigna-speaking children because they represent the dominant language group in Eritrea and the study would have become too big if all language groups were included.

Six hundred and five children completed the practice test, which was completed a week before the actual test administration, but 28 children were absent on the day of the actual test administration. Therefore the data of 577 children were collected. Children from different economic and social strata and geographic regions were included so that the results could be generalized to the general Tigrigna-speaking population. Specifically, 200 children were selected from four primary schools in the capital (Asmara), 203 children from the schools of three small towns (Adi-Tekelezan, Dibarwa and Decemhare), and 202 children from four village schools.

3.2.1 Participant selection criteria

Four criteria were considered in selecting the participants for the study. These criteria are set out in the paragraphs that follow.

3.2.1.1 Grade

The PPG Level 2 is appropriate for children in Grades 2 and 3 (Claassen, 1996). Grade 3 children were preferred for the present study for convenience of test administration. Eritrean children at the elementary school level are unfamiliar with standardized tests like the PPG Level 2. Against this background it was believed that more useful results would be obtained with the more mature Grade 3 children.

3.2.1.2 Location

Eritrea is divided into two geographic regions, namely the highland plateau region and the lowland coastal region. The Tigrigna-speaking population is mostly concentrated in the highland region of the country. Hence, schools and participants were selected from this region.

3.2.1.3 Schools

Only government-subsidized primary schools were considered for the study. The medium of instruction in all sample schools is Tigrigna. The use of participants from government schools was to ensure a representative cross-section of different socio-economic groups in Eritrea.

3.2.1.4 Language

At the elementary school level (Grade 1 to Grade 5), all children attend their classes using the language spoken at home. In order to maintain homogeneity of the sample, the participants came from homes where Tigrigna is the home language.

3.2.2 Participant selection procedure

Due to the centralized nature of administration in Eritrean school systems, the sample schools were approached in the following manner.

3.2.2.1 Permission from the Ministry of Education

Permission to test children from the different schools was obtained from the Director General of the Ministry of Education, Eritrea. This was done after a brief review of the aim of the study, the implications of participation and the nature of the test. In addition, the proposal of the study was submitted to the Director at the first meeting with the researcher.

3.2.2.2 Principals briefing

A letter requesting permission to collect data was distributed to the principals' of 11 target schools. The researcher and research assistants explained the participant selection criteria to the principal of each school. Within each school Grade 3 children were to be selected by means of a table of random numbers.

3.3 Sampling technique

The following sampling techniques were applied.

3.3.1 Strata formation

Three geographical strata groups were drawn from (a) the capital, Asmara (urban), (b) small towns (semi-urban), and (c) villages (rural areas). The towns are considered to be in between the urban and rural areas vis-à-vis technological access and other socio-economic indicators.

3.3.2 Cluster formation

Four elementary schools from the capital, three schools from towns of approximately equal distance from the capital (one elementary school from each town), and four from villages having one school each, were selected, giving a total of 11 participating schools. Although the schools were not selected at random, the children in the schools are thought to reflect the general Tigrigna-speaking population.

3.2.3 Subject selection

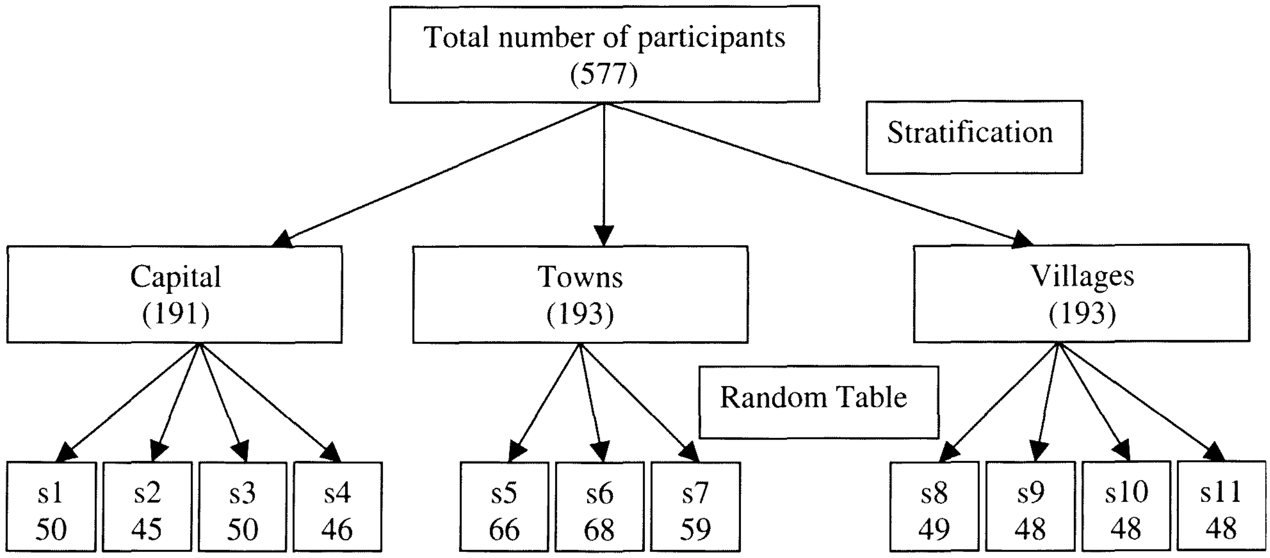
Some schools have morning and afternoon school sessions. In most instances it was impractical to collect data in the afternoon sessions and therefore most of the participants attended the morning school shifts. Within this constraint participants within each school were randomly selected by means of a table of random numbers. A detailed description of the children, their schools and school location is presented below for clarification purposes.

In the capital, four schools from different corners of the city were selected, namely Salina elementary school (50 children), Godaif elementary school (50 children), Hamad Idris Awate elementary school (50 children), and Model elementary school (50 children). Nine children were absent on the actual test administration day.

Three small towns within an approximately equal distance (40 kilometers) from the capital were selected. In the north, Denden elementary school from the town of Aditekelezan was selected. Sixty-seven children participated in the practice test, but one child was not present for the actual test administration. The second town, south of Asmara, is Dibarwa. There are two elementary schools in the town. Mereb elementary school was selected for the study. Sixty-eight children, sat for the practice and actual test administration. Decemhare, north-east of the capital, was the third town. There are four elementary schools in Decemhare, of which three are private and missionary schools. The government school, Erdi Awet elementary school, was selected for the study. Seventy children took the practice test, but 11 were absent for the actual test.

The four village schools that participated in the study were Adinifas elementary school, Kushet elementary school, Adiabeto elementary school and Sesah elementary school. Fifty children from each school participated in the practice test. However, 49 children from Adinifas and 48 each from the other three schools attended the actual test administration.

A diagrammatic representation of the participants and sampling procedures is presented in Figure 1. A summary of the gender breakdown in the 11 schools is presented in Table 1.



Note. s1 = Salina, s2 = Godaif, s3 = Hamed Idris Awate, s4 = Model, s5 = Aditekelezan (Denden), s6 = Dibarwa (Mereb), s7 = Dekemhare (Erdi Awet), s8 = Adinifas, s9 = Kushet, s10 = Adiabeto, s11 = Sesah.

Figure 1. Diagrammatic representation of participants and sampling techniques

Table 1. *Summary table for gender breakdown*

Gender	**Capital**				**Towns**			**Village**				Total
	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10	s11	
Male	22	29	26	28	31	34	35	23	25	25	29	307
Female	28	16	24	18	35	34	24	26	23	23	19	270
Total	50	45	50	46	66	68	59	49	48	48	48	577

Note. s1 = Salina, s2 = Godaif, s3 = Hamed Idris Awate, s4 = Model, s5 = Aditekelezan (Denden), s6 = Dibarwa (Mereb), s7 = Dekemhare (Erdi Awet), s8 = Adinifas, s9 = Kushet, s10 = Adiabeto, s11 = Sesah.

3.4 Research design

This investigation was conducted following a non-experimental research design. The aim was to investigate the psychometric properties of the PPG Level 2 and to this end univariate and multivariate correlational techniques were used.

3.5 Instruments

3.5.1 Demographic information sheet

A demographic questionnaire was printed on the front page of the PPG Level 2 booklet and included the following variables: Gender, age, school location, parental occupation and academic background and also parents' present status (living together, divorced, deceased).

3.5.2 Scholastic achievement

Ratings of academic performances of all participants were obtained from home-room teachers in each school. Ratings were given for five school subjects: Tigrigna, English, Mathematics, Science, and Moral. The ratings ranged between 0 and 100 and may be interpreted as percentages obtained for the particular subjects. Note that within a particular school one teacher would have been responsible for teaching a particular school subject, say Mathematics, to all the Grade 3 children. Hence, within each school ratings for each particular subject was done by one teacher only.

3.5.3 The Paper and Pencil Games

The Paper and Pencil Games was developed in South Africa and is available in 11 South African languages. The fact that it was specifically designed in a multi-cultural and multi-lingual context makes it potentially attractive in the multi-cultural Eritrean context. In addition, it is cost and time effective (Boon, 2000). However, it cannot be assumed that the psychometric properties of the PPG will be similar in Eritrea and in South Africa. Therefore it is necessary to empirically investigate the reliability and validity thereof in Eritrea.

Claassen (1996) described the rationale of the PPG as follows:

> The Paper and Pencil Games were based on the following premise: To learn new things, pupils must be able to perceive accurately and to recognize and recall what has been perceived. They also have to think logically, understand relationships, to abstract from a set of particulars, and to apply generalizations to new and different contexts. It is assumed that these processes are measured through performance on individual test items with pictorial, verbal figural, and quantitative content. The variety of cognitive skills necessary to perform well on the test could be described as general scholastic reasoning ability (p.2).

The PPG is available in two levels, namely Level 2 and Level 3. The PPG uses a minimum of language in the presentation of the items and this might help in minimizing the biases that could arise from culture, education and language which is the main problem associated with verbal scales of intelligence tests. PPG Level 2 is used with children in their second and third school years, whereas Level 3 is used with children in their fourth and fifth school years. PPG Level 2 was selected for the present study.

The PPG measures figural, quantitative and verbal skills related to scholastic achievement. The PPG is assumed to differentiate between pupils at all levels of ability, but is more appropriate at lower levels of functioning. As with all group tests, the PPG serves for screening purposes (Claassen, 1996).

The 84 items in the PPG level 2 are grouped into five sections or subtests, namely (a) Figure Classification (12 items), (b) Verbal and Quantitative Reasoning (26 items), (c) Figure Series (14 items), (d) Comprehension (16 items), and (e) Pattern Completion (16 items). The Figure Classification, Figure Series, and Pattern Completion subtests are combined into a Non-verbal Scale; the Verbal and Quantitative Reasoning and Comprehension subtests are combined into a Verbal Scale. Each of the Verbal and Non-verbal Scales consists of 42 items. The subtests are grouped into Verbal or Non-verbal Scales on the basis of the level of language comprehension that is needed to answer the items. For instance, testees should be able to answer the Non-verbal items without a thorough understanding of the language used by the tester. Arithmetic reasoning questions were classified as Verbal item types because a testee's understanding of the words on the question posed is required for a correct answer (Claassen, 1996).

All the items are presented in multiple-choice format. Items in the Figure Classification, and Verbal and Quantitative Reasoning subtests present the testee with five alternatives each. Items in Figure Series and Pattern Completion subtests present the testee with four alternatives. Items in the Comprehension subtest have only two choices, namely true or false.

The actual administration of the PPG is preceded by a practice test, called the Picture Games, which are usually completed about a week before. The Picture Games contain 18 representative practice examples. The aim thereof is to make testees familiar with the material and test procedures (Claassen, 1996).

3.5.4. Translation of the instruction material for PPG Level 2

Many researchers have recognized the problems associated with the application of standardized testing instruments in different cultural settings (Parry, 1996). In addition to the general issues of validity, there are two major threats to the validity of standardized instruments when applied in a non-western context. These are problems that arise from cultural differences and the language translation process (Parry, 1996).

Nevertheless, the adaptations of psychometric tools are indispensable undertakings in Eritrea where the lack of experts in many areas is highly magnified, particularly in psychometrics. For the sake of optimal communication and minimizing bias effects of translation, test instructions were presented to pupils in their mother tongue. This was done by literal translation of the instruction material. This method is easy and inexpensive and allows a researcher to make comparisons between or among different cultural groups. Moreover, it enables a high level of equivalence to be maintained whereby direct comparisons can be done using the same instrument (van de Vijver & Leung, 1997).

Bilingual interpreters from Eritrea were involved in the translation of the test material. Here the culture's frame of reference and local conceptualizations of some terms and ideas were considered to a certain extent while retaining the original meaning. The translation of the test instructions was done in three stages. First, a certain group did an independent translation of the original English version into Tigrigna. Because back-translation is a popular and well-known judgmental method for evaluating the equivalence of two language forms (van de Vijver & Leung, 1997), another group did a back-translation to English. Finally, to negotiate the differences between the original English version and the back-translated version, the researcher and two other individuals who did not participate on the two former tasks did some minor adjustments and modifications to the Tigrigna version.

3.6 Test administration

3.6.1 Procedure of test administration

Six fourth-year educational psychology students from the University of Asmara were recruited as research assistants to help with the administration of the PPG Level 2. The testing was done in groups of approximately 50 children per group in the biggest classrooms or auditorium halls of the respective schools. Two testers were assigned to each room.

Due to the lack of exposure to standardized tests in the Eritrean school system, the practice test was an invaluable undertaking to familiarize children with the test material. This was done one week before the actual test administration, as per the PPG manual (1996).

3.6.2 Setting

Although, the physical setting for testing varied for each school, the participants at each of the schools were tested in a similar classroom at their respective schools. The setting in each school was quiet and well lit.

3.6.3 Time limits

According to the PPG manual, the time allowed to complete the actual test is approximately two and a half to three hours. The time used varied considerably from group to group, because the examinees are led through the test item by item. However, the general report was that on average they finished after between two hours and twenty minutes to two hours and forty-five minutes, with a break of 15 minutes included.

3.6.4 Period of test administration

Research assistants were trained on how to administer the PPG practice and actual tests for five days, from 25 February to 1 March 2002. The topics included were standardization on test administration and its effect on test reliability and validity, preparing a room for test administration, controlling cheating, and sample selection. This was also augmented by providing relevant reading materials to the research assistants such as chapters on test administration. On the last two days of training, the instruction materials of the practice and actual tests were read one by one. Thereafter minor corrections such as spelling errors were made.

On 11 March 2002, a meeting was held with the research assistants to discuss the challenges and problems encountered in practice test administration so that each participant would be able to learn and make the necessary corrections, preparations and adjustments for the actual test administration. All in all, the actual test and practice test were administered over a period of three weeks, namely between 5 March and 23 March 2002.

3.7 Ethics

Although participants were probably too young to refuse involvement in the study, they were informed beforehand about what they were going to do, the purposes and the implications of the research. Twenty-eight children, who took the practice test, did not attend the actual test administration and were not forced to do so (because participation was voluntary). In addition, the test does not harm participants physically or psychologically. Test results are also kept confidential, that is only the researcher and the supervisor have access to individual testee's results.

3.8 Data analysis

The data was analyzed using the following statistical packages: SPSS, WINSTEPS and AMOS. The descriptive statistics, classical item analysis and correlation coefficients were calculated with the SPSS package. The one-parameter logistic item response theory analysis (Rasch model) was done using WINSTEPS. Finally, confirmatory factor analysis was done with the AMOS.

3.8.1 Classical item analysis

Corrected item-total correlations were calculated for each item so as to identify the specific contribution of each item to the reliability of the PPG scales. Item difficulties were also calculated.

3.8.2 Rasch analysis (1-parameter logistic item response theory analysis)

According to the Rasch model, the probability that an individual will succeed on an item depends on two aspects, namely (a) the ability of the individual (B) and (b) the difficulty of the item (D). In Rasch analysis person ability and item difficulty are expressed on the same logit scale. If an individual's ability matches the difficulty of an item, he or she will have a 50% chance of answering the item correctly. If, however, the individual's ability exceeds the item difficulty, there is a greater than 50% chance that he or she will answer the item correctly. Similarly, if the item difficulty exceeds the individual's ability, there is a less than 50% chance that he or she will answer the item correctly (Bond & Fox, 2001). These relationships can be mathematically expressed in the following formula:

$$P_{ni}(x = 1) = f(B_n - D_i)$$

This formula states that the probability of person *n* correctly answering item *i,* is a function of the difference between person *n*'s ability and item *i*'s difficulty. The function "expressing the probability of a successful response consist of a logarithmic transformation of the person B_n and item D_i estimates" (Bond & Fox, 2001), and is expressed mathematically as follows:

$$P_{ni}\,(x_{ni} = 1/B_n,\, D_i) = \frac{e^{(Bn-Di)}}{1 + e^{(Bn-Di)}}$$

where P_{ni} (x_{ni} = 1/Bn, Di) are the probability of person *n* on item *i* scoring a correct (x = 1) response given person ability (B_n) and item difficulty (D_i), and e is the natural log function (2.7183).

Person ability and item difficulty is estimated using maximum likelihood procedures. One of the attractive theoretical features of the Rasch model is that the raw scores for persons and items are sufficient statistics for the estimation of person and item parameters. This property leads to a condition called specific objectivity, which holds that person ability can be estimated separately from item difficulty and vice versa. This means, in theory, that an individual's ability estimate is independent of the particular sample of items that were chosen and that an item's difficulty estimate is independent of the particular persons that were chosen for the calibration of the items. This feature is desirable for the analysis of data sets with missing data, for building item banks, for computerized testing, and for test equating (Bond & Fox, 2001).

The estimated person and item parameters can be used in turn to estimate for each individual his or her probability of success on each item. These probabilities may then be compared with the actual data; on the basis of this comparison the fit of the items and persons to the Rasch model may be computed. Commonly used fit statistics are the INFIT mean square and the OUTFIT mean square. The INFIT is sensitive to deviations from the measurement model for on-target items, whereas OUFTIT is more sensitive to deviations from the measurement model for off-target items. INFIT and

OUTFIT mean squares have an expected value of 1.0. Values below 1.0 indicate that the person or item overfits the model (less variation in the observed responses than were modelled), whereas values above 1.0 indicate a less than desirable fit (more variation in the observed responses than was modelled). Generally, fit values below 1.0 are of less concern than fit values greater than 1.0. It is of course unrealistic to expect that items or persons will fit the model exactly. Hence, following the recommendations of Linacre and Wright (1994), items with INFIT and OUTFIT mean squares between 0.7 and 1.3 were regarded as demonstrating adequate fit.

Note that the Rasch model does not include item discrimination as a parameter. The Rasch model is a unidimensional measurement model and requires non-crossing item characteristic curves. Hence, the model proceeds on the assumption that all items discriminate equally well. Items that do not satisfy this requirement are usually multidimensional and will not fit the model properly (as indicated by INFIT and OUTFIT). The Rasch model also does not include a guessing parameter. From the Rasch perspective guessing is seen as a characteristic of people and not of items. The responses of individuals who guess correctly on items that exceed their ability will not fit the model and this will be reflected in the INFIT and OUTFIT mean squares. It is then up to the researcher to decide how to deal with the unexpected responses (Bond & Fox, 2001). In the present study the focus fell more on the item parameters than on the person parameters and therefore guessing was not explicitly considered to be a problem.

A further attractive feature of the Rasch model is that it provides a relatively simple approach to the investigation of differential item functioning (DIF). An item is said to function differently if individuals with equal ability, but from different groups, do not have an equal probability of succeeding on the item. In the Rasch model DIF is detected if the item difficulty estimates for two groups differ, because this would mean that individuals with the same level of ability, but from different groups, will not have the same probability of succeeding on the item. The presence of DIF is simply investigated by comparing the item difficulty estimates for two groups by means of a *t*-test. Alternatively, the item difficulty values for the two groups may be plotted against each other and those items whose coordinates deviate from the diagonal may be flagged for DIF (Bond & Fox, 2001).

3.8.3 Descriptive statistics

Measures of central tendency (mean, median, mode) and measures of dispersion (variance, standard deviation, kurtosis, skewness and range) were calculated for the three scales (PPG Total, Verbal and Non-verbal) for the group. The aim is to visualize the general distribution and trend of the scores. Moreover, the means and standard deviations of PPG Total, Verbal, Non-verbal, and academic performance scores for the 11 schools were computed independently.

3.8.4 Pearson correlation coefficients

The correlation coefficients between the three scales (PPG Total, Verbal and Non-verbal) and academic performance composite of the 11 schools were calculated. In addition, the pooled correlation coefficients between the three scales of PPG Level 2 and academic performance composite were obtained using HETCOR. This was done using the formula by Charter and Alexander (1993, in Silver & Hittner, 1997).

3.8.4 Confirmatory factor analysis

The validity of the PPG was assessed by means of confirmatory factor analysis. Two measurement models were specified and compared. Model 1 simply specified that the five subtests measure one general common factor, and that each subtest measures a specific factor. This model is congruent with Spearman's two-factor theory of intelligence. Model 2 specified two separate, but correlated verbal and non-verbal factors, with two subtests defining the verbal factor and three subtests defining the non-verbal factor. This model is consistent with the practice of obtaining separate verbal and non-verbal scores.

3.8.6 Comparison of means

The mean differences of the scores of the five subtests of the PPG Level 2, PPG Total score, Verbal and Non-verbal Scales scores, between males and females, and among the three regional groups were statistically tested using an independent samples *t*-test and one-way analysis of variance, respectively.

3.9 Summary

The PPG Level 2 was administered to 577 Tigrigna-speaking Grade 3 children. Participants were selected from the capital (Asmara), three small towns, and four village schools using a multi-stage sampling strategy (stratification, cluster formation and random sampling). A total of 11 schools participated. Classical and Rasch item analyses were used to test how the items function in Eritrean context. Scores for the PPG scales were correlated with teacher ratings of academic achievement. Furthermore, confirmatory factor analysis was done to examine whether the structural meaning of PPG level 2 was retained or not. Finally, mean PPG scores were compared between sexes and among regional groups.

The next chapter presents the results of the study.

Chapter Four

Results

4.1 Introduction

This chapter presents the results of the data analyses in seven broad sections. Section 1 focuses on the classical item analysis of the PPG Non-verbal and Verbal Scales, while section 2 focuses on the Rasch analysis of the Non-verbal and Verbal Scales. Differential item functioning on the Non-verbal Scale items for boys and girls will be presented in section 3, followed by differential item functioning on the Verbal Scale items for boys and girls in section 4. Section 5 presents the descriptive statistics of the PPG Level 2, Non-verbal Scale and Verbal Scale; and the correlations between Non-verbal Scale, Verbal Scale, the PPG Level 2 and the academic performance composite. The results of a confirmatory factor analysis of the PPG Level 2 are presented in section 6. Finally, section 7 focuses on the comparisons of mean scores of (a) boys and girls and (b) the three geographic regions.

4.2 Classical item analysis of the PPG Non-verbal Scale

The 42 items of the PPG Non-verbal Scale were subjected to a classical item analysis, with a specific focus on (a) each item's difficulty value, (b) each item's corrected point-biserial correlation with the total score, and (c) each item's contribution to the reliability of the total score. The results of the item analysis are summarized in Table 2.

Table 2. *Item-total Statistics of the PPG Non-verbal Scale*

Items	*p*	Corrected item total correlation	Alpha if item deleted
a5	.98	-.03	.86
a6	.97	.26	.85
a7	.93	.18	.86
a8	.50	.12	.86
a9	.91	.01	.86
a10	.89	.10	.86
a11	.83	.24	.85
a12	.40	.23	.86
a13	.83	.21	.85
a14	.48	.25	.85
a15	.21	.15	.86
a16	.80	.22	.85
c7	.90	.44	.85
c8	.61	.44	.85
c9	.84	.50	.85
c10	.88	.47	.85
c11	.77	.44	.85
c12	.83	.47	.85
c13	.83	.58	.85
c14	.60	.39	.85
c15	.70	.48	.85
c16	.75	.49	.85
c17	.50	.22	.86
c18	.51	.31	.85
c19	.37	.31	.85
c20	.42	.27	.85
e5	.83	.37	.85
e6	.90	.29	.85
e7	.68	.26	.85
e8	.74	.45	.85
e9	.80	.41	.85

Table 2. *Item-total Statistics of the PPG Non-verbal Scale (continued)*

Items	*p*	Corrected item total correlation	Alpha if item deleted
e10	.89	.29	.83
e11	.94	.29	.85
e12	.93	.33	.85
e13	.60	.49	.85
e14	.69	.39	.85
e15	.68	.53	.85
e16	.68	.49	.85
e17	.55	.34	.85
e18	.54	.35	.85
e19	.46	.31	.85
e20	.69	.28	.85

Note. Items a5 to a15 (Figure Classification subtest); items c7 to c20 (Figure Series subtest); items e5 to e20 (Pattern Completion subtest).

4.2.1 Item difficulties of the PPG Non-verbal Scale items

As can be seen in the second column of Table 2, the *p* values range between .21 (item a15) and .98 (item a5). From a classical test theory point of view items are useful to the extent that they contribute to the variance, and therefore also the reliability, of the test. Items that are very easy ($p > .80$) or very difficult ($p < .20$) have small variances and contribute little to the variance of the total test score. Furthermore, the correlation of such items with the total score will be attenuated due to their restricted range. Hence, such items are possible candidates for revision or for elimination from the test.

Seventeen items had *p* values > .80, namely a5, a6, a7, a9, a10, a11, a13, c7, c9, c10, c12, c13, e5, e6, e10, e11, and e12. The majority of these items are at the beginning of each of the subtests. Overall, it appears that many items are too easy for this group. In contrast, none of the items had *p* values < .20.

4.2.2 Corrected item-total correlations of the PPG Non-verbal Scale items with the Scale score

The third column of Table 2 presents the corrected point-biserial correlations of the non-verbal items with the total score of the PPG Non-verbal Scale. The criterion for a meaningful corrected correlation was set at $r \geq .20$. The item-total correlations ranged between -.03 and .58. Items a5, a7, a8, a9, a10, and a15 failed to meet the criterion for a meaningful correlation. Note that these items, except items a8 and a15, were also identified as being very easy items.

4.2.3 Contribution of each PPG Non-verbal item to the reliability of the Scale score

Items that contribute to the reliability of the total score are considered to be useful items, whereas items that lower the reliability of the total score are definite candidates for revision or elimination from the scale. The reliability (Cronbach's alpha) of the PPG Non-verbal Scale was estimated to be .85. The results in the fourth column show that no item's inclusion led to a meaningful reduction in the reliability of the Scale score.

4.2.4 Synthesis of the classical item-analysis of the PPG Non-verbal Scale items

In the section above it was shown that no item's inclusion lead to a reduction in the reliability of the Scale. The observed reliability (α = .85) can be described as satisfactory and compares well with the reliabilities reported by Claassen (1996) for South African children, as indicated below. However, several items had very high *p* values, demonstrating that these items were too easy for this group of participants. This is also likely to have contributed to the low item-total correlations of these items, reflecting the problem in classical test theory that item difficulty and item discrimination cannot be separated.

In the original study by Claassen (1996) the following **KR20** reliability coefficients were calculated:

Verbal Scale	**KR20** = .80
Non-verbal Scale	**KR20** = .90
PPG Level 2	**KR20** = .91

The reliability coefficients of the Non-verbal (.85) and Total PPG (.86) Scales for the present study are comparable to those reported by Claassen (1996).

4.3 Classical item analysis of the PPG Verbal Scale

The 42 items of the PPG Verbal Scale were subjected to a classical item analysis, with particular emphasis on (a) each item's difficulty value, (b) each item's corrected point-biserial correlation with the total score, and (c) each item's contribution to the reliability of the total score. The results of the item analysis are shown in Table 3 overleaf.

Table 3. *Item-total statistics of the PPG Verbal Scale*

Items	*p*	Corrected item total correlation	Alpha if item deleted
b5	.99	.05	.72
b6	.93	.25	.71
b7	.87	.22	.71
b8	.94	.25	.71
b9	.65	.13	.72
b10	.83	.21	.71
b11	.95	.30	.71
b12	.82	.28	.71
b13	.88	.32	.71
b14	.97	.28	.71
b15	.59	.27	.71
b16	.91	.27	.71
b17	.55	.23	.71
b18	.94	.24	.71
b19	.72	.40	.70
b20	.65	.27	.71
b21	.86	.19	.71
b22	.65	.25	.71
b23	.84	.28	.71
b24	.77	.16	.71
b25	.58	.22	.71
b26	.89	.20	.71
b27	.95	.17	.71
b28	.77	.29	.71
b29	.58	.23	.71
b30	.97	.08	.72
d4	.92	.26	.71
d5	.97	.13	.71
d6	.82	.16	.71
d7	.74	.24	.71
d8	.78	.10	.72

Table 3. *Item-total statistics of the PPG Verbal Scale (continued)*

Items	*p*	Corrected item total correlation	Alpha if item deleted
d9	.75	.14	.72
d10	.93	.23	.71
d11	.95	.11	.71
d12	.88	.23	.71
d13	.70	.27	.71
d14	.83	.13	.71
d15	.83	.12	.72
d16	.71	.08	.72
d17	.83	.19	.71
d18	.53	.15	.72
d19	.33	.15	.72

Note. Items b5 to b30 (Verbal and Quantitative Reasoning subtest); items d4 to d19 (Comprehension subtest).

4.3.1 Item difficulties of the PPG Verbal Scale items

The second column of Table 3 shows that the items' *p* values range between .33 (item d19) and .99 (item b5). More than half of the items, namely 24 items, had *p*-values > .80, namely b5, b6, b7, b8, b10, b11, b12, b13, b14, b16, b18, b21, b23, b26, b27, b30, d5, d6, d10, d11, d12, d14, d15 and d17. In contrast, none of the items had *p* values < .20.

4.3.2 Corrected item-total correlations of the PPG Verbal Scale items with the Scale score

Column three of Table 3 presents the corrected point-biserial correlations of the verbal items with the total score of the PPG Verbal Scale. The criterion for a meaningful corrected correlation was set at $r \geq .20$. The results show that the item-total correlations ranged between .05 and .40. Eighteen items: b5, b9, b21, b24, b26, b27, b30, d5, d6, d8, d9, d11, d14, d15, d16, d17, d18, and d19 failed to meet the criterion for a meaningful correlation. Note that these items, except items d18 and d19, were also identified among the easy items.

4.3.3 Contribution of each PPG Verbal item to the reliability of the Scale score

The reliability (Cronbach's alpha) of the PPG Verbal Scale was estimated to be .72. The results in the fourth column show that no item's inclusion led to a meaningful reduction in the reliability of the Scale score.

4.3.4 Synthesis of the classical item analysis of the PPG Verbal Scale items

In the section above it was shown that no item's inclusion lead to a reduction in the reliability of the Scale. The observed reliability (α = .72) can be described as unsatisfactory and is lower than the reliabilities reported by Claassen (1996) for South African children. Many of the items had very high *p* values, demonstrating that these items were too easy for this group of participants. This is also likely to have contributed substantially to the very low item-total correlations of these items, reflecting the problem in classical test theory that item difficulty and item discrimination cannot be separated.

4.4 Rasch analysis of the PPG Non-verbal Scale

The items of the PPG Non-verbal Scale were subjected to a Rasch analysis (also referred to as the one-parameter logistic item response theory model). The major advantage of the Rasch model over classical item analysis is (a) the estimates of item difficulty are independent of the particular group of participants, (b) the estimates of person ability are independent of the particular set of items, and (c) item difficulty estimates are not confounded with item discrimination estimates (which in the case of the Rasch model are assumed to be constant for all items).

The estimated item difficulties (in logits) and their associated standard errors, INFIT and OUTFIT statistics are presented in Table 4. The items are presented in their order of difficulty (items at the top of the table are more difficult than items at the bottom). The item difficulties range between -3.35 logits (item a5) and 2.93 logits (item a15). (Note that the average item difficulty logit is set to zero.)

The same information is presented in Table 5, but here the items are presented in their INFIT mean square order. Following the recommendations of Linacre and Wright (1994) the criterion for acceptable fit was set at INFIT and OUTFIT mean square values ranging between .70 and 1.30. Inspection of the INFIT mean squares show that they ranged between .76 (item c13) and 1.24 (item a8). However, five items had OUTFIT mean squares above 1.3, namely items a5 (2.55), a9 (2.38), a10 (1.43), a8 (1.34), and a15 (1.33).

Table 4. *PPG Non-verbal Scale items measure order*

Entry Number	Score Raw	Count	Measure	Error	INFIT mnsq	INFIT zstd	OUTFIT mnsq	OUTFIT zstd	Ptbis Corr.	Items
11	123	577	2.93	.11	1.15	2.4	1.33	2.7	.22	a15
51	211	577	2.03	.10	.98	-.5	1.08	1.3	.41	c19
8	232	577	1.84	.09	1.10	2.8	1.16	2.7	.32	a12
52	241	577	1.76	.09	1.07	2.0	1.11	1.9	.36	c20
83	264	577	1.56	.09	1.03	.9	1.08	1.4	.39	e19
10	276	577	1.46	.09	1.09	2.6	1.16	3.0	.34	a14
49	286	577	1.37	.09	1.14	3.7	1.17	3.3	.32	c17
4	290	577	1.34	.09	1.24	6.5	1.34	6.2	.21	a8
50	292	577	1.32	.09	1.04	1.3	1.09	1.8	.39	c18
82	312	577	1.15	.09	1.02	.61	.03	.5	.41	e18
81	317	577	1.11	.09	1.02	.71	.03	.6	.41	e17
46	344	577	.87	.09	.99	-.3	.98	-.4	.44	c14
77	349	577	.83	.09	.87	-3.5	.82	-3.3	.54	e13
40	353	577	.79	.09	.93	-1.9	.87	-2.3	.50	c8
71	393	577	.42	.10	1.10	2.2	1.15	2.0	.33	e7
80	394	577	.41	.10	.87	-2.9	.79	-3.1	.53	e16
79	395	577	.40	.10	.83	-3.8	.77	-3.4	.56	e15
84	397	577	.38	.10	1.09	2.0	1.08	1.1	.34	e20
78	400	577	.35	.10	.97	-.6	.98	-.2	.44	e14
47	406	577	.29	.10	.89	-2.5	.93	-.9	.50	c15
72	426	577	.08	.10	.91	-1.8	.83	-2.0	.49	e8
48	432	577	.01	.11	.86	-2.7	.81	-2.3	.51	c16

Table 4. *PPG Non-verbal Scale items measure order (continued)*

Entry Number	Score Raw	Count	Measure	Error	INFIT mnsq	INFIT zstd	OUTFIT mnsq	OUTFIT zstd	Ptbis Corr.	Items
43	443	577	-.11	.11	.92	-1.5	.82	-1.9	.47	c11
12	459	577	-.31	.11	1.12	1.9	1.18	1.5	.28	a16
73	460	577	-.32	.11	.93	-1.1	.93	-.7	.43	e9
9	476	577	-.53	.12	1.12	1.6	1.12	.9	.27	a13
44	478	577	-.56	.12	.85	-2.3	.78	-1.8	.48	c12
7	479	577	-.58	.12	1.08	1.1	1.08	.6	.29	a11
45	479	577	-.58	.12	.76	-3.7	.56	-4.0	.57	c13
69	481	577	-.61	.12	.96	-.6	.93	-.5	.40	e5
41	486	577	-.68	.12	.83	-2.4	.68	-2.6	.50	c9
42	506	577	-1.01	.14	.84	-1.9	.68	-2.2	.46	c10
74	512	577	-1.13	.14	1.00	.0	.94	-.3	.32	e10
6	513	577	-1.15	.14	1.14	1.4	1.43	2.1	.15	a10
70	519	577	-1.27	.15	.98	-.2	.93	-.4	.31	e6
39	520	577	-1.29	.15	.85	-1.5	.81	-1.0	.43	c7
5	527	577	-1.45	.16	1.16	1.42	.38	4.7	.05	a9
3	534	577	-1.63	.17	1.04	.3	1.14	.6	.21	a7
76	537	577	-1.72	.17	.95	-.4	.64	-1.7	.33	e12
75	540	577	-1.81	.18	.95	-.3	.70	-1.3	.30	e11
2	558	577	-2.56	.24	.92	-.4	.63	-1.1	.26	a6
1	568	577	-3.35	.34	1.05	.22	.55	2.0	-.01	a5
MEAN	410.	577.	.00	.12	.99	.01	.04	.1		
S.D.	109.	0.	1.31	.05	.11	2.1	.38	2.3		

Table 5. *PPG Non-verbal Scale items misfit order*

Entry Number	Score Raw	Count	Measure	Error	INFIT		OUTFIT		Items
					mnsq	zstd	mnsq	zstd	
4	290	577	1.34	.09	1.24	6.5	1.34	6.2	a8
5	527	577	-1.45	.16	1.16	1.4	2.38	4.7	a9
11	123	577	2.93	.11	1.15	2.4	1.33	2.7	a15
6	513	577	-1.15	.14	1.14	1.4	1.43	2.1	a10
49	286	577	1.37	.09	1.14	3.7	1.17	3.3	c17
12	459	577	-.31	.11	1.12	1.9	1.18	1.5	a16
9	476	577	-.53	.12	1.12	1.6	1.12	.9	a13
8	232	577	1.84	.09	1.10	2.8	1.16	2.7	a12
71	393	577	.42	.10	1.10	2.2	1.15	2.0	e7
84	397	577	.38	.10	1.09	2.0	1.08	1.1	e20
10	276	577	1.46	.09	1.09	2.6	1.16	3.0	a14
7	479	577	-.58	.12	1.08	1.1	1.08	.6	a11
52	241	577	1.76	.09	1.07	2.0	1.11	1.9	c20
1	568	577	-3.35	.34	1.05	.2	2.55	2.0	a5
50	292	577	1.32	.09	1.04	1.3	1.09	1.8	c18
3	534	577	-1.63	.17	1.04	.3	1.14	.6	a7
83	264	577	1.56	.09	1.03	.9	1.08	1.4	e19
81	317	577	1.11	.09	1.02	.7	1.03	.6	e17
82	312	577	1.15	.09	1.02	.6	1.03	.5	e18
74	512	577	-1.13	.14	1.00	.0	.94	-.3	e10
46	344	577	.87	.09	.99	-.3	.98	-.4	c14
51	211	577	2.03	.10	.98	-.5	1.08	1.3	c19

Table 5. *PPG Non-verbal Scale items misfit order (Continued)*

Entry Number	Score Raw	Count	Measure	Error	INFIT		OUTFIT		Items
					mnsq	zstd	mnsq	zstd	
70	519	577	-1.27	.15	.98	-.2	.93	-.4	e6
78	400	577	.35	.10	.97	-.6	.98	-.2	e14
69	481	577	-.61	.12	.96	-.6	.93	-.5	e5
75	540	577	-1.81	.18	.95	-.3	.70	-1.3	e11
76	537	577	-1.72	.17	.95	-.4	.64	-1.7	e12
73	460	577	-.32	.11	.93	1.1	.93	-.7	e9
40	353	577	.79	.09	.93	-1.9	.87	-2.3	c8
2	558	577	-2.56	.24	.92	-.4	.63	-1.1	a6
43	443	577	-.11	.11	.92	-1.5	.82	-1.9	c11
72	426	577	.08	.10	.91	-1.8	.83	-2.0	e8
47	406	577	.29	.10	.89	-2.5	.93	-.9	c15
80	394	577	.41	.10	.87	-2.9	.79	-3.1	e16
77	349	577	.83	.09	.87	-3.5	.82	-3.3	e13
48	432	577	.01	.11	.86	-2.7	.81	-2.3	c16
39	520	577	-1.29	.15	.85	-1.5	.81	-1.0	c7
44	478	577	-.56	.12	.85	-2.3	.78	-1.8	c12
42	506	577	-1.01	.14	.84	-1.9	.68	-2.2	c10
79	395	577	.40	.10	.83	-3.8	.77	-3.4	e15
41	486	577	-.68	.12	.83	-2.4	.68	-2.6	c9
45	479	577	-.58	.12	.76	-3.7	.56	-4.0	c13
MEAN	410.	577.	.00	.12	.99	.0			
S.D.	109.	0.	1.31	.05	.11	2.1			

An attractive feature of the Rasch model is that it allows individuals and items to be mapped on the same logit scale. This allows one to visually judge whether the items are sufficiently spread out to define a linear variable and whether the items allow discrimination between individuals in terms of ability. Furthermore, the visual map allows for an evaluation of whether the items are appropriate for the group of participants in terms of the items' difficulties. The item-person map of the PPG Non-verbal Scale items is presented in Figure 2 where persons are indicated by an #.

The map shows the distribution of the persons and items along the trait level. The variables are laid out vertically with the most able persons and most difficult items at the top, and the less able persons and the easiest items at the bottom of the map. The left-hand column locates the persons' ability measure along the latent variable. The right-hand column locates the item difficulty calibrations along the variable. An "M" marker represents the location of the mean measure. "S" markers are placed one standard deviation away from the mean. "Q" markers are placed two standard deviations away from the mean.

```
---------------------------------------------------------------
       PERSONS            MAP OF ITEMS
                  <frequ>|<less>
  5                      +
                         |
                         |
                       . |
                         |
  4                      +
                         |
                   .##   |
                         |
                        Q|
                .#####   |
  3                      +
                         |   a15
                .#####   |
                         |
              .#######   |Q
                        S|
              ########   |
  2           ########   +
         .############   |   c19
                         |   a12   c20
           ###########   |
             .########   |   e19
             .######## M|   a14   a8    c17   c18
              .#######   |S
  1            .######   +   e17   e18
              .#######   |
                .#####   |   c14   c8    e13
               .######   |
                  .###   |
                  .###   |   e14   e15   e16   e20   e7
                  .### S|   c15
  0               .###   +M  c16   e8
               .######   |   c11
                    ##   |   a16
                    .#   |   e9
                   .##   |   a13   e5
                   .##   |   a11   c12   c13   c9
                    .# Q|
 -1                  .   +   c10
                     .   |S
                     .   |   a10   e10
                     .   |   a9    c7    e6
                         |
                         |
                     .   |   a7    e11   e12
 -2                      +
                         |
                         |   a6
                         |Q
                         |
 -3                      +   a5
                  <rare>|<more>
---------------------------------------------------------------
```

Figure 2. Map of item difficulty and person ability estimates for the PPG Non-verbal Scale

Figure 2 shows that the individuals and the items are sufficiently spread out along the logit scale. However, the map also shows that there are too few difficult items to allow reliable discrimination between individuals at the upper end of the ability scale. Furthermore, there appear to be many items that are too easy for the present group of participants.

4.4.1 Synthesis of the Rasch analysis of the PPG Non-verbal Scale

According to the results of INFIT mean squares, all items functioned well for this group of children, but the OUTFIT mean squares suggested that the fit of five items were less than satisfactory. Although it seems as if the Non-Verbal scale measures an essentially unidimensional construct, many items appear to be too easy to discriminate among learners of higher cognitive ability. However, the primary purpose of PPG Level 2 is to identify children with low ability. In this sense the non-verbal items served the intended purpose.

4.5 Rasch analysis of the PPG Verbal Scale

The items of the PPG Verbal Scale were also subjected to a Rasch analysis. The estimated item difficulties (in logits) and their associated standard errors, INFIT and OUTFIT statistics are presented in Table 6. The items are presented in their order of difficulty (items at the top of the table are more difficult than items at the bottom of the table). The item difficulties range between -3.40 logits (item b5) and 2.67 logits (item d19).

The same data are presented in Table 7, but here the items are presented in their INFIT mean square order. Inspection of the INFIT and OUTFIT mean squares show that they ranged between .89 (b19) and 1.11 (d16) for the INFIT mean squares, and between .71 (b11) and 1.15 (d19) for the OUTFIT mean squares. All the items were positioned within the acceptable range of fit. Hence, it is concluded that the items meet the requirements set by the Rasch model and that they measure an essentially unidimensional construct (namely verbal ability).

Table 6. *PPG Verbal Scale item measure order*

Entry Number	Raw Score	Count	Measure	Error	INFIT		OUTFIT		Ptbis Corr.	Items
					mnsq	zstd	mnsq	zstd		
68	193	576	2.67	.09	1.07	1.7	1.15	2.6	.14	d19
67	309	576	1.72	.09	1.08	2.7	1.08	2.0	.15	d18
25	320	576	1.63	.09	1.02	.8	1.04	1.0	.23	b17
33	336	576	1.50	.09	1.02	.6	1.01	.2	.23	b25
37	338	576	1.48	.09	1.01	.3	1.00	-.1	.23	b29
23	341	576	1.46	.09	.99	-.3	1.01	.2	.26	b15
28	371	576	1.20	.09	.97	-.8	.98	-.4	.28	b20
30	376	576	1.16	.09	1.00	-.1	.98	-.4	.25	b22
17	377	576	1.15	.09	1.09	2.3	1.13	2.2	.12	b9
62	401	576	.94	.10	.97	-.6	.93	-1.1	.27	d13
65	408	576	.87	.10	1.10	2.3	1.08	1.1	.09	d16
27	420	576	.75	.10	.89	-2.5	.86	-2.0	.40	b19
56	431	576	.64	.10	1.00	.0	.94	-.8	.24	d7
58	436	576	.59	.10	1.07	1.3	1.06	.7	.13	d9
57	442	576	.53	.10	1.09	1.6	1.11	1.3	.10	d8
32	444	576	.51	.10	1.05	.9	1.04	.5	.16	b24
36	447	576	.47	.11	.95	-.8	.92	-1.0	.29	b28
55	472	576	.18	.11	1.04	.6	1.13	1.2	.16	d6
20	473	576	.16	.11	.95	-.7	.89	-1.1	.28	b12
31	477	576	.11	.12	.96	-.5	.92	-.7	.27	b23
63	477	576	.11	.12	1.05	.7	1.10	.9	.14	d14
18	479	576	.08	.12	.99	-.1	.99	-.1	.22	b10
66	479	576	.08	.12	1.02	.3	1.04	.4	.18	d17

Table 6. *PPG Verbal Scale item measure order (continued)*

Entry Number	Raw Score	Count	Measure	Error	INFIT mnsq	INFIT zstd	OUTFIT mnsq	OUTFIT zstd	Ptbis Corr.	Items
64	481	576	.06	.12	1.06	.9	1.04	.4	.11	d15
29	493	576	-.12	.12	1.01	.1	1.04	.3	.18	b21
15	501	576	-.24	.13	.99	-.2	.88	-.9	.22	b7
34	506	576	-.33	.13	.99	-.1	1.11	.7	.18	b26
61	510	576	-.40	.14	.98	-.2	.88	-.8	.23	d12
21	511	576	-.42	.14	.92	-.8	.82	-1.3	.32	b13
24	523	576	-.66	.15	.94	-.5	.91	-.6	.26	b16
53	524	576	-.68	.15	.96	-.3	.87	-.8	.25	d4
59	536	576	-.98	.17	.96	-.3	1.03	.1	.24	d10
14	538	576	-1.04	.17	.95	-.4	.77	-1.2	.25	b6
26	541	576	-1.13	.18	.95	-.4	.79	-1.0	.24	b18
16	542	576	-1.17	.18	.94	-.4	.73	-1.3	.25	b8
19	545	576	-1.27	.19	.92	-.5	.71	-1.4	.29	b11
35	547	576	-1.34	.19	.99	-.1	.79	-.9	.16	b27
60	547	576	-1.34	.19	1.02	.1	1.08	.3	.10	d11
22	556	576	-1.74	.23	.92	-.4	.59	-1.6	.27	b14
38	556	576	-1.74	.23	1.03	.1	.92	-.3	.07	b30
54	561	576	-2.04	.26	1.00	.0	.92	-.2	.13	d5
13	572	576	-3.40	.50	1.01	.0	1.04	.1	.05	b5
MEAN	460.	576.	.00	.14	1.00	.1	.96	-.1		
S.D.	84.	0.	1.20	.07	.05	1.0	.13	1.0		

Table 7. *PPG Verbal Scale items misfit order*

Entry Number	Raw Score	Count	Measure	Error	INFIT mnsq	INFIT zstd	OUTFIT mnsq	OUTFIT zstd	Items
65	408	576	.89	.10	1.11	2.5	1.08	1.1	d16
57	447	576	.49	.11	1.09	1.6	1.11	1.3	d8
17	375	576	1.19	.09	1.09	2.3	1.13	2.2	b9
67	307	576	1.76	.09	1.08	2.7	1.08	2.0	d18
68	189	576	2.73	.10	1.07	1.7	1.15	2.6	d19
58	434	576	.63	.10	1.07	1.3	1.06	.7	d9
64	479	576	.10	.12	1.06	.9	1.04	.4	d15
63	478	576	.12	.12	1.05	.7	1.10	.9	d14
32	443	576	.54	.10	1.05	.9	1.04	.5	b24
55	472	576	.20	.11	1.04	.6	1.13	1.2	d6
33	333	576	1.54	.09	1.03	.8	1.01	.2	b25
38	560	576	-1.96	.26	1.02	.1	.92	-.3	b30
25	317	576	1.67	.09	1.02	.7	1.04	1.0	b17
66	480	576	.09	.12	1.02	.2	1.04	.4	d17
60	549	576	-1.40	.20	1.01	.1	1.08	.3	d11
37	334	576	1.54	.09	1.01	.2	1.00	-.1	b29
29	494	576	-.11	.12	1.00	.0	1.04	.3	b21
13	573	576	-3.67	.58	1.00	.0	1.04	.1	b5
30	372	576	1.22	.09	1.00	.0	.98	-.4	b22
56	426	576	.72	.10	1.00	.0	.94	-.8	d7
18	480	576	.09	.12	1.00	.0	.99	-.1	b10
54	561	576	-2.02	.26	1.00	.0	.92	-.2	d5
23	337	576	1.51	.09	.99	-.3	1.01	.2	b15

Table 7. *PPG Verbal Scale items misfit order (continued)*

Entry Number	Raw Score	Count	Measure	Error	INFIT mnsq	INFIT zstd	OUTFIT mnsq	OUTFIT zstd	Items
15	501	576	-.22	.13	.99	-.1	.88	-9	b7
35	548	576	-1.36	.20	.99	-.1	.79	-9	b27
34	510	576	-.38	.14	.98	-.2	1.11	.7	b26
28	373	576	1.21	.09	.98	-.5	.98	-.4	b20
62	400	576	.97	.10	.98	-.6	.93	-1.1	d13
61	509	576	-.36	.13	.98	-.3	.88	-.8	d12
59	536	576	-.96	.17	.96	-.2	1.03	.1	d10
31	481	576	.08	.12	.96	-.6	.92	-.7	b23
53	528	576	-.76	.15	.96	-.4	.87	-.8	d4
36	446	576	.51	.11	.95	-.8	.92	-1.0	b28
20	473	576	.18	.11	.95	-.7	.89	-1.1	b12
14	538	576	-1.02	.17	.95	-.4	.77	-1.2	b6
26	544	576	-1.21	.19	.94	-.4	.79	-1.0	b18
16	542	576	-1.15	.18	.94	-.4	.73	-1.3	b8
24	527	576	-.73	.15	.93	-.6	.91	-.6	b16
21	509	576	-.36	.13	.92	-.9	.82	-1.3	b13
22	558	576	-1.83	.24	.92	-.4	.59	-1.6	b14
19	546	576	-1.28	.19	.92	-.5	.71	-1.4	b11
27	416	576	.82	.10	.89	-2.6	.86	-2.0	b19
MEAN	460	576	.00	.14	1.00	.1			
S.D.	86	0	1.25	.08	.05	1.0			

The item-person map of the PPG Verbal Scale items is presented in Figure 3.

```
----------------------------------------------------------
      PERSONS              MAP OF ITEMS
             <frequ>|<less>
   5               .  +
                      |
                  .#  |
                      |
   4                  +
                      |
                      |
                .### Q|
                      |
              .#####  |
   3                  +
                      |
            .####### S|    d19
           .########  |
                      |Q
        .###########  |
   2      ##########  +
            .####### M|
             .######  |    b17    d18
              .#####  |    b15    b25    b29
              .#####  |
              .#####  |S   b20    b22    b9
   1           ##### S+    d13
             .######  |    b19    d16
                  .#  |    d7     d9
                  .#  |    b24    b28    d8
                  .#  |
                   . Q|    b12    b23    d14    d6
   0               .  +M   b10    d15    d17
                   .  |    b21    b7
                   .  |    b26    d12
                      |    b13
                      |    b16    d4
                      |
  -1               .  +    b6     d10
                      |S   b18    b8
                      |    b11    b27    d11
                      |
                      |    b14    b30
                      |
  -2                  +    d5
                      |
                      |Q
                      |
                      |
  -3                  +
                      |
                      |    b5
                      |
  -4                  +
              <rare>|<more>
----------------------------------------------------------
```

*Figure 3.*Map of item difficulty and person ability estimates for the PPG Verbal Scale

Figure 3 shows that the individuals and the items are sufficiently spread out along the logit scale. However, the map also shows that there are too few difficult items to allow reliable discrimination between individuals at the upper end of the ability scale. Furthermore, there appear to be too many verbal items that are too easy for the present group of participants.

4.5.1 Synthesis of the Rasch analysis of the PPG Verbal Scale

The assumption of unidimensionality is satisfied from the results of the INFIT and OUTFIT mean squares in the Rasch analysis. All the items have an INFIT and OUTFIT mean square of close to 1.00. However, many items are too easy to discriminate between learners with different abilities. Good tests usually have the items targeted to (lined up with) the persons. In Figure 2 (Non-verbal Scale), however, 10 items are located below the logit of -1 without any person located in that range. Similarly, in Figure 3 (Verbal Scale), 20 items are placed below the logit of 0 with no testees on the trait level. However, as mentioned in the above section, the test was designed to discriminate among children at low ability levels.

4.6 Differential item functioning on the PPG Non-verbal items for boys and girls

One of the most useful properties of the Rasch model is its ability to detect and quantify differential item functioning (DIF). In the Rasch model (where item discrimination is assumed to be constant across all items), DIF is present to the extent that item difficulty estimates are dissimilar for the different groups of interest. The results of the DIF analysis are presented in Table 8 overleaf.

Table 8. *DIF results on the PPG Non-verbal Scale items for boys and girls*

Person Group[a]	DIF Added	DIF s.e.	Person Group[b]	DIF Added	DIF s.e.	DIF contrast	Joint s.e	t	d.f.	Item Number	Name
1	-.15	.51	2	.13	.46	-.28	.68	-.41	575	1	a5
1	.14	.32	2	-.16	.37	.30	.48	.62	575	2	a6
1	.25	.21	2	-.32	.27	.58	.34	1.68	575	3	a7
1	.42	.13	2	-.47	.14	.89	.19	4.74**	575	4	a8
1	.20	.20	2	-.25	.24	.45	.32	1.41	575	5	a9
1	-.02	.20	2	.02	.20	-.04	.28	-.15	575	6	a10
1	.03	.17	2	-.04	.17	.07	.24	.29	575	7	a11
1	.18	.13	2	-.20	.14	.39	.19	2.06*	575	8	a12
1	-.16	.17	2	.15	.16	-.30	.24	-1.26	575	9	a13
1	-.14	.13	2	.16	.14	-.30	.19	-1.61	575	10	a14
1	.13	.15	2	-.15	.16	.28	.22	1.28	575	11	a15
1	.10	.15	2	-.11	.16	.20	.23	.91	575	12	a16
1	-.05	.21	2	.04	.21	-.09	.30	-.29	575	39	c7
1	-.08	.13	2	.08	.14	-.16	.19	-.84	575	40	c8
1	.05	.17	2	-.06	.18	.11	.25	.43	575	41	c9
1	-.08	.19	2	.07	.19	-.15	.27	-.57	575	42	c10
1	-.05	.15	2	.05	.15	-.10	.22	-.46	575	43	c11
1	-.01	.17	2	.01	.17	-.02	.24	-.07	575	44	c12
1	-.08	.17	2	.08	.17	-.16	.24	-.67	575	45	c13
1	.13	.13	2	-.15	.14	.28	.19	1.50	575	46	c14
1	.04	.14	2	-.04	.15	.08	.20	.39	575	47	c15
1	-.02	.15	2	.02	.15	-.03	.21	-.16	575	48	c16
1	.16	.13	2	-.17	.13	.33	.19	1.78	575	49	c17

Table 8. *DIF results on the PPG Non-verbal Scale items for boys and girls (continued)*

Person Group[a]	DIF Added	DIF s.e.	Person Group[b]	DIF Added	DIF s.e.	DIF contrast	Joint s.e	t	d.f.	Item Number	Name
1	.00	.13	2	.00	.13	-.01	.19	-.04	575	50	c18
1	-.02	.13	2	.03	.14	-.05	.19	-.24	575	51	c19
1	-.04	.13	2	.04	.14	-.08	.19	-.42	575	52	c20
1	-.02	.17	2	.02	.17	-.05	.24	-.19	575	69	e5
1	.10	.20	2	-.12	.22	.22	.29	.73	575	70	e6
1	-.17	.14	2	.17	.14	-.34	.20	-1.73	575	71	e7
1	-.04	.15	2	.04	.15	-.08	.21	-.38	575	72	e8
1	-.04	.16	2	.04	.16	-.07	.23	-.33	575	73	e9
1	-.12	.20	2	.11	.19	-.24	.28	-.85	575	74	e10
1	-.26	.27	2	.22	.23	-.49	.36	-1.35	575	75	e11
1	.03	.24	2	-.04	.25	.07	.34	.22	575	76	e12
1	-.10	.13	2	.10	.14	-.20	.19	-1.04	575	77	e13
1	-.14	.14	2	.14	.14	-.29	.20	-1.43	575	78	e14
1	.04	.14	2	-.05	.14	.09	.20	.45	575	79	e15
1	-.12	.14	2	.12	.14	-.24	.20	-1.23	575	80	e16
1	-.15	.13	2	.17	.13	-.32	.19	-1.70	575	81	e17
1	-.05	.13	2	.05	.13	-.10	.19	-.51	575	82	e18
1	-.03	.13	2	.04	.14	-.07	.19	-.37	575	83	e19
1	-.01	.14	2	.01	.14	-.03	.20	-.14	575	84	e20

Note. Negative value of DIF added implies item biased against the particular group. Positive DIF size is a higher item difficulty measure.

[a] Person group 1 is boys. [b] Person group 2 is girls.

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

The t-values in Table 8 show that only two non-verbal items met the criterion of statistically significant DIF ($t > 1.96$), namely items a8 ($t = 4.74$) and a12 ($t = 2.06$). The difference in item difficulty estimates for item a8 is .89 logits, which can be considered meaningful. The difference in item difficulty estimates for item a12 is .39 logits, which probably will not make a meaningful difference in practical measurement situations. Hence, it is concluded that with the possible exception of item a8, there is little evidence of differential item functioning for boys and girls on the Non-verbal Scale items. Note that item a8 was also the item with the highest INFIT (1.24) mean square and an OUTFIT (1.34) mean square greater than 1.3, suggesting that this item functions in a way that is inconsistent with the other items in the scale.

4.7 Differential item functioning on the PPG Verbal items for boys and girls

The items of the Verbal Scale were also subjected to a DIF analysis. The results of the DIF analysis are presented in Table 9 overleaf.

Table 9. *DIF results on the PPG Verbal Scale items for boys and girls*

Person Group[a]	DIF Added	DIF s.e.	Person Group[b]	DIF Added	DIF s.e.	DIF contrast	Joint s.e	t	d.f.	Item Number	Name
1	.70	.58	2			maximum score				13	b5
1	-.08	.25	2	.08	.24	-.16	.34	-.47	574	14	b6
1	-.02	.18	2	.02	.18	-.04	.26	-.14	574	15	b7
1	-.16	.27	2	.15	.24	-.30	.36	-.84	574	16	b8
1	.02	.13	2	-.02	.14	.04	.19	.21	574	17	b9
1	.06	.16	2	-.07	.17	.14	.23	.58	574	18	b10
1	-.35	.31	2	.29	.24	-.64	.40	-1.61	574	19	b11
1	-.05	.16	2	.06	.16	-.11	.23	-.50	574	20	b12
1	-.01	.19	2	.02	.19	-.03	.27	-.11	574	21	b13
1	.20	.31	2	-.25	.39	.45	.50	.91	574	22	b14
1	-.06	.12	2	.07	.13	-.13	.18	-.74	574	23	b15
1	-.10	.22	2	.10	.21	-.19	.31	-.63	574	24	b16
1	.16	.12	2	-.18	.13	.33	.18	1.85	574	25	b17
1	-.02	.26	2	.02	.26	-.04	.37	-.11	574	26	b18
1	-.06	.14	2	.07	.14	-.12	.20	-.63	574	27	b19
1	.02	.13	2	-.02	.14	.04	.19	.20	574	28	b20
1	.19	.16	2	-.23	.19	.42	.25	1.68	574	29	b21
1	-.09	.13	2	.10	.13	-.19	.19	-1.02	574	30	b22
1	.25	.15	2	-.31	.19	.56	.24	2.33*	574	31	b23
1	-.06	.15	2	.07	.15	-.13	.21	-.62	574	32	b24
1	-.06	.12	2	.07	.13	-.14	.18	-.76	574	33	b25
1	-.07	.19	2	.07	.19	-.14	.27	-.51	574	34	b26
1	.06	.27	2	-.06	.29	.12	.40	.30	574	35	b27

Table 9. *DIF results on the PPG Verbal Scale items for boys and girls (continued)*

Person Group[a]	DIF Added	DIF s.e.	Person Group[b]	DIF Added	DIF s.e.	DIF contrast	Joint s.e	t	d.f.	Item Number	Name
1	-.12	.15	2	.12	.15	-.24	.21	-1.13	574	36	b28
1	.02	.12	2	-.02	.13	.04	.18	.24	574	37	b29
1	.11	.34	2	-.13	.39	.24	.52	.46	574	38	b30
1	.11	.21	2	-.13	.23	.24	.31	.76	574	53	d4
1	-.08	.39	2	.08	.36	-.17	.53	-.31	574	54	d5
1	-.02	.16	2	.02	.16	-.04	.23	-.16	574	55	d6
1	.08	.14	2	-.09	.15	.17	.20	.82	574	56	d7
1	.07	.14	2	-.07	.16	.14	.21	.66	574	57	d8
1	-.20	.15	2	.21	.14	-.41	.21	-1.98*	574	58	d9
1	-.02	.24	2	.02	.24	-.05	.34	-.14	574	59	d10
1	-.14	.30	2	.14	.27	-.28	.40	-.69	574	60	d11
1	.09	.18	2	-.10	.20	.19	.27	.70	574	61	d12
1	.01	.13	2	-.01	.14	.01	.19	.07	574	62	d13
1	.09	.16	2	-.10	.17	.19	.23	.79	574	63	d14
1	-.03	.16	2	.03	.17	-.06	.23	-.25	574	64	d15
1	-.12	.14	2	.13	.14	-.24	.20	-1.25	574	65	d16
1	-.04	.16	2	.04	.17	-.08	.23	-.36	574	66	d17
1	.27	.12	2	-.31	.13	.58	.18	3.22**	574	67	d18
1	-.15	.13	2	.19	.14	-.34	.19	-1.76	574	68	d19

Note. Negative value of DIF added implies item biased against the particular group Positive DIF size is a higher item difficulty measure. All girls answered b5 correctly.

[a] Person group 1 is boys. [b] Person group 2 is girls.

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

The t-values in Table 9 show that only three verbal items met the criterion for statistically significant DIF ($t > 1.96$), namely items b23 ($t = 2.33$), d9 ($t = -1.98$) and d18 ($t = 3.22$). The difference in item difficulty estimates for item b23 and d18 are .56 and .58 logits, which can be considered meaningful. The difference in item difficulty estimates for item d9 is -.41 logits, which will probably not make a meaningful difference in practical measurement situations. Item d9 (Are all pencils shorter than pens?) is the only item that favoured females. The other two verbal items that functioned differently namely b23 in the Verbal and Quantitative Reasoning subtest and d18 in the Comprehension subtest, are related to play and play materials that might be attributed to socialization effects of gender differences. Hence, it is concluded that with the possible exception of item b23 and d18, there is little evidence of differential item functioning for boys and girls on the Verbal Scale items.

4.8 Descriptive statistics of the PPG Total, Non-verbal, and Verbal Scales scores

In the previous sections it was established that the Non-verbal and Verbal Scales are essentially unidimensional and that almost all items contribute to the quality of the respective scales. The averages, standard deviations, minimum and maximum scores, skewness, and kurtosis values of the three scales are presented in Table 10.

Table 10. *Descriptive statistics summary (N = 577)*

	Non-verbal	Verbal	PPG Total
Items	42	42	84
Mean	29.82	33.57	63.40
Median	31	34	65
Mode	34	36	62
Standard deviation	6.6	4.39	9.43
Minimum score	9	13	34
Maximum score	41	42	82
Skewness	-0.727	-0.768	-0.643
Kurtosis	-0.086	0.658	-0.115

The skewness coefficients for all three Scales showed that the scores are negatively skewed, suggesting that most items were easy for this group of participants.

To clarify the distribution and location of the obtained scores for the three scales, Figures 4, 5, and 6 present their frequency distributions. For further detail of the frequency distributions, see Appendix A.

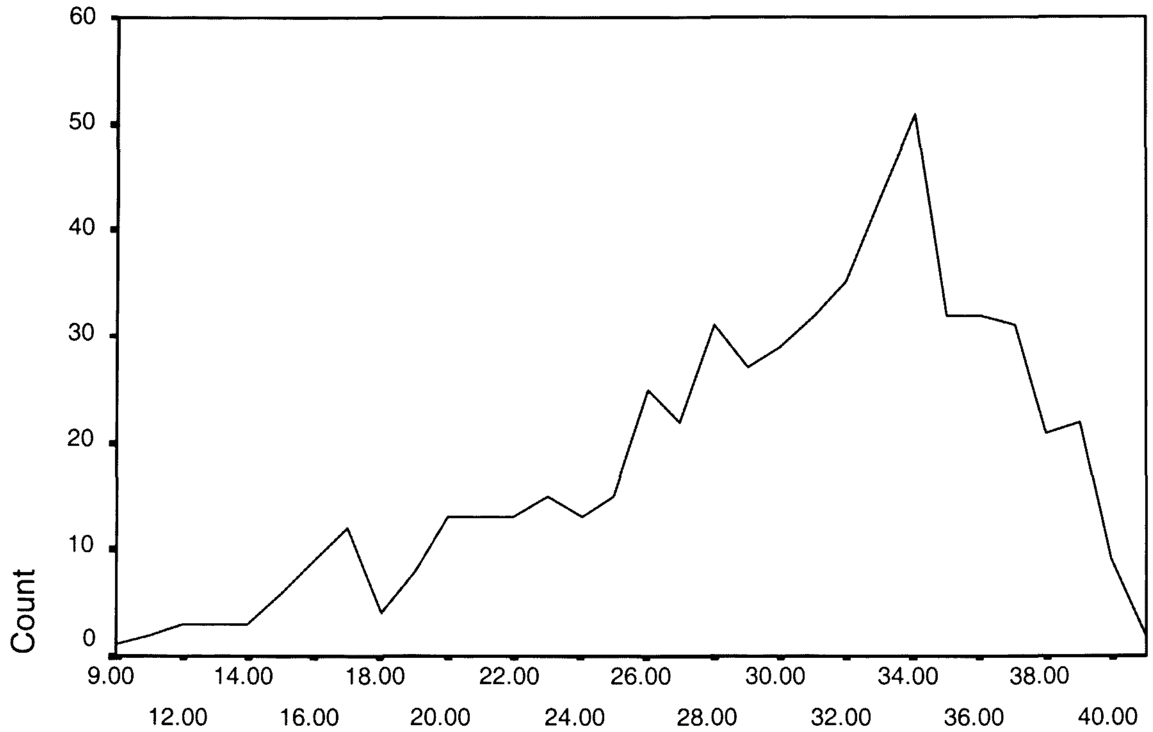


Figure 4. Frequency distribution of the Non-verbal Scale scores (N = 577)

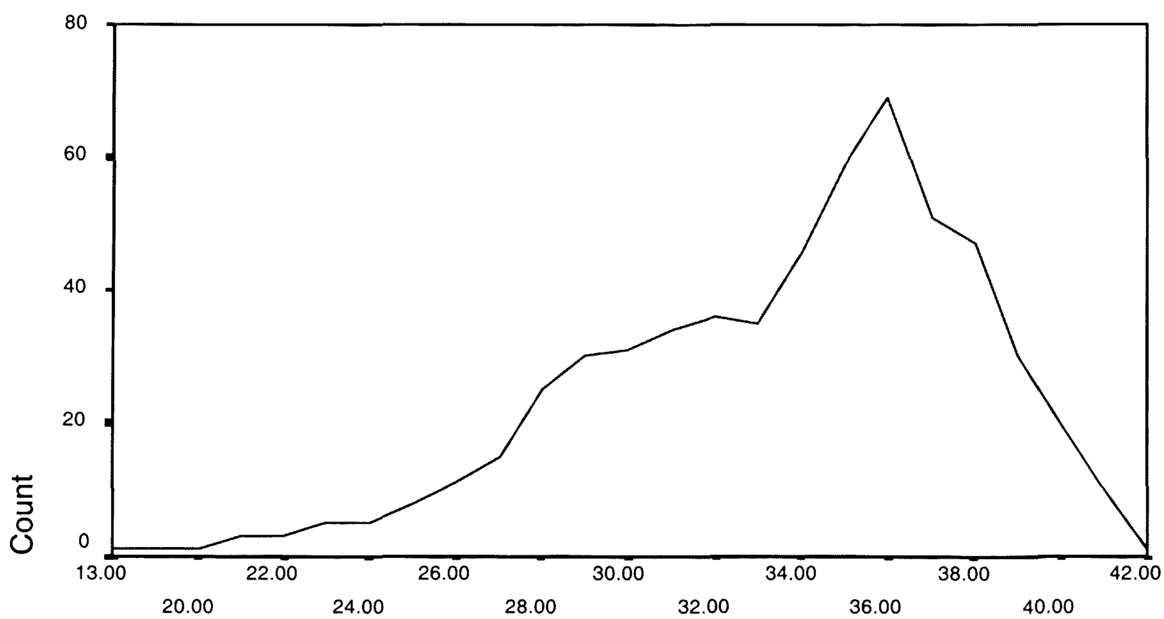


Figure 5. Frequency distribution of the Verbal Scale scores (N = 577)

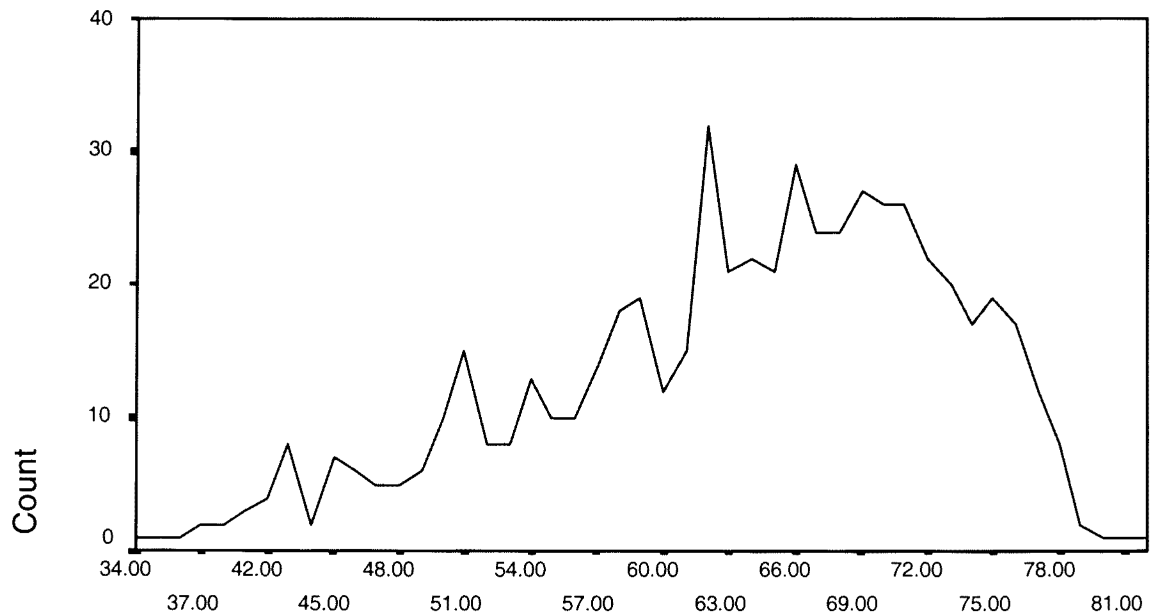


Figure 6. Frequency distribution of the PPG Total Scale scores (*N* = 577)

The means and standard deviations, and the correlations between the three scales of the PPG Level 2, and an academic performance composite for each of the 11 schools are presented in Table 11. The correlations of the PPG Total score with the academic performance composite ranges between .35 (School 10) and .67 (School 5). For the Non-verbal Scale the correlations range between .34 (School 7) and .63 (School 1), whereas for the Verbal Scale the correlations range between .20 (School 10) and .63 (School 5).

Table 11. *Summary statistics for the PPG scales and the academic performance composite*

Variable	M	SD	*r*	n
School 1 PPG total	66.98	8.90	.64	50
Non-verbal	32.86	5.44	.63	
Verbal	34.12	4.23	.54	
Academic	71.87	12.39		
School 2 PPG total	68.29	6.52	.47	45
Non-verbal	32.07	5.67	.38	
Verbal	36.22	2.47	.36	
Academic	71.47	10.84		
School 3 PPG total	69.26	6.17	.42	50
Non-verbal	32.12	5.04	.38	
Verbal	37.14	2.31	.30	
Academic	69.93	12.32		
School 4 PPG total	66.09	8.34	.51	46
Non-verbal	32.15	6.11	.56	
Verbal	33.93	3.57	.24	
Academic	71.52	12.59		
School 5 PPG total	64.20	8.25	.67	66
Non-verbal	30.12	6.07	.51	
Verbal	34.90	3.78	.63	
Academic	68.66	11.66		
School 6 PPG total	61.41	9.39	.61	68
Non-verbal	28.48	7.05	.54	

Table 11. *Summary statistics for the PPG scales and the academic performance composite (continued)*

Variable	M	SD	*r*	n
School 6 Verbal	32.93	3.90	.48	
Academic	62.70	10.12		
School 7 PPG total	60.85	9.83	.45	59
Non-verbal	28.88	7.17	.34	
Verbal	31.97	4.04	.49	
Academic	68.70	11.34		
School 8 PPG total	57.61	9.28	.39	49
Non-verbal	25.51	6.60	.35	
Verbal	32.10	3.72	.37	
Academic	59.89	14.28		
School 9 PPG total	59.88	10.16	.60	48
Non-verbal	28.44	6.38	.61	
Verbal	31.44	5.46	.40	
Academic	64.07	13.32		
School 10 PPG total	64.67	9.14	.35	48
Non-verbal	31.50	5.32	.35	
Verbal	33.17	6.45	.20	
Academic	74.31	10.94		
School 11 PPG total	59.46	9.36	.45	48
Non-verbal	26.71	6.92	.42	
Verbal	32.75	3.83	.35	
Academic	57.50	10.98		

In order to compute pooled correlations for the 11 subgroups, the HETCOR program was used (Silver & Hittner, 1997). The obtained pooled correlations between the academic performance composite and the three PPG scales were as follows: PPG Total Scale (r = .56), PPG Non-verbal Scale (r = .53), and PPG Verbal Scale (r = .41).

Overall, the findings support the validity of the PPG Level 2 as a predictor of academic achievement. Note that the lower correlations of the PPG Verbal Scale with the academic criterion may possibly be ascribed to the lower reliability of the PPG Verbal Scale (α = .72) in comparison with the PPG Total Scale (α = .86) and PPG Non-verbal Scale (α = .85).

4.9 Confirmatory factor analysis of the PPG Level 2 Scales

The five subtests of the PPG Level 2 were subjected to two confirmatory factor analyses in order to determine if it is best to view the test as measuring a unidimensional intelligence construct (possibly Spearman's *g* factor), or two separate dimensions of intelligence, namely non-verbal and verbal intelligence.

The correlations among the five subtests are presented in Table 12. Table 12 shows that all the correlations are positive and statistically significant ($p < .01$).

Table 12. *Intercorrelations among the five subtests PPG Level 2 (N = 577)*

Subtest	VQ	FS	CP	PC
FC	.29*	.38*	.19*	.39*
VQ		.35*	.29*	.30*
FS			.34*	.59*
CP				.26*

Note. FC= Figure Classification, VQ= Verbal and Quantitative Reasoning, FS= Figure Series, CP= Comprehension, and PC= Pattern Completion.

* $p < .01$.

The correlations range between .19 and .59 suggesting that the different pairs of subtests share between 4% (Figure Classification and Comprehension) and 35% (Figure Series and Pattern Completion) of their variance.

4.9.1 Postulated factor models

Two factor models were postulated for the PPG Level 2. Model 1 postulated that the five subtests of the PPG Level 2 (Figure Classification, Verbal and Quantitative Reasoning, Figure Series, Comprehension, and Pattern Completion) measure a single common factor. Each subtest is also influenced by a specific factor (that reflects measurement error and reliable variance unique to the particular subtest). This model represents Spearman's two-factor theory of intelligence.

Model 2 postulated that the two Verbal subtests (Verbal and Quantitative Reasoning and Comprehension) measure a common verbal factor, and the three Non-verbal subtests (Figure Classification, Figure Series, and Pattern Completion) measure a common non-verbal factor. Again, each subtest is also influenced by a specific factor. The verbal and non-verbal factors are allowed to correlate.

In order to identify the two models, both the variances of the common factors and the regressions of the specific factors on the subtests were fixed at unity. The factor loadings (or regression coefficients of the common factors on the subtests) and the correlations between the factors (in Model 2 only) were freely estimated from the data.

Three types of goodness-of-fit indices were used to assess the models. The first index is the chi-square statistic. A model is considered to have acceptable fit if the difference between the variance-covariance matrices generated by the original data and by the hypothesized solution is small, yielding a non-significant chi-square (Grimm & Yarnold, 1995). The χ^2 statistic is dependent on sample size and often results in a statistically significant difference when large samples are used, even when fit appears satisfactory using other indices (Terence & Tanka, 1989). Because χ^2 is one of the most frequently used fit indices in a structural analysis, it was nevertheless included (McDonald & Marsh, 1990).

The next cluster of fit indices is an aggregate of similar indices known as single sample fit indices. These are: Goodness-of-fit index (GFI), Adjusted Goodness-of-fit index (AGFI), Normed fit index, Tucker-Lewis fit index, and Comparative fit index. Despite their differences in calculations and the variables, the common attribute for these indices is that coefficient values range from zero to 1.00. Values close to 1.00 indicate a satisfactory fit between the model and the data (Byrne, 2001).

The third index is the root mean square error of approximation (RMSEA). The RMSEA takes into account the number of free parameters needed and error of approximation in the population in order to achieve a given level of fit. Browne and Cudek (in Byrne, 2001) suggested that a RMSEA less than .05 indicates a close fit, whereas RMSEA values between .05 and .08 indicate a satisfactory fit.

The expected cross-validation index (ECVI) was used to compare the relative fit of the two models. The model with the smallest ECVI value shows the greater potential for replication.

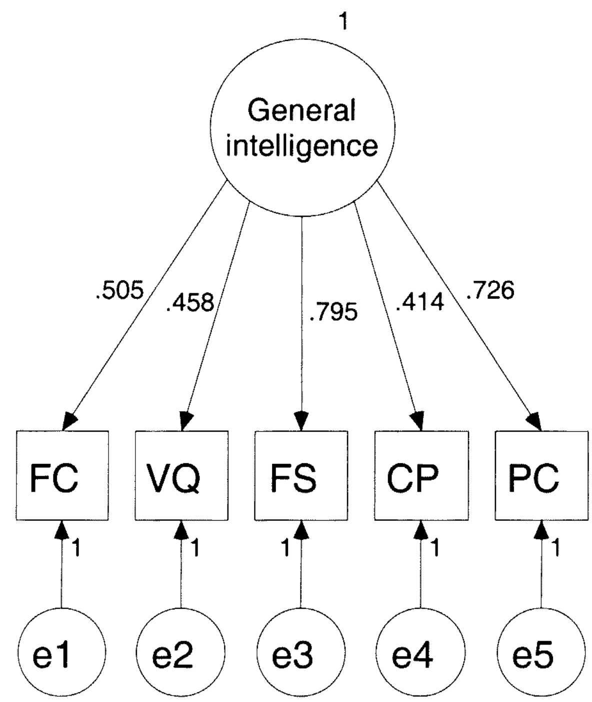
4.9.2 Confirmatory factor analysis results for Model 1

The covariance matrix of the five PPG Level 2 subtests was subjected to a maximum-likelihood confirmatory factor analysis. The goodness of fit indices for Model 1 are presented in Table 13. The chi-square statistic suggested that the fit between the model and the data was inadequate from a statistical point of view (χ^2 (5) = 20.52, p = .001). However, the point estimates of the Goodness-of-fit index (.99), Adjusted Goodness-of-fit index (.96), Normed fit index (.96), Tucker-Lewis fit index (.94), and the Comparative fit index (.97), suggested a satisfactory fit between the model and the data from a practical perspective. Furthermore, the point estimate of the RMSEA (.07) meets the suggested criterion for satisfactory fit. Hence it is concluded that Model 1 gives a satisfactory explanation of the covariances between the five subtests of the PPG Level 2.

Table 13. *Goodness-of-fit indices for Models 1and 2*

Fit measure (Indices)	Model 1	Model 2
Chi-square (χ^2)	20.52	10.03
Degrees of freedom	5	4
p	.001	.040
RMSEA	.07	.05
Goodness of fit index (GFI)	.99	.99
Adjusted GFI (AGFI)	.96	.97
Normed fit index	.96	.98
Tucker-Lewis index	.94	.97
Comparative fit index	.97	.99
Expected Cross-Validation index	.07	.06

The factor loadings (or standardized regression coefficients) in Figure 7 show that the Non-verbal subtests are better indicators than the Verbal subtests of the general factor. However, all the factor loadings can be described as of at least moderate strength - ranging from .41 for the Comprehension subtest to .80 for the Figure Series subtest. The approximate amount of variance that the general factor explains in each subtest respectively is (a) Figure Classification (26%), Verbal and Quantitative Reasoning (21%), Figure Series (63%), Comprehension (17%), and Pattern Completion (52%).



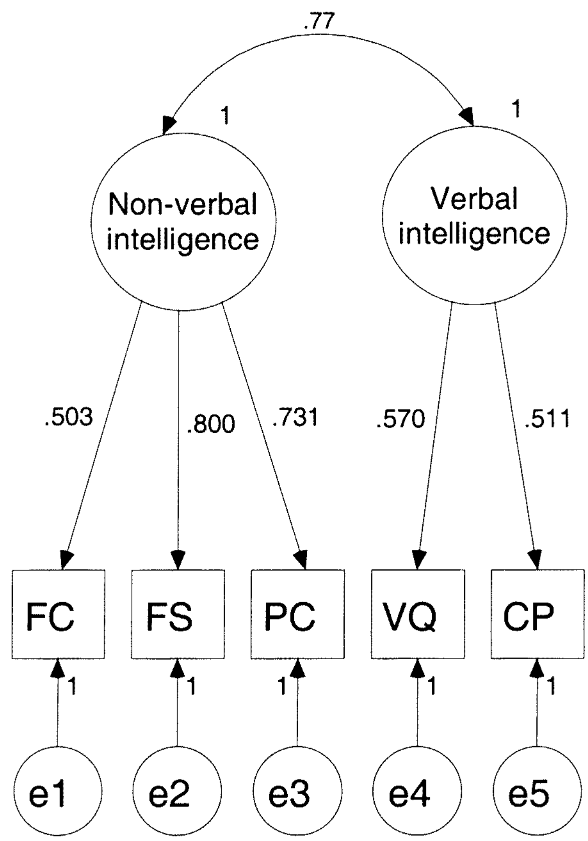
FC = Figure Classification, VQ = Verbal and Quantitative Reasoning,
FS = Figure Series, CP = Comprehension, PC = Pattern Completion

Figure 7. Standardized estimated parameters of Model 1

4.9.3 Confirmatory factor analysis results for Model 2

The goodness-of-fit indices for Model 2 are also presented in Table 13. The χ^2 statistic suggested that the fit between the model and the data was not adequate from a statistical point of view (χ^2 (4) = 10.03, p = .040). However, the point estimates of the Goodness-of-fit index (.99), Adjusted Goodness-of-fit index (.97), Normed fit index (.98), Tucker-Lewis fit index (.97), and the Comparative fit index (.99), suggested a satisfactory fit between the model and the data from a practical perspective. Furthermore, the point estimate of the RMSEA (.05) meets the suggested criterion for satisfactory fit. Hence it is concluded that Model 2 gives a satisfactory explanation of the covariances between the subtests of the PPG Level 2.

The factor loadings (or standardized regression coefficients) in Figure 8 show that Figure Series (.80) and Pattern Completion (.73) are better indicators than Figure Classification (.50) of the Non-verbal factor. The factor loadings on the Verbal Scale can be described as moderately strong: .57 for Verbal and Quantitative Reasoning and .51 for Comprehension. The approximate amount of variance that the Non-verbal factor explains in each subtest respectively is (a) Figure Classification (25%), Figure Series (64%), and Pattern Completion (54%), whereas the Verbal factor explains 33% and 26% of the variance in the Verbal and Quantitative Reasoning and Comprehension subtests, respectively.



FC= Figure Classification, FS= Figure Series, PC= Pattern Completion, VQ= Verbal ands Quantitative Reasoning, and CP= Comprehension

Figure 8. Standardized estimated parameters of Model 2

4.9.4 Synthesis of the confirmatory factor analyses

Inspection of Table 13 indicates that the two models display reasonably satisfactory to very satisfactory fits with the observed data in practical terms. Comparison of the fit indices further indicates that, relatively speaking, Model 2 displays the best fit. Hence, according to the results, the oblique two-factor model (Model 2) provided a slight improvement in fit over the one-factor model (Model 1).

4.10 Analysis of variance

One of the secondary objectives of the study was to examine whether there were significant differences in the means of the five subtests, Verbal, Non-verbal, and the Total Scales scores for boys and girls, and for the three regional groups.

4.10.1 Differences between boys and girls in terms of the PPG Level 2

The scores of the boys (n = 307) and the girls (n = 270) for the five subtests, the Verbal Scale, the Non-verbal Scale, and the Total Scale were subjected to ANOVA. The results are summarized in Table 14. The criterion for significance was set at $p < .01$, rather than the commonly used $p < .05$, giving an overall error rate of .08 for the combined set of eight (five subtests and three Scales of PPG Level 2) variables.

Table 14. *ANOVA results of gender differences for the PPG Level 2*

		Sum of Squares	*df*	Mean Squares	*F*	*p*
FC	Between Groups	.83	1	.83	.32	.57
	Within Groups	1479.90	575	2.57		
	Total	1480.74	576			
VQ	Between Groups	10.19	1	10.18	.95	.33
	Within Groups	6166.18	575	10.72		
	Total	6176.36	576			
FS	Between Groups	15.97	1	15.97	1.57	.21
	Within Groups	5842.25	575	10.16		
	Total	5858.22	576			
CP	Between Groups	3.08	1	3.08	.68	.41
	Within Groups	2605.10	575	4.53		
	Total	2608.18	576			
PC	Between Groups	72.86	1	72.86	6.93	.01
	Within Groups	6043.04	575	10.51		
	Total	6115.91	576			
NV	Between Groups	135.02	1	135.02	3.11	.08
	Within Groups	24928.95	575	43.36		
	Total	25063.97	576			
VR	Between Groups	24.47	1	24.47	1.27	.26
	Within Groups	11098.19	575	19.30		
	Total	11122.66	576			

Table 14. *ANOVA results of gender differences for the PPG Level 2*

		Sum of Squares	*df*	Mean Squares	*F*	*p*
TOTAL	Between Groups	274.45	1	274.45	3.10	.08
	Within Groups	50952.27	575	88.61		
	Total	51226.72	576			

Note. FC = Figure Classification, VQ = Verbal and Quantitative Reasoning, FS = Figure Series, CP = Comprehension, PC = Pattern Completion, NV = Non-verbal Scale, and VR = Verbal Scale.

The mean scores of boys and girls differed significantly for only one subtest, namely Pattern Completion, $F(1,575) = 6.93$, $p = .01$. However, comparison of the mean scores for boys (11.94) and girls (11.24) and the squared point–biserial correlation between gender and Pattern Completion ($r^2 = .01$), showed that in practical terms the difference between boys and girls for the subtest was trivial.

4.10.2 Differences among the three regional groups in terms of the PPG Level 2

The scores of students from the three regional groups, namely the capital ($n = 191$), towns ($n= 193$), and villages ($n = 193$) were also subjected to an analysis of variance (ANOVA). The results are summarized in Table 15. With the criterion for significance set at $p < .01$, the mean scores of the three groups differed for the five subtests, the Verbal Scale, the Non-verbal Scale, and the Total Scale.

Table 15. *ANOVA results of the regional groups for the PPG Level 2*

		Sum of squares	*df*	Mean Squares	*F*	*p*
FC	Between Groups	66.72	2	33.36	13.54	.00
	Within Groups	1414.02	574	2.46		
	Total	1480.74	576			
VQ	Between Groups	589.89	2	290.96	29.85	.00
	Within Groups	5594.47	574	9.75		
	Total	6176.36	576			
FS	Between Groups	324.10	2	162.05	16.81	.00
	Within Groups	5534.12	574	9.64		
	Total	5858.22	576			
CP	Between Groups	45.27	2	22.64	5.07	.00
	Within Groups	2562.91	574	4.47		
	Total	2608.18	576			
PC	Between Groups	345.61	2	172.81	17.19	.00
	Within Groups	5770.29	574	10.05		
	Total	6115.91	576			
NV	Between Groups	1888.30	2	944.15	23.38	.00
	Within Groups	23175.67	574	40.38		
	Total	25063.97	576			
VR	Between Groups	950.16	2	475.08	26.81	.00
	Within Groups	10172.50	574	17.72		
	Total	11122.66	576			

Table 15. ANOVA *results of the regional groups for the PPG Level 2 (Continued)*

		Sum of squares	*df*	Mean Squares	*F*	*p*
TOTAL	Between Groups	5514.736	2	2757.37	34.62	.00
	Within Groups	45711.98	574	79.64		
	Total	51226.72	576			

Note. FC = Figure Classification, VQ = Verbal and Quantitative Reasoning, FS = Figure Series, CP = Comprehension, PC = Pattern Completion, NV = Non-verbal Scale, and VR = Verbal Scale.

Tukey post hoc tests showed that the children from the capital scored significantly better than the children from the villages for all subtests and scales (see Table 16). In addition children from the capital scored significantly better than those from the towns for the Figure Classification, Verbal and Quantitative reasoning, and Figure Series subtests, but not for Comprehension and Pattern Completion subtests. Furthermore, the post hoc tests revealed that the students from the capital scored significantly better than the children from the towns on the Verbal, Non-verbal, and Total Scales.

The mean subtest scores of children from the towns and the villages were not significantly different, except for Pattern Completion where the town children performed better. Also, the mean scores of the town children did not differ significantly from those of the village students for the Verbal, Non-verbal, and Total Scales.

Table 16. *Tukey's post hoc analysis results for the three regional groups*

Dependent Variable	(I) SL	(J) SL	Mean Difference (I-J)	*p*
FC	1	2	.71	.00
		3	.73	.00
	2	1	-.71	.00
		3	.02	.99
	3	1	-.73	.00
		2	-.02	.99
VQ	1	2	1.79	.00
		3	2.36	.00
	2	1	-1.79	.00
		3	.58	.17
	3	1	-2.36	.00
		2	-.58	.17
FS	1	2	1.52	.00
		3	1.65	.00
	2	1	-1.52	.00
		3	.13	.91
	3	1	-1.65	.00
		2	-.13	.91
CP	1	2	.54	.03
		3	.64	.009
	2	1	-.54	.03
		3	.09	.90
	3	1	-.64	.009
		2	-.09	.90

Table 16. *Tukey's post hoc analysis results for the three regional groups (continued)*

Dependent Variable	(I) SL	(J) SL	Mean Difference (I-J)	*p*
PC	1	2	.91	.01
		3	1.90	.00
	2	1	-.91	.01
		3	.98	.00
	3	1	-1.90	.00
		2	-.98	.00
NV	1	2	3.15	.00
		3	4.28	.00
	2	1	-3.15	.00
		3	1.13	.19
	3	1	-4.28	.00
		2	-1.13	.19
VR	1	2	2.33	.00
		3	2.99	.00
	2	1	-2.33	.00
		3	.67	.26
	3	1	-2.99	.00
		2	-.67	.26
TOTAL	1	2	5.48	.00
		3	7.28	.00
	2	1	-5.48	.00
		3	1.80	.12
	3	1	-728	.00
		2	-1.8	.12

4.11 Summary

Classical and Rasch item analyses were carried out on the Non-verbal and Verbal Scales separately. The internal consistency of the Non-verbal Scale can be considered satisfactory for a screening instrument (Cronbach's α = .85). Furthermore, the Non-verbal items showed satisfactory fit to the Rasch model (INFIT values less than 1.3 for all items and OUTFIT values less than 1.3 for all except five items), suggesting that they essentially measure a unidimensional construct. The verbal items also showed satisfactory fit to the Rasch model (INFIT and OUTFIT values less than 1.3 for all items). The internal consistency of the Verbal Scale was lower (Cronbach's α = .72), which can possibly be attributed to the relative easiness of the items for the participants. The five subscales of the PPG (two Verbal and three Non-verbal) were also subjected to confirmatory factor analyses. Two models were postulated and compared. Model 1 postulated a single general factor, whereas Model 2 postulated two correlated factors, namely a Verbal and a Non-verbal factor. Model 2 fit the data reasonably well, but a high correlation between the factors was observed (r = .77; r^2 = .59), suggesting the presence of a general factor. Model 1, with a single general factor, provided a more frugal description of the data and also showed a satisfactory fit with the data, although a bit weaker than that of Model 2.

Comparison of mean scores showed that boys and girls differed significantly only on the Pattern Completion subtest, but this difference was small in practical terms. Children from the capital generally obtain higher scores than children from towns and villages. There was not much difference between the scores of children from the towns and the villages.

The following chapter presents a discussion of the results, the limitations of the study, the implications of the results and the recommendations for future research.

Chapter Five

Discussion, Limitations, Recommendations and Conclusion

5.1 Introduction

This study investigated the psychometric properties of the PPG Level 2 for Grade 3 Tigrignia-speaking children in Eritrea. The PPG was developed in South Africa as a group test of general cognitive ability for children in Grades 2 and 3 (PPG Level 2), and Grades 4 and 5 (PPG Level 3). The PPG provides Total, Verbal and Non-verbal scores. The Verbal Scale consists of two subscales, namely (a) Verbal and Quantitative Reasoning and (b) Comprehension. The Non-verbal Scale consists of three subscales, namely (a) Figure Classification, (b) Figure Series, and (c) Pattern Completion. Although the PPG has the appearance of a standardized intelligence test, Claassen (1996) emphasized that it is best used as a screen for academic difficulties. Hence, decisions about individual children should not be made on the basis of PPG scores alone. Instead, the PPG should be used to identify children who are possibly at risk for academic difficulties or failures.

The PPG is available in 11 South African languages. The fact that it was specifically devised in a multi-cultural and multi-lingual context makes it potentially attractive in the multi-cultural Eritrean context. The psychometric properties of the PPG Level 2 were investigated for 577 Grade 3 Tigrigna-speaking children in Eritrea. The children were selected from schools in the capital, Asmara (n = 191), from three small towns (n = 193), and from four schools in rural villages (n = 193). Eleven schools participated. The participants can be considered representative of the Grade 3 Tigrigna-speaking population in Eritrea.

This chapter presents a discussion on the results presented in the previous chapter. The results are related to findings previously obtained in similar studies. The results of the classical and the Rasch analyses will be discussed first, followed by the confirmatory factor analysis. The mean differences between males and females and among the three regional groups for the PPG Level 2 scales and subtests, will be discussed as well as the correlation between academic performance and PPG Level 2 scores. Furthermore, some of the limitations of the study will be outlined. The chapter will conclude by answering a series of questions pertinent to the study.

5.2 Classical item analysis

5.2.1 Reliabilities

Internal consistency coefficients of the Verbal score, Non-verbal score and Total score of the PPG Level 2 for the total South African sample groups were .80, .90 and .91, respectively (Claassen, 1996). For the Tigrigna-speaking children Cronbach's alpha coefficients for the three scales were as follows: Verbal Scale (α = .72), Non-verbal Scale (α = .85), and Total Scale (α = .86). The coefficients of the Total PPG Level 2 and Non-verbal Scales can be considered as satisfactory for screening purposes, whereas that of the Verbal Scale is lower and unsatisfactory. The relatively low internal consistency of the Verbal Scale can possibly be ascribed to the relative easiness of the items for the participants.

Note that the reliability of the Verbal Scale in comparison to the reliabilities of the Non-verbal and Total Scales was also relatively low for the South African group. It is not unexpected for the reliability of test scores to be lower when a test is applied in a context that differs from the standardization sample. In this light, the relatively small drop in the reliabilities of the Non-verbal and Total Scales is not disturbing.

5.2.2 Item difficulties and discriminations

More than half of the verbal items had difficulty indices of $p > .80$ (see Table 3), showing that the items were too easy. Such items have small variances and contribute little to the total variance as well as the reliability of the test. Items that contribute to the reliability of the total score are considered to be useful items, however, no item's inclusion (see Tables 3 and 4) led to meaningful reductions in the reliability of the Non-verbal and Verbal Scale scores, respectively.

The item-total correlations also showed that several items do not discriminate well among individuals with different ability levels. A problem with classical item analysis, however, is that item difficulty is confounded with item discrimination. Therefore it is not surprising that the item-total correlations (discrimination index) of several items were weak.

5.3 Rasch analysis

5.3.1 Fit between the data and the Rasch model

The satisfactory fit of the Verbal Scale and Non-Verbal Scale items to the Rasch model, as expressed in the INFIT and OUTFIT statistics (see Tables 5 and 7), suggests that the items of the two scales measure essentially unidimensional constructs.

It is of interest to note some differences between the classical and Rasch item analyses. For this purpose, the focus will fall on item a15, which forms part of the Figure Classification subtest. Five figures were presented and children were asked to identify the odd-one-out. The figures for this particular item were a glove, a shoe, a pair of trousers, a handbag, and a necktie. The majority (454 out of 577) of the children answered the item incorrectly, and thus it can be considered to be a very difficult item. A likely explanation is that gloves are not common in Eritrea and that many children, especially those from the semi-urban and rural areas, interpreted the glove as a hand. As a result many children chose the glove as the odd-one-out rather than the handbag (which is the correct answer according to the scoring key). The classical analysis identified this item as a poorly discriminating item. This is mainly due to the confounding of item difficulty with item discrimination. In contrast, the Rasch analysis identified item a15 as a potentially good item. It was calibrated as the most difficult non-verbal item. The Rasch analysis showed that item a15 discriminates well at the top end of the ability range and that it probably should be retained rather than discarded as suggested by the classical analysis.

5.3.2 Item difficulty calibrations

An important outcome of the Rasch analysis is the item difficulty calibrations, which allows for a sample-free hierarchical ordering of the items. It is useful to compare this order with the serial order of the items as they are presented in the scale. Table 6 shows that the item difficulties (in logit units) of the Verbal Scale ranged between -3.40 logits (item b5) and 2.67 logits (item d19). The item difficulties for the Non-verbal Scale ranged between 3.35 logits (item a5) and 2.93 logits (item a15) (see Table 4). By convention the mean of the item difficulties was set to zero.

It is of interest to note that the hierarchical order of the verbal items (in terms of difficulty) did not correspond well with the serial order in which the items are presented. Several difficult items are presented early and some easy items presented late in the scale. This is potentially problematic, because it seems desirable to present testees with easy items first and more difficult items later. For instance, items such as b30, b27 and d11 are amongst the easiest items, but were placed towards the end in their respective subtests. In contrast, the correspondence between item difficulties and serial order of the non-verbal items was much better.

5.3.3 Item-person map

The item-person map (see Figures 2 and 3) shows simultaneously the ability of the participants and the item difficulties on a common logit scale. The Rasch analysis showed that the verbal items were too easy for the majority of the children. Similarly, several non-verbal items were too easy. Good tests usually have the items lined up with the persons. It is not surprising, however, to find that many items were easy, because they are basically designed to screen children at low cognitive ability level and children in Grade 3 would be expected to score relatively high on the PPG. Because most of the items were relatively easy, one would expect the PPG scales to provide more accurate measures at the low end of the ability trait rather than at the high end. This feature is consistent with the aim of screening for children with potential academic difficulties.

5.3.4 Differential item functioning

Only five biased items in terms of gender (items with a significant DIF value) were identified. Halpern (2000) pointed out that sex role stereotypes influence cognitive sex differences. In general, females are thought to be more concerned with relationships and males more concerned with objects. However, many studies find no significant differences between men and women with the exception of a few studies of tests of quantitative ability and of non-verbal tests of divergent thinking (Jensen, 1980).

The two biased non-verbal items were related to the recognition and manipulation of geometric shapes (circles, squares and triangles) and favoured the boys. This suggests that girls might not be familiar with such shapes and that boys may have an advantage when confronted with such items. In addition, two of the verbal items favoured the boys. These items were related to sports activities and playing materials mostly associated with boys and therefore gave the boys an advantage. No reasonable explanation could be found for the single verbal item that favoured the girls (Are all pencils shorter than pens?). Given the few items that displayed DIF, it can be concluded that the PPG Level 2 essentially measures the same constructs for boys and girls and that the items function similarly for the two groups.

5.4 Confirmatory factor analysis

The results of the confirmatory factor analysis provide support for the hypothesis that the subtests measure general cognitive ability. All the subtests are strong indicators of the general factor. It also seems appropriate to distinguish between verbal and non-verbal ability. These two abilities are highly correlated. The three non-verbal subtests are strong indicators of non-verbal ability and the two verbal subtests are strong indicators of verbal ability.

The single factor model provides a more parsimonious description of the relationships between the five subtests, but does not explain the relationships as well as the correlated two-factor model in a statistical sense. In this study the position is taken that both models have value and that the results provide support for two levels of interpretation, namely on a Total score level and on a Verbal and Non-verbal score level.

5.5 Comparison of mean scores

Environmental factors, such as occupation, schooling, interventions, family environment, lack of parental interest in education, restricted parental intellectual stimulation (no travel, television or books), poor medical care and nutrition, overcrowded living conditions, and other factors contribute to the overall variance of intelligence test scores (Neisser et al., 1996; Vernon, 1973). A secondary objective of this study was to determine whether or not there are statistically significant differences in the PPG Level 2 means of the various groupings formed in terms of gender and school location.

The ANOVA showed statistically significant difference between boys and girls for Pattern Completion only; Claassen (1996) reported a similar finding for South African children. Claassen incorporated the Pattern Completion subtest intentionally (as it is measuring an important cognitive component) despite the fact that males perform better than females on the subtest. Hence, we might infer boys are better than girls in visual reasoning ability. There are some types of cognitive abilities that vary as a function of gender (Halpern, 2000). Jensen (1980) reported that analytic spatial-visualization ability is one of the two abilities that consistently show significant gender difference in favour of males. The other ability is mathematical or quantitative reasoning ability.

Even though it has been reported consistently that there are no sex differences in overall intelligence test scores (Gage & Berliner, 1992; Halpern, 2000; Neisser et al, 1996) some studies offer a possible explanation for long-observed ability differences between men and women on specific types of cognitive tasks. For instance, the University of Pennsylvania Medical Centre (1999) reported that women surpass men on verbal tasks, while men do better on spatial tasks due to differences in brain chemistry. The persistent gender difference in spatial cognition in favour of boys, the case in this study, has been linked to physical play and movement. That is, young children want to have a variety of play activities and breaks from seatwork that in turn enhances cognitive growth according to the type of activities boys and girls engaged in (Bjorkland & Brown, 1998). Nevertheless it should be born in mind that the above-mentioned cognitive differences are relatively small and that cognitive similarities between the sexes are greater than the differences (Halpern, 2000).

A research report on intelligence differences among rural-urban residents in United States of America by Miner (1956) indicated that people living in rural areas scored reliably below residents of large cities. A recent study on intelligence and achievement also reported that students in metropolitan areas do better than those outside metropolitan areas (Gage & Berliner, 1992). A possible reason might be that tests measure factors that are developed through experience of city life (Mittner, 1998, in Boon, 2000). In the present study children from the capital performed significantly higher than children from the villages in all the scales and subtests of the PPG Level 2. Children from the capital also scored significantly better than children from the towns for all the tests except for Comprehension and Pattern Completion. In comparing children from towns and villages, there were no mean differences except for the Pattern Completion subtest.

The differences mentioned in the previous paragraph might be linked to environmental factors. Some of the factors that might contribute to the observed differences are: (a) children from the capital are in a better position to have access to educational opportunities offered by television and educational play materials; (b) children in the capital and towns may receive better parental support in terms of education because many parents in the villages are illiterate; (c) better schooling also influence test scores (Neisser et al., 1996) and teachers in the capital are generally more experienced and have access to better teaching facilities; and (d) children from villages are often required to contribute heavily to household or agricultural activities, such as looking after cattle and fetching water and wood, which reduce their play time.

Although the gap between rural and urban areas is closing in industrialized countries, it remains important to study the effects of geographical area on intelligence in countries like Eritrea. In Greece, for example, consistent differences in children's intelligence test scores (in favour of urban children) were recently reported (Alexopoulos, 1997).

5.6 The correlation between PPG Level 2 and academic achievement

One of the frequently used criteria for general intelligence tests is academic achievement (Gottfredson, 1998). In this study the correlations between the PPG scores and teacher ratings of academic achievement were examined. The pooled correlation for the PPG Total score with academic achievement was .56, for the Non-Verbal score the correlation was .53, and for the Verbal score the correlation was .42.

The correlations for the Total and Non-verbal scores are similar to those typically reported in the educational psychology literature (Gottfredson, 1998; Neisser et al., 1996) and provide support for the validity of these scales as a screen for academic difficulties. The results show that the PPG Level 2 explains approximately 31% of the variation in academic performance. Hence, based on the estimated value it can be inferred that the PPG Level 2 can probably be used as a valid predictor of the children's academic performance.

5.7 Limitations of the present study

An important limitation of this study is that the participants are not truly representative of the total Tigrigna-speaking Grade 3 population in a statistical sense. Specifically, children from remote village were not included in the study due to time and financial constraints. However, aside from this limitation, attempts were made to sample participants as widely and as representative as possible.

The academic achievement measures are based on teachers' ratings of school performances and are not standardized. Jensen (1980) noted that teachers' marks in elementary school are not a very solid criterion for studies of predictive validity. Hence, using standardized tests of academic achievement could strengthen the study. However, no such tests are available in Eritrea.

A further limitation is that the PPG Level 2 was not related to an established cross-cultural measure of intelligence, such as Ravens Progressive Matrices (RPM). If it could be shown that the PPG scores are significantly related to scores on the RPM, it would provide further evidence in support of the construct validity of the PPG.

5.8 Conclusion and recommendations for further research

This section will conclude the study by answering a series of questions that address the major outcomes of the study.

1. Does the PPG Level 2 provide reliable scores for Tigrigna-speaking Grade 3 children?

Yes and no. The reliability coefficients for the PPG Level 2 Total Scale and the Non-verbal Scale are satisfactory. However, the reliability coefficient for the Verbal Scale was relatively lower and should be improved.

2. Why is the reliability of the Verbal Scale lower?

The low reliability of the Verbal Scale might be attributed to the following factors: (a) The items are appropriate to differentiate children at low ability level, and hence are too easy for the diversified group (learners of different ability levels) and therefore do not contribute much to the variance of the Verbal Scale, (b) the unsatisfactory arrangement of items in the Verbal Scale in terms of difficulty might contribute to low reliability, (c) the children in this study took the PPG Level 2 in the second semester of Grade 3 and are therefore expected to perform relatively well on a test that is designed to be used by Grade 2 and Grade 3 learners. The PPG is best used at the beginning of an academic year.

3. Do the items of the Verbal Scale define a unidimensional construct?

Yes, according to the results of the Rasch analysis the items measure an essentially unidimensional attribute.

4. Do the items of the Non-verbal Scale define a unidimensional construct?

Yes, according to the results of the Rasch analysis the items measure an essentially unidimensional attribute.

5. Does the PPG Level 2 retain its structural meaning in the Eritrean context?

Yes, the construct validity of both the PPG Level 2 Total scale and the Verbal and Non-verbal Scales for Tigrigna-speaking Grade 3 Eritrean children received support through confirmatory factor analyses. This suggests that the theoretical construct measured by the PPG Level 2 retained its meaning for Tigrigna-speaking children who come from a socio-political, economical and cultural context very different from South Africa.

6. Do the items function similarly for boys and girls?

Five out of 84 items functioned differently for boys and girls. Two non-verbal items, and three verbal items were found to function differently for boys and girls. The non-verbal items, which were related to the recognition and manipulation of geometric shapes, favoured the boys. In addition, two of the verbal items favoured the boys. These items were related to sporting activities and playing materials mostly associated with boys. No reasonable explanation could be found for the item that favoured the girls.

7. Do boys and girls differ in terms of mean scores for the PPG scales?

Boys and girls did not differ in terms of mean scores, except for the Pattern Completion subtest, where boys performed better than girls. However, in a practical sense the difference was trivial.

8. Do children from different regional groups differ in terms of mean scores for the PPG scales?

Yes, the results of this investigation revealed that children from the capital generally scored higher than children from the towns and villages. Not much difference was observed between children from the towns and villages.

9. Is the PPG Level 2 related to academic performance?

Yes, the pooled correlation for the PPG Total score with the academic achievement composite was .56, which is similar to previous findings. The PPG explains approximately 31% of the variance in teacher-rated academic performance.

10. Should the PPG Level 2 be used for academic screening in Eritrea?

The correlation between the PPG scales and ratings of academic achievement provide support for the validity of the PPG as a screen of academic difficulties. The results show that the PPG, which was developed in South Africa, may be fruitfully exported to Eritrea as a screening device for slow learners. It is recommended, however, that before the PPG Level 2 is routinely used for screening purposes with Tigrigna-speaking children, the functioning of the Verbal Scale should be re-examined and that some of the verbal items should possibly be rewritten.

11. What is the way forward?

Although the PPG Level 2 seems to work relatively well in measuring the cognitive abilities of Tigrigna-speaking children, it will be worthwhile to replicate the study by making some modifications in the items and incorporating a more representative sample. It is, however, hoped that results from this study will serve as an impetus for further research.

There is much that still needs to be explored in terms of adequately and fairly assessing the cognitive functioning of learners in Eritrea. The following suggestions are made for further studies:

(a) In order to minimize classroom teachers' bias, it is recommended that standardized tests of academic achievement are used. Such tests still have to be developed.

(b) It is recommended that some explicit measure of socio-economic status be included in future studies. In the present study socio-economic status was indirectly indicated by geographical area.

(c) It is recommended that the serial order of the verbal items be arranged to correspond with the difficulties of the items.

(d) It is recommended that more difficult verbal items be included in the scale.

(e) The concurrent validity of the PPG Level 2 should be assessed by also including another intelligence test, perhaps the RPM, in a future study.

In conclusion, this study has demonstrated that the general cognitive ability of Grade 3 Tigrigna-speaking children can be measured relatively well by the PPG Level 2. However, the test may be improved by modifying some of the items and by including some more difficult items. Researchers now need to take up the challenge to implement these modifications and to investigate the validity of the PPG Level 2 and other similar instruments with children in other age groups.

Bibliography

Alexopoulos, D. S. (1997). Urban and rural residence and IQ. Psychological Reports, 80, 851-860.

Asuni, T. (1970). Problems of child guidance of the Nigerian school child. West African Journal of Education, 14, 49-54.

Bjorkland, D. F., & Brown, R. D. (1998). Physical play and cognitive development: Integrating activity, cognition, and education. Child Development, 69 (3), 604-606.

Blasco, P. (2001). Early intervention services for infants, toddlers and their families. Boston: Allyn & Bacon.

Bond, T., & Fox, C. (2001). Applying Rasch model: fundamental measurement in the human sciences. London: Lawrence Erlbaum Associates.

Boon, M. (2000). The validity and reliability of the Paper and Pencil Games Level 3 for Xhosa speaking Pupils. Unpublished M. A. Thesis, University of Stellenbosch.

Byrne, B. M. (2001). Structural equation modeling with AMOS: Basic concepts, applications, and programming. London: Lawrence Erlbaum Associates.

Carnine, D. (1994). Introduction to the mini-series: diverse learners and prevailing, emerging and research-based educational approaches and their tools. School Psychology Review, 23, 341-350.

Cattell, R. B. (1966). Handbook of multivariate exiperimental psychology. Chigaco: Rand McNally.

Claassen, N. C. W. (1996). Paper and Pencil Games (PPG): Manual. Pretoria: Human Sciences Research Council.

Deary, I. J. (2001). Intelligence: A very short introduction. USA: Oxford University Press.

Eysenck, H. J. (1998). Intelligence: A new look. New Brunswick: Translation Publishers.

Foxcroft, C., & Roodt, G. (2001). (Eds.) An introduction to psychological assessment: In the South African context. Cape Town: Oxford University Press.

Gage, N. L., & Berliner, D. C. (1992). Educational psychology. Boston: Houghton Mifflin.

Gage, N. L., & Berliner, D. C. (1988). Educational psychology. Boston: Houghton Mifflin.

Gage, N. L., & Berliner, D. C. (1984). Educational psychology. Boston: Houghton Mifflin.

Gardner, H. (1983). Frames of mind: the theory of multiple intelligences. New York: Basic Books.

Gardner, H., Kornhaber, M. L., & Wake, W. K. (1996). Intelligence: Multiple perspectives. Fort Worth: Harcourt Brace College Publishers.

Ghenet, T. (2000). An analysis of the need for school-based teacher support for learners with special needs in Eritrean Schools. Unpublished M. A. Thesis, University of Western Cape.

Gottfredson, L. S. (1998). The general intelligence factor. Scientific American Presents, 9 (4), 24-29.

Graham, J., & Lilly, R. (1984). Psychological testing. New Jersey: Prentice-Hall.

Grimm, R., & Yarnold, P. (1995). Reading and understanding multivariate statistics. Washington, DC: American psychological Association.

Halpern, D. F. (2000). Sex differences in cognitive abilities. London: Lawrence Erlbaum Associates.

Jensen, A. R. (1980). Bias in mental testing. Great Britain: Methuen.

Jensen, A. R. (1998). The *g* factor: The science of mental ability. USA: Praeger.

Jensen, A.R. (1998). The g factor and the design of education. In Sternberg, R. & Williams, W. (Eds.) Intelligence, instruction, and assessment. New Jersey: Lawrence Erlbaum Associates, Publishers.

Jonassen, D., & Grabowski, B. (1993). Individual differences, learning and instruction. New Jersey: Lawrence Erlbaum Associates.

Kaplan, R., & Saccuzzo, D. (2001). Psychological testing: Principles, applications and issues. USA: Wadsworth Thomson Learning.

Kline, P. (1991). Intelligence: The psychometric view. London: Routledge.

Linacre, J. M., & Wright, B. D. (1994). Reasonable mean-square values. Rasch measurement transactions, 8 (3), 370. Available: http://www.rasch.org/rmt/rmt83.htm. Accessed: November 6, 2002.

McDonald, R. P., & Marsh, H.W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. Psychological Bulletin, 107, 247-255.

Mcintosh, D., Gibney, L., Quinn, K., & Kundert, D. (2000). Concurrent validity of the early screening profiles and the differential ability scales with an at-risk preschool sample. Psychology in the Schools, 37, 201-207.

Meisels, S., & Atkins-Burnett, S. (2001). The elements of childhood assessment. In Shonkoff, J., & Meisels, S. (Eds.) Handbook of early childhood intervention. New York: Cambridge University Press.

Miner, J. B. (1956). Intelligence in the United States. Westport, Connecticut: Greenwood Press.

Neisser, U., Boodoo, G., Bouchard, T. J., Wade Boykin, A., Broady, N., Ceci, S. J, et al., (1996.). Intelligence: Knowns and unknowns. American Psychologist, 51, 77-101.

Parry, C. D. H. (1996). A review of psychiatric epidemiology in Africa: Strategies for increasing validity when using instruments transculturally. Transcultural Psychiatric Research Review, 33, 173-188.

Rust, J., & Golombok, S. (1989). Modern psychometrics: The science of psychological assessment. London: Routledge.

Sattler, J. M. (2001). Assessment of children cognitive applications. San Diego, CA: Sattler.

Schaefer, B., & McDermott, P. (1999). Learning behaviour and intelligence as explanations for children's scholastic achievement. Journal of School Psychology, 37, 299-311.

Silver, N. C., & Hittner, J. B. (1997). A Microsoft FORTRAN 77 Program for pooling subgroup correlation coefficients. Educational and Psychological Measurement, 57, 876-878.

Terence, J., & Tanka, J. (1989). Influence of sample size, estimation method, and model specification on Goodness-of-fit assessments in structural equation models. Journal of Applied Psychology, 74, 625-635.

van de Vijver, F. J. R., & Leung, K. (1997). Methods and data analysis for cross-cultural research. Thousand Oaks: Sage Publications

van de Vijver, F. J. R., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. European Review of Applied Psychology, 47, 263-280.

Vernon, P. E. (1973). Intelligence and cultural environment. London: Butler & Tanner.

Ministry of Education Eritrea (1999). Education for all in Eritrea: policies, strategies, and prospects preliminary draft report. Available: http://www2.unesco.org/wef/countryreports/eritrea/rapport_1.html. Accessed: 19 November 2002.

University of Pennsylvania Medical Center (1999). Sex differences found in proportions of gray and white matter in the brain: links to differences in cognitive performance seen. Available: http://www.sciencedaily.com/releases/1999/05/990518072823.htm. Accessed: 9 November 2002.

Appendix A. Frequency Tables of PPG Scales

Non-verbal Scale

		Frequency	Percent	Valid	Cumulative Percent
Valid	9.00	1	.2	.2	.2
	11.00	2	.3	.3	.5
	12.00	3	.5	.5	1.0
	13.00	3	.5	.5	1.6
	14.00	3	.5	.5	2.1
	15.00	6	1.0	1.0	3.1
	16.00	9	1.6	1.6	4.7
	17.00	12	2.1	2.1	6.8
	18.00	4	.7	.7	7.5
	19.00	8	1.4	1.4	8.8
	20.00	13	2.3	2.3	11.1
	21.00	13	2.3	2.3	13.3
	22.00	13	2.3	2.3	15.6
	23.00	15	2.6	2.6	18.2
	24.00	13	2.3	2.3	20.5
	25.00	15	2.6	2.6	23.1
	26.00	25	4.3	4.3	27.4
	27.00	22	3.8	3.8	31.2
	28.00	31	5.4	5.4	36.6
	29.00	27	4.7	4.7	41.2
	30.00	29	5.0	5.0	46.3
	31.00	32	5.5	5.5	51.8
	32.00	35	6.1	6.1	57.9
	33.00	43	7.5	7.5	65.3
	34.00	51	8.8	8.8	74.2
	35.00	32	5.5	5.5	79.7
	36.00	32	5.5	5.5	85.3
	37.00	31	5.4	5.4	90.6
	38.00	21	3.6	3.6	94.3
	39.00	22	3.8	3.8	98.1
	40.00	9	1.6	1.6	99.7
	41.00	2	.3	.3	100.0
	Total	577	100.0	100.0	

Verbal Scale

		Frequency	Percent	Valid	Cumulative Percent
Valid	13.00	1	.2	.2	.2
	18.00	1	.2	.2	.3
	20.00	1	.2	.2	.5
	21.00	3	.5	.5	1.0
	22.00	3	.5	.5	1.6
	23.00	5	.9	.9	2.4
	24.00	5	.9	.9	3.3
	25.00	8	1.4	1.4	4.7
	26.00	11	1.9	1.9	6.6
	27.00	15	2.6	2.6	9.2
	28.00	25	4.3	4.3	13.5
	29.00	30	5.2	5.2	18.7
	30.00	31	5.4	5.4	24.1
	31.00	34	5.9	5.9	30.0
	32.00	36	6.2	6.2	36.2
	33.00	35	6.1	6.1	42.3
	34.00	46	8.0	8.0	50.3
	35.00	59	10.2	10.2	60.5
	36.00	69	12.0	12.0	72.4
	37.00	51	8.8	8.8	81.3
	38.00	47	8.1	8.1	89.4
	39.00	30	5.2	5.2	94.6
	40.00	20	3.5	3.5	98.1
	41.00	10	1.7	1.7	99.8
	42.00	1	.2	.2	100.0
	Total	577	100.0	100.0	

PPG Total Scale

		Frequency	Percent	Valid	Cumulative Percent
Valid	34.00	1	.2	.2	.2
	35.00	1	.2	.2	.3
	36.00	1	.2	.2	.5
	37.00	2	.3	.3	.9
	39.00	2	.3	.3	1.2
	40.00	3	.5	.5	1.7
	42.00	4	.7	.7	2.4
	43.00	8	1.4	1.4	3.8
	44.00	2	.3	.3	4.2
	45.00	7	1.2	1.2	5.4
	46.00	6	1.0	1.0	6.4
	47.00	5	.9	.9	7.3
	48.00	5	.9	.9	8.1
	49.00	6	1.0	1.0	9.2
	50.00	10	1.7	1.7	10.9
	51.00	15	2.6	2.6	13.5
	52.00	8	1.4	1.4	14.9
	53.00	8	1.4	1.4	16.3
	54.00	13	2.3	2.3	18.5
	55.00	10	1.7	1.7	20.3
	56.00	10	1.7	1.7	22.0
	57.00	14	2.4	2.4	24.4
	58.00	18	3.1	3.1	27.6
	59.00	19	3.3	3.3	30.8
	60.00	12	2.1	2.1	32.9
	61.00	15	2.6	2.6	35.5
	62.00	32	5.5	5.5	41.1
	63.00	21	3.6	3.6	44.7
	64.00	22	3.8	3.8	48.5
	65.00	21	3.6	3.6	52.2
	66.00	29	5.0	5.0	57.2
	67.00	24	4.2	4.2	61.4
	68.00	24	4.2	4.2	65.5
	69.00	27	4.7	4.7	70.2
	70.00	26	4.5	4.5	74.7
	71.00	26	4.5	4.5	79.2
	72.00	22	3.8	3.8	83.0
	73.00	20	3.5	3.5	86.5
	74.00	17	2.9	2.9	89.4
	75.00	19	3.3	3.3	92.7
	76.00	17	2.9	2.9	95.7
	77.00	12	2.1	2.1	97.7
	78.00	8	1.4	1.4	99.1
	79.00	2	.3	.3	99.5
	80.00	1	.2	.2	99.7
	81.00	1	.2	.2	99.8
	82.00	1	.2	.2	100.0
	Total	577	100.0	100.0	