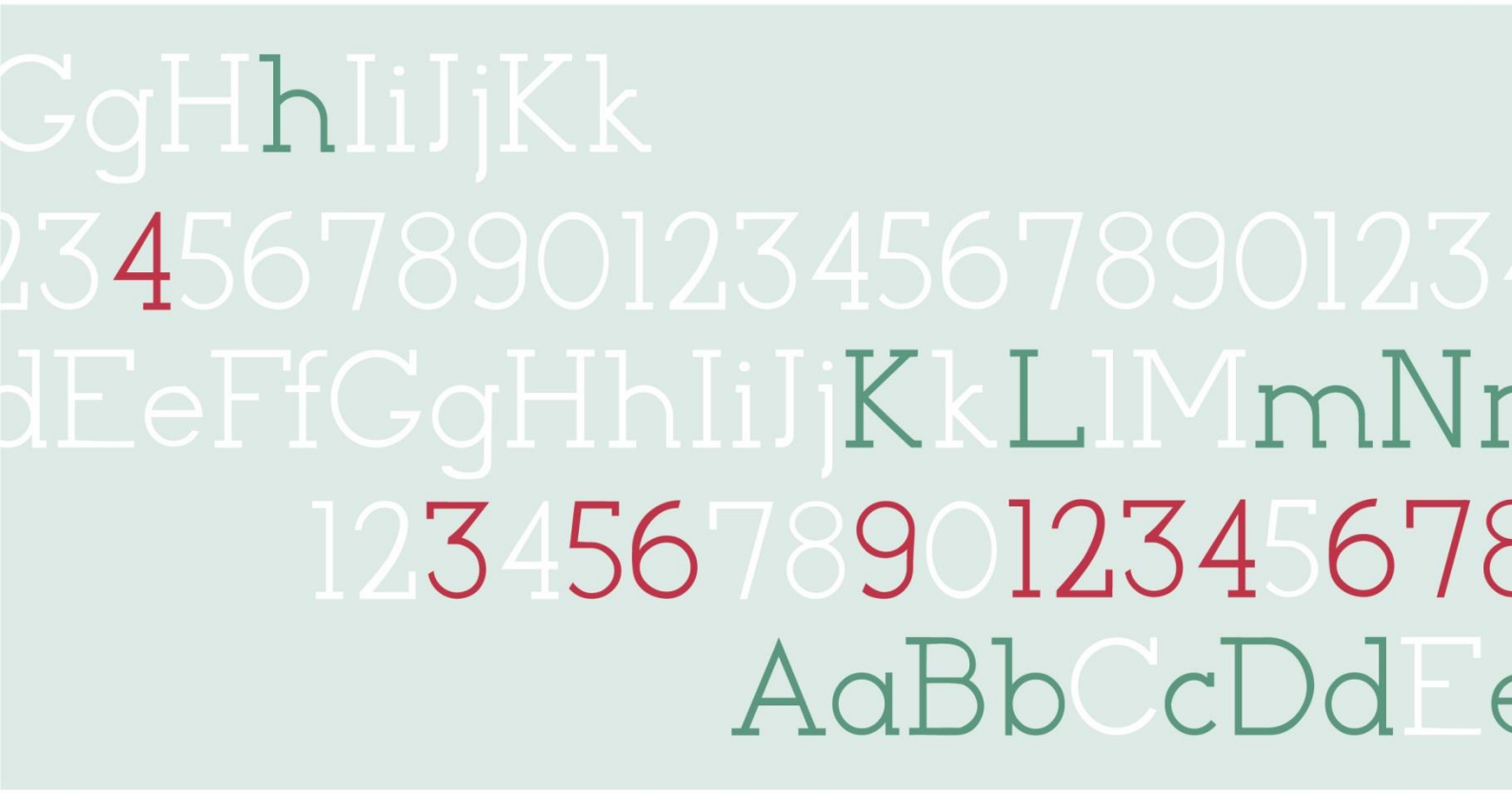
verbatim
verbal ability assessment

numeratum
numerical ability assessment



TECHNICAL MANUAL

Verbatim & Numeratum

Developed by CJ van Zyl & Dr N Taylor

PSYCHOMETRICS

JvR Psychometrics
Johannesburg Head Office
15 Hunter Avenue, Ferndale, Randburg, 2194
P.O. Box 2560, Pinegowrie, 2123, Johannesburg, South Africa
+27 11 781 3705/6/7 (tel), +27 11 781 3703 (fax), Email: info@jvrafrica.co.za

Cape Town Regional Office
3rd Floor, Imperial Bank Terraces, Tyger Waterfront, Carl Cronje Drive, Belville
P.O. Box 5534, Tygervalley, 7536, Cape Town, South Africa
+27 86 133 3916 (tel), +27 86 100 4225 (fax), Email: capeinfo@jvrafrica.co.za

Free State Regional Office
8 Hugenote Street, Jordania, Bethlehem, 9700, South Africa
+27 58 303 3843 (tel), Email: freestate-info@jvrafrica.co.za

(Website) www.jvrafrica.co.za
(Blog) www.psychologyafrica.com

ABOUT THE AUTHOR(S)

The Verbatim and Numeratum were designed and developed by Casper J.J. van Zyl and Nicola Taylor of JvR Psychometrics' (hereafter referred to as JvR) Research Department. JvR was established in 1993 with the focus of importing, researching, and distributing psychological tests. Today, JvR represents highly respected international test publishers and distributors for the Southern African region. JvR also actively sources, develops, and promotes psychological tests of high quality developed in Southern Africa. JvR takes great pride in being South African, and is privileged to have shown significant growth in the past 20 years. Distributing and publishing in excess of 200 tests, JvR is able to provide assessments in all the specialist fields of psychology including corporate, clinical, educational, neurological, counselling, and forensic use. Psychology as a science is dependent on continuous research. What has become known as being *evidence based* is the cornerstone of JvR's business and values. The JvR Research Department continuously evaluates the statistical integrity of the assessments distributed, and also develops psychometric instruments.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

Introduction

1.1 Main Features/Brief Overview

The Verbatim and Numeratum are designed to provide an assessment of verbal and numerical reasoning ability. They can be used for screening, competency-based selection, and training, and are recommended for jobs that require a Grade 11 English and/or Mathematical proficiency. The Verbatim consists of 42 scored items and can be used to assess whether the person has sufficient skill in processing and understanding written information in English. The Numeratum consists of 28 scored items and can be used to assess whether the person has sufficient reasoning skills to work with data or numerical ability. The Verbatim and Numeratum should take between 35-60 minutes to complete.

1.2 Purpose and Rationale

The primary purpose of the Verbatim and Numeratum is to assess a person's ability to understand and accurately problem-solve using English verbal and numerical information. The need for assessments of specific abilities and skills that are not necessarily measured by comprehensive mental ability assessments was identified. The most common request for a special skills assessment is for verbal and numerical assessments appropriate for use in South Africa's unique context. The Verbatim (verbal reasoning) and Numeratum (numerical reasoning) were developed to meet this need.

1.3 Administration Overview

The Verbatim and Numeratum are available in paper and pencil versions and online via the JvR Online Portal. Accessibility to these tests is managed through the JvR Client Services. The Verbatim and Numeratum can be used in any industry—including the corporate and educational setting—for screening of applicants entering the organisation. It may also be used for development purposes. These assessments are not suitable for clinical, counselling, or correctional purposes. The tests should be administered by a qualified psychometrist or psychologist as per the Professional Board for Psychology of the Health Professions Council of South Africa Scope of Practice for psychological practitioners.

CHAPTER TWO

The Conceptual Framework

2.1 History of the Product

The Verbatim and Numeratum were developed by JvR in 2012 and released in 2013 as a research version. In 2015, data on both assessments were analysed and norms generated. In addition, research items were included to allow for future updates to the two assessments. The research items are not included in the overall score. The assessments were developed due to the extensive demand for assessments of specific abilities and skills that are not necessarily measured by comprehensive mental ability assessments.

2.2 Theory

With the exception of a few additional references listed in the text, the following discussion is primarily based on the work of Arthur Jensen (1998) and John Carrol (1993).

"No other term has proved harder to define than "Intelligence". Though [psychologists] have been attempting to define intelligence for at least a century, even the experts in the field still cannot agree on a definition" (Jensen, 1998, p. 46).

Francis Galton and Herbert Spencer hypothesised that a general type of mental ability is necessary for all cognitive activities that require mental effort. Even though Galton, who is

considered the father of differential psychology, correctly assumed that ability would be normally distributed in the population, he was never successful in measuring individual differences in intelligence. One of the main reasons for this failure was his belief that information is gained through the senses and provided all that was necessary for the development of ideas, impressions, knowledge, and intelligence. Influenced by Darwin's theory of natural selection, Galton was of the opinion that the more perceptive an individual's senses were, the larger the canvass upon which intelligence could develop would be. Thus, Galton then assumed that human intelligence could be understood by measuring fine sensory discrimination and reaction time to auditory and visual stimuli.

Alfred Binet built upon some of Galton's more successful work by creating tests that were cognitively more complex. These tests tapped into higher mental processes that are associated with intelligence; for example, reasoning, verbal comprehension, and the acquisition of knowledge. Unlike Galton, Binet's tests functioned well and could be used to identify children with mental retardation and to determine school readiness of children. Although Binet offered intuitive reasons for why his tests worked, a thorough theoretical explanation was only offered later by Charles Spearman.

Using a new analytical method called factor analysis, Spearman was able to investigate the notion that intelligence consists of a single *general factor* (*g*), based on the finding that people who perform well on one cognitive test tend to perform well on other similar tests. This analytic method demonstrated that a general mental ability was indeed part of all cognitive tasks requiring mental effort. Spearman considered *g* a type of 'mental energy' that could be applied to different cognitive tasks. His development and use of factor analysis provided empirical support to Galton and Spencer's original idea that there is a general trait or attribute underlying cognitive abilities.

Interest in *g* is still widespread today, and estimates of an individual's standing on this attribute are commonly measured using assessments such as the Wechsler Adult Intelligence Scale (WAIS) and Raven's Progressive Matrices (RPM). Although both assessments yield estimates of *g*, there are important differences between them. These differences can best be understood with Raymond Cattell's concepts of fluid (*Gf*) and crystallised (*Gc*) intelligence, both of which are considered to be sub-factors of general intelligence.

Fluid intelligence includes our ability to reason and make sense of abstract and novel information, to decouple information from present contexts, and to engage working memory to form new mental representations. This ability is considered independent of learning, experience, and education. Fluid intelligence is used in problem solving strategies and solving puzzles. Crystallised intelligence, in contrast, is related to learning, knowledge and skills. It involves knowledge that comes from prior learning and past experiences. Crystallised intelligence relies on accessing information stored in long term memory and includes reading comprehension, vocabulary exams, and numerical literacy. This type of intelligence is therefore based upon facts and is rooted in individual experiences.

Fluid and crystallised intelligence form the overall capacity to learn and solve problems that most people refer to as intelligence. Both are equally important and may work either in unison or independently. For example, when solving a mathematical problem, fluid intelligence will assist in selecting a strategy to find the solution while crystallized intelligence might assist in the recall of an appropriate formula (Postlethwaite, 2011).

The WAIS contains many subtests, some of which measure *Gc* and others *Gf*. These subtests in combination provide a total IQ score, which is considered an indicator of *g*. In contrast, the RPM is essentially a measure of G*f* as it requires abstract reasoning with novel content. However, research has shown that individuals scoring high on *Gf* also typically score higher on measures of *Gc*, when the population groups are homogenous and the content of the *Gc* test is

culturally relevant and appropriate. In this respect, the RPM could also be considered an indicator of *g.* It is important to recognise that in heterogeneous populations, individuals scoring high on *Gf* will not automatically score higher on *Gc*.

The Verbatim and Numeratum were developed to measure fairly basic levels of verbal and numerical reasoning. Both assessments contain a number of subtests that could be said to measure aspects of *Gf* and *Gc*. For example, on the Numeratum, the Number Problem section would tap primarily *Gc* since the items represent fairly basic school level mathematical problems. Similarly, the Synonyms and Opposites section on the Verbatim would tap *Gc*, whereas the Reasoning section on both assessments would primarily be tapping G*f*. Both assessments would be weighed more heavily toward *Gc*, because the main concern is to determine if a candidate has acquired a relatively basic level of English verbal and numerical skills required for employment. It is also important to note that the Verbatim and Numeratum are not measures of *g*, but rather can be said to measure specific subcomponents of the larger *g* construct.

CHAPTER THREE

The Verbatim & Numeratum Scales

3.1 Verbatim

The Verbatim measures verbal ability, which is important for most jobs where the *lingua franca* is English. Respondents with high scores on this instrument could be considered at least minimally proficient in understanding and reading English. Respondents with low scores are likely to find it difficult to understand and define basic English words. The Verbatim consists of 42 questions divided into five sections. The first four sections require that the respondent selects the correct answer from a choice of four multiple choice options. The final section requires that the candidate selects between 'true', 'false' or 'cannot say'.

Synonyms (8 Items): Respondents are asked to identify words that are the same or similar in meaning.

Opposites (8 Items): Respondents are asked to identify words with opposite meanings.

Analogies (8 Items): Respondents are instructed to identify the relationship between a pair of words and to identify equivalent or similar relationships in different pairs of words.

Reasoning (8 Items): Respondents are required to identify an individual's ability to reason with letters and other verbal content.

Interpretation (10 Items): Respondents are tested on their ability to read and accurately comprehend verbal content.

3.2 Numeratum

The Numeratum assesses numerical reasoning. Candidates with high scores on this scale might be regarded as having the ability to do simple mathematical calculations and have basic numeracy skills. Candidates who score low on this scale are likely to be unable to do simple mathematical calculations, and are unlikely to be sufficiently numerate. The Numeratum consists of 28 questions divided into 3 sections. The candidate is required to select the correct response from a multiple choice format. Electronic calculators may **NOT** be used to assist candidates in answering the questions.

Number Problems (10 Items): Respondents are asked to complete mathematical problems that are composed of mostly simple addition, subtraction, multiplication and division.
Patterns (8 Items): Respondents are asked to identify patterns in numerical content.
Interpretation (10 Items): Respondents are requested to identify, read, and interpret basic numerical information.

3.3 Administration and Scoring

The Verbatim and Numeratum assessments can be administered as paper and pencil assessments or online via JvR Online. Scoring is only done through JvR Online, and can be done by JvR's Client Services department (for an additional fee) or by the test user themselves. Test users must have their own portal set up on JvR Online to manage the scoring and reporting on the Verbatim and Numeratum assessments. Once the test user has their own portal, they will have free access to this manual and to answer sheets for each assessment that they can download and print when administering paper and pencil versions. Additional resources are also available in the form of research and case studies as they are completed.

3.3.1 Online administration

Links for the candidate are created from the JvR Online user portal, with an email that gets sent to the candidate with instructions on how to complete the assessment. A full *How To Guide* is available to demonstrate how to set up projects and manage assessments on JvR Online.

The candidate receives the link for the assessment, clicks on it, and is taken to a registration page. Here they enter their username and password, agree to the terms and conditions and give consent to take the assessment. It is recommended that the Verbatim and Numeratum assessments are conducted in a proctored setting to ensure that online dictionaries and calculators are not used.

Each of the subtests of the Verbatim and Numeratum is timed. This is done in such a way that most test takers will be able to answer all the questions in the allocated time. If the individual completes the subtest before the allotted time, this is recorded and reported on in the report.

The Verbatim subtests have the following time limits:

- Section 1: SYNONYMS (10 minutes)
- Section 2: OPPOSITES (10 minutes)
- Section 3: ANALOGIES (10 minutes)
- Section 4: REASONING (20 minutes)
- Section 5: INTERPRETATION (20 minutes)

The Numeratum subtests have the following time limits:

- Section 1: NUMBER PROBLEMS (20 minutes)
- Section 2: PATTERNS (20 minutes)
- Section 3: INTERPRETATION (20 minutes)

Each section for each test has practice examples which must be successfully completed before the section can begin. The candidate will work through each practice example before the timed sections begin. Once they have completed the test and clicked Submit, they will either be taken to the next assessment to be completed on JvR Online or prompted to close the window and log off. Reports will be available immediately for generation on JvR Online.

3.3.2 Paper and pencil administration

Test booklets will need to be ordered from JvR Psychometrics prior to administration from info@jvrafrica.co.za. Please allow at least 24 hours for processing of the order, and longer if it is a large order. Ensure that you allow for enough time for delivery before placing an order for assessment material.

When administering the Verbatim and Numeratum, ensure that the candidate has a question booklet, answer sheet, pencil and eraser, and additional paper for working out. The instructions for each test are in the question booklet. Ensure that each test taker reads them carefully. Each of the subtests of the Verbatim and Numeratum is timed. This is done in such a way that most candidates will be able to answer all the questions in the allocated time.

The Verbatim subtests have the following time limits:

- Section 1: SYNONYMS (10 minutes)
- Section 2: OPPOSITES (10 minutes)
- Section 3: ANALOGIES (10 minutes)
- Section 4: REASONING (20 minutes)
- Section 5: INTERPRETATION (20 minutes)

The Numeratum subtests have the following time limits:

• Section 1: NUMBER PROBLEMS (20 minutes)

• Section 2: PATTERNS (20 minutes)

• Section 3: INTERPRETATION (20 minutes)

Each section for each test has practice examples that must be successfully completed before the section can begin. Work through each practice example and ensure that the candidate understands what they need to do for that section. Once they are ready to begin the next section, start them with the instruction to complete just the section and indicate when they are finished.

In order to score the Verbatim and Numeratum, the candidate's responses must be captured on the JvR Online portal. A full *How to Guide* is available to explain how to capture the data and generate reports. Reports will be available immediately after the scores have been captured.

3.4 Interpretation and Use of Results

The individual report for both the Verbatim and the Numeratum provides a breakdown of the individual's performance in each of the assessment areas, as well as an overall indication of verbal and numerical reasoning. At present normative scores are only provided for the overall Verbatim or Numeratum score, not for the individual scales. However, the report indicates whether the individual found the sections Easy, Moderately Hard, or Difficult compared to others, based on their performance on the scales.

The feedback report for the Verbatim provides the following breakdown:

Synonyms: This result indicates the individual's overall performance in correctly identifying an alternative word to the provided anchor word.

Opposites: This result indicates the individual's overall performance in correctly identifying an opposite word to the provided anchor word.

Analogies: This result indicates the individual's overall performance in correctly identifying one relationship among combinations of words that have a similar relationship to the pair of words presented.

Reasoning: This result indicates the individual's overall performance in correctly deducing the next letter or word that forms a series.

Interpretation: This result indicates the individual's overall performance in correctly answering questions based on information contained in a paragraph.

Overall Verbal Performance: This result is based on the number of correct items that the individual obtained in the Verbatim test. A stanine norm score is provided. The number of items correct, as well as the number of items attempted is provided in the counsellor's section in the report.

The feedback report for the Numeratum provides the following breakdown:

Number Problems: This result indicates the individual's overall performance in accurately performing arithmetic calculations such as addition, subtraction, multiplication, and division.

Patterns: This result indicates the individual's overall performance in accurately deducing the correct number in a series of numbers or in matching a particular numerical relationship.

Interpretation: This result indicates the individual's overall performance in accurately answering questions based on the information contained in a chart.

Overall Numerical Performance: This result is based on the number of correct items that the individual obtained in the Numeratum test. A stanine norm score is provided. The number of items correct, as well as the number of items attempted is provided in the counsellor's section in the report.

The time taken is included in the report as an indication as to the speed that the individual completed each section. For paper-and-pencil administered assessments, this merely reflects the full time allowed, unless completed by the administrator.

CHAPTER FOUR

Development and Standardisation

The Verbatim and Numeratum each measure various aspects of verbal and numerical reasoning. The items were selected based upon preliminary research, which identified items that showed the most promise. Data is currently being collected to further investigate item functioning. At a later stage this data will be used to refine the assessment over time.

4.1 Preliminary Research

Items were selected for the Verbatim and Numeratum using research conducted with a group of 210 respondents ranging in age from 17 to 62 years. The average age of the respondents was 26. Women were overrepresented in this sample (*n* = 145, 69.01%). The majority of the respondents identified themselves as first language Afrikaans speakers (49.05%), and 31.43% selected English as their first language. Respondents with an African language as first language had smaller representations. A full language breakdown is shown in Table 1.

Table 1: Home Language of Development Sample

Language	Frequency	Percent
Afrikaans	103	49.05
English	66	31.43
English and Afrikaans	2	0.95
Sepedi	2	0.95
Sesotho	5	2.38
Setswana	6	2.86
Swati	1	0.48
Tsonga	2	0.95
Xhosa	1	0.48
Zulu	6	2.86
Other	16	7.62
Total	210	100

With regard to ethnicity, 70% of the 210 respondents identified themselves as White, and 12.86% of the respondents identified themselves as Black African. Of the remaining respondents, 3.81% selected Coloured and 6.67% selected Asian/Indian. Data relating to ethnicity was not provided by 6.67% of the respondents. The complete ethnic breakdown is available in Table 2.

Table 2: Ethnicity of the Development Sample

Ethnicity	Frequency	Percent
Asian/Indian	14	6.67
Black	27	12.86
Coloured	8	3.81
White	147	70.00
Missing	14	6.67
Total	210	100

The data presented in Table 3 show that, in this sample, the majority of respondents (43.33%) indicated that they had a high school education, while 9 (4.29%) indicated that they only had some high school learning. Within the sample, 8 respondents (3.81%) indicated they had completed a Bachelor's degree and 52 of the respondents (24.76%) indicated that they had an honour's degree. Additionally, 13 respondents had a Master's degree (6.19%) and 3 (1.43%) had a doctorate degree.

Table 3: Education Level of the Development Sample

Education Level	Frequency	Percent
Lower than Grade 12	9	4.29
Grade 12	91	43.33
National Diploma	17	8.10
Bachelor's Degree	8	3.81
Honours Degree	52	24.76
Master's Degree	13	6.19
Doctorate Degree	3	1.43
Other	17	8.10
Total	210	100

The majority of the respondents indicated that they lived in Gauteng (84.76%). A few respondents indicated that they resided in Kwa-Zulu Natal (0.95%), the Free State (1.43%) and the Western Cape (4.29%).

Table 4: Province of Residence of Development Sample

Province	Frequency	Percent
Free State	3	1.43
Gauteng	178	84.76
KwaZulu Natal	2	0.95
Western Cape	9	4.29
Missing	18	8.57
Total	210	100

In terms of their employment, 32.86% of respondents indicated that they were employed full time. Eleven percent indicated that they were employed part time, while 20.95% selected 'other'. A complete breakdown is listed in Table 5.

Table 5: Employment Level of the Development Sample

Employment Level	Frequency	Percent
Contractual	14	6.67
Full-time	69	32.86
Full-time with flexible hours	9	4.29
Part-time	23	10.95
Self-employed/Own business	6	2.86
Other	44	20.95
Missing	45	21.43
Total	210	100

4.2 Normative Sample Verbatim

The Verbatim was administered to 439 respondents ranging in age from 19 to 63 years (mean = 34.52, median = 33, SD = 9.65). Respondents from the development sample were included in this sample group. The normative sample consisted of 292 (66.51%) women and 147 (33.49%) men. The majority of the respondents identified English (*n* = 96, 21.87%) as their home language. A full language breakdown is provided in Table 6.

Table 6: Home Language of Verbatim Normative Sample

Language	Frequency	Percent
Afrikaans	62	14.12
English	96	21.87
Ndebele	5	1.14
Sepedi	43	9.79
Sesotho	32	7.29
Setswana	54	12.30
Swati	7	1.59
Tsonga	18	4.10
Venda	19	4.33
Xhosa	33	7.52
Zulu	66	15.03
Other	1	0.23
Missing	3	0.68
Total	439	100

With regard to ethnicity, most of the respondents identified themselves as belonging to the Black/African ethnic group (*n* = 289, 65.83%). The complete ethnic breakdown is available in Table 7.

Table 7: Ethnicity of the Verbatim Normative Sample

Ethnicity	Frequency	Percent
Asian/Indian	26	5.92
Black	289	65.83
Coloured	7	1.59
White	113	25.74
Missing	4	0.91
Total	210	100

The education level of the respondents is provided in Table 8. It can be seen that 270 (61.50%) of the respondents who provided information on their highest qualification had tertiary education. A National Diploma was the highest qualification level that was most often indicated (*n* = 122, 27.79%).

Table 8: Education Level of the Verbatim Normative Sample

Education Level	Frequency	Percent
Lower than Grade 12	1	0.23
Grade 12	28	6.38
National Diploma	122	27.79
Bachelor's Degree	61	13.90
Honours Degree	50	11.39
Master's Degree	33	7.52
Doctorate Degree	4	0.91
Missing	140	31.89
Total	439	100

The province of residence of the respondents is provided in Table 9. It can be seen that most of the respondents indicated that they resided in the Gauteng province (*n* = 238, 54.21%).

Table 9: Province of Residence of the Verbatim Normative Sample

Province	Frequency	Percent
Eastern Cape	2	0.46
Free State	5	1.14
Gauteng	238	54.21
Kwa-Zulu Natal	19	4.33
Limpopo	13	2.96
Mpumalanga	11	2.51
Northern Cape	1	0.23
North-West	10	2.28
Western Cape	9	2.05
Missing	131	29.8
Total	439	100

Table 10 indicates the employment level of the respondents. The majority of the respondents (*n* = 176, 40.09%) indicated that they had full-time employment.

Table 10: Employment Level of the Verbatim Normative Sample

Employment Level	Frequency	Percent
Contract	44	10.02
Full-time	176	40.09
Missing	161	36.67
Part-time	12	2.73
Retired	1	0.23
Unemployed	45	10.25
Missing	161	36.67
Total	439	100

4.3 Normative Sample Numeratum

The Numeratum was administered to 426 respondents ranging in age from 19 to 63 years (mean = 36.11, median = 34, SD = 8.82). Respondents from the development sample were included in this sample group. The sample consisted of 240 (56.34%) women and 185 (43.43%) men. The majority of the respondents identified English (*n* = 96, 22.54%) as their home language. A full language breakdown is provided in Table 11.

Table 11: Home Language of Numeratum Normative Sample

Language	Frequency	Percent
Afrikaans	83	19.48
English	96	22.54
Ndebele	4	0.94
Sepedi	32	7.51
Sesotho	26	6.10
Setswana	44	10.33
Swati	6	1.41
Tsonga	14	3.29
Venda	21	4.93
Xhosa	29	6.81
Zulu	65	15.26
Other	0	0.00
Missing	6	1.41
Total	426	100

With regard to ethnicity, most of the respondents identified themselves as belonging to the Black/African ethnic group (*n* = 255, 59.86%). The complete ethnic breakdown is available in Table 12.

Table 12: Ethnicity of the Numeratum Normative Sample

Ethnicity	Frequency	Percent
Asian/Indian	28	6.57
Black	255	59.86
Coloured	10	2.35
White	130	30.52
Missing	3	0.70
Total	426	100

The education level of the respondents is provided in Table 13. It can be seen that 290 (68.08%) of the respondents who provided information on their highest qualification had tertiary education. A National Diploma was the highest qualification level that was most often indicated (*n* = 130, 30.52%).

Table 13: Education Level of the Numeratum Normative Sample

Education Level	Frequency	Percent
Lower than Grade 12	10	2.35
Grade 12	46	10.80
National Diploma	130	30.52
Bachelor's Degree	72	16.90
Honours Degree	54	12.68
Master's Degree	32	7.51
Doctorate Degree	2	0.47
Missing	80	18.78
Total	426	100

The province of residence of the respondents is provided in Table 14. It can be seen that most of the respondents indicated that they resided in the Gauteng province (*n* = 251, 59.76%).

Table 14: Province of Residence of the Numeratum Normative Sample

Province	Frequency	Percent
Eastern Cape	9	2.14
Free State	8	1.90
Gauteng	251	59.76
Kwa-Zulu Natal	33	7.86
Limpopo	11	2.62
Mpumalanga	13	3.10
Northern Cape	3	0.71
North-West	12	2.86
Western Cape	15	3.57
Missing	71	16.90
Total	426	100

Table 15 indicates the employment level of the respondents. The majority of the respondents (*n* = 207, 49.29%) indicated that they had full-time employment.

Table 15: Employment Level of the Numeratum Normative Sample

Employment Level	Frequency	Percent
Contract	50	11.90
Full-time	207	49.29
Part-time	14	3.33
Retired	1	0.24
Unemployed	51	12.14
Missing	103	24.52
Total	426	100

CHAPTER FIVE

Psychometric Properties

5.1 Verbatim

5.1.1 Data Screening

Outliers for the total Verbatim score for each respondent were investigated graphically using a histogram and bean plot (Kampstra, 2008). Inspection of the bean plot indicated that there were two potential outliers at the lower end of the distribution. A one-sided Grubbs test was therefore conducted to determine if the value was a statistically significant outlier (Grubbs, 1950). It was found that the lowest value was a statistically significant outlier (G = 3.7724, U = .9674, p = .03). This value was therefore removed and the Grubbs test was re-run. The next lowest value was not identified as an outlier (G = 3.4591, U = .9726, p = .11). No more outliers appeared when re-investigating the histograms and bean plots. Multivariate outliers across all the scales were subsequently investigated by plotting robust Mahalanobis distances against the quantiles of the χ^2 distribution (Garrett, 1989). Minimal multivariate outliers were detected. Multivariate normality was investigated using Mardia's coefficient (Mardia, 1970). The results indicated that the scales as a whole deviated from multivariate normality.

5.1.2 Descriptive Statistics

Descriptive statistics for each of the Verbatim scales and the total Verbatim scores are provided in Table 16. The mean total Verbatim score was 26.18 (median= 26, SD = 5.26). For

the most part the variables did not demonstrate particularly problematic skewness or kurtosis values. The standard error values were all generally low. Histograms and bean plots of the Verbatim scales and total Verbatim score are provided in Figure 1 to Figure 6. Inspection of these plots suggested that the Opposites, Analogies, and Reasoning scales deviated from normality.

Table 16: Descriptive Statistics for the Verbatim Scales and Total Score

Scale	Mean	SD	Med	Trim	Min	Max	Skew	Kurt	SE
Synonyms	4.93	1.23	5	4.91	1	8	0.07	-0.13	0.06
Opposites	5.39	1.21	5	5.41	0	8	-0.38	1.05	0.06
Analogies	5.07	1.59	5	5.11	0	8	-0.16	-0.35	0.08
Reasoning	5.20	1.74	5	5.28	0	8	-0.51	0.02	0.08
Interpretation	5.59	1.74	6	5.62	0	10	-0.15	-0.19	0.08
Total	26.18	5.26	26	26.17	8	41	-0.03	-0.24	0.25

Note. SD = Standard Deviation, Med = Median, Trim = Trimmed Mean, Skew = Skewness, Kurt = Kurtosis, SE = Standard Error.

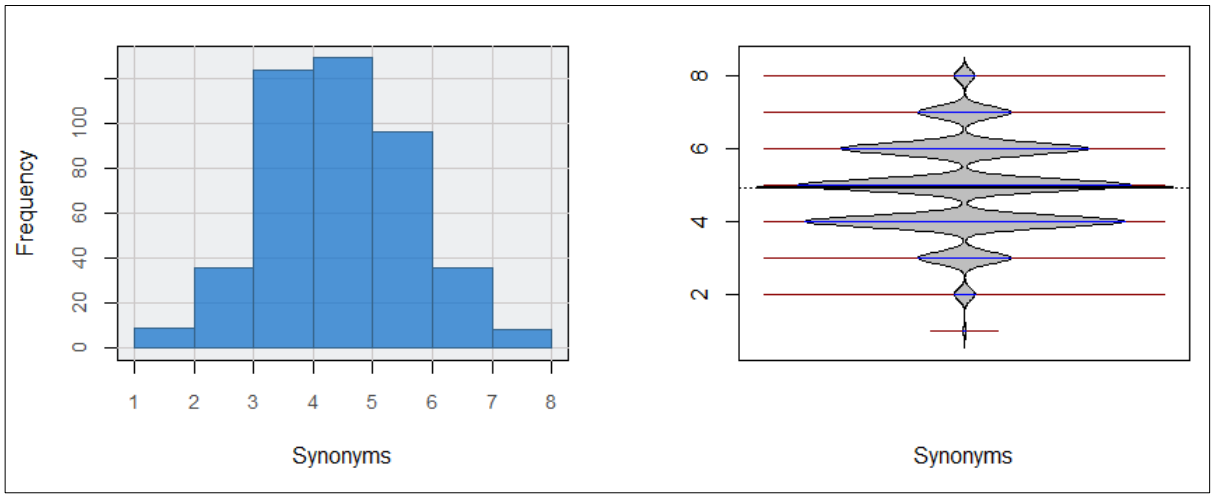


Figure 1: Histogram and bean plot for the Synonyms scale

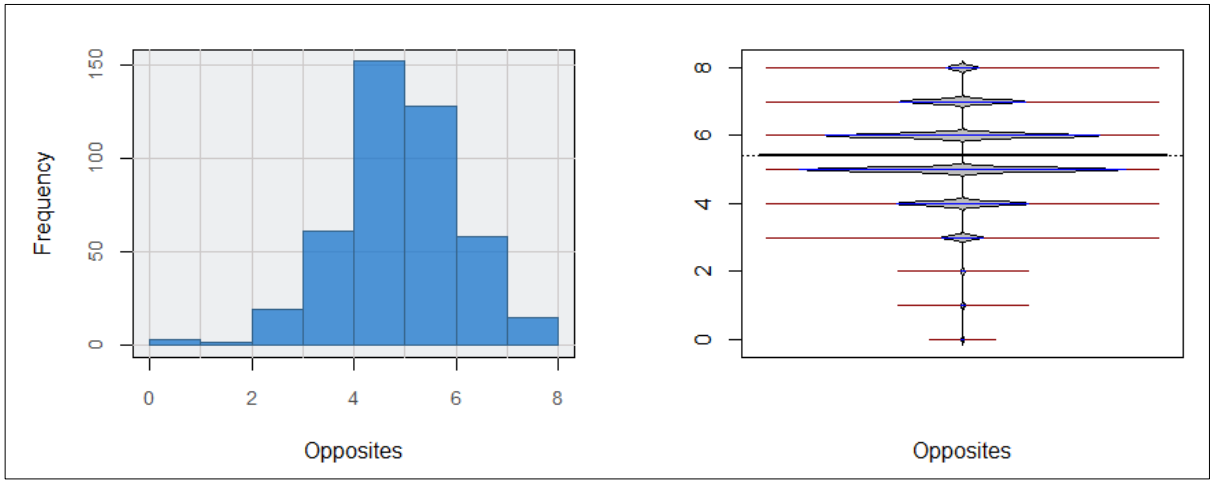


Figure 2: Histogram and bean plot for the Opposites scale

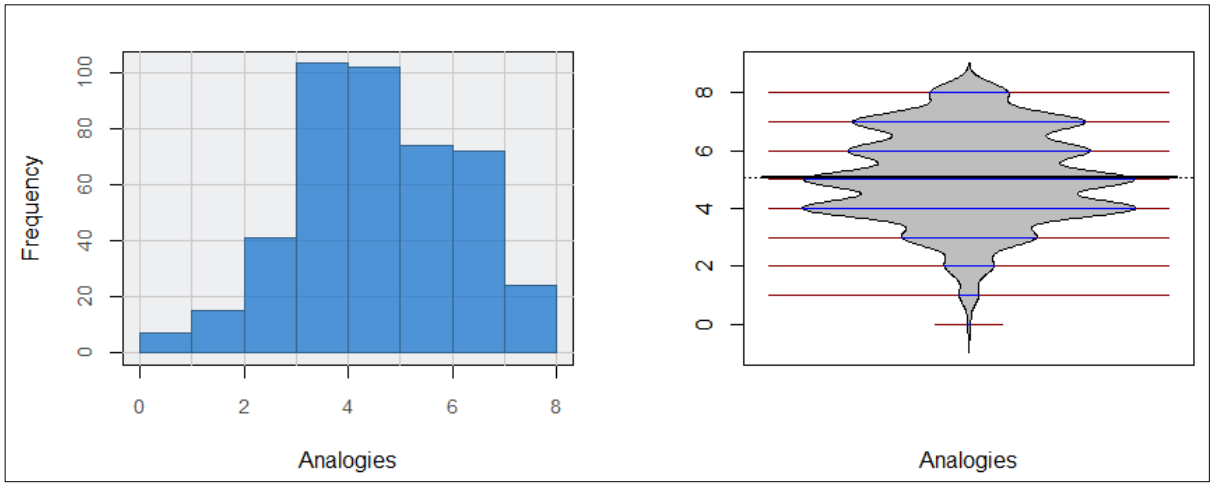


Figure 3: Histogram and bean plot for the Analogies scale

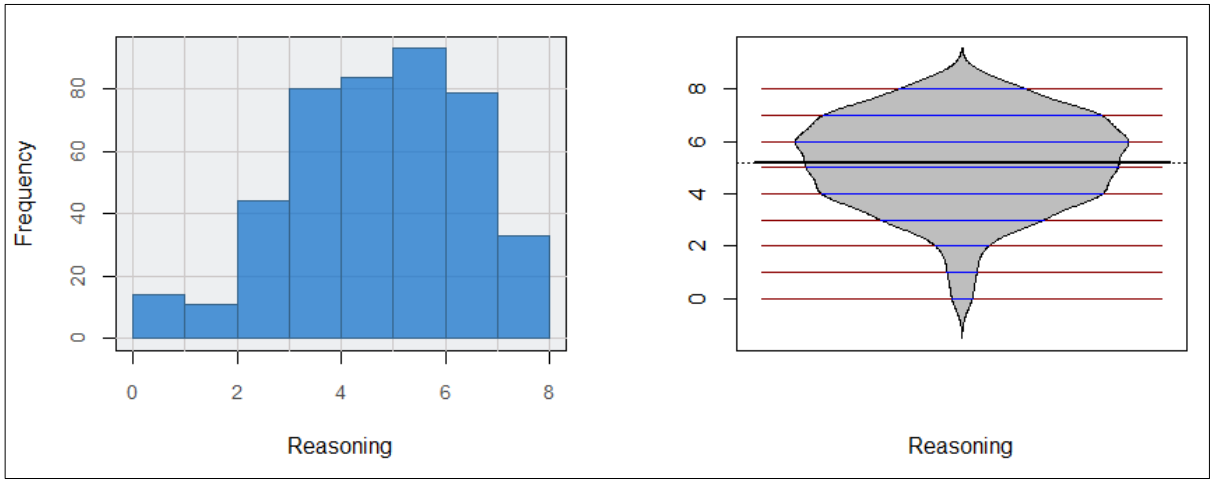


Figure 4: Histogram and bean plot for the Reasoning scale

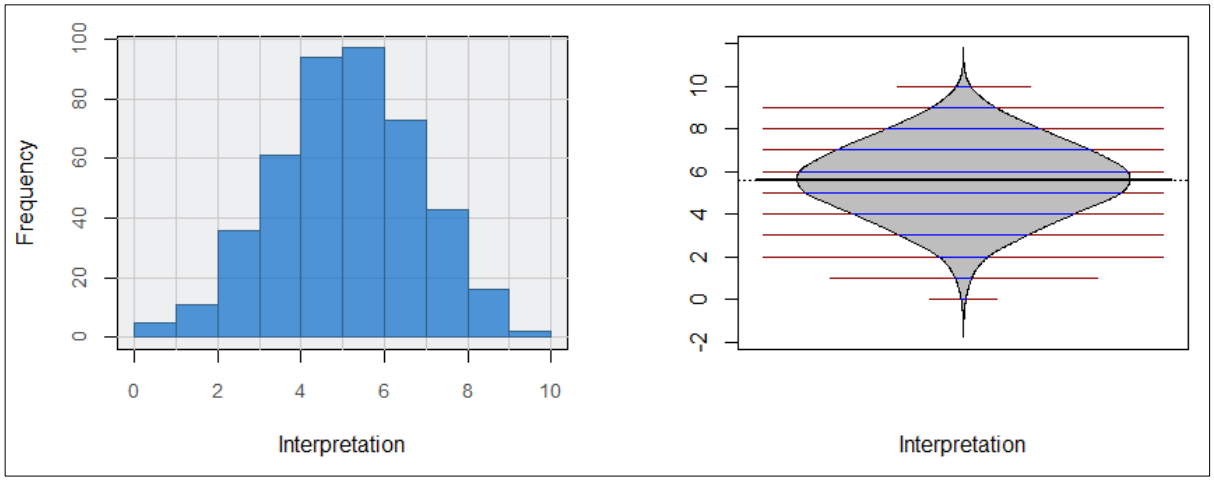


Figure 5: Histogram and bean plot for the Interpretation scale

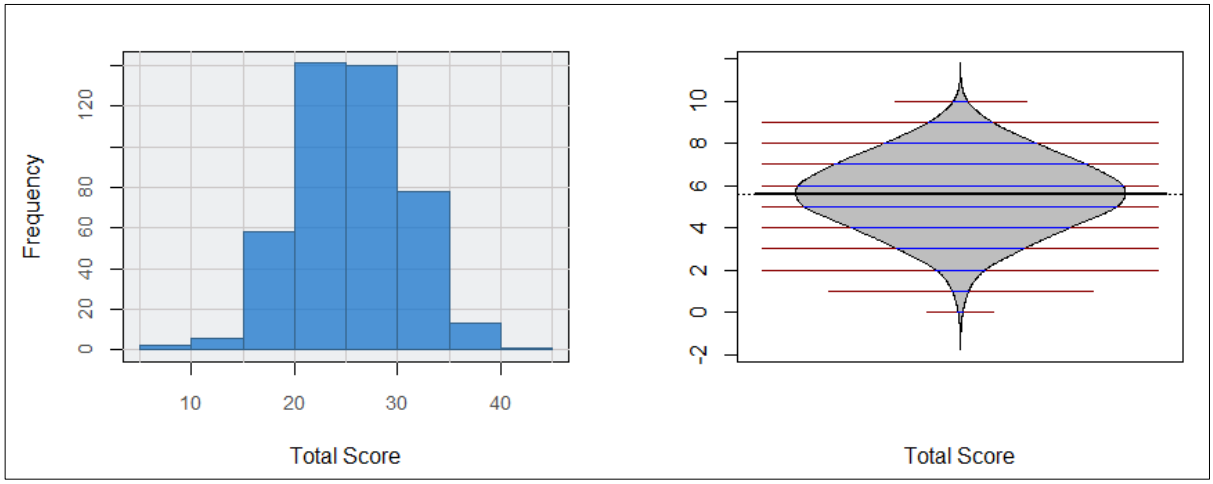


Figure 6: Histogram and bean plot for the total Verbatim score

5.1.3 Reliability

Cronbach alpha coefficients (Cronbach, 1951), Robust Cronbach alpha coefficients (Zhang & Yuan, 2014), Guttman's Lambda 2 (Guttman, 1945), and Ordinal Reliability (Gadermann, Guhn, & Zumbo, 2012) for the Verbatim scales and total Verbatim score are provided in Table 17. The scale reliabilities for the Verbatim scales were mostly unsatisfactory, with Cronbach alpha coefficients ranging from .26 to .61. The Cronbach alpha coefficient for the total Verbatim score was .75 and is therefore satisfactory. Given this alpha coefficient caution must be used in interpreting the results of the total Verbatim score (Nunnally, 1970; Urbina, 2004). The item separation index values indicated that the item locations were generally stable. The person separation index values for the Verbatim scales indicated that the scales may not be sensitive enough to distinguish between low and high scorers on the scale (i.e., more items may be required). For the total Verbatim score the person separation reliability was higher, indicating that there may be more value interpreting the total Verbatim score rather than the scale scores.

Table 17: Reliability Coefficients for the Verbatim Scales and Total Verbatim Score

Scale	α	R α	O α	λ2	IR	PR
Synonyms	.29 (.20 - .37)	.23	.58	.31	.99	.23
Opposites	.32 (.24 - .40)	.19	.72	.34	.98	.18
Analogies	.52 (.46 - .58)	.51	.70	.55	.99	.44
Reasoning	.61 (.56 - .65)	.61	.79	.62	.99	.48
Interpretation	.26 (.17 - .34)	.26	.36	.32	.97	.28
Total	.75 (.72 - .77)	.73	.87	.77	.99	.76

Note. α = Cronbach's alpha (90% confidence intervals in parenthesis), R α = Robust alpha, O α = Ordinal alpha, λ2 = Guttman's Lambda 2, IR = Item Reliability Index (Rasch), PR = Person Reliability Index (Rasch).

Haberman's (2008) subscale scoring test based on the proportional reduction in mean squared error and McDonald's omega (McDonald, 1999) were used to investigate whether or not interpretation should be conducted at the scale score level or total Verbatim score level (i.e., verbal reasoning; Reise, Bonifay, & Haviland, 2013). Application of Haberman's (2008) approach indicated that interpretation of a total Verbatim score rather than Verbatim scale scores may be warranted (Table 18). An unrestricted bifactor model was therefore run using a Schmid-Leiman transformation (see Reise et al., 2013 and Zinbarg, Yovel, Revelle, & McDonald, 2006). The tetrachoric correlation matrix was not positive definite and therefore Pearson correlation coefficients were used as input instead. In order to determine the relative merits of interpretation at the general factor level (i.e., a total score), Omega total (ω_t), Omega hierarchical (ω_h), and Omega sub-scale (ω_s) were calculated (Reise et al., 2013).

Omega total (ω_t) for the five factor model with a general factor was .80, while Omega hierarchical (ω_h) was .57. The explained common variance (ECV) was used to index the degree of unidimensionality (Reise, 2012). In this analysis the ECV was .36, indicating that 36% of the common variance was due to the general factor. While there are not specific criteria for an acceptable ECV value, the higher the value (i.e., closer to 1), the stronger the general factor (Reise, 2012). Omega sub-scale (ω_s) indicates what the reliability of the scales would be after controlling for the general factor (Reise et al., 2013). The ω_s values for the five scales were .32,

.42, .23, .06, and .06. The five factors were not well-defined and it is therefore not possible to label them. The ω_s values were generally lower than ω_h indicating that the subscales appear to explain limited reliable variance beyond that due to the general factor, but the ECV value was small. It therefore appears that there is some evidence for interpretation of a total Verbatim score rather than Verbatim scale scores. However, given the limitations of the Schmid-Leiman transformation with exploratory factor analysis (see Reise, Moore, & Haviland, 2010) caution must be used in interpretation of the results[1].

Table 18: Haberman's (2008) Subscale Scoring Test Results

Scale	$PRMSE_s$	$PRMSE_T$
Synonyms	.29	.89
Opposites	.32	.82
Analogies	.52	.78
Reasoning	.61	.68
Interpretation	.26	.82
Total	.75	

PRMSE = proportional reduction in mean squared error.

5.1.4 Item difficulty and item discrimination

Item difficulty[2] and item discrimination values were estimated within a Classical Test Theory (CTT) framework (*cf.* Lord & Novick, 1968*;* Raykov & Marcoulides, 2011). The item difficulty index is the proportion of respondents who correctly answered the item to the total number of respondents and the item discrimination index is the ability of an item to discriminate between respondents who scored high and low on the scale/test (Kerlinger & Lee, 2000; Nunnally,

[1] A confirmatory bifactor model was also investigated, but the model failed to converge.
[2] Item difficulty/locations within the Rasch framework are available later in the document.

1970). According to Kerlinger and Lee (2000) item difficulties should range between .50 and .70, where a value of 1 indicates that all respondents obtained the correct answer (i.e., too easy) while a value of 0 indicates that none of the respondents obtained the correct answer (i.e., too difficult) (Raykov & Marcoulides, 2011). However, for an ability test it would be expected that the item difficulties have a larger range. The inter-item correlation coefficient, item-total score correlation, item difficulty, and item discrimination values for the Verbatim scales are provided in Table 19.

Inspection of Table 19 indicates that most of the items had item difficulties that were not in the range of .50 to .70. There were also a fair number of items that were potentially too easy (S1, S2, O1, O2, O3, A1, A2, and R4). The average item difficulties for the five Verbatim scales were: Synonyms = .62, Opposites = .68, Analogies = .63, Reasoning = .65, and Comprehension = .56. The item discrimination values were generally quite low across the different scales, indicating that items at the scale level may not be optimal in discriminating between respondents who scores high and low on the latent trait. The average item discrimination values for the five Verbatim scales were: Synonyms = .33, Opposites = .30, Analogies = .43, Reasoning = .47, and Comprehension = .38.

The inter-item correlation coefficient, item-total score correlation, item difficulty, and item discrimination values for the total Verbatim score are provided in Table 20. The item difficulties remained the same (as expected). The item discrimination values were again generally quite low. Item I7 had a 0 discrimination value indicating that it was unable to discriminate between respondents who scored high and low on the Verbatim. The average item difficulty was .62 and the average item discrimination was .28.

Table 19: Average Inter-Item Correlation Coefficient, Correlation with Scale Score, Item Difficulty, and Item Discrimination for the Verbatim Scale Items.

Item	Average R	Raw R	R Cor	Dif	Disc
Synonym					
S1	.05	.25	.18	.95	.09
S2	.05	.24	.22	.98	.06
S3	.05	.38	.23	.88	.27
S4	.04	.45	.27	.76	.38
S5	.04	.52	.28	.55	.64
S6	.06	.44	.09	.41	.48
S7	.04	.49	.29	.22	.47
S8	.05	.39	.18	.18	.32
Opposites					
O1	.09	.27	.31	.96	.08
O2	.06	.33	.66	.99	.04
O3	.06	.31	.63	.98	.04
O4	.09	.48	.30	.81	.34
O5	.10	.51	.16	.60	.54
O6	.08	.58	.31	.55	.71
O7	.12	.38	.01	.27	.29
O8	.11	.42	.08	.24	.37
Analogies					
A1	.13	.26	.16	.94	.11
A2	.11	.35	.30	.93	.15
A3	.13	.35	.18	.81	.26
A4	.10	.59	.42	.45	.71
A5	.11	.48	.34	.76	.42
A6	.09	.62	.48	.52	.73
A7	.10	.59	.43	.42	.66
A8	.11	.48	.31	.24	.42
Reasoning					
R1	.20	.32	.13	.80	.23
R2	.15	.60	.55	.82	.45
R3	.15	.60	.48	.60	.68
R4	.16	.46	.40	.91	.23
R5	.16	.56	.39	.41	.69
R6	.16	.58	.44	.61	.65
R7	.17	.52	.39	.22	.49
R8	.16	.50	.42	.84	.35
Interpretation					
I1	.05	.22	-.03	.29	.22
I2	.03	.37	.22	.62	.39
I3	.03	.43	.30	.54	.49
I4	.02	.46	.37	.57	.53
I5	.03	.43	.33	.71	.38
I6	.02	.47	.38	.51	.55
I7	.05	.18	-.13	.69	.10
I8	.04	.27	.00	.52	.28
I9	.03	.39	.24	.44	.42
I10	.03	.39	.27	.71	.39

Note. Average R = average inter-item correlation coefficient, Raw R = item-total correlation not corrected for item overlap, R Cor = item-total correlation corrected for item overlap and scale reliability, Dif = item difficulty, Disc = item discrimination.

Table 20: Average Inter-Item Correlation Coefficient, Correlation with Scale Score, Item Difficulty, and Item Discrimination for the Total Verbatim Score Items.

Item	Average R	Raw R	R Cor	Dif	Disc
S1	.08	.25	.30	.95	.09
S2	.08	.24	.28	.97	.06
S3	.08	.32	.30	.88	.21
S4	.08	.28	.26	.76	.25
S5	.07	.44	.38	.54	.53
S6	.08	.24	.17	.41	.23
S7	.08	.22	.16	.22	.22
S8	.08	.20	.14	.18	.18
O1	.07	.25	.32	.96	.10
O2	.07	.25	.38	.98	.04
O3	.07	.27	.40	.98	.05
O4	.07	.34	.34	.81	.29
O5	.07	.37	.32	.60	.40
O6	.07	.49	.46	.55	.57
O7	.08	.11	.03	.27	.07
O8	.08	.17	.10	.23	.16
A1	.08	.19	.21	.93	.08
A2	.08	.28	.29	.93	.12
A3	.08	.27	.25	.81	.22
A4	.07	.45	.39	.45	.55
A5	.07	.35	.34	.76	.32
A6	.07	.49	.46	.52	.58
A7	.07	.48	.45	.42	.58
A8	.08	.31	.26	.24	.28
R1	.08	.20	.16	.80	.16
R2	.07	.43	.43	.82	.34
R3	.07	.39	.35	.60	.46
R4	.07	.39	.43	.91	.23
R5	.07	.43	.39	.41	.49
R6	.07	.47	.43	.60	.53
R7	.07	.44	.40	.22	.38
R8	.07	.39	.39	.84	.33
I1	.08	.14	.07	.29	.14
I2	.07	.38	.35	.62	.42
I3	.08	.17	.10	.54	.14
I4	.07	.46	.41	.57	.54
I5	.08	.24	.18	.71	.24
I6	.07	.40	.34	.51	.47
I7	.08	.06	-.01	.69	.00
I8	.08	.08	-.01	.52	.08
I9	.08	.27	.19	.44	.33
I10	.08	.21	.16	.71	.23

Note. Average R = average inter-item correlation coefficient, Raw R = item-total correlation not corrected for item overlap, R Cor = item-total correlation corrected for item overlap and scale reliability, Dif = item difficulty, Disc = item discrimination. S = Synonyms, O = Opposites, A = Analogies, R = Reasoning, I = Interpretation.

5.1.5 Correlation Coefficients

Pearson correlation coefficients and Spearman-rho rank order correlation coefficients for the five Verbatim scales are provided in Table 21. Disattenuated correlations lead to overcorrection due to the low Cronbach alpha coefficients and are therefore not provided. Inspection of the non-parametric Loess regression lines (Cleveland, 1979) indicated that for the most part the relationships between the variables were linear. Inspection of multivariate normality using Mardia's coefficient (Mardia, 1970) found that bivariate normality was violated across most of the variables. For the most part the correlation coefficients had small to medium effect sizes (Cohen, 1988).

Table 21: Pearson and Spearman-Rho Rank Order Correlations for the Verbatim Scales

	S	O	A	R	I
Synonyms	.	.36	.41	.36	.27
Opposites	.38	.	.39	.34	.29
Analogies	.43	.40	.	.49	.34
Reasoning	.35	.36	.48	.	.38
Interpretation	.28	.28	.34	.37	.

Note. Pearson correlations below the diagonal, Spearman rho rank-order correlations above the diagonal.

5.1.6 Rasch Analysis

A Rasch (1960) analysis was conducted on the total Verbatim score to inspect item fit statistics and item locations (difficulties). Items with Infit and Outfit mean square values (IMNSQ and OMNSQ) > 1.40 were considered to be underfitting items, and items with IMNSQ and OMNSQ values < .60 to be overfitting items. OMNSQ investigates unexpected responses to items that are either too easy or too difficult for the respondent, whereas IMNSQ investigates unexpected responses on items that are targeted at the respondents underlying latent ability measure (Linacre, 2015). Item and person separation reliabilities are provided in Table 17. The item fit statistics and item locations for the total Verbatim score are provided in Table 22 and Figure 7 below.

Inspection of Table 22 and Figure 7 indicates that the item locations ranged between approximately -4 to 3 logits, and mostly covered the underlying ability trait level of the respondents. However, for the most part the items appeared to be too easy. No items demonstrated underfit, while three items demonstrated overfit (S2, O2, and O3).

Table 22: Total Verbatim Score Item Location and Item Fit Statistics (Sorted by OMNSQ)

Item	Measure	SE	Infit MNSQ	Z	Outfit MNSQ	Z	PT-Measure Corr	Exp
I1	1.89	.11	1.14	2.6	1.36	4.2	.14	.33
S7	2.27	.12	1.04	.7	1.33	3.1	.24	.32
I8	.75	.10	1.22	6.4	1.32	6.2	.09	.35
I7	-.07	.11	1.21	4.2	1.31	3.9	.07	.32
O7	1.98	.11	1.15	2.6	1.31	3.5	.13	.33
O8	2.19	.12	1.11	1.7	1.28	2.8	.17	.32
S8	2.57	.13	1.07	.9	1.23	1.8	.21	.30
I9	1.14	.10	1.05	1.5	1.20	3.8	.27	.35
I3	.67	.10	1.15	4.4	1.19	3.7	.17	.35
A3	-.77	.13	1.00	.0	1.18	1.5	.26	.28
I5	-.16	.11	1.06	1.2	1.18	2.2	.24	.32
R1	-.69	.12	1.07	1.0	1.12	1.1	.20	.29
S6	1.29	.10	1.10	2.5	1.10	2.0	.24	.35
I1	-.16	.11	1.09	1.9	1.09	1.2	.22	.32
A1	-2.07	.20	.98	-.1	1.05	.3	.19	.19
S4	-.44	.12	1.02	.4	1.02	.3	.27	.30
A8	2.18	.12	.99	-.2	1.00	.0	.33	.32
O5	.40	.10	.98	-.5	.95	-.8	.37	.34
I2	.30	.10	.97	-.9	.97	-.5	.37	.34
I6	.82	.10	.96	-1.2	.97	-.7	.39	.35
R3	.39	.10	.96	-1.1	.97	-.6	.38	.34
A5	-.46	.12	.96	-.7	.97	-.3	.34	.30
O4	-.76	.13	.96	-.5	.90	-.9	.33	.28
A2	-2.07	.20	.96	-.2	.75	-1.1	.27	.19
S3	-1.34	.15	.96	-.4	.80	-1.3	.31	.24
A4	1.10	.10	.91	-2.8	.94	-1.2	.45	.35
S2	-3.13	.31	.94	-.1	.54	-1.3	.24	.13
R5	1.29	.10	.93	-1.9	.92	-1.5	.43	.35
O1	-2.54	.24	.93	-.3	.70	-1.0	.25	.17
S1	-2.43	.23	.93	-.3	.79	-.7	.25	.17
S5	.65	.10	.92	-2.3	.93	-1.6	.43	.35
R8	-1.00	.14	.92	-.9	.82	-1.4	.37	.27
I4	.54	.10	.90	-3.0	.91	-1.9	.45	.34
R6	.37	.10	.90	-2.9	.85	-2.9	.46	.34
R2	-.84	.13	.90	-1.3	.76	-2.1	.41	.28
R7	2.27	.12	.90	-1.6	.81	-2.1	.44	.32
A7	1.21	.10	.89	-3.3	.87	-2.8	.48	.35
R4	-1.67	.17	.89	-.9	.63	-2.1	.38	.22
O2	-3.60	.39	.89	-.2	.43	-1.3	.26	.11
O6	.64	.10	.89	-3.6	.85	-3.2	.48	.35
A6	.76	.10	.88	-3.9	.88	-2.7	.48	.35
O3	-3.46	.36	.88	-.3	.45	-1.4	.27	.12
Mean	.00	.14	.99	-.1	.97	.1		
SD	1.61	.07	.09	2.2	.23	2.3		

Note. OMNSQ ≥ 1.4 or ≤ .60 in bold. S = Synonyms, O = Opposites, A = Analogies, R = Reasoning, I = Interpretation.

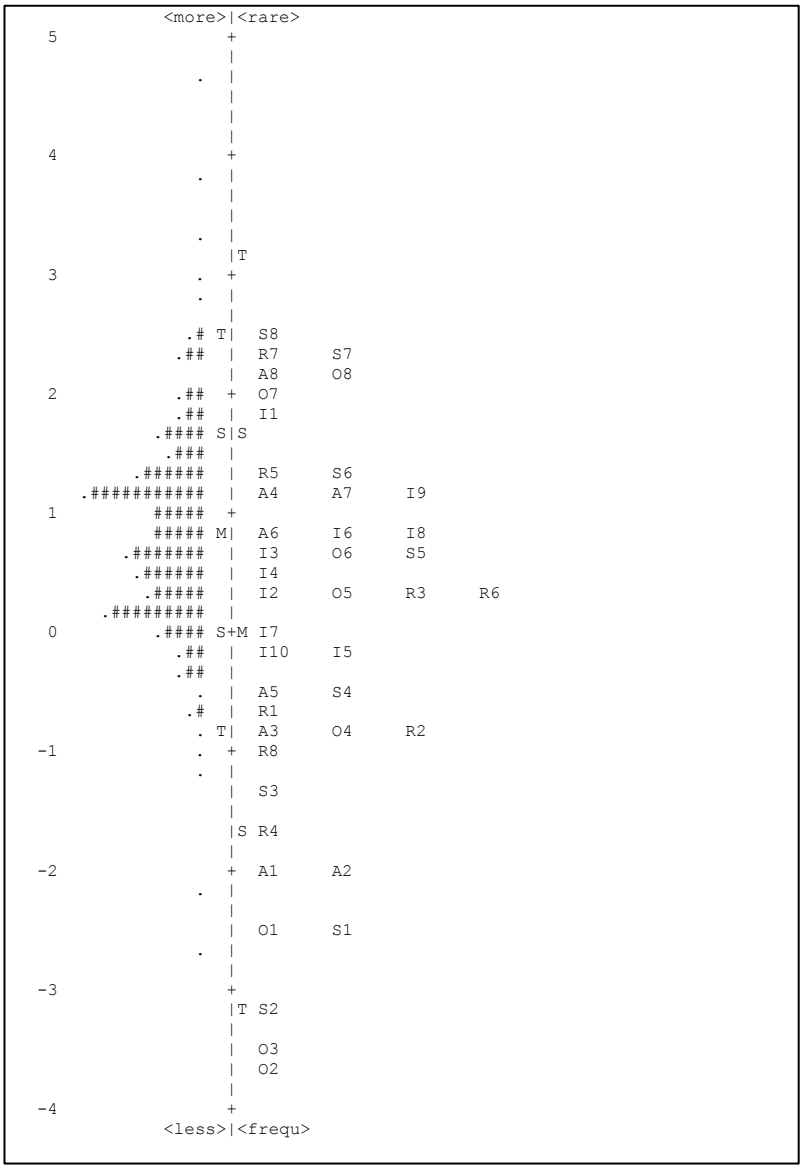


Figure 7: Wright map for the total Verbatim score

5.1.7 Construct Validity

Construct validity of the Verbatim was determined using bifactor analysis (Holzinger & Swineford, 1937) with an orthogonal bifactor rotation (Jenrich & Bentler, 2011). The bifactor model models both a general factor that accounts for the covariance amongst the indicators and domain specific or group factors that account for unique variance among clusters of items not accounted for by the general factor (Brown, 2015; Jennrich & Bentler, 2011). A confirmatory bifactor model was stipulated but the model failed to converge. An unrestricted (exploratory) bifactor analysis was thus stipulated using minimum residual ordinary least square extraction and orthogonal bifactor rotation. Pearson correlation coefficients were used as input rather than tetrachoric correlations because the tetrachoric correlation matrix was not positive definite. The RMSEA for the model was .02 (.01 - .02) and the TLI was .93.

The results in Table 23 indicate that there is evidence for a general factor, with the other group factors weakly defined. The absolute loadings on the general factor ranged from .03 to .63. It is noteworthy that only 20 of the 42 items (48%) had an absolute loading ≥ .30 on the general factor. Furthermore, 19 of the 42 items (45%) did not have a meaningful loading (|≥.30|) on any factor. For most of the items the item unique variances outweighed the communality values. The general factor accounted for 44% of the common variance whereas the common variance of the scales ranged from 7% to 23%. The total variance accounted for by the general factor and scales were 21%, indicating that there was 79% unique variance in the model. Overall, the factor structure does not support a five factor group model with a general factor but provides some evidence for interpretation at the general factor level rather than at the scale level. The general factor was, however, weakly defined.

Table 23: Standardised Factor Pattern Coefficients for the Verbatim

Item	F1	F2	F3	F4	F5	F6	h^2	u^2
S1	0.34	-0.14	0.14	0.01	0.22	0.21	0.25	0.75
S2	0.32	-0.02	-0.01	-0.05	-0.09	0.04	0.12	0.89
S3	0.25	0.16	0.00	-0.07	0.03	0.04	0.09	0.91
S4	0.27	-0.01	0.13	-0.03	0.17	0.01	0.12	0.88
S5	0.29	0.34	0.17	0.05	0.06	-0.12	0.25	0.75
S6	0.12	0.12	-0.04	-0.25	-0.08	0.17	0.13	0.87
S7	0.12	0.20	0.13	-0.09	-0.28	-0.27	0.23	0.77
S8	0.07	0.15	0.17	0.14	-0.18	0.04	0.11	0.89
O1	0.40	-0.17	-0.10	0.08	-0.09	0.13	0.24	0.76
O2	0.63	-0.52	0.06	0.02	-0.11	-0.10	0.69	0.31
O3	0.63	-0.49	0.02	-0.09	-0.02	0.00	0.64	0.36
O4	0.33	0.04	0.26	-0.03	0.22	-0.02	0.23	0.77
O5	0.26	0.19	0.03	-0.07	0.01	0.06	0.12	0.89
O6	0.39	0.26	0.21	0.09	-0.04	-0.15	0.30	0.70
O7	0.00	0.05	0.01	-0.12	-0.11	0.03	0.03	0.97
O8	0.08	0.07	0.00	-0.12	0.05	0.09	0.04	0.96
A1	0.23	-0.11	0.03	0.02	0.10	0.12	0.09	0.91
A2	0.30	-0.01	0.05	0.04	0.15	-0.05	0.12	0.88
A3	0.25	0.04	-0.03	-0.07	-0.02	0.09	0.08	0.92
A4	0.26	0.41	0.24	0.10	0.05	-0.08	0.31	0.69
A5	0.34	0.04	0.09	0.15	0.14	-0.03	0.17	0.83
A6	0.37	0.33	0.06	0.07	-0.05	-0.05	0.26	0.74
A7	0.38	0.29	0.01	0.07	-0.01	0.03	0.24	0.76
A8	0.18	0.20	0.14	0.10	-0.14	0.25	0.18	0.82
R1	0.12	0.03	-0.11	0.04	0.03	0.22	0.08	0.92
R2	0.45	0.18	-0.46	0.16	0.03	-0.12	0.49	0.51
R3	0.31	0.22	-0.36	0.21	-0.09	-0.03	0.33	0.67
R4	0.46	0.04	-0.08	0.00	0.18	0.04	0.26	0.74
R5	0.35	0.24	-0.11	-0.14	-0.12	0.12	0.24	0.76
R6	0.34	0.32	-0.11	-0.15	-0.08	0.10	0.27	0.73
R7	0.34	0.32	0.07	-0.11	-0.18	-0.10	0.28	0.73
R8	0.38	0.09	-0.12	0.03	0.08	0.09	0.18	0.82
I1	0.06	0.10	-0.14	-0.10	0.19	-0.12	0.09	0.91
I2	0.30	0.20	-0.10	-0.32	0.15	-0.01	0.27	0.73
I3	0.06	0.00	0.05	0.10	0.05	0.13	0.04	0.97
I4	0.30	0.34	0.11	-0.09	-0.11	0.18	0.27	0.73
I5	0.14	0.07	-0.02	0.11	0.22	0.02	0.09	0.91
I6	0.26	0.27	0.13	0.07	0.03	0.09	0.17	0.83
I7	-0.01	-0.09	0.01	0.17	-0.18	0.29	0.15	0.85
I8	-0.03	0.06	0.11	0.32	0.05	-0.02	0.12	0.88
I9	0.11	0.28	-0.07	-0.21	0.09	-0.15	0.17	0.83
I10	0.12	0.03	0.03	0.25	-0.02	0.02	0.08	0.92
SS	3.78	1.99	.82	.73	.64	.61		
PV	9%	5%	2%	2%	2%	1%		
CV	44%	23%	10%	9%	8%	7%		

Note. SS = sum of squared loadings, PV = % variance explained by each factor, CV = % common variance explained. S = Synonyms, O = Opposites, A = Analogies, R = Reasoning, I = Interpretation.

5.2 Numeratum

5.2.1 Data Screening

Outliers for the total Numeratum score for each respondent were investigated graphically using a histogram and bean plot (Kampstra, 2008). Inspection of the bean plot indicated that there were no potential outliers. A one-sided Grubbs test was conducted on the lowest total Numeratum score value to determine if the value was a statistically significant outlier (Grubbs, 1950). It was found that the lowest value was not a statistically significant outlier (G = 2.4141, U = .9861, p = 1.00). Multivariate outliers across all the scales were subsequently investigated by plotting robust Mahalanobis distances against the quantiles of the χ^2 distribution (Garrett, 1989). Minimal multivariate outliers were detected. Multivariate normality was investigated using Mardia's coefficient (Mardia, 1970). The results indicated that the sub-scales as a whole deviated from multivariate normality.

5.2.2 Descriptive Statistics

Descriptive statistics for each of the Numeratum scales and the total Numeratum score are provided in Table 24. The mean total Numeratum score was 19.67 (median = 20, SD = 4.40). For the most part the variables did not demonstrate particularly problematic skewness or kurtosis values. The standard error values were all generally low. Histograms and bean plots of the Numeratum scales and total Numeratum score are provided in Figure 8 to Figure 11. Inspection of these plots suggested that the three scale scores and the total Numeratum score deviated from normality.

Table 24: Descriptive Statistics for the Numeratum Scales and Total Score

Scale	Mean	SD	Med	Trim	Min	Max	Skew	Kurt	SE
Number Problems	6.47	1.65	6	6.44	1	10	0.14	-0.04	0.08
Patterns	5.61	2.12	6	5.83	0	8	-0.74	-0.43	0.10
Interpretation	7.59	1.70	8	7.72	3	10	-0.61	-0.12	0.08
Total	19.67	4.40	20	19.83	9	28	-0.31	-0.70	0.21

Note. SD = Standard Deviation, Med = Median, Trim = Trimmed mean, Skew = Skewness, Kurt = Kurtosis, SE = Standard Error.

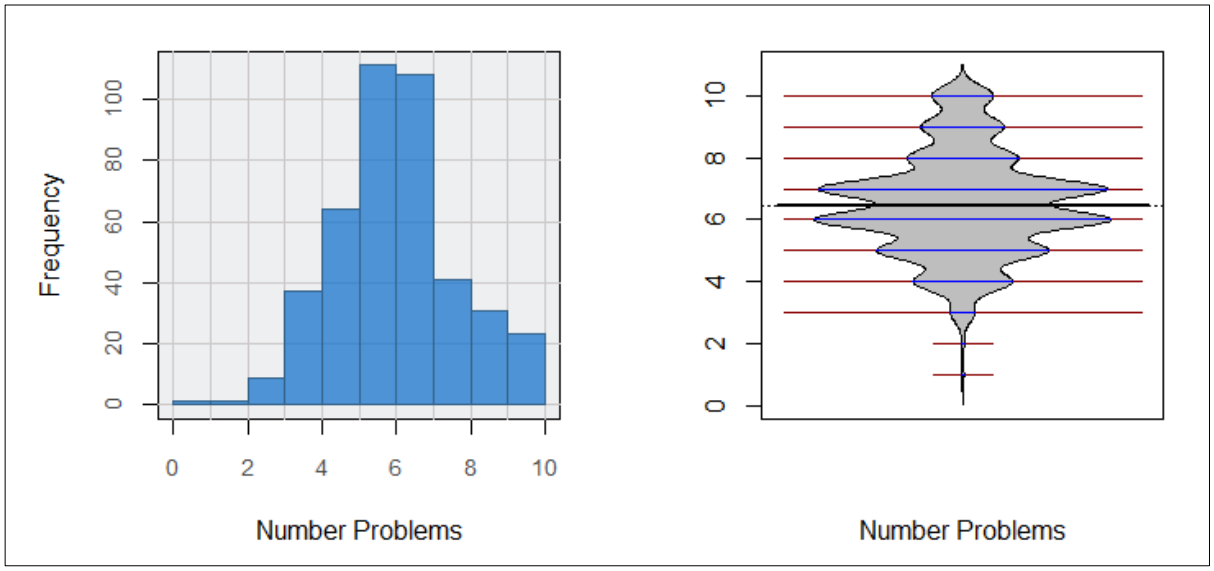


Figure 8: Histogram and bean plot for the Number Problems scale

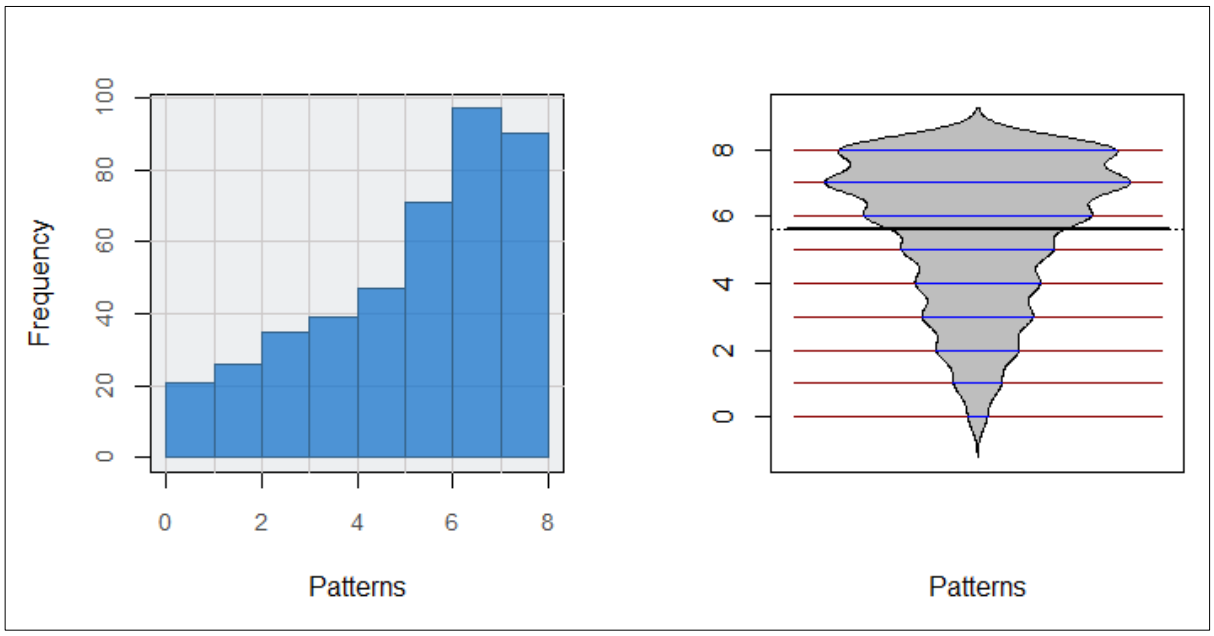


Figure 9: Histogram and bean plot for the Patterns scale

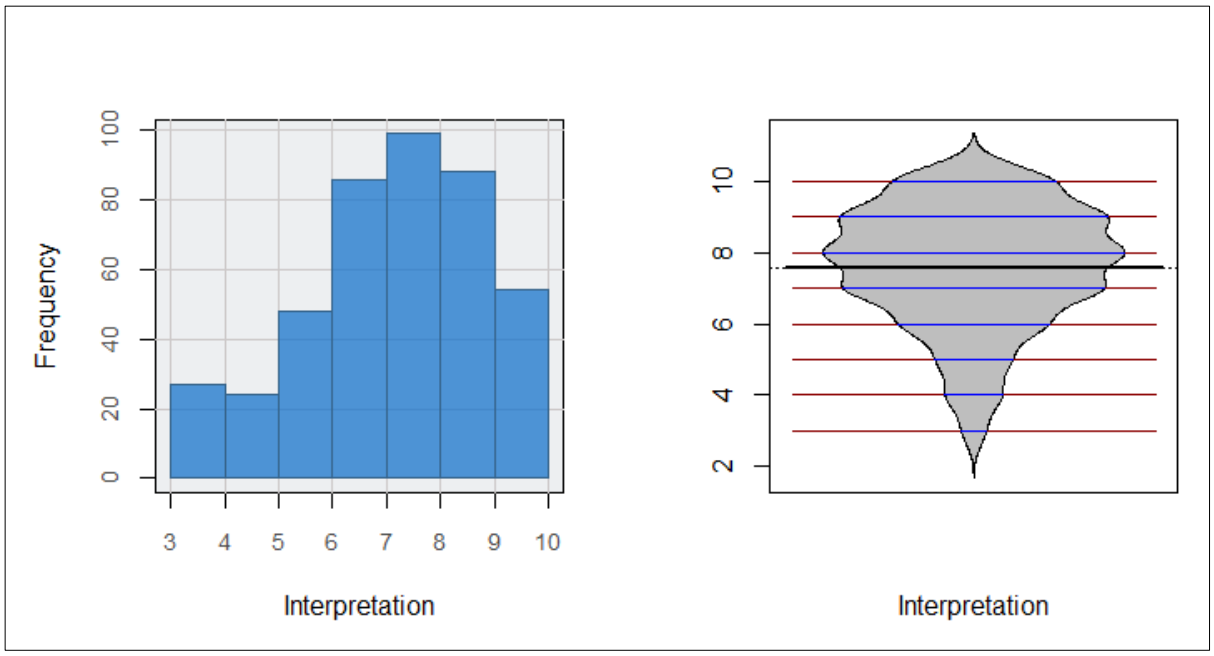


Figure 10: Histogram and bean plot for the Interpretation scale

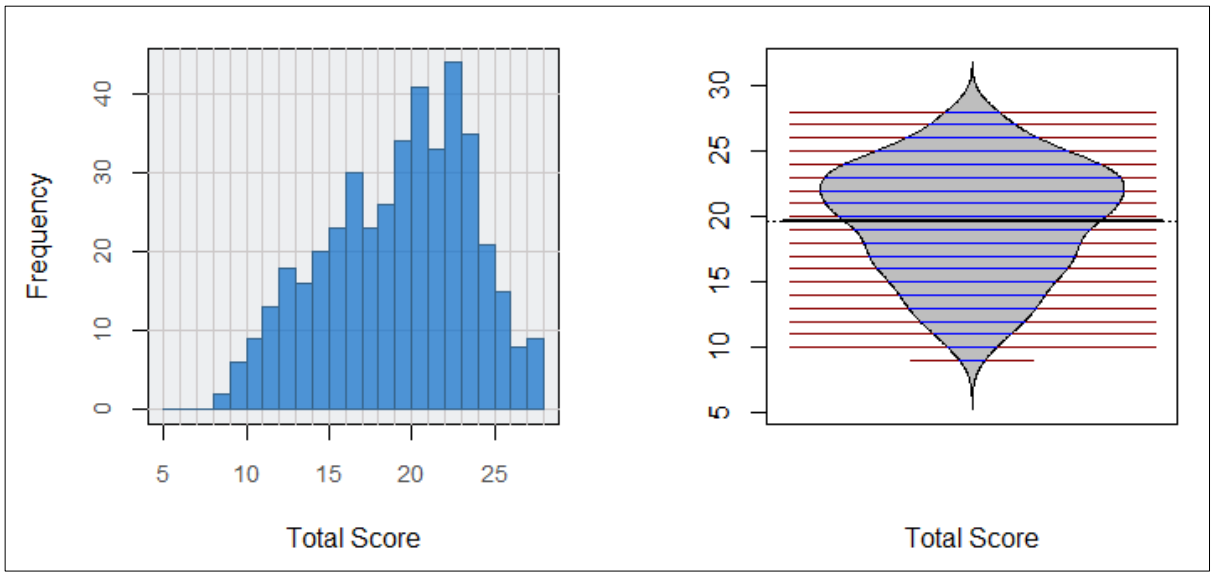


Figure 11: Histogram and bean plot for the total Numeratum score

5.2.3 Reliability

Cronbach alpha coefficients (Cronbach, 1951), Robust Cronbach alpha coefficients (Zhang & Yuan, 2014), Guttman's Lambda 2 (Guttman, 1945), and Ordinal Reliability (Gadermann, Guhn, & Zumbo, 2012) for the Numeratum scales and total Numeratum score are provided in Table 25. The scale reliabilities for the Numeratum scales were mostly unsatisfactory, with Cronbach alpha coefficients ranging from .52 to .75. The Cronbach alpha coefficient for the total Numeratum score was .80. Given this alpha coefficient, caution must be used in interpreting the results of the total Numeratum score (Nunnally, 1970; Urbina, 2004). The item separation index values indicated that the item locations were generally stable. The person separation index values for the Numeratum scales indicated that the scales may not be sensitive enough to distinguish between low and high scorers on the scale (i.e., more items may be required).

For the total Numeratum scores the person separation reliability was higher, indicating that there may be more value interpreting the total score rather than the scale scores.

Table 25: Reliability Coefficients for the Numeratum Scales and Total Score

Scale	α	R α	O α	λ2	IR	PR
Number Problems	.52 (.46 - .58)	.52	.71	.57	.99	.48
Patterns	.75 (.72 - .78)	.73	.87	.76	.98	.50
Interpretation	.57 (.52 - .62)	.48	.77	.60	.99	.39
Total	.80 (.78 - .82)	.79	.89	.81	.99	.76

Note. α = Cronbach's alpha (90% confidence intervals in parenthesis), R α = Robust alpha, O α = Ordinal alpha, λ2 = Guttman's Lambda 2, IR = Item Reliability Index (Rasch), PR = Person Reliability Index (Rasch).

Haberman's (2008) subscale scoring test based on the proportional reduction in mean squared error and McDonald's omega (McDonald, 1999) were used to investigate whether or not interpretation should be conducted at the scale score level or total Numeratum score level (Reise et al., 2013). Application of Haberman's (2008) approach indicated that interpretation of a total Numeratum score rather than Numeratum scale scores may be warranted (Table 26). An unrestricted bifactor model was therefore run using a Schmid-Leiman transformation (see Reise et al., 2013 and Zinbarg et al., 2006). The tetrachoric correlation matrix was not positive definite and therefore Pearson correlation coefficients were used as input instead. In order to determine the relative merits of interpretation at the general factor level (i.e., a total score), Omega total (ω_t), Omega hierarchical (ω_h), and Omega sub-scale (ω_s) were calculated (Reise et al., 2013).

Omega total (ω_t) for the three factor model with a general factor was .82, while Omega hierarchical (ω_h) was .71. The explained common variance (ECV) was used to index the degree of unidimensionality (Reise, 2012). In this analysis the ECV was .60, indicating that 60% of the common variance was due to the general factor. While there are no specific criteria for an acceptable ECV value, the higher the value (i.e., closer to 1), the stronger the general factor (Reise, 2012). Omega sub-scale indicates what the reliability of the scales would be after

controlling for the general factor (Reise et al., 2013). The ω_s values for the three scales were .00, .09, and .30. The three factors were not well-defined and it is therefore not possible to label them. The ω_s values were lower than ω_h, but the ECV value was small. It therefore appears that there is some evidence for interpretation of a total Numeratum score rather than Numeratum scale scores. However, such an interpretation must be made with caution. Furthermore, given the limitations of the Schmid-Leiman transformation with exploratory factor analysis (see Reise et al., 2010), caution must be used in interpretation of the results[3].

Table 26: Haberman's (2008) Subscale Scoring Test Results

Scale	$PRMSE_s$	$PRMSE_T$
Number Problems	.52	.63
Patterns	.75	.78
Interpretation	.57	.72
Tot	.80	

PRMSE = proportional reduction in mean squared error.

5.2.4 Item difficulty and item discrimination

Item difficulty[4] and item discrimination values were estimated within a CTT framework (*cf.* Lord & Novick, 1968*;* Raykov & Marcoulides, 2011). The item difficulty index is the proportion of respondents who correctly answered the item to the total number of participants and the item discrimination index is the ability of an item to discriminate between respondents who scored high and low on the scale/test (Kerlinger & Lee, 2000; Nunnally, 1970). According to Kerlinger and Lee (2000), item difficulties should range between .50 and .70, where a value of 1 indicates that all respondents obtained the correct answer (i.e., too easy) while a value of 0 indicates that none of the respondents obtained the correct answer (i.e., too difficult) (Raykov

[3] A confirmatory bifactor model was also investigated, but the model failed to converge.
[4] Item difficulty/locations within the Rasch framework are available later in the document.

& Marcoulides, 2011). However, in an ability test it would be expected that there is a greater range of item difficulties. The inter-item correlation coefficient, item-total score correlation, item difficulty, and item discrimination values for the Numeratum scales are provided in Table 27.

Inspection of Table 27 indicates that most of the items had item difficulties that were not in the range of .50 to .70. There were also a fair number of items that were potentially too easy (NP1, NP2, NP3, P1, I1, I3, and I6). The average item difficulty across the Numeratum scales were: Number Problems = .65, Patterns = .70, and Interpretation = .76. The item discrimination values were generally quite low across the Number Problems and Interpretation scales, indicating that items in these scales may not be optimal in discriminating between respondents who score high and low on the latent trait. The average item discrimination across the three scales was: Number Problems = .34, Patterns = .57, Interpretation = .37.

Table 27: Average Inter-Item Correlation Coefficient, Correlation with Scale Score, Item Difficulty, and Item Discrimination for the Numeratum Scale Items.

Item	Average R	Raw R	R Cor	Dif	Disc
Number Problems					
NP1	0.10	0.32	0.27	0.94	0.15
NP2	0.10	0.32	0.25	0.93	0.16
NP3	0.11	0.19	0.15	0.98	0.06
NP4	0.09	0.46	0.34	0.85	0.36
NP5	0.13	0.19	-0.06	0.79	0.18
NP6	0.10	0.46	0.26	0.58	0.47
NP7	0.08	0.58	0.49	0.70	0.66
NP8	0.09	0.54	0.43	0.23	0.44
NP9	0.08	0.65	0.65	0.20	0.49
NP10	0.09	0.54	0.44	0.27	0.48
Patterns					
P1	0.30	0.44	0.40	0.91	0.20
P2	0.26	0.63	0.60	0.81	0.44
P3	0.26	0.67	0.63	0.78	0.56
P4	0.27	0.63	0.55	0.70	0.63
P5	0.24	0.78	0.76	0.68	0.81
P6	0.31	0.46	0.29	0.64	0.49
P7	0.27	0.68	0.58	0.52	0.80
P8	0.30	0.54	0.39	0.56	0.65
Interpretation					
I1	0.10	0.34	0.29	0.94	0.14
I2	0.10	0.30	0.19	0.87	0.22
I3	0.10	0.27	0.22	0.94	0.11
I4	0.10	0.51	0.42	0.82	0.36
I5	0.10	0.52	0.35	0.71	0.54
I6	0.10	0.36	0.35	0.95	0.11
I7	0.10	0.52	0.42	0.82	0.41
I8	0.10	0.52	0.33	0.32	0.63
I9	0.10	0.53	0.33	0.44	0.65
I10	0.10	0.57	0.51	0.78	0.51

Note. Average R = average inter-item correlation coefficient, Raw R = item-total correlation not corrected for item overlap, R Cor = item-total correlation corrected for item overlap and scale reliability, Dif = item difficulty, Disc = item discrimination.

The inter-item correlation coefficient, item-total score correlation, item difficulty, and item discrimination values for the total Numeratum score are provided in Table 28. The item difficulties remained the same (as expected). The item discrimination values were again generally quite low, but did show improvement. Items NP3 and NP5 had small discrimination values, indicating that they were unable to discriminate between respondents who scored high and low on the Numeratum. The average item difficulty was .70 and the average item discrimination was .35.

Table 28: Average Inter-Item Correlation Coefficient, Correlation with Scale Score, Item Difficulty, and Item Discrimination for the Total Numeratum Score Items.

Item	Average R	Raw R	R Cor	Dif	Disc
NP1	0.12	0.25	0.23	0.94	0.13
NP2	0.12	0.24	0.23	0.93	0.13
NP3	0.12	0.14	0.14	0.98	0.04
NP4	0.12	0.43	0.39	0.85	0.36
NP5	0.13	0.08	-0.01	0.79	0.06
NP6	0.12	0.28	0.20	0.58	0.33
NP7	0.11	0.56	0.54	0.70	0.60
NP8	0.12	0.32	0.27	0.23	0.30
NP9	0.12	0.47	0.45	0.20	0.39
NP10	0.12	0.34	0.28	0.27	0.34
P1	0.12	0.33	0.32	0.91	0.18
P2	0.12	0.50	0.49	0.81	0.39
P3	0.11	0.57	0.56	0.78	0.51
P4	0.11	0.57	0.55	0.70	0.57
P5	0.11	0.70	0.71	0.68	0.76
P6	0.12	0.42	0.35	0.64	0.46
P7	0.11	0.62	0.59	0.52	0.75
P8	0.12	0.46	0.40	0.56	0.51
I1	0.12	0.23	0.22	0.94	0.13
I2	0.12	0.19	0.15	0.87	0.17
I3	0.12	0.21	0.19	0.94	0.09
I4	0.12	0.38	0.34	0.82	0.30
I5	0.12	0.41	0.35	0.71	0.44
I6	0.12	0.29	0.31	0.95	0.12
I7	0.12	0.39	0.37	0.82	0.32
I8	0.12	0.44	0.37	0.32	0.49
I9	0.12	0.43	0.37	0.44	0.51
I10	0.12	0.50	0.48	0.78	0.46

Note. Average R = average inter-item correlation coefficient, Raw R = item-total correlation not corrected for item overlap, R Cor = item-total correlation corrected for item overlap and scale reliability, Dif = item difficulty, Disc = item discrimination. NP = Number Problems, P = Patterns, I= Interpretation.

5.2.5 Correlation Coefficients

Pearson correlation coefficients and Spearman-rho rank order correlation coefficients for the three Numeratum scales are provided in Table 29. Disattenuated correlations potentially lead to overcorrection due to the low Cronbach alpha coefficients (see Osborne, 2003) and are therefore not provided. Inspection of the non-parametric Loess lines (Cleveland, 1979) indicated that for the most part the relationships between the variables were linear. Inspection of multivariate normality using Mardia's coefficient (Mardia, 1970) found that

bivariate normality was violated across all the scales. For the most part the correlation coefficients had medium to large effect sizes (Cohen, 1988).

Table 29: Pearson and Spearman-Rho Rank Order Correlations for the Numeratum Scales

Scale	Number Problems	Patterns	Interpretation
Number Problems	.	.48	.35
Patterns	.46	.	.59
Interpretation	.35	.56	.

Note. Pearson correlations below the diagonal, Spearman rho rank-order correlations above the diagonal. NP = Number Problems, P = Patterns, I= Interpretation.

5.2.6 Rasch Analysis

A Rasch (1960) analysis was conducted on the total Numeratum score to inspect item fit statistics and item locations (difficulties). Items with Infit and Outfit mean square values (IMNSQ and OMNSQ) > 1.40 were considered to be underfitting items, and items with IMNSQ and OMNSQ values < .60 to be overfitting items. OMNSQ investigates unexpected responses to items that are either too easy or too difficult for the respondent, whereas IMNSQ investigates unexpected responses on items that are targeted at the respondents underlying latent ability measure (Linacre, 2015). Item and person separation reliabilities are provided in Table 25. The item fit statistics and item locations for the total Numeratum scale are provided in Table 30 and Figure 12 below.

Table 30: Total Numeratum Score Item Location and Item Fit Statistics

Item	Measure	SE	Infit MNSQ	Z	Outfit MNSQ	Z	PT-Measure Corr	Exp
NP5	-.29	.13	1.35	4.7	2.12	5.8	.09	.37
I2	-1.01	.16	1.14	1.4	1.86	3.2	.18	.30
I3	-1.94	.21	1.02	.2	1.56	1.6	.18	.22
NP8	2.96	.13	1.08	1.2	1.47	3.1	.39	.48
NP10	2.68	.13	1.07	1.2	1.43	3.3	.40	.49
NP6	.95	.11	1.25	4.9	1.31	3.8	.30	.46
I4	-.54	.14	.99	-.1	1.29	1.6	.33	.35
NP2	-1.73	.20	1.00	.0	1.21	.8	.22	.23
I9	1.72	.11	1.05	1.1	1.12	1.6	.45	.49
I8	2.36	.12	.99	-.2	1.12	1.2	.48	.49
I5	.25	.12	1.04	.8	1.11	1.0	.39	.42
P6	.66	.11	1.06	1.2	1.09	1.0	.41	.45
I7	-.52	.14	.99	-.1	1.07	.5	.35	.35
P8	1.08	.11	1.04	.9	1.06	.8	.44	.47
P3	-.21	.13	.81	-3.1	1.02	.2	.49	.38
I1	-1.90	.21	1.02	.2	.93	-.1	.21	.22
NP3	-2.94	.32	1.00	.1	.89	-.1	.13	.14
NP1	-1.90	.21	1.00	.1	.85	-.4	.23	.22
P1	-1.49	.18	.95	-.3	.80	-.7	.29	.26
I6	-2.14	.23	.95	-.2	.65	-1.0	.25	.20
NP4	-.74	.14	.94	-.7	.70	-1.7	.39	.33
NP9	3.19	.14	.92	-1.0	.78	-1.4	.53	.47
I10	-.23	.13	.89	-1.6	.86	-1.0	.45	.38
P2	-.49	.14	.88	-1.6	.83	-1.1	.44	.36
NP7	.31	.12	.87	-2.6	.82	-1.8	.51	.43
P4	.26	.12	.85	-2.8	.82	-1.7	.52	.42
P7	1.29	.11	.82	-4.3	.75	-4.0	.60	.48
P5	.39	.12	.70	-6.3	.58	-5.0	.63	.43
Mean	.00	.15	.99	-.3	1.08	.3		
SD	1.59	.05	.13	2.3	.35	2.3		

Note. OMNSQ ≥ 1.4 or ≤ .60 in bold. NP = Number Problems, P = Patterns, I= Interpretation.

Inspection of Table 30 and Figure 12 indicates that the item locations ranged between approximately -2.94 to 3.19 logits. The items mostly covered the underlying ability trait, except for at the upper ability trait. Several items also appeared to be too easy (NP3, I6, I3, NP1, I1, and P1). Five items demonstrated underfit (NP5, PP2, PP3, NP8, and NP10) and one item demonstrated overfit (P5).

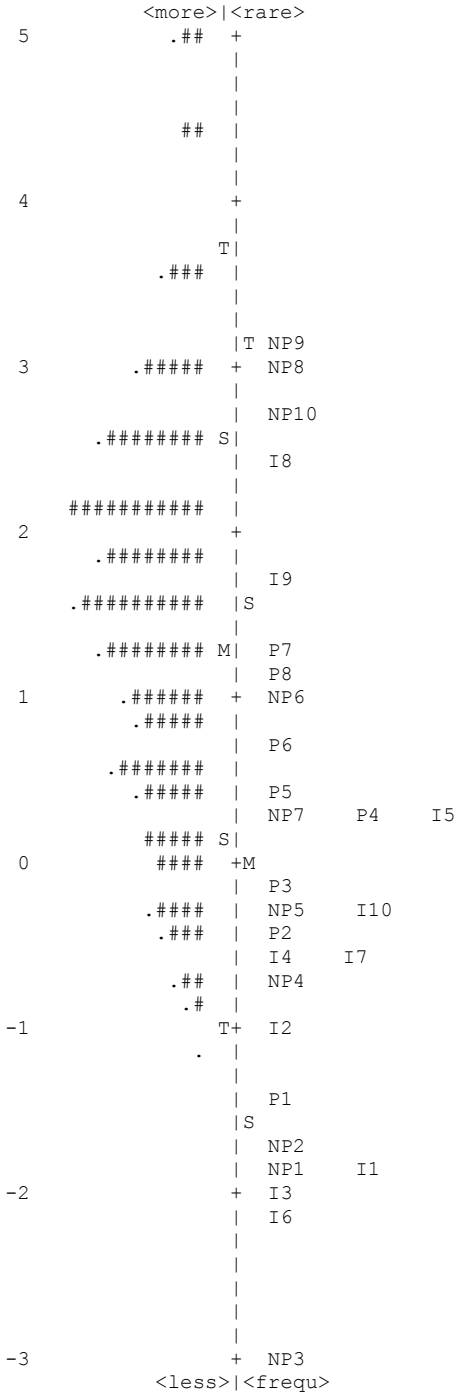


Figure 12: Wright map for the total Numeratum score

5.2.7 Construct Validity

Construct validity of the Numeratum was determined using bifactor analysis (Holzinger & Swineford, 1937) with an orthogonal bifactor rotation (Jenrich & Bentler, 2011). The bifactor model models both a general factor that accounts for the covariance amongst the indicators and domain specific or group factors that account for unique variance not accounted for by the general factor (Brown, 2015; Jennrich & Bentler, 2011). A confirmatory bifactor model was stipulated but the model failed to converge. An unrestricted (exploratory) bifactor analysis was thus stipulated using minimum residual ordinary least square extraction with Pearson correlations used as input. Tetrachoric correlations could not be used as input because the matrix was not positive definite. The RMSEA for the model was .03 (.03 - .04) and the TLI was .90.

The results in Table 31 indicate that there is evidence for a general factor, with the other group factors weakly defined. The absolute loadings on the general factor ranged from .10 to .73. It is noteworthy that only 16 of the 28 items (57%) had an absolute loading ≥ .30 on the general factor. Furthermore, 10 of the 28 items (36%) did not have a meaningful loading (|≥.30|) on any factor. For most of the items the item unique variances outweighed the communality values. The general factor accounted for 55% of the common variance whereas the common variance of the scales ranged from 8% to 26%. The total variance accounted for by the general factor and scales were 25%, indicating that there was 75% unique variance in the model. Overall, the factor structure does not support a three factor group model with a general factor but provides some evidence for interpretation at the general factor level rather than at the scale level.

Table 31: Standardised Factor Pattern Coefficients for the Numeratum

Item	F1	F2	F3	F4	h^2	u^2
NP1	0.19	0.10	0.09	-0.13	0.07	0.93
NP2	0.15	0.14	0.09	0.15	0.07	0.93
NP3	0.12	0.04	0.06	-0.02	0.02	0.98
NP4	0.36	0.17	0.27	0.10	0.24	0.76
NP5	-0.10	0.18	0.11	-0.01	0.05	0.95
NP6	0.16	0.03	0.12	-0.07	0.05	0.96
NP7	0.48	0.19	0.29	0.03	0.35	0.65
NP8	0.43	-0.39	-0.04	0.09	0.35	0.65
NP9	0.73	-0.54	-0.03	0.02	0.82	0.18
NP10	0.46	-0.41	-0.05	-0.05	0.39	0.61
P1	0.27	0.28	-0.44	0.04	0.35	0.66
P2	0.44	0.37	-0.38	-0.08	0.47	0.53
P3	0.50	0.39	-0.12	-0.02	0.41	0.59
P4	0.51	0.29	0.08	-0.21	0.39	0.61
P5	0.65	0.39	-0.02	-0.12	0.58	0.42
P6	0.32	0.09	0.13	-0.11	0.14	0.86
P7	0.54	0.22	0.09	-0.07	0.35	0.65
P8	0.35	0.18	0.00	0.02	0.16	0.84
I1	0.15	0.19	-0.20	0.13	0.12	0.88
I2	0.11	0.08	-0.08	0.10	0.03	0.97
I3	0.15	0.04	-0.01	0.26	0.09	0.91
I4	0.27	0.19	0.05	0.20	0.16	0.85
I5	0.32	0.15	0.26	-0.03	0.20	0.80
I6	0.22	0.13	0.04	0.33	0.17	0.83
I7	0.30	0.25	0.00	0.24	0.21	0.79
I8	0.35	0.11	0.13	-0.17	0.18	0.82
I9	0.29	0.18	0.07	-0.12	0.13	0.87
I10	0.38	0.37	0.09	0.28	0.36	0.64
SS	3.83	1.79	0.74	0.57		
PV	14%	6%	3%	2%		
CV	55%	26%	11%	8%		

Note. SS = sum of squared loadings, PV = % variance explained by each factor, CV = % common variance explained. NP = Number Problems, P = Patterns, I= Interpretation.

CHAPTER SIX

Correlation between the Verbatim and Numeratum

Pearson correlation coefficients and Spearman-rho rank order correlation coefficients for Verbatim and Numeratum scales and total scores are provided in Table 32. Inspection of the non-parametric Loess lines (Cleveland, 1979) indicated that for the most part the relationships between the variables were linear. Inspection of multivariate normality using Mardia's coefficient (Mardia, 1970) found that bivariate normality was violated across most of the variables. For the most part the correlation coefficients had medium to large effect sizes (Cohen, 1988).

Table 32: Pearson and Spearman-Rho Rank Order Correlations for the Verbatim and Numeratum Scales

	S	O	A	R	I	VT	NP	P	NI	NT
S	.	.	.	.	.	.	0.24	0.31	0.27	0.35
O	.	.	.	.	.	.	0.20	0.25	0.20	0.28
A	.	.	.	.	.	.	0.25	0.42	0.39	0.44
R	.	.	.	.	.	.	0.33	0.50	0.40	0.51
I	.	.	.	.	.	.	0.25	0.37	0.33	0.39
VT	.	.	.	.	.	.	0.37	0.54	0.46	0.57
NP	0.22	0.20	0.26	0.31	0.24	0.36	.	.	.	.
P	0.31	0.27	0.40	0.47	0.34	0.53	.	.	.	.
NI	0.25	0.22	0.39	0.34	0.32	0.45	.	.	.	.
NT	0.33	0.29	0.44	0.48	0.38	0.56	.	.	.	.

Note. *N* = 336. Pearson correlation coefficients below the diagonal, Spearman rho rank-order correlations above the diagonal. S = Synonyms, O = Opposites, A = Analogies, R = Reasoning, I = Verbatim Interpretation, VT = total Verbatim score, NP = Number Problems, P = Patterns, NI = Numeratum Interpretation, NT = total Numeratum score.

CHAPTER SEVEN

Concluding Comments

The Verbatim and Numeratum tests were developed to address the need for specific numerical and verbal reasoning skills that are not necessarily measured by other mental ability assessments. These tests aim to assess a person's ability to understand and problem-solve accurately using verbal and numerical information and can be used for screening, competency-based selection and training. At this stage it is important to be aware of the scored descriptors in the interpretation section. Selection decisions should not be made using these scores alone – they must be done in conjunction with other evidence.

CHAPTER EIGHT

R Packages

The following R (R Core Team, 2015) packages were used in the analysis:

Psych (Revelle, 2015), lessR (Gerbing, 2015), beanplot (Kampstra, 2014), outliers (Komsta, 2011), mvoutlier (Filzmoser & Gschwandtner, 2015), MVN (Korkmaz, Goksuluk, & Zararsiz, 2015), coefficientalpha (Zhang & Yuan, 2015), cocron (Diedenhofen, 2013), Lambda4 (Hunt, 2013), sirt (Robitzsch, 2015), psychometric (Fletcher, 2010), lavaan (Rosseel, 2012), and car (Fox & Weisberg, 2011), as well as all associated dependencies.

CHAPTER NINE

References

Brown, T. A. (2015). Confirmatory factor analysis for applied research (2nd edition). New York: Guilford Press.

Carrol, J. B. (1993). Human cognitive abilities: A survey of factor analytic studies. Cambridge: Press Syndicate.

Cleveland, William S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association, 74*(368), 829–836. doi:10.2307/2286407.JSTOR2286407.MR 0556476.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334.

Diedenhofen, B. (2013). cocron: Statistical comparisons of two or more alpha coefficients (Version 1.0-0). Retrieved from http://r.birkdiedenhofen.de/pckg/cocron/

Filzmoser, P., & Gschwandtner, M. (2015). mvoutlier: Multivariate outlier detection based *on robust methods* (Version = 2.0.6). Retrieved from http://www.statistik.tuwien.ac.at/public/filz/

Fletcher, T. D. (2010). *psychometric: Applied Psychometric Theory* (Version 2.2). Retrieved from https://cran.r-project.org/web/packages/psychometric/index.html.

Fox, F., & Weisberg, S. (2011). *An {R} Companion to Applied Regression* (2nd edition). Thousand Oaks CA: Sage. Retrieved from http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-

type and ordinal item response data: A conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation, 17*(3), 1-13.

Garrett, R. G. (1989). The chi-square plot: a tool for multivariate outlier recognition. *Journal* of Geochemical Exploration, 32 (1-3), 319–341.

Gerbing, D. W. (2015). *lessR: Less Code, More Results* (R package version 3.3.1. Retrieved from http://cran.r-project.org/web/packages/lessR/index.html.

Grubbs, F. E. (1950). Sample Criteria for testing outlying observations. *The Annals of Mathematical Statistics, 21*(1), 27-58.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*(4), 255-282.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, 33(2), 204-229.

Hunt, T. (2013). Lambda4: Collection of Internal Consistency Reliability Coefficients (Version 3.0). Retrieved from https://cran.r-project.org/web/packages/Lambda4/index.html

Holzinger, K. J., & Swineford, F. (1937). The bi-factor method. *Psychometrika, 2*(1), 41-54.

Jennrich, R. I., & Bentler, P. M. (2011). Exploratory bi-factor analysis. *Psychometrika, 76*(4), 537-549.

Jensen, A. R. (1998). *The g Factor: The Science of Mental Ability*. Westport CT: Praeger.

Kampstra, P. (2008). Beanplot: A boxplot alternative for visual comparison of distributions. Journal of Statistical Software, 28(1), 1-9.

Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of Behavioral Research*. Wadsworth, Thomson Learning: Northridge, CA.

Komsta, L. (2011). *outliers: Tests for outliers* (Version 0.14.). Retrieved from http://www.r-project.org/

Korkmaz, S., Goksuluk, D., Zararsiz, G. (2014). MVN: An R Package for Assessing Multivariate Normality. *The R Journal, 6*(2), 151-162.

Linacre, J.M. (2015). *Misfit diagnosis: infit outfit mean-square standardized***.** Retrieved from http://www.winsteps.com/winman/diagnosingmisfit.htm

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Oxford, England: Addison-Wesley.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57*(3), 519-530.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Hillsdale: Erlbaum.

Nunnally Jr, J. C. (1970). *Introduction to psychological measurement*. McGraw-Hill, New York.

Osborne, J. W. (2003). Effect sizes and the disattenuation of correlation and regression coefficients: lessons from educational psychology. *Practical Assessment, Research and Evaluation, 8*(11).

Postlethwaite, B. E. (2011). Fluid ability, crystallised ability, and performance across multiple domains: a meta-analysis. Unpublished doctoral thesis.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen, Denmark: Danmarks Paedogogische Institute.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. New York: Routledge Taylor and Francis Group.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, 47(5), 667-696.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of personality assessment, 95*(2), 129-140.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of personality assessment, 92*(6), 544-559.

Revelle, W. (2015) psych: Procedures for Personality and Psychological Research (Version = 1.5.6.). Retrieved from http://CRAN.R-project.org/package=psych.

Robitzsch, A. (2015). sirt: Supplementary Item Response Theory Models. Retrieved from https://sites.google.com/site/alexanderrobitzsch/software version 1.8-9

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1-36. Retrieved from http://www.jstatsoft.org/v48/i02/

Urbina, S. (2004). *Essentials of psychological testing*. Hoboken, NJ: John Wiley.

Zhang, Z. & Yuan, K. H. (2014). *Robust Coefficients Alpha and Omega and their Confidence Intervals: Methods and Software*. Retrieved from https://cran.r-project.org/web/packages/coefficientalpha/coefficientalpha.pdf

Zhang, Z., & Yuan, K. H. (2015). coefficientalpha: Robust Coefficient Alpha and Omega with *Missing and Non-Normal Data* (Version 0.5). Retrieved from https://cran.r-project.org/web/packages/coefficientalpha/index.html

Zinbarg, R., Yovel, I., Revelle, W. & McDonald, R. (2006). Estimating generalizability to a universe of indicators that all have one attribute in common: A comparison of estimators for omega. *Applied Psychological Measurement, 30*, 121-144.

APPENDIX A

Verbatim Feedback Report

verbatim
verbal ability assessment

VERBATIM

FEEDBACK REPORT

DEVELOPED BY:
JvR PSYCHOMETRICS

NAME:	John Sample
GENDER:	Male
REPORT DATE:	13.03.2013

CONFIDENTIAL REPORT

The information in this report is confidential and must not be made known to anyone other than authorised personnel, unless released by the expressed written permission of the person taking the assessment. The information should be considered together with all other information gathered in the assessment process.

www.jvrpsychometrics.co.za

VERBATIM FEEDBACK REPORT

This report is based on the candidate's responses to the assessment and provides the results of their Verbatim test. It gives an indication of the way they understand and work with verbal information.

INTERPRETATION GUIDELINES

The results are presented on a stanine scale, which has a mean of 5 and a standard deviation of 2. This means that most people score between 3 and 7 on the scale. The interpretation for each range is provided below:

HIGH SCORE: Scores of 7 and above indicate the candidate found it easy to work with verbal content, and would work well in a position that has a lot of English verbal content and requires decisions to be made on that content.

MIDRANGE SCORE: Scores between 3 and 7 fall into the midrange. This means that the candidate possibly found some elements of working with verbal content easy and others difficult. They would function well in positions that have a moderate amount of English verbal content.

LOW SCORE: Scores of 3 and below indicate that the candidate found it difficult to work with verbal content, and would be better suited for positions that do not rely on a large amount of English verbal content.

All results are compared to a South African benchmark.

VERBATIM RESULTS

VERBAL REASONING

The following sections make up the scales of the Verbatim. The results below show which of the sections the individual found easy, moderately hard, and difficult to complete compared to others.

SYNONYMS
The ability to identify words that are the same or similar in meaning.

OPPOSITES
The ability to identify words with opposite meanings.

ANALOGIES
The ability to identify the relationship between a pair of words and to identify equivalent or similar relationships in different pairs of words.

REASONING
The ability to reason with letters and other verbal content.

INTERPRETATION
The ability to read and comprehend verbal content accurately.

COUNSELLOR'S SECTION

The table below contains the percentage of items the candidate responded to correctly under each scale measure by Verbatim. It also shows the number of items they omitted and the time they took to complete each section.

VERBATIM SCALE	PERCENTAGE CORRECTLY ANSWERED	OMITTED ITEMS	TIME ALLOWED	TIME TAKEN TO COMPLETE
SYNONYMS			10 minutes	
OPPOSITES			10 minutes	
ANALOGIES			10 minutes	
REASONING			20 minutes	
INTERPRETATION			20 minutes	

The report also includes information that can provide an indication of the candidate's speed and accuracy while completing the test. The "Number Attempted" refers to the number of questions the candidate has attempted to answer during the tests. This includes only items that the candidate provided a response to. When considered against the total number of items presented to the candidate, "Number of Items", it provides an indication of the speed at which the candidate completed the assessment.

The "Number Correctly Answered", refers to those items that the candidate attempted and answered correctly. When taken into consideration with the "Number Attempted", it provides you with an indication of the accuracy of the candidate while completing the assessment.

Overall Time to Complete Verbatim: 70 minutes
Number of Items:
Number Attempted:
Number Correctly Answered:

Numeratum Feedback Report

numeratum
numerical ability assessment

NUMERATUM

FEEDBACK REPORT

DEVELOPED BY:
JvR PSYCHOMETRICS

NAME:	John Sample
GENDER:	Male
REPORT DATE:	13.03.2013

CONFIDENTIAL REPORT

The information in this report is confidential and must not be made known to anyone other than authorised personnel, unless released by the expressed written permission of the person taking the assessment. The information should be considered together with all other information gathered in the assessment process.

www.jvrpsychometrics.co.za

NUMERATUM FEEDBACK REPORT

This report is based on the candidate's responses to the assessment and provides the results of their Numeratum test. It gives an indication of the way they understand and work with numbers and solves problems that involve numbers.

INTERPRETATION GUIDELINES

The results are presented on a stanine scale, which has a mean of 5 and a standard deviation of 2. This means that most people score between 3 and 7 on the scale. The interpretation for each range is provided below:

HIGH SCORE: Scores of 7 and above indicate the candidate found it easy to work with numerical content, and would work well in a position that has a lot of numerical content and requires decisions to be made on that content.

MIDRANGE SCORE: Scores between 3 and 7 fall into the midrange. This means that the candidate possibly found some elements of working with numerical content easy and others difficult. They would function well in positions that have a moderate amount of numerical content that needs to be interpreted.

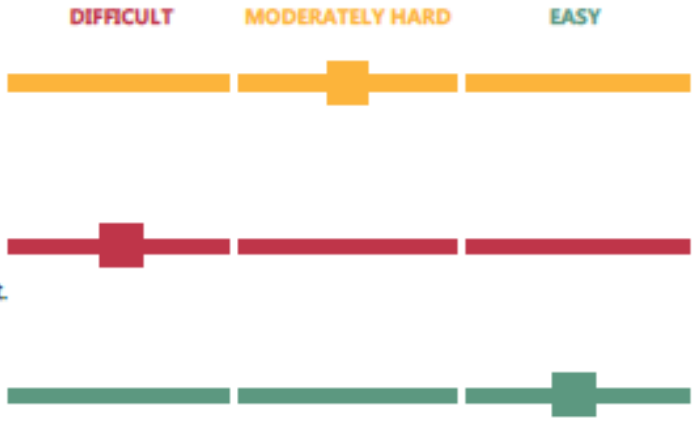
LOW SCORE: Scores of 3 and below indicate that the candidate found it difficult to work with numbers, and would be better suited for positions that do not rely on a large amount of numerical content.

All results are compared to a South African benchmark.

NUMERATUM RESULTS

NUMERICAL REASONING

The following sections make up the scales of the Numeratum. The results below show which of the sections the individual found easy, moderately hard, and difficult to complete compared to others.



NUMBER PROBLEMS
The ability to add, subtract, multiply and divide across a range of numerical problems.

PATTERNS
The ability to solve problems that requires identification of patterns in numerical content.

INTERPRETATION
The ability to read and interpret basic numerical information contained in a graph.

COUNSELLOR'S SECTION

The table below contains the percentage of items the candidate responded to correctly under each scale measure by Numeratum. It also shows the number of items they omitted and the time they took to complete each section.

NUMERATUM SCALE	PERCENTAGE CORRECTLY ANSWERED	OMITTED ITEMS	TIME ALLOWED	TIME TAKEN TO COMPLETE
NUMBER PROBLEMS			20 minutes	
PATTERNS			20 minutes	
INTERPRETATION			20 minutes	

The report also includes information that can provide an indication of the candidate's speed and accuracy while completing the test. The "Number Attempted" refers to the number of questions the candidate has attempted to answer during the tests. This includes only items that the candidate provided a response to. When considered against the total number of items presented to the candidate, "Number of Items", it provides an indication of the speed at which the candidate completed the assessment.

The "Number Correctly Answered", refers to those items that the candidate attempted and answered correctly. When taken into consideration with the "Number Attempted", it provides you with an indication of the accuracy of the candidate while completing the assessment.

Overall Time to Complete Numeratum: 60 minutes
Number of Items:
Number Attempted:
Number Correctly Answered: