JVR PSYCHOMETRICS

VERBATIM AND NUMERATUM TECHNICAL MANUAL 2ND EDITION

V VERBATIM

N NUMERATUM

VERBATIM and NUMERATUM

Technical Manual 2nd Edition

DEVELOPED BY:

CJ van Zyl and Dr N Taylor on behalf of JVR Psychometrics (Pty) Ltd

MANUAL PREPARED BY:

JVR Psychometrics (Pty) Ltd

AUTHOR'S NOTE:

Correspondence regarding this technical manual should be addressed to the Product and Research team, JVR Psychometrics.

Email: info@jvrafrica.co.za

This product has been developed and is solely distributed by JVR Psychometrics (Pty) Ltd.

JVR Psychometrics forms part of the JVR Africa Group

JOHANNESBURG HEAD OFFICE

15 Hunter Street, Ferndale, Randburg, 2194

P.O. Box 2560, Pinegowrie, 2123, Johannesburg, South Africa

(Tel) +27 11 781 3705/6/7

(Email) info@jvrafrica.co.za

(Website) https://jvrafricagroup.co.za/psychometrics

(Blog) https://jvrafricagroup.co.za/blog/

ACKNOWLEDGEMENTS

A special thank you to each member of the Product and Research team who assisted in any shape or form throughout the process of shortening the Verbatim and Numeratum. Thank you for your valuable insights, eagerness to assist where required, and your contagious enthusiasm for the work we do. A special thank you also to Morné Stander for the design elements and Elena Rust for the proofreading and editing. Our sincere gratitude also goes out to every organisation and employee, even though we cannot mention them by name, who completed the assessments, as this enabled us to critically evaluate the psychometric properties of the questionnaires and make the necessary modifications.

CONTENTS

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Employee performance continues to form a focal area of inquiry in the management domain and adjacent fields (Van Iddekinge et al., 2018). “Conceptual models and considerable empirical evidence suggest that two key determinants of performance are cognitive ability and motivation” (Van Iddekinge et al., 2018, p. 250). The Verbatim and Numeratum both focus on the cognitive ability aspect, as they have been designed to assess verbal and numerical reasoning ability, respectively. The Verbatim comprises 28 scored items that can be used to assess whether an individual has sufficient skill in processing and understanding written information in English. The Numeratum comprises 16 scored items that can be used to assess whether an individual has sufficient reasoning skills to work with numerical data. It should take approximately ± 20 minutes to complete either of these questionnaires. The Verbatim and Numeratum and their theoretical foundation are discussed in greater depth below.

Purpose and Rationale

The Verbatim and Numeratum’s primary purpose 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.

User Qualifications

According to the Health Professions Act, No. 56 of 1974, measures of cognitive ability are considered psychological assessments. Therefore, only trained psychology professionals, who are registered with the Health Professions Council of South Africa may gain access to use the Verbatim and Numeratum in South Africa. No certification training is required for registered professionals in South Africa to use these questionnaires.

Appropriate Use

The Verbatim and Numeratum can be used for screening, competency-based selection, and training, and are recommended for entry-level or supervisory jobs that require Grade 12 (NQF level 4) English and/or Mathematical proficiency. It is critical to note that the Verbatim and Numeratum are NOT designed to assess cognitive deficits or learning disabilities and should under no circumstances be used in any shape or form as a diagnostic tool. Ideally, test-takers should not have completed the Verbatim and/or Numeratum within the previous six months. This requirement is designed to minimise practice effects should anyone retake the Verbatim and/or the Numeratum.

CHAPTER 2: THEORETICAL FOUNDATIONS

The Evolvement of the Verbatim and Numeratum

The Verbatim and Numeratum were developed due to the extensive demand for assessments of specific abilities and skills that are not necessarily measured by comprehensive mental ability assessments. JVR developed them in 2012 and released their research versions in 2013. In 2015, data on both assessments were analysed and norms were generated. In addition, research items were included to allow for future updates to the two assessments. The research items were not included in the calculation of the overall score. In 2023, the feasibility of shortening both questionnaires were examined. Findings suggested that by using a mixture of the original and research items, shortened versions of the Verbatim and Numeratum with respectable psychometric properties could be developed. Chapter 3 elaborates on the preceding process.

Theoretical Underpinning

"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).

Apart from a few additional references listed in the text, the following discussion is primarily based on the work of Jensen (1998) and Carroll (1993).

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 believed 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[1] 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 factor analysis, Spearman (1904) 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.

Elaborating on the prevailing view of general intelligence, Cattell (1963) introduced the concepts of fluid (G*f*) and crystallised (G*c*) intelligence, both of which were considered subfactors 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

[1] This was the terminology they used during those times. Since then, due to the negative connotation attached to the term mental retardation, the latter has been replaced with intellectual disability.

example, when solving a mathematical problem, fluid intelligence will assist in selecting a strategy to find the solution while crystallised intelligence might assist in recalling an appropriate formula (Postlethwaite, 2011).

The Verbatim and Numeratum were developed to measure fairly basic levels of verbal and numerical reasoning. Both assessments contain several subtests that could be said to measure aspects of *Gf* and *Gc*. For example, on the Numeratum, the Number Problems 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 items related to reasoning 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 should be noted that the intent of the Verbatim and Numeratum is not to exclusively measure *g*, but rather to measure specific subcomponents of the larger *g* construct. This is an important distinction to make due to the ongoing debate regarding the amount of variance specific abilities traditionally explain beyond *g* (e.g., Eid et al., 2018; Kell & Lang, 2017; Ree & Carretta, 2022).

Assessment Scales and Sections

VERBATIM

The Verbatim consists of 28 scored questions divided into five sections. The first four sections require that the respondent selects the correct response from a multiple-choice format. The final section requires that the candidate selects between 'true', 'false', or 'cannot say'. A short description of each section is listed below.

Synonyms (5 Items): Respondents are asked to identify words that are the same or similar in meaning.
Opposites (6 Items): Respondents are asked to identify words with opposite meanings.
Analogies (6 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 (6 Items): Respondents are required to identify an individual's ability to reason with letters and other verbal content.
Interpretation (5 Items[2]): Respondents are tested on their ability to read and accurately comprehend verbal content.

[2] The Verbatim's Interpretation section includes 8 items, 5 scorable items and 3 research items.

NUMERATUM

The Numeratum consists of 16 scored questions divided into three sections. The candidate is required to select the correct response from a multiple-choice format. A short description of each section is listed below.

Number Problems (5 Items): Respondents are asked to complete mathematical problems that are composed of mostly simple addition, subtraction, multiplication, and division.

Patterns (6 Items): Respondents are asked to identify patterns in numerical content.

Interpretation (5 Items[3]): Respondents are requested to identify, read, and interpret basic numerical information.

[3] The Numeratum's Interpretation section includes 8 items, 5 scorable items and 3 research items.

CHAPTER 3: REVISING THE VERBATIM AND NUMERATUM

Data analytic approach, data screening, and data cleaning

During the data cleaning process, several duplicates were flagged. The first entry was kept if the participant completed the assessment more than once within a timeframe shorter than six months. When the same participant completed the assessment with a time lag of at least six months, their last entry was used[4]. Additionally, participants aged below 17 or above 69 years were also removed. Only participants who completed all the Verbatim (N = 5188) and Numeratum's (N = 2627) original and research items respectively from September 2015 to August 2023 were retained. Using the *sample.split* function from the *caTools* (Tuszynski, 2021) package, samples were randomly split in half according to gender[5], allowing the demographic representation in the training and test samples to be as close as possible. Regarding the aforementioned, "Machine learning models are often fit on one data set (the 'training set'), and predictions are made and evaluated using a new data set (the 'training set')" (Rosenbusch et al., 2021, p. 2). The rationale behind this is to avoid double dipping. "Double dipping is a term for overfitting a model through both building and evaluating the model on the same data-set, yielding inappropriately high statistical significance and circular logic" (Ball et al., 2020, p. 261). Consequently, the initial analyses were conducted on the training samples to identify potentially problematic items. An evaluation of the psychometric properties of the potentially problematic items and the ant colony optimisation (ACO) approach, all discussed in detail below, was used to guide the selection of items for shortened versions of the Verbatim and Numeratum. To avoid double dipping, as alluded to earlier, the test samples were used to assess the psychometric properties of these shortened versions of the Verbatim and Numeratum. Table 1 reports the sociodemographic composition of the training and test samples.

[4] Unfortunately, the number of participants who completed the assessments more than once was not sufficient to calculate test-retest reliability.
[5] Gender was the only demographic variable with no missing data.

Table 1. *Sociodemographic Composition of the Samples*

Variable	Training Sample (*N* = 2594)		Test Sample (*N* = 2594)	
	Verbatim			
Gender	*n*	%	*n*	%
Women	1392	53.7%	1392	53.7%
Men	1202	46.3%	1202	46.3%
Ethnicity	*n*	%	*n*	%
Black African	1252	62.6%	1220	59.9%
White	327	16.4%	363	17.8%
Coloured	141	7.0%	160	7.9%
Indian/Asian	196	9.8%	217	10.7%
Other	83	4.2%	76	3.7%
Language	*n*	%	*n*	%
English	906	35.2%	916	35.6%
Zulu	315	12.2%	310	12.1%
Afrikaans	366	14.2%	393	15.3%
Xhosa	173	6.7%	201	7.8%
Sotho	159	6.2%	123	4.8%
Venda	85	3.3%	90	3.5%
Pedi	214	8.3%	209	8.1%
Tsonga	100	3.9%	101	3.9%
Tswana	183	7.1%	163	6.3%
Ndebele	12	0.5%	23	0.9%
Swati/Swazi	52	2.0%	37	1.4%
Other	9	0.3%	5	0.2%
Education	*n*	%	*n*	%
Grade 12	288	12.4%	297	12.7%
Diploma/certificate	247	10.6%	244	10.4%
Bachelor's degree	329	14.2%	317	13.5%
Honours degree	226	9.7%	255	10.9%
Mismatch [a]	1146	49.4%	1148	48.9%
Other [b]	83	3.6%	85	3.6%

Numeratum				
Variable	Training Sample (*N* = 1313)		Test Sample (*N* = 1314)	
Gender	*n*	%	*n*	%
Women	639	48.6%	639	48.6%
Men	674	51.4%	675	51.4%
Ethnicity	*n*	%	*n*	%
Black African	613	59.7%	588	58.4%
White	199	19.4%	188	18.7%
Coloured	75	7.3%	66	6.6%
Indian/Asian	104	10.1%	125	12.4%
Other	34	3.3%	39	3.9%
Language	*n*	%	*n*	%
English	513	39.2%	501	38.3%
Zulu	136	10.4%	142	10.8%
Afrikaans	198	15.1%	224	17.1%
Xhosa	89	6.8%	74	5.7%
Sotho	61	4.7%	64	4.9%
Venda	52	4.0%	44	3.4%
Pedi	99	7.6%	100	7.6%
Tsonga	57	4.4%	45	3.4%
Tswana	80	6.1%	79	6.0%
Ndebele	7	0.5%	7	0.5%
Swati/Swazi	16	1.2%	25	1.9%
Other	1	0.1%	3	0.2%
Education	*n*	%	*n*	%
Grade 12	177	14.8%	166	14.0%
Diploma/certificate	139	11.6%	135	11.4%
Bachelor's degree	164	13.7%	156	13.1%
Honours degree	123	10.1%	117	9.8%
Mismatch [a]	559	46.6%	583	49.0%
Other [b]	36	3.0%	31	2.6%

Note. Missing data are not reported and were not considered in the calculation of percentages. [a] Various participants indicated that their highest educational level was below Grade 8. This however did not coincide with the qualifications they listed and the scores they obtained. Due to the potential threat these cases posed

to the reliability of the data, they were not used for purposes of mean difference testing, measurement invariance testing, and differential item functioning when comparing different educational groups.

[b] The highest level of education of those grouped in the other educational category included the following: below Grade 12, Master's degree, Doctoral degree. These cases were also not included for purposes of mean difference testing, measurement invariance testing, or differential item functioning due to the low number of participants in each category.

In the following paragraphs, the items identified as potentially problematic, together with the ACO approach, are discussed shortly to provide some context as to how the Verbatim and Numeratum were shortened.

A short summary of potentially problematic Verbatim and Numeratum items

Table 2 pinpoints the items that performed the least satisfactorily in different psychometric categories. These categories include (a) the easiest items regarding CTT difficulty statistics and Rasch item locations, (b) the items that were least capable of discriminating between low and high scorers, (c) items displaying the most problematic underfit, (d) items with low general factor loadings, (e) items with no significant factor loadings on any factor, (f) the highest correlations between item pairs, (g) low item-rest correlations, and (h) items that displayed moderate to large DIF across gender, and specific ethnic and language groups.

Table 2. *Potentially Problematic Verbatim and Numeratum Items*

Scale	Item difficulty	Item disc.	Underfit	Gen. Factor	No sig. loading	Correlation	Item-rest cor.	DIF
Verbatim Original	S1, S2, S3, O1, O2, O3, A1, A2, R4	S1, S2, O2, O3, O7	O7	S6, S7, S8, O7, O8, A3, R1, VI1, VI3, VI5, VI7, VI8, VI9, VI10	S6, S7, S8, O7, O8, A3, R1, VI5, VI8	O2-O3	S1, S6, S7, S8, O7, O8, A3, VI1, VI3, VI5, VI7, VI8, VI9, VI10	S2, S4, S6, O3, O5, A1, A4
Verbatim Original + Research	A11	VI11	VI11, VI12	S9, S12, R10, R11, VI11, VI12, VI15, VI16	S12, R10, R11, VI12, VI15, VI16	-	S9, S12, A11, R10, R11, VI11, VI12, VI15, VI16	S9, A9
Numeratum Original	NP1, NP2, NP3, P1, NI1, NI3, NI6	NP3, NI3	NP5, NI2, NI3	NP5	NP5	NP8-NP9, NP9-NP10	NP5, NI3	NP3, NI1, NI5, NI6
Numeratum Original + Research	-	-	NI12, NI15, NI19	-	-	NP8-NP10, NI11-NI13	NI19	-

Note. Items mentioned in Original were not repeated in Original and Research. In the 'Scale' column, Original refers to the original items of the Verbatim and Numeratum, whereas Original and Research refer to the original and research items of the Verbatim and Numeratum, respectively.

As per Table 2, numerous items served as potential candidates to be discarded due to their undesirable performance on various metrics. In conjunction with the findings of Table 2 and other metrics (e.g., cross-loadings that were discovered), the Ant Colony Optimisation (ACO) approach was used as a guiding framework to shorten the Verbatim and Numeratum.

The ACO Approach

The ACO approach is one of many ant-based optimisation algorithms in existence (Dorigo & Stützle, 2010). This approach is "a metaheuristic optimization procedure that is capable of solving complex combinatorial problems in an efficient way" (Olaru & Danner, 2021, p. 200). From a psychological assessment perspective, one of these combinatorial optimisation problems is selecting items for a short scale (Olaru & Danner, 2021). To aid item selection for short scales, ACO may serve as a tool to select and evaluate item combinations to find the best results through optimisation criteria defined by the researcher (Olaru & Jankowsky, 2022).

To facilitate the process, a function was written in *R* version 4.3.0 (R Core Team, 2023), enabling the researcher to specifically optimise the following criteria: (a) mean slope parameter (discrimination parameter), (b) residuals, (c) mean RMSEA fit, (d) Cronbach's alpha coefficient, (e) maximum test information at the desired theta, and (f) the omega explained common variance. These metrics were considered important to provide a balanced representation of the constructs. The following *R* packages were used to achieve the preceding objective: *Cronbach* (Tsagris & Frangos, 2020), *fungible* (Waller, 2023), *mirt* (Chalmers, 2012), and *psych* (Revelle, 2023). As the ACO algorithm produces different combinatorial solutions, the *R* script was run 10 times on the Verbatim and Numeratum items, with the aim of selecting six items per subscale for the full scales respectively (e.g., Verbatim: 5 subscales, 5*6 = 30; Numeratum: 3 subscales, 3*6 = 18). Table 3 and Table 4 provide the combination of items per ACO run.

Table 3. *ACO Item Combinations for the Verbatim*

Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10
S1	S1	S1	S2	S1	S2	S3	S1	S3	S1
S4	S2	S2	S3	S2	S5	S5	S3	S5	S3
S5	S3	S3	S6	S5	S7	S7	S4	S7	S5
S6	S4	S4	S7	S6	S8	S9	S7	S9	S9
S10	S5	S8	S8	S8	S10	S10	S8	S10	S10
S11	S11	S11	S9	S11	S11	S11	S10	S11	S11
O2	O1	O2	O4	O2	O1	O1	O4	O2	O3
O4	O4	O4	O5	O4	O6	O4	O6	O4	O5
O6	O5	O6	O6	O6	O7	O6	O8	O5	O7
O9	O6	O10	O8	O9	O9	O9	O9	O6	O10
O11	O8	O11	O10	O10	O11	O10	O10	O9	O11
O12	O10	O12	O12	O12	O12	O12	O12	O12	O12
A3	A5	A1	A1	A1	A2	A3	A1	A4	A1
A6	A6	A2	A4	A4	A4	A6	A3	A6	A3
A7	A7	A6	A5	A5	A6	A7	A6	A7	A4
A8	A8	A7	A7	A6	A7	A9	A7	A9	A5
A9	A9	A8	A9	A7	A8	A10	A9	A10	A8
A10	A10	A9	A10	A10	A9	A11	A10	A11	A9
R3	R2	R5	R2	R3	R2	R3	R3	R1	R2
R4	R4	R6	R3	R5	R3	R5	R5	R2	R4
R5	R5	R7	R4	R6	R4	R6	R6	R3	R5
R6	R8	R8	R5	R7	R6	R7	R7	R7	R6
R7	R9	R10	R9	R9	R7	R9	R9	R9	R7
R8	R10	R11	R11	R11	R8	R10	R10	R11	R11
VI1	VI2	VI1	VI1	VI2	VI2	VI2	VI5	VI6	VI2
VI2	VI4	VI2	VI2	VI4	VI3	VI4	VI6	VI7	VI4
VI4	VI5	VI3	VI5	VI6	VI6	VI6	VI7	VI8	VI5
VI9	VI10	VI4	VI8	VI9	VI7	VI9	VI9	VI9	VI10
VI10	VI11	VI6	VI14	VI12	VI10	VI11	VI13	VI13	VI13
VI15	VI13	VI9	VI15	VI13	VI14	VI12	VI16	VI16	VI14

As per Table 3, the following items were the most consistently chosen for the Verbatim: Synonyms (S1, S2, S3, S5, S7, S8, S10, and S11), Opposites (O2, O4, O5, O6, O9, O10, O11, and O12), Analogies (A1, A4, A6, A7, A8, A9, and A10), Reasoning (R2, R3, R4, R5, R6, R7, R9, and R11), and Interpretation (VI2, VI4, VI5, VI6, VI9, VI10, and VI13).

Table 4. *ACO Item Combinations for the Numeratum*

Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10
NP1	NP5	NP1	NP3	NP3	NP2	NP5	NP2	NP1	NP2
NP3	NP6	NP4	NP4	NP5	NP4	NP6	NP3	NP3	NP6
NP4	NP7	NP6	NP5	NP7	NP6	NP9	NP4	NP6	NP7
NP6	NP8	NP10	NP7	NP9	NP7	NP11	NP6	NP7	NP10
NP12	NP11	NP11	NP8	NP11	NP12	NP12	NP7	NP10	NP11
NP13	NP13	NP13	NP11	NP13	NP13	NP13	NP10	NP11	NP12
P1	P1	P4	P4	P3	P5	P1	P4	P1	P2
P3	P2	P5	P5	P5	P6	P5	P5	P6	P5
P7	P4	P7	P7	P7	P7	P7	P6	P7	P6
P9	P5	P8	P9	P8	P8	P8	P7	P9	P7
P10	P9	P10	P10	P9	P9	P9	P9	P10	P8
P11	P11	P11	P11	P10	P11	P11	P11	P11	P10
NI2	NI4	NI8	NI2	NI5	NI2	NI4	NI5	NI5	NI8
NI5	NI9	NI12	NI8	NI8	NI8	NI5	NI6	NI9	NI9
NI8	NI10	NI14	NI9	NI9	NI9	NI7	NI7	NI12	NI13
NI14	NI14	NI15	NI10	NI10	NI10	NI9	NI8	NI13	NI16
NI17	NI16	NI17	NI15	NI12	NI15	NI12	NI16	NI14	NI17
NI18	NI17	NI18	NI18	NI18	NI18	NI18	NI18	NI18	NI18

As per Table 4, the following items were the most consistently chosen for the Numeratum: Number Problems (NP3, NP4, NP6, NP7, NP11, and NP13), Patterns (P5, P7, P8, P9, P10, and P11), and Interpretation (NI5, NI8, NI9, NI10, NI12, NI14, NI17, and NI18).

Although the previous paragraphs outlined the most consistently chosen items, it did not mean that these items formed the best combinations by default. This was something that required further judgment. For example, as DIF was not part of the optimisation criteria, Table 2 was consulted to see if any of the aforementioned items were flagged for DIF. Items S2, O5, A1, A4, and A9 were flagged for DIF

on the Verbatim. Only item A9 was kept for further consideration as its DIF effect size was fairly close to negligible and its performance regarding other psychometric criteria was satisfactory. Additionally, items NP3 and NI5 were flagged for DIF on the Numeratum. Item NI5 was kept as its DIF effect size was fairly close to negligible. Furthermore, low general factor loadings, no significant factor loading on any factor, or low item-rest correlations as seen in Table 2 suggest that items S7, S8, R11, VI5, and VI10 of the Verbatim may be candidates for deletion and that certain subscales may be reduced to five items if suitable replacements cannot be found. In instances where the items featured the same amount of time per subscale and more than six probable items could be selected per subscale based on the results of the ACO iterations (see Table 3 and Table 4), decisions had to be made with regard to which item fit the best within a specific combination. For example, in the Opposites subscale of the Verbatim, items O2, O5, and O11 surfaced the same number of times. As O5 was discarded earlier due to DIF, O2 and O11 remained. As O2 cross-loaded on the Analogies subscale, O11 seemed a better item to select. Similar approaches were followed for the other subscales. The continuous examination of the ACO combinations (see Table 3 and Table 4) and important metrics not covered by the algorithm enabled the researcher to select either five or six items per subscale to sufficiently represent the shortened versions of the Verbatim and Numeratum. The following items were selected for the Verbatim: Synonyms (S1, S3, S5, S10, and S11), Opposites (O4, O6, O9, O10, O11, and O12), Analogies (A5, A6, A7, A8, A9, and A10), Reasoning (R3, R5, R6, R7, R8, and R9), and Interpretation (VI2, VI4, VI6, VI13, and VI14). The following items were selected for the Numeratum: Number Problems (NP4, NP6, NP7, NP12, and NP13), Patterns (P5, P7, P8, P9, P10, and P11), and Interpretation (NI5, NI8, NI9, NI10, and NI17). Table 5 compares the original, original and research, and the updated short forms of the questionnaires on which this manual is based.

Table 5. *Comparison of Items Across the Different Format Lengths of the Verbatim and Numeratum*

Verbatim	# Items (Original)	# Items (Orig. and Research)	# Items (Short)
Synonyms	8	12	5
Opposites	8	12	6
Analogies	8	11	6
Reasoning	8	11	6
Interpretation	10	16	5
Total	42	62	28 [a]
Numeratum	**# Items (Original)**	**# Items (Orig. and Research)**	**# Items (Short)**
Number problems	10	14	5
Patterns	8	11	6
Interpretation	10	19	5
Total	28	44	16 [a]

Note. [a] Refers to the number of scored items.

CHAPTER 4: ADMINISTRATION

The Verbatim and Numeratum can be administered online. Each section for each test has practice examples which must be successfully completed before the scored items can be attempted. Both assessments are timed (the time limits can be found **here**). This is done in such a way that most test-takers will be able to answer all the questions in the allocated time. Information regarding completion times and accuracy can also be found in the feedback report.

Online administration

It is recommended that the Verbatim and Numeratum are conducted in a proctored setting to ensure that online dictionaries, calculators, or any other programmes (e.g., Excel, Word) are not used, as the latter may provide an unfair advantage to the candidate and provide a skewed representation of their performance on the test/s. Further items like cameras, phones, or any recording devices are also prohibited to prevent the copying and distribution of materials by unauthorised persons and to preserve the integrity of the test. We recommend that candidates are clearly informed that such activity might lead to their test results being deemed invalid.

The Verbatim and Numeratum are available on the OneJVR platform and accessibility to these assessments is managed through JVR's Client Services. OneJVR is an online administration platform that was developed to host local and self-published assessments as well as several international assessments. Individual users can set up their workspaces by completing the following form (https://tinyurl.com/56ceayfs). For more information about OneJVR or workspace-related queries, please contact Client Services (clientservices@jvrafrica.co.za).

CHAPTER 5: INTERPRETATION AND FEEDBACK

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. Normative scores (stens, stanines, and percentiles) are provided for the overall Verbatim or Numeratum score. Although the number of attempted/correct items are indicated at subscale level, interpretation and decision-making should not be conducted at this level. The analyses discussed in Chapter 6 (Verbatim) and Chapter 7 (Numeratum) add support for this.

In a nutshell, the feedback report for the Verbatim provides the following information:

- A brief introduction about the assessment and what it gives an indication of.
- The candidate's normative scores (sten, stanine, percentile).
- How the candidate's score compared to the norm group and what that means.
- The candidate's decision-making process, including the speed and accuracy in which items were answered.
- Items response insights, which include the number of attempted items and the number of correct answers, accompanied by the time taken to complete the assessment.

A sample feedback report is provided below.

Verbatim Interpretation Guidelines

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 synonyms, antonyms, and analogies, as well as reason and interpret verbal information.

VERBAL REASONING

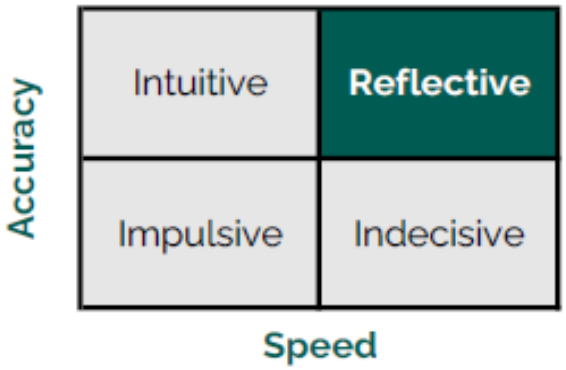
Stanine

Below Average *Mid Range* *Above Average*

Sten 8 **Percentile** 88

Compared to others in the norm group, Test Sample found it easy to work with verbal information. Given that they have displayed a high level of verbal reasoning, they would likely work well in a position that requires processing and decision-making of written/verbal English content.

VERBAL INFORMATION DECISION-MAKING APPROACH



- Takes time to consider all available verbal information
- Evaluates all verbal information before making decisions

RESPONSE INSIGHTS

	Attempted	Correct
Synonyms	5	5
Opposites	6	6
Analogies	6	3
Reasoning	6	5
Interpretation	8	6

Time taken to complete Verbatim: 00:16:27

In a nutshell, the feedback report for the Numeratum provides the following information:

- A brief introduction about the assessment and what it gives an indication of.
- The candidate's normative scores (sten, stanine, percentile).
- How the candidate's score compared to the norm group and what that means.
- The candidate's decision-making process, including the speed and accuracy in which items were answered.
- Items response insights, which include the number of attempted items and the number of correct answers, accompanied by the time taken to complete the assessment.

A sample feedback report is provided below.

Numeratum Interpretation Guidelines

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 through number problems, number patterns, and interpret graphical and tabulated numerical information.

NUMERICAL REASONING

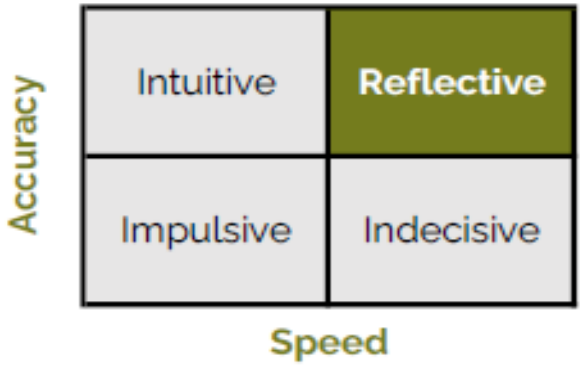
Stanine

Below Average *Mid Range* *Above Average*

Sten 9 **Percentile** 85

Test Sample found some elements of working with numerical information easier than others, similar to others in the norm group. Given their score is within the midrange, they would likely function well in positions that have a moderate amount of numerical processing activities.

NUMERICAL INFORMATION DECISION-MAKING APPROACH



- Takes time to consider all available numerical information
- Evaluates all numerical information before making decisions

RESPONSE INSIGHTS

	Attempted	Correct
Number Problems	5	5
Patterns	6	6
Interpretation	8	4

Time taken to complete Numeratum: 00:28:19

CHAPTER 6: PSYCHOMETRIC PROPERTIES-VERBATIM

In this section, the current, shortened Verbatim's psychometric properties are discussed in the following order: (a) testing assumptions of normality and the presence of outliers, (b) descriptive statistics, (c) correlation coefficients, (d) reliability coefficients, (e) Rasch analysis, (f) construct validity, (g) item difficulty and discrimination, (h) differential item functioning, (i) measurement invariance, and (j) mean differences across groups. The reader is reminded that these analyses were carried out on the test sample (see **Table 1** for the sample composition and **Chapter 3** for an explanation on why the samples were split).

Testing assumptions of normality and the presence of outliers

Checking for normality or other assumptions and outliers is essential for ensuring the reliability and validity of statistical analyses, making informed decisions, and understanding the characteristics of the data under investigation. It helps to make appropriate choices in selecting statistical methods and interpreting results. Generally, specific assumptions accompany statistical tests (e.g., normality, homogeneity of variance), and when these assumptions are met, the use of the parametric version of the test is preferable (Erceg-Hurn & Mirosevich, 2008). However, if some or all these assumptions are violated, alternative statistical approaches or nonparametric tests may be more appropriate (Hoekstra et al., 2012). Apart from assumption violations, outliers (i.e., data points that differ significantly from others in a dataset) may distort statistical findings (Osborne & Overbay, 2019). Outliers may be indicative of errors in data collection or measurement, or they might represent genuine extreme values (Osborne & Overbay, 2019). Hence, it is important to assess them before decisions are made on how they should be dealt with.

A one-sided Grubbs test was conducted on the highest and lowest total Verbatim score values to determine if they were statistically significant outliers (Grubbs, 1950). Using the *grubbs.test* function

from the *outliers* (Komsta, 2022) package, neither the highest (G = 1.82, U = 0.9987, p = 1) nor the lowest value (G = 2.96, U = 0.9966, p = 1) was statistically significant. Multivariate outliers across all the scales were subsequently investigated by plotting robust Mahalanobis distances against the quantiles of the χ2 distribution (Garrett, 1989). Minimal multivariate outliers were detected. Hence, for the most part, the data points were not significantly different from the rest. Results (p < 0.001) from formal statistical univariate normality tests (Shapiro-Wilk, Anderson-Darling, and Lilliefors[6]) as obtained through the *mvn* function in the *MVN* (Korkmaz et al., 2014) package showed deviations from normality. Multivariate normality was investigated using Mardia's coefficient (Mardia, 1970). The results indicated that the Verbatim scales deviated from multivariate normality. This implies that the joint distribution of multiple variables did not follow a multivariate normal distribution. It should be noted that in large samples, as in the current sample, violations of normality may be less of a concern compared to smaller samples due to the Central Limit Theorem (Gao et al., 2017). Therefore, depending on the statistical test, other assumptions (e.g., homogeneity of variance) largely dictated whether parametric tests, tests that require less restrictive assumptions, or if nonparametric tests were used in subsequent analyses.

[6] Similar to the Kolmogorov-Smirnov test.

Descriptive statistics

Table 6 provides descriptive statistics for each of the Verbatim scales and the total Verbatim score. These were calculated with the *describe* function from the *psych* (Revelle, 2023) package.

Table 6. *Descriptive Statistics for the Verbatim Scales and Total Verbatim Score*

Scale	M	SD	Med	Trim	Mad	Min	Max	Skew	Kurt	SE
Synonyms	3.94	1.07	4	4.09	1.48	0	5	-0.86	0.09	0.02
Opposites	4.57	1.46	5	4.78	1.48	0	6	-1.01	0.37	0.03
Analogies	3.23	1.62	3	3.23	1.48	0	6	-0.05	-0.92	0.03
Reasoning	3.32	1.52	3	3.34	1.48	0	6	-0.10	-0.77	0.03
Interpretation	3.05	1.45	3	3.13	1.48	0	5	-0.38	-0.76	0.03
Total	18.11	5.44	19	18.41	5.93	2	28	-0.41	-0.64	0.11

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

As per Table 6, the mean total Verbatim score was 18.11 (median = 19, SD = 5.44). Regarding univariate normality, the skewness and kurtosis values fell within acceptable ranges (-2 to 2; Koh, 2014). This suggests that each variable's distribution was reasonably symmetric and that the tails of the distribution were not excessively heavy or light compared to a normal distribution. The standard error values were all generally low. Low standard errors are normally desirable as it suggests that the sample statistic (e.g., the sample mean) is likely to be a more accurate reflection of the population parameter (Harding et al., 2014).

Correlation coefficients

Inspection of multivariate normality using Mardia's coefficient (Mardia, 1970) found that bivariate normality was violated across most variables. Although Pearson correlation coefficients do not necessitate bivariate normality, Spearman's rank correlation coefficients were also calculated as a nonparametric alternative. Table 7 provides the Pearson correlation coefficients and Spearman's rank correlation coefficients for the five Verbatim scales. These were calculated with the *rcorr* function in the *Hmisc* (Harrell, 2023) package. The correlations predominantly had medium to large effect sizes (Cohen, 1988). This confirmed the relatedness, yet uniqueness of the scales, which is to be expected as they all measure aspects of verbal ability. Inter-factor correlations are provided later.

Table 7. *Pearson and Spearman's Rank Correlations for the Verbatim Scales*

Scale	S	O	A	R	VI
Synonyms	-	0.49*	0.45*	0.38*	0.38*
Opposites	0.51*	-	0.52*	0.43*	0.46*
Analogies	0.46*	0.51*	-	0.54*	0.49*
Reasoning	0.40*	0.45*	0.54*	-	0.47*
Interpretation	0.40*	0.47*	0.49*	0.47*	-

Note. Pearson correlations are below the diagonal, Spearman's rank correlations are above the diagonal. Values of 0.10, 0.30, and 0.50 correspond to small-, medium-, and large effects. *$p < 0.001$. S = Synonyms, O = Opposites, A = Analogies, R = Reasoning, VI = Interpretation.

Reliability

Cronbach's alpha coefficient α; (Cronbach, 1951) is arguably the most commonly used measure of reliability in psychological science (Hayes & Coutts, 2020). One of its major criticisms however revolves around the assumption of tau-equivalence (e.g., all items in the scale have equal factor loadings, all test items have the same true score), as data seldom adhere to this assumption (Teo & Fan, 2013). Consequently, in the absence of tau-equivalence, Cronbach's alpha may underestimate true reliability (Teo & Fan, 2013). Therefore, many suggest the use of McDonald's omega ω; (McDonald, 1999) as it is less reliant on the tau-equivalence assumption. To offer a more comprehensive view of the measurement properties of the scales, both the aforementioned reliability coefficients were analysed in addition to Rasch reliability coefficients. Table 8 provides the reliability coefficients for the Verbatim scales and total Verbatim score. These were calculated with the *ci.reliability* function in the *MBESS* (Kelley, 2022) package. The Rasch reliability coefficients were calculated in Winsteps. Model-based reliability coefficients are provided later.

Table 8. *Reliability Coefficients for the Verbatim Scales and Total Verbatim Score*

Scale	α	ω	PR	IR
Synonyms	0.50	0.53	-	-
Opposites	0.61	0.61	-	-
Analogies	0.58	0.59	-	-
Reasoning	0.54	0.54	-	-
Interpretation	0.55	0.56	-	-
Total	0.84	0.84	0.81	1.00

Note. α = Cronbach's Alpha Coefficient, ω = Coefficient Omega, PR = Person Reliability Index (Rasch), Item Reliability Index (Rasch).

As per Table 8, the reliability coefficients for the Verbatim subscales were mostly unsatisfactory, with coefficients (α and ω) ranging from 0.50 to 0.61. The reliability of the subscales is however less concerning as the total score is meant to be interpreted (see the last paragraph of the Construct Validity section). The reliability coefficients for the total Verbatim score were deemed acceptable according to conventional guidelines (> 0.70; Nunnally, 1978). The item separation index values indicated that the item locations were generally stable. The person separation index (PSI) values for the Verbatim subscales indicated that the subscales may not be sensitive enough to distinguish between low and high scorers. For the total Verbatim score, the PSI value (2.04) was slightly higher than the generally preferable score of 2 (Combrinck, 2020), implying that the total Verbatim scale is sensitive enough to distinguish between low and high scorers.

Additionally, the reliability coefficients for different gender, ethnic, language, and educational groups were examined. Table 9 reports these results.

Table 9. *Reliability Coefficients for Different Gender, Ethnic, Language, and Educational Groups*

Gender							
Female				Male			
α		ω		α		ω	
0.84		0.84		0.84		0.84	
Ethnicity							
Black African		White		Coloured		Indian	
α	ω	α	ω	α	ω	α	ω
0.84	0.84	0.80	0.80	0.83	0.83	0.82	0.82
Language							
English		Zulu		Afrikaans		Xhosa	
α	ω	α	ω	α	ω	α	ω
0.82	0.82	0.83	0.84	0.82	0.82	0.84	0.84
Education							
Grade 12		Diploma		Bachelor's		Honours	
α	ω	α	ω	α	ω	α	ω
0.84	0.84	0.81	0.81	0.78	0.79	0.79	0.80

Note. α = Cronbach's Alpha Coefficient, ω = Coefficient Omega.

As per Table 9, the reliability coefficients appeared fairly consistent within and across the different groups. All reliability coefficients were deemed acceptable according to conventional guidelines (> 0.70; Nunnally, 1978).

Furthermore, Haberman's (2008) subscale scoring test based on the proportional reduction in mean squared error (PRMSE) was used to investigate whether interpretation should be conducted at the scale score level or total Verbatim score level. Meijer et al. (2017) found that "subscores provided added value over the total score if and only if $PRMSE_s$[7] is larger than $PRMSE_x$" (p. 3). When using the *prmse.subscores.scales* function in the *sirt* (Robitzsch, 2022) package, the symbol X denotes the subscale and Z the full scale. Table 10 reports these values.

[7] $PRMSE_s$ refer to the subtest score.

Table 10. *Haberman's Subscale Scoring Test Results*

Scale	$PRMSE_X$	$PRMSE_Z$
Synonyms	0.50	0.71
Opposites	0.61	0.76
Analogies	0.58	0.82
Reasoning	0.54	0.78
Interpretation	0.55	0.74

Note. PRMSE = Proportional reduction of mean squared error. Meijer et al. (2017) refer to the sub scores/subscales as $PRMSE_s$ and the total score/full scale as $PRMSE_X$. The sirt R package refer to the subscales as $PRMSE_X$ and the total score as $PRMSE_Z$.

As per Table 10, none of the $PRMSE_X$ values exceeded the $PRMSE_Z$ values, implying that the Verbatim's total score should rather be interpreted than its scale scores.

Rasch Analysis

A Rasch (1960) analysis was conducted on the total Verbatim score to inspect item fit statistics and item locations (difficulties) in Winsteps version 4.6.1 (Linacre, 2020a). Depending on the circumstances, different Infit (IMNSQ) and Outfit (OMNSQ) mean square values may signal underfitting or overfitting items (Aryadoust et al., 2020). 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). As criteria to assess item fit, items with mean square (infit/outfit) values ≥ 1.40 were indicative of potential underfit, whereas items with mean square (infit/outfit) values ≤ 0.60 signalled potential overfit (Bond & Fox, 2015). However, as overfit is typically deemed less worrisome than underfit (Tesio et al., 2023), greater focus was placed on mean square (infit/outfit) values > 1.00. Consequently, as additional criteria, OMNSQ values ≥ 1.30 were inspected first, followed by an inspection of IMNSQ values ≥ 1.10 to identify misfitting items. Table 11 provides the item fit statistics and item locations for the total Verbatim score. Item and person reliabilities were provided earlier (see Table 8).

Table 11. *Total Verbatim Score Item Location and Item Fit Statistics*

Item	Location	SE	IMNSQ	Z	OMNSQ	Z	PT Corr.	Exp.
S1	-3.20	.12	0.96	-0.4	0.79	-1.15	0.21	0.18
S3	-2.06	.08	0.97	-0.62	0.90	-0.82	0.31	0.28
S5	0.05	.05	1.00	-0.05	0.98	-0.45	0.45	0.45
S10	-1.48	.06	0.88	-3.18	**0.60**	-5.23	0.44	0.33
S11	0.84	.04	1.01	0.68	1.01	0.5	0.47	0.47
O4	-1.16	.06	1.01	0.32	1.37	4.47	0.33	0.36
O6	0.31	.05	1.01	0.63	1.04	1.25	0.45	0.46
O9	-0.65	.05	1.02	0.71	1.07	1.28	0.39	0.41
O10	-0.79	.05	0.87	-4.60	0.77	-4.07	0.49	0.39
O11	-1.09	.06	0.86	-4.24	0.77	-3.51	0.47	0.37
O12	-0.24	.05	0.98	-0.92	0.91	-2.11	0.46	0.43
A5	-0.53	.05	1.02	0.70	1.06	1.18	0.40	0.41
A6	0.43	.05	0.98	-0.93	0.99	-0.39	0.48	0.47
A7	0.92	.04	0.93	-4.21	0.92	-2.87	0.53	0.47
A8	2.12	.05	1.08	3.26	**1.43**	8.37	0.37	0.45
A9	1.23	.05	0.86	-7.83	0.83	-6.1	0.57	0.47
A10	0.07	.05	1.09	4.19	1.05	1.28	0.40	0.45
R3	0.01	.05	1.00	-0.20	1.01	0.18	0.45	0.45
R5	0.87	.04	1.03	1.60	1.06	1.99	0.45	0.47
R6	0.27	.05	1.01	0.52	0.97	-0.85	0.46	0.46
R7	2.31	.05	0.99	-0.33	1.02	0.45	0.44	0.44
R8	-1.21	.06	0.99	-0.31	0.97	-0.43	0.36	0.36
R9	1.31	.05	1.13	6.81	1.24	7.34	0.37	0.47
VI2	0.10	.05	1.05	2.41	1.15	3.93	0.41	0.45
VI4	0.45	.05	0.97	-1.72	0.96	-1.41	0.49	0.47
VI6	0.67	.04	1.08	4.45	1.15	4.89	0.41	0.47
VI13	-0.11	.05	0.97	-1.33	0.94	-1.46	0.46	0.44
VI14	0.56	.05	1.05	2.46	1.06	1.87	0.44	0.47

Note. OMNSQ ≥ 1.40 or ≤ 0.60 in bold. Location = Item location, SE = Standard Error, IMNSQ = Infit Mean Square Values, Z = z-standardised statistics, OMNSQ = Outfit Mean Square Values, PT Corr. = Point-Measure Correlation, Exp. = Expected value. S = Synonyms, O = Opposites, A = Analogies, R = Reasoning, VI = Interpretation.

As per Table 11, the item locations ranged between -3.20 and 2.31 logits, mostly covering the underlying ability trait level of the respondents. One item (A8) demonstrated underfit, whereas one item (S10) demonstrated overfit as per the OMNSQ ≥ 1.40 or ≤ 0.60 guidelines. Regarding OMNSQ ≥ 1.30 and IMNSQ ≥ 1.10 values, no items breached this threshold, although items A8, O4, and R9 came fairly close.

To assess unidimensionality, principal component analysis was conducted on the standardised residuals. The Eigenvalue of the first (1.48) contrast did not exceed 2, indicating evidence of unidimensionality (e.g., Raîche, 2005). Furthermore, the local independence of items was assessed by looking at the largest standardised residual correlations. Items R6 and R7 had the largest standardised residual correlation (0.11), which is considerably lower than the typical 0.70 guideline (Linacre, 2020b). Yen's Q3 statistic for the correlation between O11 and O12 was 0.14, which is lower than typical suggestions of 0.30 (Aryadoust et al., 2020). Consequently, there were no obvious indications of local dependence (i.e., participants' responses to one item seemed independent to their responses to other items).

Construct Validity

Regarding the factor structure of the Verbatim, findings from the previous technical manual (van Zyl & Taylor, 2015) suggested that a bifactor exploratory structural equation model (bifactor ESEM) offers the best representation of the data. Therefore, a bifactor ESEM model was specified with the weighted least square mean and variance adjusted (WLSMV) estimator in Mplus version 8.4 (Muthén & Muthén, 2012–2019). The model's performance was assessed through the following commonly reported fit metrics: comparative fit index (CFI), Tucker-Lewis index (TLI), the root mean square error of approximation (RMSEA), and the standardised root mean square residual (SRMR). Values close to 0.95 (CFI and TLI), 0.06 (RMSEA), and 0.08 (SRMR) generally indicate good model fit (Hu & Bentler, 1999). Additionally, a 1-factor, correlated 5-factor, bifactor confirmatory factor analytic (bifactor CFA) model, and an exploratory structural equation model (ESEM) were specified for comparative rather than interpretive purposes. Table 12 reports the results of the specified models.

Table 12. *Fit Statistics of Different Factor Models*

Model	χ2	*df*	CFI	TLI	RMSEA	90% CI	SRMR
1 Factor	920.865*	350	0.971	0.969	0.025	0.023, 0.027	0.044
5 Factor	679.480*	340	0.983	0.981	0.020	0.017, 0.022	0.038
Bifactor CFA	566.679*	322	0.988	0.985	0.017	0.015, 0.019	0.036
ESEM	283.888*	248	0.998	0.997	0.007	0.000, 0.011	0.024
Bifactor ESEM	223.599*	225	1.000	1.000	0.000	0.000, 0.008	0.021

Note. χ*2* = Chi-square, *df* = Degrees of Freedom, CFI = Comparative Fit Index, TLI = Tucker-Lewis Index, RMSEA = Root Mean Square Error of Approximation with 90% Confidence Intervals, SRMR = Standardised Root Mean Square Residual.

As per Table 12, the bifactor ESEM model's fit statistics comfortably exceeded CFI and TLI values of 0.95, and also comfortably fell beneath the RMSEA and SRMR thresholds of 0.06 and 0.08, respectively. As the inter-factor correlations of bifactor models are constrained to zero, the correlated 5-factor CFA model and ESEM model's inter-factor correlations were compared. Table 13 reports these correlations.

Table 13. *Standardised Inter-Factor Correlations for the Verbatim Scales*

Scale	S	O	A	R	VI
Synonyms	-	0.60*	0.45*	0.52*	0.53*
Opposites	0.89*	-	0.36*	0.46*	0.60*
Analogies	0.83*	0.89*	-	0.53*	0.46*
Reasoning	0.75*	0.80*	0.93*	-	0.59*
Interpretation	0.75*	0.82*	0.85*	0.85*	-

Note. The correlated 5-factor CFA model is below the diagonal, inter-factor correlations from the ESEM model are above the diagonal. *$p < 0.001$. S = Synonyms, O = Opposites, A = Analogies, R = Reasoning, VI = Interpretation.

On average, the sizes of the ESEM model's ($M_r = 0.51$) inter-factor correlations were considerably lower than the correlated 5-factor CFA model ($M_r = 0.84$). Howard et al. (2018) proposes that "ESEM tends to provide more exact estimates of true factor correlations" (p. 2649) compared to CFA and "that ESEM should be retained whenever the results show a discrepant pattern of factor correlations" (p. 2650). To decide between the bifactor ESEM and ESEM model, the former did not display significantly better fit than the latter, although the bifactor ESEM model had perfect CFI (1.000), TLI (1.000), and RMSEA (0.000) fit. Therefore, cross-loadings between the models were compared. The ESEM model generally

displayed higher cross-loadings than the bifactor ESEM model, providing a potential indication of an unmodelled general factor (Howard et al., 2018). This offered some support for using the bifactor ESEM model. Table 14 reports the standardised factor loadings, standard errors, item uniqueness or bifactor standardised residual variance, and the item explained common variance (IECV) for the bifactor ESEM model.

Table 14. *Bifactor ESEM Model Statistics*

	General		Synonyms		Opposites		Analogies		Reasoning		Interpretation			
Item	λ	S.E.	λ	S.E.	λ	S.E.	λ	S.E.	λ	S.E.	λ	S.E.	δ	IECV
S1	**0.46***	0.05	**0.19**	0.10	0.28*	0.09							0.65	0.60
S3	**0.52***	0.03	**0.31***	0.06									0.62	0.72
S5	**0.53***	0.03	**0.45***	0.06									0.50	0.57
S10	**0.70***	0.03	**0.28***	0.06			-0.15*	0.05					0.40	0.81
S11	**0.54***	0.02	**0.24***	0.05									0.65	0.81
O4	**0.44***	0.03	0.13*	0.05	**0.29***	0.06							0.69	0.62
O6	**0.55***	0.03	0.16*	0.05	**0.03**	0.06	-0.14*	0.05					0.65	0.86
O9	**0.49***	0.03			**0.15***	0.05							0.72	0.88
O10	**0.69***	0.02			**0.32***	0.05							0.42	0.81
O11	**0.70***	0.03			**0.42***	0.07	-0.13*	0.05					0.32	0.71
O12	**0.58***	0.02			**-0.07**	0.06	0.16*	0.05			-0.10*	0.04	0.62	0.89
A5	**0.50***	0.03			0.17*	0.05	**-0.04**	0.06					0.72	0.87
A6	**0.58***	0.02					**0.05**	0.05					0.66	0.99
A7	**0.61***	0.02					**0.31***	0.05					0.52	0.78
A8	**0.36***	0.03			-0.14*	0.05	**0.27***	0.06	0.19*	0.04			0.74	0.50
A9	**0.70***	0.02	-0.08*	0.04			**0.21***	0.05					0.45	0.90
A10	**0.44***	0.03	0.10*	0.05			**0.05**	0.05					0.79	0.92
R3	**0.53***	0.03							**0.13***	0.04	0.09*	0.04	0.68	0.86
R5	**0.51***	0.02							**0.03**	0.04			0.72	0.94

R6	**0.51***	0.03	-0.09*	0.04					**0.59***	0.09			0.37	0.42
R7	**0.50***	0.03	0.10*	0.04					**0.45***	0.07			0.53	0.53
R8	**0.47***	0.03			0.16*	0.05			**0.18***	0.06	0.21*	0.05	0.66	0.65
R9	**0.36***	0.03					0.35*	0.05	**0.02**	0.05			0.74	0.51
VI2	**0.49***	0.03	-0.14*	0.04			-0.13*	0.05			**0.33***	0.06	0.62	0.62
VI4	**0.60***	0.02			-0.15*	0.05					**0.06**	0.06	0.61	0.92
VI6	**0.51***	0.03	-0.16*	0.05	-0.11*	0.05			-0.11*	0.04	**-0.14***	0.07	0.67	0.78
VI13	**0.57***	0.03									**0.37***	0.05	0.54	0.69
VI14	**0.48***	0.03									**0.43***	0.05	0.58	0.55

Note. *$p < 0.05$. λ = Standardised factor loadings, S.E. = standard error, δ = item uniqueness/bifactor standardised residual variance, IECV = item explained common variance. Statistically significant cross-loadings are underlined. Standardised factor loadings for specific factors are indicated in bold. Standardised factor loadings that were not statistically significant, together with their standard errors were removed. IECV values were derived from Dueber's (2017) Bifactor Indices Calculator in Excel. S = Synonyms, O = Opposites, A = Analogies, R = Reasoning, VI = Interpretation.

As per Table 14, the general factor loadings were all statistically significant ($p < 0.001$), ranging from 0.36 to 0.70. These were considered acceptable as per Spector's (1992) suggestion of a minimum value of 0.30 to 0.35 for an item to load onto a factor. The standard errors were also generally low (≤ 0.05). Except for item R6, all items had larger general than specific factor loadings. The specific factors were fairly weakly defined compared to the general factor. Statistically significant standardised factor loadings were found for four of the five Synonyms items, four of the six Opposites items, three of the six Analogies items, four of the six Reasoning items, and four of the six Interpretation items. Hence, 19 of the 28 Verbatim items loaded significantly on their intended target factor, although only ten items had standardised factor loadings above 0.30. All item uniqueness values fell within an acceptable range (> 0.10 δ < 0.90; van Zyl & ten Klooster, 2022). One item (R6 = 0.42) had an IECV value < 0.50. Items A8 and R9 had the lowest general factor loadings (λ=0.36). Statistically significant cross-loadings ranged from-0.08 to 0.28 (24 items), except for item R9 which had a reasonable cross-loading of 0.35. When items significantly loaded onto their target construct, lower accompanying cross-loadings were generally observed.

Furthermore, the orthogonally rotated factor loadings obtained from Mplus were used to calculate other bifactor indices as reported in Table 15. Cross-loadings were ignored in calculating the specific factors' reliability (Morin et al., 2020). The Bifactor Indices Calculator in Excel (Dueber, 2017) was used for this purpose.

Table 15. *Bifactor Indices for the Bifactor ESEM Model*

Factor	ECV	ωh	ωRel.	H	FD
General Factor	0.73	0.90	0.96	0.93	0.96
Synonyms	0.06	0.17	0.22	0.42	0.71
Opposites	0.06	0.08	0.10	0.41	0.72
Analogies	0.05	0.05	0.07	0.35	0.66
Reasoning	0.06	0.14	0.19	0.48	0.75
Interpretation	0.05	0.10	0.14	0.39	0.69

Note. ECV=Explained Common Variance, ωh=Coefficient Omega Hierarchical, ωRel.=Relative Omega, H=Construct Replicability, FD = Factor Determinacy.

As per Table 15, the general factor explained 73% of the common variance. The group factors' explained variance ranged from 5 to 6%. Coefficients omega hierarchical and relative omega were 0.90 and 0.96, respectively. The general factor was the only well-defined factor ($H > 0.80$; Rodriguez et al., 2016a) with

a coefficient of 0.93. The percentage of uncontaminated correlations (PUC[8]) was 0.83. Rodriguez et al. (2016a) propose that “when ECV is > .70 and PUC > .70, relative bias will be slight and the common variance can be regarded as essentially unidimensional" (p. 232). Furthermore, the absolute relative parameter bias (ARPB) was 3.5%, implying that the items’ unidimensional factor loadings did not substantially differ from their general factor loadings (ARPB < 10-15%; Rodriguez et al., 2016b).

To gain additional insights into the validity of the bifactor ESEM model, the data were analysed in FACTOR version 12.04.01 (Lorenzo-Seva & Ferrando, 2023) with the following model specifications: matrix analysed = polychoric matrix (tetrachoric) with sweet smoothing; estimation = Robust diagonally weighted least squares (RDWLS); and rotation = Orthogonal Procrustean rotation. The adequacy of the polychoric correlation matrix was as follows: Bartlett’s statistic = 23540.9 (*df* = 378, $p < 0.001$); Kaiser-Meyer-Olkin (KMO) test = 0.95 (which is considered very good). Furthermore, results showed that none of the items should be removed based on their Measure of Sampling Adequacy (MSA) values. MSA values below 0.50 suggest that the item does not measure the same domain as the remaining items in the pool and should probably be removed (Lorenzo-Seva & Ferrando, 2021). Goodness-of-fit metrics indicated a close fit to the specified model: Root Mean Square Error of Approximation (RMSEA) = 0.000; Comparative Fit Index (CFI) = 0.999; Non-Normed Fit Index (NNFI) = 1.000.

Overall, to ascertain whether scale scores, a total score, or both should be interpreted, results as gathered from reliability indicators, Haberman’s test, Rasch analysis, and bifactor analysis were examined and suggest that a total Verbatim score should be interpreted. More research is needed to determine the value-add of the specific factors beyond the general factor. This is therefore reflected in the feedback reports, which only provide standard scores and interpretation for the total Verbatim score.

Item difficulty and item discrimination

Item difficulty 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 in relation 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, 1970). According to Kerlinger and Lee (2000) item difficulties should range between 0.50 and 0.70, where a

[8] Cross-loadings were excluded to calculate the PUC.

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). For an ability test, the item difficulties would be expected to have a larger range. Table 16 provides the item-rest correlation as calculated in jamovi version 2.3.28 (The jamovi project, 2023) as well as item difficulty and item discrimination values for the total Verbatim score using the *item.exam* function in the *psychometric* (Fletcher, 2022) package.

Table 16. *Item-Rest Correlations, Item Difficulty, and Item Discrimination for the Total Verbatim Score Items*

Item	Item-rest correlation	Difficulty	Discrimination
S1	0.20	0.97	0.06
S3	0.29	0.92	0.17
S5	0.39	0.66	0.49
S10	0.42	0.87	0.31
S11	0.39	0.52	0.54
O4	0.29	0.84	0.25
O6	0.39	0.62	0.51
O9	0.34	0.77	0.36
O10	0.46	0.79	0.43
O11	0.45	0.83	0.37
O12	0.41	0.71	0.50
A5	0.35	0.76	0.38
A6	0.42	0.59	0.54
A7	0.46	0.50	0.64
A8	0.26	0.28	0.35
A9	0.50	0.44	0.67
A10	0.32	0.66	0.44
R3	0.39	0.67	0.46
R5	0.37	0.51	0.53
R6	0.39	0.62	0.53
R7	0.34	0.25	0.42
R8	0.32	0.85	0.28
R9	0.27	0.42	0.42
VI2	0.35	0.65	0.45
VI4	0.43	0.59	0.58
VI6	0.33	0.55	0.48
VI13	0.41	0.69	0.49
VI14	0.36	0.57	0.50

As per Table 16, the average item difficulty was 0.65 and the average item discrimination was 0.43. Item S1 was the least capable of discriminating between respondents who scored high and low on the

Verbatim. No items had item-rest correlation values below a minimally acceptable benchmark of 0.20 (Zijlmans et al., 2018). The average item-rest correlation was 0.37.

Differential item functioning

Differential item functioning (DIF) through ordinal logistic regression was investigated with the *rundif* function in the *lordif* (Choi et al., 2016) package. The Rasch Person measures, as exported from Winsteps were used as the conditioning variable. Three models were compared (baseline, uniform DIF, and non-uniform DIF). The first and third models were compared first to establish an overall DIF effect size. Thereafter, the DIF was examined to determine whether it was uniform or non-uniform. The statistical significance value was set to $p < 0.001$ (as opposed to $p < 0.05$) with consideration for Type I errors. A change in Nagelkerke's pseudo R-squared (R^2) across the models was assessed to establish the magnitude of DIF. The effect size guidelines of Jodoin and Gierl (2001) were used in this regard: negligible ($R^2 < 0.035$), moderate ($R^2 = 0.035$ to 0.070), and large ($R^2 > 0.070$). DIF was investigated in a pairwise manner for gender (female vs. male), ethnicity (Black African vs. White), language (English vs. Zulu, and English vs. Afrikaans), and education (Grade 12 vs. Diploma/Certificate, and Bachelor's degree vs Honours degree)[9]. Table 17 to Table 22 report these results.

[9] For DIF between groups not mentioned here, see Appendix A.

Table 17. *Differential Item Functioning Across Gender Groups*

	p-values for χ2 difference tests			Change in Nagelkerke's R^2		
Item	M1-M2	M1-M3	M2-M3	M1-M2	M1-M3	M2-M3
			Gender			
S1	0.001	0.002	0.467	0.019	0.019	0.001
S3	**0.000**	**0.000**	0.031	0.024	0.027	0.004
S5	0.658	0.718	0.494	0.000	0.000	0.000
S10	**0.000**	**0.000**	0.085	0.008	0.010	0.002
S11	0.019	0.060	0.713	0.002	0.002	0.000
O4	0.297	0.471	0.518	0.001	0.001	0.000
O6	0.002	0.002	0.101	0.004	0.005	0.001
O9	0.114	0.043	0.052	0.001	0.003	0.002
O10	0.031	0.077	0.502	0.002	0.002	0.000
O11	0.854	0.927	0.731	0.000	0.000	0.000
O12	0.070	0.194	0.994	0.001	0.001	0.000
A5	0.474	0.723	0.711	0.000	0.000	0.000
A6	0.546	0.821	0.859	0.000	0.000	0.000
A7	0.010	0.009	0.090	0.002	0.004	0.001
A8	0.058	0.097	0.300	0.002	0.002	0.000
A9	0.263	0.262	0.232	0.000	0.001	0.000
A10	0.354	0.577	0.625	0.000	0.000	0.000
R3	0.031	0.086	0.628	0.002	0.002	0.000
R5	0.641	0.037	0.011	0.000	0.003	0.003
R6	0.065	0.130	0.407	0.001	0.002	0.000
R7	0.278	0.310	0.280	0.000	0.001	0.000
R8	0.162	0.044	0.038	0.001	0.004	0.002
R9	0.396	0.288	0.183	0.000	0.001	0.001
VI2	0.094	0.176	0.410	0.001	0.002	0.000
VI4	0.029	0.029	0.128	0.002	0.003	0.001
VI6	0.547	0.404	0.229	0.000	0.001	0.001
VI13	0.033	0.060	0.290	0.002	0.002	0.000
VI14	0.015	0.025	0.231	0.002	0.003	0.001

As per **Table 17**, although some statistically significant chi-squared (χ2) values were observed, their effect sizes were negligible ($R^2 < 0.035$).

Table 18. *Differential Item Functioning Across Ethnic Groups*

	p-values for χ2 difference tests			Change in Nagelkerke's R^2		
Item	M1-M2	M1-M3	M2-M3	M1-M2	M1-M3	M2-M3
			Ethnicity			
S1	0.353	0.143	0.082	0.002	0.009	0.007
S3	**0.000**	0.002	0.651	0.017	0.017	0.000
S5	0.666	0.880	0.791	0.000	0.000	0.000
S10	**0.000**	**0.000**	0.055	0.016	0.019	0.003
S11	**0.000**	**0.000**	0.035	0.012	0.015	0.003
O4	0.118	0.031	0.034	0.002	0.006	0.004
O6	0.118	0.018	0.018	0.002	0.005	0.004
O9	0.003	0.008	0.282	0.007	0.008	0.001
O10	0.009	0.010	0.121	0.005	0.007	0.002
O11	0.006	0.015	0.318	0.006	0.006	0.001
O12	**0.000**	**0.000**	0.026	0.010	0.013	0.004
A5	0.188	0.413	0.861	0.001	0.001	0.000
A6	0.524	0.007	0.002	0.000	0.006	0.006
A7	0.001	0.005	0.593	0.006	0.006	0.000
A8	0.461	0.204	0.105	0.000	0.002	0.002
A9	**0.000**	**0.000**	0.330	0.028	0.029	0.001
A10	0.530	0.001	**0.000**	0.000	0.010	0.010
R3	0.568	0.444	0.255	0.000	0.001	0.001
R5	0.117	0.122	0.185	0.002	0.003	0.001
R6	0.237	0.068	0.046	0.001	0.004	0.003
R7	0.377	0.500	0.435	0.001	0.001	0.000
R8	0.547	0.830	0.914	0.000	0.000	0.000
R9	0.024	0.078	0.851	0.004	0.004	0.000
VI2	0.001	**0.000**	0.023	0.007	0.011	0.004
VI4	0.090	0.186	0.482	0.002	0.002	0.000
VI6	0.008	0.025	0.535	0.005	0.005	0.000
VI13	0.262	0.491	0.684	0.001	0.001	0.000
VI14	0.601	0.808	0.694	0.000	0.000	0.000

Note. M1-M2 = Model 1 vs. Model 2 (uniform DIF), M1-M3 = Model 1 vs. Model 3 (total DIF), M2-M3 = Model 2 vs. Model 3 (non-uniform DIF).

As per Table 18, no notable differences were observed on any items between Black African and White participants. Differences between Black African and Indian, as well as Indian and White participants,

were also evaluated (see **Appendix A**). Although some statistically significant chi-square (χ2) values were observed, their effect sizes were negligible ($R^2 < 0.035$).

Table 19. *Differential Item Functioning Across Language Groups*

	p-values for χ2 difference tests			Change in Nagelkerke's R^2		
Item	M1-M2	M1-M3	M2-M3	M1-M2	M1-M3	M2-M3
			Language (English vs. Zulu)			
S1	0.102	0.238	0.658	0.010	0.011	0.001
S3	0.746	0.820	0.589	0.000	0.001	0.000
S5	0.656	0.779	0.584	0.000	0.000	0.000
S10	0.984	0.605	0.316	0.000	0.001	0.001
S11	0.040	0.020	0.057	0.004	0.006	0.003
O4	0.002	0.007	0.493	0.013	0.014	0.001
O6	0.213	0.210	0.210	0.001	0.003	0.001
O9	0.938	0.895	0.642	0.000	0.000	0.000
O10	0.001	0.002	0.723	0.013	0.013	0.000
O11	0.313	0.205	0.143	0.001	0.004	0.002
O12	0.005	0.014	0.431	0.007	0.008	0.001
A5	0.217	0.392	0.554	0.002	0.002	0.000
A6	0.600	0.012	0.003	0.000	0.008	0.007
A7	0.789	0.607	0.336	0.000	0.001	0.001
A8	0.133	0.040	0.040	0.002	0.006	0.004
A9	0.192	0.014	0.009	0.001	0.006	0.005
A10	0.575	0.002	0.000	0.000	0.012	0.011
R3	0.739	0.240	0.098	0.000	0.003	0.003
R5	0.230	0.112	0.087	0.001	0.004	0.002
R6	0.487	0.217	0.109	0.000	0.003	0.002
R7	0.385	0.684	0.952	0.001	0.001	0.000
R8	0.097	0.177	0.397	0.003	0.004	0.001
R9	0.254	0.060	0.038	0.001	0.005	0.004
VI2	0.003	0.006	0.276	0.008	0.009	0.001
VI4	0.630	0.815	0.674	0.000	0.000	0.000
VI6	0.341	0.348	0.272	0.001	0.002	0.001
VI13	0.121	0.200	0.368	0.002	0.003	0.001
VI14	0.482	0.733	0.722	0.000	0.001	0.000

Note. M1-M2 = Model 1 vs. Model 2 (uniform DIF), M1-M3 = Model 1 vs. Model 3 (total DIF), M2-M3 = Model 2 vs. Model 3 (non-uniform DIF).

As per Table 19, although some statistically significant chi-square (χ2) values were observed, their effect sizes were negligible ($R^2 < 0.035$).

Table 20. *Differential Item Functioning Across Language Groups*

	p-values for χ2 difference tests			Change in Nagelkerke's R^2		
Item	M1-M2	M1-M3	M2-M3	M1-M2	M1-M3	M2-M3
Language (English vs. Afrikaans)						
S1	0.780	0.917	0.758	0.000	0.001	0.000
S3	0.056	0.126	0.483	0.006	0.007	0.001
S5	0.844	0.525	0.264	0.000	0.001	0.001
S10	**0.000**	**0.000**	0.173	0.024	0.027	0.002
S11	0.032	0.101	0.971	0.004	0.004	0.000
O4	0.687	0.853	0.692	0.000	0.000	0.000
O6	**0.000**	**0.000**	0.083	0.018	0.020	0.003
O9	0.035	0.090	0.531	0.005	0.005	0.000
O10	0.166	0.377	0.855	0.002	0.002	0.000
O11	0.451	0.288	0.166	0.001	0.003	0.002
O12	0.992	0.996	0.929	0.000	0.000	0.000
A5	0.075	0.146	0.408	0.003	0.004	0.001
A6	0.082	0.220	0.915	0.003	0.003	0.000
A7	0.059	0.147	0.598	0.003	0.003	0.000
A8	0.001	0.002	0.368	0.010	0.011	0.001
A9	0.557	0.249	0.119	0.000	0.002	0.002
A10	0.288	0.400	0.401	0.001	0.002	0.001
R3	0.046	0.099	0.427	0.004	0.004	0.001
R5	0.190	0.402	0.748	0.001	0.002	0.000
R6	0.906	0.753	0.456	0.000	0.000	0.000
R7	0.902	0.807	0.520	0.000	0.000	0.000
R8	0.036	0.097	0.593	0.005	0.006	0.000
R9	0.073	0.081	0.179	0.003	0.004	0.002
VI2	0.380	0.001	**0.000**	0.001	0.012	0.012
VI4	0.484	0.467	0.310	0.000	0.001	0.001
VI6	0.024	0.069	0.616	0.004	0.005	0.000
VI13	0.725	0.893	0.750	0.000	0.000	0.000
VI14	0.553	0.772	0.684	0.000	0.000	0.000

Note. M1-M2 = Model 1 vs. Model 2 (uniform DIF), M1-M3 = Model 1 vs. Model 3 (total DIF), M2-M3 = Model 2 vs. Model 3 (non-uniform DIF).

As per Table 20, no notable differences were observed on any items between English and Afrikaans participants. In addition to English and Zulu, and English and Afrikaans, differences between English and Xhosa, and English and Pedi participants, were also evaluated (see Appendix A). Although some statistically significant chi-square (χ2) values were observed, their effect sizes were negligible (R^2 < 0.035).

Table 21. *Differential Item Functioning Across Educational Groups*

	p-values for χ2 difference tests			Change in Nagelkerke's R^2		
Item	M1-M2	M1-M3	M2-M3	M1-M2	M1-M3	M2-M3
Education (Grade 12 vs. Diploma/Certificate)						
S1	0.478	0.777	0.999	0.003	0.003	0.000
S3	0.172	0.080	0.075	0.006	0.015	0.010
S5	0.024	0.012	0.054	0.010	0.018	0.007
S10	0.004	0.009	0.283	0.018	0.021	0.003
S11	0.036	0.058	0.253	0.009	0.012	0.003
O4	0.768	0.958	0.991	0.000	0.000	0.000
O6	0.572	0.219	0.100	0.001	0.006	0.006
O9	0.030	0.088	0.710	0.011	0.011	0.000
O10	0.138	0.327	0.862	0.004	0.004	0.000
O11	0.559	0.503	0.309	0.001	0.003	0.002
O12	0.736	0.330	0.147	0.000	0.004	0.004
A5	0.436	0.476	0.349	0.001	0.003	0.002
A6	0.671	0.657	0.417	0.000	0.002	0.001
A7	0.636	0.451	0.242	0.000	0.003	0.002
A8	0.764	0.777	0.519	0.000	0.001	0.001
A9	0.027	0.070	0.517	0.008	0.009	0.001
A10	0.478	0.438	0.284	0.001	0.004	0.002
R3	0.583	0.681	0.494	0.001	0.002	0.001
R5	0.131	0.028	0.027	0.005	0.015	0.010
R6	0.110	0.209	0.448	0.005	0.007	0.001
R7	0.996	0.928	0.699	0.000	0.000	0.000
R8	0.742	0.930	0.848	0.000	0.000	0.000
R9	0.035	0.109	0.907	0.010	0.010	0.000
VI2	0.593	0.250	0.115	0.001	0.006	0.005
VI4	0.747	0.704	0.439	0.000	0.001	0.001
VI6	0.938	0.095	0.030	0.000	0.010	0.010
VI13	0.448	0.650	0.595	0.001	0.002	0.000
VI14	0.619	0.883	0.996	0.000	0.000	0.000

Note. M1-M2 = Model 1 vs. Model 2 (uniform DIF), M1-M3 = Model 1 vs. Model 3 (total DIF), M2-M3 = Model 2 vs. Model 3 (non-uniform DIF).

As per Table 21, no notable differences were observed on any items between Grade 12 and Diploma/Certificate participants ($R^2 < 0.035$).

Table 22. *Differential Item Functioning Across Educational Groups*

	p-values for χ2 difference tests			Change in Nagelkerke's R^2		
Item	M1-M2	M1-M3	M2-M3	M1-M2	M1-M3	M2-M3
Education (Bachelor's degree vs. Honours degree)						
S1	0.779	0.116	0.040	0.001	**0.043**	**0.042**
S3	0.515	0.149	0.066	0.002	0.019	0.017
S5	0.814	0.968	0.920	0.000	0.000	0.000
S10	0.474	0.762	0.861	0.003	0.003	0.000
S11	0.136	0.286	0.592	0.004	0.005	0.001
O4	0.987	0.073	0.022	0.000	0.015	0.015
O6	0.841	0.980	0.995	0.000	0.000	0.000
O9	0.896	0.113	0.037	0.000	0.012	0.012
O10	0.827	0.871	0.633	0.000	0.001	0.001
O11	0.122	0.292	0.801	0.008	0.008	0.000
O12	0.201	0.363	0.532	0.004	0.004	0.001
A5	0.643	0.892	0.907	0.001	0.001	0.000
A6	0.256	0.488	0.701	0.002	0.003	0.000
A7	0.664	0.696	0.464	0.000	0.001	0.001
A8	0.796	0.014	0.004	0.000	0.017	0.017
A9	0.894	0.828	0.549	0.000	0.001	0.001
A10	0.649	0.880	0.826	0.000	0.000	0.000
R3	0.988	0.761	0.460	0.000	0.001	0.001
R5	0.584	0.859	0.948	0.001	0.001	0.000
R6	0.926	0.219	0.082	0.000	0.006	0.006
R7	0.020	0.065	0.824	0.011	0.011	0.000
R8	0.526	0.732	0.638	0.001	0.002	0.001
R9	0.778	0.526	0.272	0.000	0.003	0.002
VI2	0.010	0.035	0.730	0.014	0.014	0.000
VI4	0.797	0.955	0.870	0.000	0.000	0.000
VI6	0.952	0.402	0.177	0.000	0.004	0.004
VI13	0.140	0.045	0.045	0.005	0.014	0.009
VI14	0.452	0.650	0.587	0.001	0.002	0.001

Note. M1-M2 = Model 1 vs. Model 2 (uniform DIF), M1-M3 = Model 1 vs. Model 3 (total DIF), M2-M3 = Model 2 vs. Model 3 (non-uniform DIF).

As per Table 22, item S1 had an effect size of R^2 of 0.043. The accompanying chi-square ($\chi 2$) values were however not significant at the $p < 0.001$ level. Furthermore, no notable differences were observed between participants with Bachelor's degrees and those with Honours degrees.

In summary, from a DIF perspective, no notable concerns were observed in terms of how the items functioned across different gender, and specific ethnic and language groups. This offers preliminary support to refrain from developing gender-, ethnic-, and language-based norms.

Measurement invariance

Putnick and Bornstein (2016) propose that before any meaningful comparisons between groups can be made, measurement invariance should be established. "Measurement invariance assesses the (psychometric) equivalence of a construct across groups …" (Putnick & Bornstein, 2016, p. 72). Widaman and Reise's (1997) steps were followed for measurement invariance testing. Hence, configural, metric (or weak factorial), scalar (or strong factorial), and strict (or residual/invariant uniqueness) models were specified and tested sequentially in Mplus with the WLSMV estimator. In each model, an additional constraint was added. De Beer and Morin's (2022) (B) ESEM invariance syntax generator for Mplus was used but had to be slightly modified to accommodate the dichotomous nature of the data. To draw parallels between the models, delta changes (Δ) in CFI, TLI, and RMSEA were assessed. Changes of -0.01 (CFI and TLI) and 0.015 (RMSEA) from one model to the next signalled noninvariance (Chen, 2007; Cheung & Rensvold, 2002). Table 23 reports the measurement invariance results for gender, and specific ethnic, language, and educational groups.

Table 23. *Bifactor ESEM Measurement Invariance Testing for Gender, Ethnicity, Language, and Education*

Model	χ^2	*df*	CFI	TLI	RMSEA	SRMR	CM	ΔCFI	ΔTLI	ΔRMSEA
Gender (Women, *n* = 1392; Men, *n* = 1202)										
M1: Configural	455.534*	450	1.000	0.999	0.003 [0.000, 0.010]	0.031	-	-	-	-
M2: Metric	634.844*	576	0.997	0.996	0.009 [0.002, 0.013]	0.037	M1	-0.003	-0.003	0.006
M3: Scalar	633.010*	582	0.997	0.996	0.008 [0.000, 0.012]	0.038	M2	0.000	0.000	-0.001
M4: Strict	677.159*	604	0.996	0.995	0.010 [0.004, 0.013]	0.039	M3	-0.001	-0.001	0.002
Ethnicity (Black African, *n* = 1220; White, *n* = 363; Indian, *n* = 217)										
M1: Configural	649.396*	675	1.000	1.000	0.000 [0.000, 0.010]	0.044	-	-	-	-
M2: Metric	979.923*	939	0.996	0.995	0.009 [0.000, 0.015]	0.063	M1	-0.004	-0.005	0.009
M3: Scalar	977.130*	955	0.998	0.997	0.006 [0.000, 0.013]	0.061	M2	0.002	0.002	-0.003
M4: Strict	1063.224*	983	0.991	0.990	0.012 [0.003, 0.017]	0.065	M3	-0.007	-0.007	0.006
Language (English, *n* = 916; Zulu, *n* = 310; Afrikaans, *n* = 393)										
M1: Configural	612.959*	675	1.000	1.000	0.000 [0.000, 0.000]	0.043	-	-	-	-
M2: Metric	914.363*	939	1.000	1.000	0.000 [0.000, 0.010]	0.059	M1	0.000	0.000	0.000
M3: Scalar	991.140*	955	0.996	0.995	0.008 [0.000, 0.015]	0.061	M2	-0.004	-0.005	0.008
M4: Strict	967.960*	983	1.000	1.000	0.000 [0.000, 0.011]	0.060	M3	0.004	0.005	-0.008
Education (Grade 12 + Diploma/Certificate, *n* = 541; Bachelor's/Honours degree, *n* = 572)										
M1: Configural	464.669*	450	0.997	0.995	0.008 [0.000, 0.017]	0.052	-	-	-	-
M2: Metric	588.441*	576	0.998	0.997	0.006 [0.000, 0.015]	0.061	M1	0.001	0.002	-0.002
M3: Scalar	604.210*	582	0.996	0.995	0.008 [0.000, 0.016]	0.061	M2	-0.002	-0.002	0.002
M4: Strict	633.973*	604	0.994	0.993	0.009 [0.000, 0.017]	0.062	M3	-0.002	-0.002	0.001

Note. χ^2 = Chi-square, *df* = Degrees of Freedom, CFI = Comparative Fit Index, TLI = Tucker-Lewis Index, SRMR = Standardised Root Mean Square Residual, RMSEA = Root Mean Square Error of Approximation with 90% Confidence Intervals, CM = Comparison Model, ΔCFI = Change in CFI, ΔTLI = Change in TLI, ΔRMSEA = Change in RMSEA.

As per Table 23, all models displayed good fit statistics (CFI, TLI > 0.95; RMSEA < 0.06; SRMR < 0.08; Hu & Bentler, 1999). Furthermore, no problematic delta changes were observed for CFI (Δ-0.01), TLI (Δ-0.01), or RMSEA (Δ 0.015) from one model to the next (Chen, 2007; Cheung & Rensvold, 2002). Hence, strict measurement invariance for the Verbatim was established across gender, and specific ethnic, language, and educational groups. This meant that meaningful group comparisons could be made at the total score level.

Mean differences across groups

Before mean group comparisons were made, the homogeneity of variance assumption was tested with the *leveneTest* function in the *car* (Fox & Weisberg, 2019) package. Results from Levene's (1960) test showed that the homogeneity of variance assumption was violated across almost all groups. Therefore, depending on the number of groups (two or more), either Welch two-sample t-tests or Welch ANOVAs were performed to assess mean group differences. These were calculated with the *t.test* function (two groups) or the *oneway.test* function (> two groups) in the *stats* (R Core Team, 2023) package. Post-hoc tests were carried out with the *posthoc_anova* function in the *biostat* (Gegzna, 2020) package. Table 24 reports the results of the Welch two-sample t-test for the gender groups.

Table 24. *Mean Differences Between Gender Groups*

	Men		Women		x̄ diff.	*t*	*df*	Effect size [a]
	M	*SD*	*M*	*SD*				
Total	18.77	5.46	17.54	5.34	1.23	5.760	2552.4	0.23*

Note. x̄ diff. = Mean difference between comparison groups, *t* = Welch statistic, *df* = Degrees of freedom, Effect size = Cohen's *d*. *$p < 0.05$. [a] Values of 0.20, 0.50, and 0.80 correspond to small-, medium-, and large effect sizes (Cohen, 1988).

As per Table 24, statistically significant mean differences were found between male and female participants. A Cohen's *d* value of 0.23 indicated a small effect size (Cohen, 1988). Other descriptive statistics showed that obtaining correct answers on all items of the test were accomplishable for both gender groups. In the following analyses, more than two groups were involved, hence Welch ANOVAs were performed.

Results from the Welch ANOVA showed statistically significant mean differences between ethnic groups [$F(3, 471.1) = 64.279$, $p < 0.001$], language groups [$F(4, 640.4) = 44.497$, $p < 0.001$], and educational

groups [$F(3, 601.6) = 65.438$, $p < 0.001$]. Post-hoc Games-Howell tests showed that statistically significant mean differences were found between different ethnic, language, and educational groups that had at least 100 observations[10]. For practical purposes, only statistically significant results were displayed. Table 25 reports these results.

[10] For practical purposes, only language groups with > 200 observations were used.

Table 25. *Mean Differences Between Ethnic, Language, and Educational Groups*

Mean differences based on Ethnicity												
Group	**Ethnicity**	***n***	**Mean**	**Variance**	***SD***	**Comp.**	**x̄ diff.**	**LCI**	**UCI**	***t***	***df***	***p***
1	Black	1220	16.883	29.841	5.463	2 vs. 1	3.294	2.548	4.039	11.38	687.13	<0.001
2	White	363	20.176	21.549	4.642	4 vs. 1	3.693	2.768	4.619	10.31	326.39	<0.001
3	Coloured	160	17.562	27.317	5.227	3 vs. 2	-2.614	-3.854	-1.374	5.45	274.23	<0.001
4	Indian	217	20.576	22.560	4.750	4 vs. 3	3.014	1.660	4.367	5.75	323.34	<0.001
Mean differences based on Home Language												
Group	**Language**	***n***	**Mean**	**Variance**	***SD***	**Comp.**	**x̄ diff.**	**LCI**	**UCI**	***t***	***df***	***p***
1	English	916	20.036	23.735	4.872	2 vs. 1	-3.604	-4.560	-2.648	10.32	486.47	<0.001
2	Zulu	310	16.432	29.741	5.454	3 vs. 1	-1.087	-1.915	-0.259	3.59	714.70	0.003
3	Afrikaans	393	18.949	25.809	5.080	4 vs. 1	-2.892	-4.040	-1.743	6.91	273.81	<0.001
4	Xhosa	201	17.144	29.944	5.472	5 vs. 1	-3.634	-4.735	-2.533	9.06	293.00	<0.001
5	Pedi	209	16.402	28.232	5.313	3 vs. 2	2.517	1.417	3.617	6.26	640.31	<0.001
						4 vs. 3	-1.805	-3.075	-0.535	3.90	377.74	0.001
						5 vs. 3	-2.547	-3.775	-1.320	5.69	408.21	<0.001
Mean differences based on Education												
Group	**Education**	***n***	**Mean**	**Variance**	***SD***	**Comp.**	**x̄ diff.**	**LCI**	**UCI**	***t***	***df***	***p***
1	Grade 12	297	15.084	31.354	5.599	3 vs. 1	4.080	3.009	5.151	9.82	574.62	<0.001
2	Diploma	244	15.791	26.915	5.188	4 vs. 1	4.971	3.859	6.083	11.52	548.20	<0.001
3	Degree	317	19.164	21.296	4.615	3 vs. 2	3.373	2.287	4.459	8.01	489.50	<0.001
4	Honours	255	20.055	20.580	4.536	4 vs. 2	4.264	3.137	5.391	9.76	481.88	<0.001

Note. *n* = Group sample size, *SD* = Standard Deviation, Comp. = Groups being compared, x̄ diff. = Mean difference between comparison groups, LCI = Lower Confidence Interval of mean difference, UCI = Upper Confidence Interval of mean difference, *t* = t-statistic as derived from Games-Howell test, *df* = Degrees of Freedom, Black = Black African.

At this stage, we only knew that statistically significant mean differences existed and between which groups. To assess the magnitude of these mean differences, effect sizes were calculated in a pairwise manner with the *cohens_d* function in the *rstatix* (Kassambara, 2023) package. Cohen's (1988) guidelines were followed in this regard to establish small- (0.20), medium- (0.50), and large (0.80) effect sizes.

Regarding ethnicity, medium effects were found between Black African and White (d = 0.65), Black African and Indian (d = 0.72), Coloured and White (d = 0.53), and Coloured and Indian groups (d = 0.60). No statistically significant differences were observed between Black African and Coloured, or between White and Indian groups ($p > 0.05$). Other descriptive statistics showed that obtaining correct answers on all items of the test were accomplishable for all the preceding ethnic groups. The Black African and Coloured groups had higher variability in their scores compared to the White and Indian groups. Furthermore, although many ethnic groups did not disclose their highest level of education, the Indian group had the largest percentage (56.84%) of participants that fell within the Bachelor's or Honours degree group. Although medium effect sizes were found, we ascribe this mainly to variability in scores across the different groups, as well as educational differences. Hence, at this stage, we believe that ethnic-based norms are not justified.

Regarding home language groups, small effect sizes were observed between Afrikaans and English (d = 0.22), Afrikaans and Zulu (d = 0.48), Afrikaans and Xhosa (d = 0.34), and Afrikaans and Pedi groups (d = 0.49). Medium effect sizes were observed between English and Zulu (d = 0.70), English and Xhosa (d = 0.56), and English and Pedi (d = 0.71) groups. No statistically significant differences were observed between Zulu and Xhosa, Zulu and Pedi, or between the Xhosa and Pedi groups ($p > 0.05$). Other descriptive statistics showed that obtaining correct answers on all items of the test (e.g., a score of 28) was accomplishable for the English and Afrikaans groups. The maximum score for the Pedi, Xhosa, and Zulu groups was 27 (i.e., one incorrect answer). It should be kept in mind that the respective English and Afrikaans samples were larger compared to the other language groups (i.e., larger sample = bigger probability of obtaining a maximum score). The Pedi, Xhosa, and Zulu groups had higher variability in their scores compared to the Afrikaans and English groups. Furthermore, although many language groups did not disclose their highest level of education, in terms of quantity, the English group had considerably more participants that fell within the Bachelor's or Honours degree group. Although medium effect sizes were found, we ascribe this mainly to variability in scores across the different groups, as well as educational differences. Hence, at this stage, we believe that language-based norms are not justified.

Regarding educational level, large effect sizes were observed between Grade 12 participants and those with an Honours degree (d = 0.98). Large effect sizes were also observed between participants with a Diploma or Certificate and those with an Honours degree (d = 0.88). A borderline medium effect size was observed between Grade 12 participants and those with a bachelor's degree (d = 0.795). Similarly, a medium effect size was found between participants with a Diploma or Certificate and those with a bachelor's degree (d = 0.69). No statistically significant differences were observed between the Grade 12 and Diploma or Certificate groups, or between the bachelor's degree or Honours degree groups (p > 0.05). Implications regarding the effect sizes are discussed in the norms section.

CHAPTER 7: PSYCHOMETRIC PROPERTIES-NUMERATUM

In this section, the current, shortened Numeratum's psychometric properties are discussed in the following order: (a) testing assumptions of normality and the presence of outliers, (b) descriptive statistics, (c) correlation coefficients, (d) reliability coefficients, (e) Rasch analysis, (f) construct validity, (g) item difficulty and discrimination, (h) differential item functioning, (i) measurement invariance, and (j) mean differences across groups. The reader is reminded that these analyses were carried out on the test sample (see **Table 1** for the sample composition and **Chapter 3** for an explanation on why the samples were split).

Testing assumptions of normality and the presence of outliers

Checking for normality or other assumptions and outliers is essential for ensuring the reliability and validity of statistical analyses, making informed decisions, and understanding the characteristics of the data under investigation. It helps to make appropriate choices in selecting statistical methods and interpreting results. Generally, specific assumptions accompany statistical tests (e.g., normality, homogeneity of variance), and when these assumptions are met, the use of the parametric version of the test is preferable (Erceg-Hurn & Mirosevich, 2008). However, if some or all of these assumptions are violated, alternative statistical approaches or nonparametric tests may be more appropriate (Hoekstra et al., 2012). Apart from assumption violations, outliers (i.e., data points that differ significantly from others in a dataset) may distort statistical findings (Osborne & Overbay, 2019). Outliers may be indicative of errors in data collection or measurement, or they might represent genuine extreme values (Osborne & Overbay, 2019). Hence, it is important to assess them before decisions are made on how they should be dealt with.

A one-sided Grubbs test was conducted on the highest and lowest total Numeratum score values to determine if they were statistically significant outliers (Grubbs, 1950). Using the *grubbs.test* function from the *outliers* (Komsta, 2022) package, neither the highest (G = 1.34, U = 0.9986, p = 1) nor the lowest value (G = 2.57, U = 0.9949, p = 1) was statistically significant. Multivariate outliers across all the scales were subsequently investigated by plotting robust Mahalanobis distances against the quantiles of the χ2 distribution (Garrett, 1989). Minimal multivariate outliers were detected. Hence, for the most part, the data points were not significantly different from the rest. Additionally, the Numeratum scales and total Numeratum score were plotted to see if they deviated from normality. Results ($p < 0.001$) from formal statistical univariate normality tests (Shapiro-Wilk, Anderson-Darling, and Lilliefors[11]) as obtained through the *mvn* function in the *MVN* (Korkmaz et al., 2014) package showed deviations from normality. Multivariate normality was investigated using Mardia's coefficient (Mardia, 1970). The results indicated that the Numeratum scales deviated from multivariate normality. This imply that the joint distribution of multiple variables did not follow a multivariate normal distribution. It should be noted that in large samples, as in the current sample, violations of normality may be less of a concern compared to smaller samples due to the Central Limit Theorem (Gao et al., 2017). Therefore, depending on the statistical test, other assumptions (e.g., homogeneity of variance) largely dictated whether parametric tests, tests that require less restrictive assumptions, or if nonparametric tests were used in subsequent analyses.

[11] Similar to the Kolmogorov-Smirnov test.

Descriptive statistics

Table 26 provides the descriptive statistics for each of the Numeratum scales and the total Numeratum score. These were calculated with the *describe* function from the *psych* (Revelle, 2023) package.

Table 26. *Descriptive Statistics for the Numeratum Scales and Total Numeratum Score*

Scale	M	SD	Med	Trim	Mad	Min	Max	Skew	Kurt	SE
Number Problems	3.77	1.27	4	3.93	1.48	0	5	-0.83	-0.08	0.03
Patterns	3.89	1.79	4	4.05	1.48	0	6	-0.59	-0.69	0.05
Interpretation	3.21	1.48	3	3.31	1.48	0	5	-0.38	-0.94	0.04
Total	10.87	3.84	12	11.12	4.45	1	16	-0.47	-0.83	0.11

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

As per Table 26, the mean total Numeratum score was 10.87 (median = 12, SD = 3.84). Regarding univariate normality, the skewness and kurtosis values fell within acceptable ranges (-2 to 2; Koh, 2014). This suggests that each variable's distribution was reasonably symmetric and that the tails of the distribution were not excessively heavy, or light compared to a normal distribution. The standard error values were all generally low. Low standard errors are normally desirable as it suggests that the sample statistic (e.g., the sample mean) is likely to be a more accurate reflection of the population parameter (Harding et al., 2014).

Correlation coefficients

Inspection of multivariate normality using Mardia's coefficient (Mardia, 1970) found that bivariate normality was violated across most variables. Although Pearson correlation coefficients do not necessitate bivariate normality, Spearman's rank correlation coefficients were also calculated as a nonparametric alternative. Table 27 provides the Pearson correlation coefficients and Spearman's rank correlation coefficients for the three Numeratum scales. These were calculated with the *rcorr* function in the *Hmisc* (Harrell, 2023) package. The correlations had large effect sizes (Cohen, 1988). This confirmed the relatedness, yet uniqueness of the scales, which is to be expected as they all measure aspects of numerical ability. Inter-factor correlations are provided later.

Table 27. *Pearson and Spearman's Rank Correlations for the Numeratum Scales*

Scale	Number Problems	Patterns	Interpretation
Number Problems	-	0.56*	0.54*
Patterns	0.56*	-	0.62*
Interpretation	0.52*	0.62*	-

Note. Pearson correlations are below the diagonal, Spearman's rank correlations are above the diagonal. Values of 0.10, 0.30, and 0.50 correspond to small-, medium-, and large effects (Cohen, 1988). *$p < 0.001$.

Reliability

Cronbach's alpha coefficient (α; Cronbach, 1951) is arguably the most commonly used measure of reliability in psychological science (Hayes & Coutts, 2020). One of its major criticisms however revolves around the assumption of tau-equivalence (e.g., all items in the scale have equal factor loadings, all test items have the same true score), as data seldom adhere to this assumption (Teo & Fan, 2013). Consequently, in the absence of tau-equivalence, Cronbach's alpha may underestimate true reliability (Teo & Fan, 2013). Therefore, many suggest the use of McDonald's omega (ω; McDonald, 1999) as it is less reliant on the tau-equivalence assumption. To offer a more comprehensive view of the measurement properties of the scales, both the aforementioned reliability coefficients were analysed in addition to Rasch reliability coefficients. Table 28 provides the reliability coefficients for the Numeratum scales and total Numeratum score. These were calculated with the *ci.reliability* function in the *MBESS* (Kelley, 2022) package. The Rasch reliability coefficients were calculated in Winsteps. Model-based reliability coefficients are provided later.

Table 28. *Reliability Coefficients for the Numeratum Scales and Total Numeratum Score*

	α	ω	PR	IR
Number Prob.	0.56	0.56	-	-
Patterns	0.72	0.72	-	-
Interpretation	0.64	0.64	-	-
Total	0.83	0.83	0.70	0.99

Note. α = Cronbach's Alpha Coefficient, ω = Coefficient Omega, PR = Person Reliability Index (Rasch), Item Reliability Index (Rasch).

As per Table 28, the reliability coefficients for the Numeratum scales were mostly unsatisfactory, with coefficients (α and ω) ranging from 0.56 to 0.72. The reliability of the subscales is however less concerning as the total score is meant to be interpreted (see the last paragraph of the Construct Validity

section). The reliability coefficients for the total Numeratum score were deemed acceptable according to conventional guidelines (> 0.70; Nunnally, 1978). The item separation index values indicated that the item locations were generally stable. The person separation index (PSI) values for the Numeratum scales indicated that the scales may not be sensitive enough to distinguish between low and high scorers on the scale. For the total Numeratum score, the PSI value (1.54) was higher, indicating that there may be more value in interpreting the total Numeratum score rather than the scale scores. This PSI value was however still lower than the generally preferable score of 2 (Combrinck, 2020). Fisher (1992) however suggests that PSI values ≥ 1.50 and/or Person Reliability ≥ 0.70 represent acceptable separation and are deemed sufficient to distinguish two strata (e.g., low and high ability) within the sample. Consequently, as per minimum acceptable guidelines, the total Numeratum score slightly equalled or breached the preceding thresholds and was deemed adequate.

Additionally, the reliability coefficients for different gender, ethnic, language, and educational groups were examined. Table 29 reports these results.

Table 29. *Reliability Coefficients for Different Gender, Ethnic, Language, and Educational Groups*

Gender							
Female				Male			
α		ω		A		ω	
0.82		0.83		0.83		0.84	
Ethnicity							
Black African				White		Indian	
α		ω		α	ω	α	ω
0.82		0.83		0.82	0.82	0.81	0.81
Language							
English		Zulu			Afrikaans		Pedi
α	ω	α	ω	α	ω	α	ω
0.82	0.82	0.78	0.78	0.82	0.82	0.82	0.82
Education							
Grade 12		Diploma			Bachelor's		Honours
α	ω	α	ω	α	ω	α	ω
0.76	0.76	0.78	0.79	0.77	0.77	0.75	0.74

Note. α = Cronbach's Alpha Coefficient, ω = Coefficient Omega.

As per Table 29, the reliability coefficients appeared fairly consistent within and across the different groups. All reliability coefficients were deemed acceptable according to conventional guidelines (> 0.70; Nunnally, 1978).

Furthermore, Haberman's (2008) subscale scoring test based on the proportional reduction in mean squared error (PRMSE) was used to investigate whether interpretation should be conducted at the scale score level or total Numeratum score level. Meijer et al. (2017) found that "subscores provided added value over the total score if and only if $PRMSE_s$[12] is larger than $PRMSE_x$" (p. 3). When using the *prmse.subscores.scales* function in the *sirt* (Robitzsch, 2022) package, the symbol X denotes the subscale and Z the full scale. Table 30 reports these values.

Table 30. *Haberman's Subscale Scoring Test Results*

Scale	$PRMSE_X$	$PRMSE_Z$
Number Problems	0.56	0.77
Patterns	0.72	0.82
Interpretation	0.64	0.81

Note. PRMSE = Proportional reduction of mean squared error. Meijer et al. (2017) refer to the subscores/subscales as $PRMSE_s$ and the total score/full scale as $PRMSE_X$. The sirt R package refers to the subscores/subscales as $PRMSE_X$ and the total score/full scale as $PRMSE_Z$.

As per Table 30, none of the $PRMSE_X$ values exceeded the $PRMSE_Z$ values, implying that the Numeratum's total score should rather be interpreted than its scale scores.

Rasch Analysis

A Rasch (1960) analysis was conducted on the total Numeratum score to inspect item fit statistics and item locations (difficulties) in Winsteps version 4.6.1 (Linacre, 2020a). Depending on the circumstances, different Infit (IMNSQ) and Outfit (OMNSQ) mean square values may signal underfitting or overfitting items (Aryadoust et al., 2020). 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). As criteria to assess item fit, items with mean square (infit/outfit) values ≥ 1.40 were indicative of potential underfit, whereas items with mean square (infit/outfit) values ≤ 0.60 signalled potential overfit (Bond & Fox,

[12] $PRMSE_s$ refer to the subtest score.

2015). However, as overfit is typically deemed less worrisome than underfit (Tesio et al., 2023), greater focus was placed on mean square (infit/outfit) values > 1.00. Consequently, as additional criteria, OMNSQ values ≥ 1.30 were inspected first, followed by an inspection of IMNSQ values ≥ 1.10 to identify misfitting items. Table 31 provides the item fit statistics and item locations for the total Numeratum score. Item and person reliabilities were provided earlier (see Table 28).

Table 31. *Total Numeratum Score Item Location and Item Fit Statistics*

Item	Location	SE	IMNSQ	Z	OMNSQ	Z	PT Corr.	Exp.
NP4	-1.36	.09	0.99	-0.21	0.94	-0.45	0.41	0.40
NP6	0.38	.07	1.13	3.88	1.16	2.95	0.49	0.55
NP7	-0.79	.08	0.84	-4.03	0.69	-3.78	0.54	0.46
NP12	-1.20	.09	1.08	1.66	1.12	1.06	0.38	0.42
NP13	0.32	.07	1.15	4.49	1.24	4.27	0.47	0.55
P5	-0.37	.07	0.77	-6.90	0.63	-5.78	0.61	0.50
P7	0.51	.07	0.91	-3.12	0.86	-2.84	0.61	0.56
P8	0.51	.07	1.16	5.01	1.22	4.13	0.48	0.56
P9	0.58	.07	0.98	-0.63	0.95	-1.10	0.58	0.56
P10	-1.12	.08	0.82	-4.15	**0.56**	-4.73	0.52	0.42
P11	1.22	.07	1.10	3.11	1.18	3.41	0.54	0.59
NI5	-0.44	.07	0.94	-1.54	0.84	-2.15	0.52	0.49
NI8	1.42	.07	0.99	-0.31	1.02	0.33	0.60	0.60
NI9	0.81	.07	1.01	0.33	1.05	1.11	0.57	0.57
NI10	-0.99	.08	0.97	-0.68	1.03	0.28	0.45	0.44
NI17	0.49	.07	1.08	2.58	1.13	2.58	0.52	0.56

Note. OMNSQ ≥ 1.40 or ≤ 0.60 in bold. Location = Item location, SE = Standard Error, IMNSQ = Infit Mean Square Values, Z = z-standardised statistics, OMNSQ = Outfit Mean Square Values, PT Corr. = Point-Measure Correlation, Exp. = Expected value. NP = Number Problems, P = Patterns, NI = Interpretation.

As per Table 31, the item locations ranged between -1.36 and 1.42 logits, mostly covering the underlying ability trait level of the respondents. No items displayed underfit, whereas one item (P10) demonstrated overfit as per the OMNSQ ≥ 1.40 or ≤ 0.60 guidelines. Regarding OMNSQ ≥ 1.30 and IMNSQ ≥ 1.10 values, no items breached these thresholds, although items NP13 and P8 came fairly close.

To assess unidimensionality, principal component analysis was conducted on the standardised residuals. The Eigenvalue for the first contrast (1.51) did not exceed 2, providing evidence of unidimensionality (e.g., Raîche, 2005). Furthermore, the local independence of items was assessed by looking at the largest standardised residual correlations. Items P5 and P10 had the largest standardised residual correlation (0.21), which is lower than the typical 0.70 guideline (Linacre, 2020b). Yen's Q3 statistic for the correlation between P5 and P10 was 0.21, which is lower than typical suggestions of 0.30 (Aryadoust et al., 2020). Consequently, there were no obvious indications of local dependence (i.e., participants' responses to one item seemed independent to their responses to other items).

Construct Validity

Regarding the factor structure of the Numeratum, findings from the previous technical manual (van Zyl & Taylor, 2015) suggested that a bifactor exploratory structural equation model (bifactor ESEM) offers the best representation of the data. Therefore, a bifactor ESEM model was specified with the weighted least square mean and variance adjusted (WLSMV) estimator in Mplus version 8.4 (Muthén & Muthén, 2012–2019). The model's performance was assessed through the following commonly reported fit metrics: comparative fit index (CFI), Tucker-Lewis index (TLI), the root mean square error of approximation (RMSEA), and the standardised root mean square residual (SRMR). Values close to 0.95 (CFI and TLI), 0.06 (RMSEA), and 0.08 (SRMR) generally indicate good model fit (Hu & Bentler, 1999). Additionally, a 1-factor, correlated 3-factor, bifactor confirmatory factor analytic (bifactor CFA) model, and an exploratory structural equation model (ESEM) were specified for comparative rather than interpretive purposes. Table 32 reports the results of the specified models.

Table 32. *Fit Statistics of Different Factor Models*

Model	$\chi 2$	*df*	CFI	TLI	RMSEA	90% CI	SRMR
1 Factor	274.217*	104	0.980	0.977	0.035	0.030, 0.040	0.050
3 Factor	218.474*	101	0.986	0.983	0.030	0.024, 0.035	0.045
Bifactor CFA	134.170*	88	0.994	0.992	0.020	0.013, 0.027	0.035
ESEM	106.768*	75	0.996	0.994	0.018	0.009, 0.025	0.029
Bifactor ESEM	69.415*	62	0.999	0.998	0.010	0.000, 0.020	0.022

Note. $\chi 2$ = Chi-square, *df* = Degrees of Freedom, CFI = Comparative Fit Index, TLI = Tucker-Lewis Index, RMSEA = Root Mean Square Error of Approximation with 90% Confidence Intervals, SRMR = Standardised Root Mean Square Residual.

As per Table 32, the bifactor ESEM model's fit statistics comfortably exceeded CFI and TLI values of 0.95, and also comfortably fell beneath the RMSEA and SRMR thresholds of 0.06 and 0.08, respectively. As the inter-factor correlations of bifactor models are constrained to zero, the correlated 3-factor CFA model and ESEM model's inter-factor correlations were compared. Table 33 reports these correlations.

Table 33. *Standardised Inter-Factor Correlations for the Numeratum Scales*

Scale	Number Problems	Patterns	Interpretation
Number Problems	-	0.56*	0.67*
Patterns	0.85*	-	0.75*
Interpretation	0.87*	0.91*	-

Note. The correlated 3-factor CFA model is below the diagonal, inter-factor correlations from the ESEM model are above the diagonal. *$p < 0.001$.

On average, the sizes of the ESEM model's ($M_r = 0.66$) inter-factor correlations were reasonably lower than the correlated 3-factor CFA model ($M_r = 0.88$). Howard et al. (2018) propose that "ESEM tends to provide more exact estimates of true factor correlations" (p. 2649) compared to CFA and "that ESEM should be retained whenever the results show a discrepant pattern of factor correlations" (p. 2650). To decide between the bifactor ESEM and ESEM model, the former did not display significantly better fit than the latter. Therefore, cross-loadings between the models were compared. The ESEM model generally displayed higher cross-loadings than the bifactor ESEM model, providing a potential indication of an unmodelled general factor (Howard et al., 2018). This offered some support for using the bifactor ESEM model. Table 34 reports the standardised factor loadings, standard errors, item uniqueness or bifactor standardised residual variance, and the item explained common variance (IECV) for the bifactor ESEM model.

Table 34. *Bifactor ESEM Model Statistics*

	General		Number Problems		Patterns		Interpretation			
Item	λ	S.E.	λ	S.E.	λ	S.E.	λ	S.E.	δ	IECV
NP4	**0.61***	0.04	**0.24***	0.10			-0.19*	0.07	0.53	0.80
NP6	**0.56***	0.03	**-0.06**	0.07					0.68	0.98
NP7	**0.74***	0.04	**0.33***	0.07			0.11*	0.06	0.33	0.82
NP12	**0.52***	0.05	**0.55***	0.11					0.41	0.47
NP13	**0.60***	0.04	**-0.29***	0.14	-0.15*	0.05	-0.12*	0.06	0.52	0.76
P5	**0.79***	0.03			**0.36***	0.06			0.24	0.82
P7	**0.69***	0.03			**0.29***	0.05			0.43	0.84
P8	**0.47***	0.04			**0.36***	0.06			0.64	0.61
P9	**0.67***	0.03			**0.14***	0.06			0.53	0.96
P10	**0.72***	0.04	-0.15*	0.05	**0.53***	0.08			0.18	0.64
P11	**0.60***	0.03			**-0.05**	0.07			0.63	0.99
NI5	**0.64***	0.03					**0.48***	0.11	0.35	0.64
NI8	**0.66***	0.03					**0.14***	0.07	0.54	0.95
NI9	**0.62***	0.03					**0.29***	0.07	0.53	0.81
NI10	**0.56***	0.04			0.31*	0.06	**0.17***	0.08	0.56	0.71
NI17	**0.55***	0.03					**0.23***	0.07	0.64	0.83

Note. *$p < 0.05$. λ = Standardised factor loadings, S.E. = standard error, R^2 = R-squared value, δ = item uniqueness/bifactor standardised residual variance, IECV = item explained common variance. Statistically significant cross-loadings are underlined. Standardised factor loadings for specific factors are indicated in bold. Standardised factor loadings that were not statistically significant, together with their standard errors were removed. IECV values were derived from Dueber's (2017) Bifactor Indices Calculator in Excel. NP = Number Problems, P = Patterns, NI = Interpretation.

As per Table 34, the general factor loadings were all statistically significant ($p < 0.001$), ranging from 0.47 to 0.79. These were considered acceptable as per Spector's (1992) suggestion of a minimum value of 0.30 to 0.35 for an item to load onto a factor. The standard errors were also generally low (≤ 0.05). Except for item NP12, all items had larger general than specific factor loadings. The specific factors were fairly weakly defined compared to the general factor. Statistically significant standardised factor loadings were found for four of the five Number Problems items, five of the six Patterns items, and all five Interpretation items. Hence, 14 of the 16 Numeratum items loaded significantly on their intended target factor, although only six items had standardised factor loadings above 0.30. All item uniqueness values fell within an acceptable range (> 0.10 δ < 0.90; van Zyl & ten Klooster, 2022). One item (NP12 = 0.47) had an IECV value < 0.50. Item P8 had the lowest general factor loading (λ=0.47). Statistically significant cross-loadings ranged from 0.11 to -0.19 (four items), except for item NI10 which had a reasonable cross-loading of 0.31. For the most part, the values of the cross-loadings were lower than the target loadings.

Furthermore, the orthogonally rotated factor loadings obtained from Mplus were used to calculate other bifactor indices as reported in Table 35. Cross-loadings were ignored in calculating the specific factors' reliability (Morin et al., 2020). The Bifactor Indices Calculator in Excel (Dueber 2017) was used for this purpose.

Table 35. *Bifactor Indices for the Bifactor ESEM Model*

Factor	ECV	ωh	ωRel.	H	FD
General Factor	0.77	0.89	0.95	0.92	0.95
Number Prob.	0.07	0.05	0.06	0.44	0.75
Patterns	0.10	0.13	0.15	0.49	0.79
Interpretation	0.06	0.13	0.16	0.37	0.69

Note. ECV=Explained Common Variance, ωh=Coefficient Omega Hierarchical, ωRel.=Relative Omega, H = Construct Replicability, FD = Factor Determinacy.

As per Table 35, the general factor explained 77% of the common variance. The group factors' explained variance ranged from 6 to 10%. Coefficients omega hierarchical and relative were 0.89 and 0.95, respectively. The general factor was the only well-defined factor ($H > 0.80$; Rodriguez et al., 2016a) with a coefficient of 0.92. The percentage of uncontaminated correlations (PUC[13]) was 0.71. Rodriguez et al. (2016a) propose that "when ECV is > .70 and PUC > .70, relative bias will be slight and the common

[13] Cross-loadings were excluded to calculate the PUC.

variance can be regarded as essentially unidimensional" (p. 232). Furthermore, the absolute relative parameter bias (ARPB) was 6.5%, implying that the items' unidimensional factor loadings did not substantially differ from their general factor loadings ARPB < 10-15% (Rodriguez et al., 2016b).

To gain additional insights into the validity of the bifactor ESEM model, the data were analysed in FACTOR version 12.04.01 (Lorenzo-Seva & Ferrando, 2023) with the following model specifications: matrix analysed = polychoric matrix (tetrachoric) with sweet smoothing; estimation = Robust diagonally weighted least squares (RDWLS); and rotation = Orthogonal Procrustean rotation. The adequacy of the polychoric correlation matrix was as follows: Bartlett's statistic = 9860.5 (df = 120, p < 0.001); Kaiser-Meyer-Olkin (KMO) test = 0.92 (which is considered very good). Furthermore, results showed that none of the items should be removed based on their Measure of Sampling Adequacy (MSA) values. MSA values below 0.50 suggest that the item does not measure the same domain as the remaining items in the pool and should probably be removed (Lorenzo-Seva & Ferrando, 2021). Goodness-of-fit metrics indicated a close fit to the specified model: Root Mean Square Error of Approximation (RMSEA) = 0.011; Comparative Fit Index (CFI) = 0.999; Non-Normed Fit Index (NNFI) = 0.998.

Overall, to ascertain whether scale scores, a total score, or both should be interpreted, results as gathered from reliability indicators, Haberman's test, Rasch analysis, and bifactor analysis were examined and suggest that a total Numeratum score should be interpreted. More research is needed to determine the value-add of the specific factors beyond the general factor.

Item difficulty and item discrimination

Item difficulty 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 in relation 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, 1970). According to Kerlinger and Lee (2000) item difficulties should range between 0.50 and 0.70, 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, the item difficulties would be expected to have a larger range. Table 36 provides the item-rest correlation as calculated in jamovi version 2.3.28 (The jamovi

project, 2023), as well as item difficulty and item discrimination values for the total Numeratum score using the *item.exam* function in the *psychometric* (Fletcher, 2022) package.

Table 36. *Item-Rest Correlations, Item Difficulty, and Item Discrimination for the Total Numeratum Score Items*

Item	Item-rest cor.	Difficulty	Discrimination
NP4	0.36	0.86	0.31
NP6	0.39	0.63	0.55
NP7	0.52	0.80	0.53
NP12	0.31	0.84	0.31
NP13	0.37	0.64	0.51
P5	0.60	0.74	0.64
P7	0.54	0.61	0.70
P8	0.37	0.61	0.55
P9	0.50	0.60	0.66
P10	0.51	0.83	0.43
P11	0.41	0.50	0.62
NI5	0.48	0.75	0.53
NI8	0.48	0.46	0.68
NI9	0.47	0.56	0.67
NI10	0.41	0.82	0.39
NI17	0.42	0.61	0.58

As per Table 36, the average item difficulty was 0.68 and the average item discrimination was 0.54. No items had item-rest correlation values below a minimally acceptable benchmark of 0.20 (Zijlmans et al., 2018). The average item-rest correlation was 0.44.

Differential item functioning

Differential item functioning (DIF) through ordinal logistic regression was investigated with the *rundif* function in the *lordif* (Choi et al., 2016) package. The Rasch Person measures, as exported from Winsteps, were used as the conditioning variable. Three models were compared (baseline, uniform DIF, and non-uniform DIF). The first and third models were compared first to establish an overall DIF effect size. Thereafter, the DIF was examined to determine whether it was uniform or non-uniform. The statistical significance value was set to $p < 0.001$ (as opposed to $p < 0.05$) with consideration for Type I

errors. A change in Nagelkerke's pseudo R-Squared (R^2) across the models was assessed to establish the magnitude of DIF. The effect size guidelines of Jodoin and Gierl (2001) were used in this regard: negligible ($R^2 < 0.035$), moderate ($R^2 = 0.035$ to 0.070), and large ($R^2 > 0.070$). DIF was investigated in a pairwise manner for gender (female vs. male), ethnicity (Black African vs. White), language (English vs. Zulu, and English vs. Afrikaans), and education (Grade 12 vs. Diploma/Certificate, and Bachelor's degree vs Honours degree)[14]. Table 37 to Table 42 report these results.

Table 37. *Differential Item Functioning Across Gender Groups*

	p-values for χ2 difference tests			Change in Nagelkerke's R^2		
Item	M1-M2	M1-M3	M2-M3	M1-M2	M1-M3	M2-M3
			Gender			
NP4	0.556	0.373	0.202	0.000	0.002	0.002
NP6	0.125	0.111	0.154	0.002	0.003	0.002
NP7	0.333	0.620	0.888	0.001	0.001	0.000
NP12	0.564	0.023	0.007	0.000	0.008	0.008
NP13	0.461	0.738	0.800	0.000	0.000	0.000
P5	0.639	0.092	0.033	0.000	0.003	0.003
P7	0.035	0.100	0.680	0.003	0.003	0.000
P8	0.455	0.753	0.926	0.000	0.000	0.000
P9	0.028	0.051	0.290	0.003	0.004	0.001
P10	0.014	0.014	0.113	0.005	0.008	0.002
P11	0.382	0.452	0.364	0.001	0.001	0.001
NI5	0.552	0.343	0.182	0.000	0.002	0.001
NI8	0.003	0.002	0.075	0.006	0.008	0.002
NI9	0.555	0.174	0.076	0.000	0.002	0.002
NI10	0.385	0.159	0.087	0.001	0.004	0.003
NI17	0.655	0.859	0.747	0.000	0.000	0.000

Note. M1-M2 = Model 1 vs. Model 2 (uniform DIF), M1-M3 = Model 1 vs. Model 3 (total DIF), M2-M3 = Model 2 vs. Model 3 (non-uniform DIF).

As per Table 37, no notable differences were observed on any items for the gender groups ($R^2 < 0.035$).

[14] For DIF between groups not mentioned here, see Appendix A.

Table 38. *Differential Item Functioning Across Ethnic Groups*

	p-values for χ2 difference tests			Change in Nagelkerke's R^2		
Item	M1-M2	M1-M3	M2-M3	M1-M2	M1-M3	M2-M3
			Ethnicity			
NP4	0.004	0.003	0.060	0.015	0.021	0.006
NP6	0.909	0.318	0.131	0.000	0.003	0.003
NP7	**0.000**	**0.000**	0.216	**0.035**	**0.037**	0.002
NP12	0.006	**0.000**	0.002	0.014	0.032	0.018
NP13	0.005	0.017	0.590	0.011	0.011	0.000
P5	0.100	0.184	0.411	0.003	0.004	0.001
P7	0.006	0.019	0.567	0.008	0.009	0.000
P8	0.057	0.074	0.210	0.005	0.007	0.002
P9	0.494	0.714	0.650	0.000	0.001	0.000
P10	0.004	0.008	0.297	0.012	0.013	0.002
P11	0.264	0.298	0.279	0.002	0.003	0.002
NI5	0.070	0.166	0.585	0.004	0.005	0.000
NI8	0.645	0.627	0.396	0.000	0.001	0.001
NI9	0.097	0.105	0.186	0.003	0.005	0.002
NI10	0.025	0.003	0.009	0.008	0.018	0.010
NI17	0.196	0.281	0.352	0.002	0.003	0.001

Note. M1-M2 = Model 1 vs. Model 2 (uniform DIF), M1-M3 = Model 1 vs. Model 3 (total DIF), M2-M3 = Model 2 vs. Model 3 (non-uniform DIF).

As per **Table 38**, item NP7 was the only item that showed indications of DIF between the Black African and White ethnic groups. The effect size of the DIF is marginally above being negligible (R^2 = 0.037). Consequently, this does not warrant serious concern. Similar results were found for item NP7 between the Indian and White ethnic groups (R^2 = 0.037). No notable differences were however observed between the Black African and Indian ethnic groups (R^2 < 0.035) (see **Appendix A**).

Table 39. *Differential Item Functioning Across Language Groups*

	p-values for χ2 difference tests			Change in Nagelkerke's R^2		
Item	M1-M2	M1-M3	M2-M3	M1-M2	M1-M3	M2-M3
			Language (English vs. Zulu)			
NP4	0.252	0.448	0.590	0.003	0.004	0.001
NP6	0.324	0.601	0.830	0.002	0.002	0.000
NP7	0.715	0.892	0.757	0.000	0.000	0.000
NP12	0.997	0.266	0.104	0.000	0.007	0.007
NP13	0.087	0.200	0.591	0.004	0.005	0.000
P5	0.197	0.259	0.308	0.002	0.004	0.002
P7	0.674	0.577	0.336	0.000	0.002	0.001
P8	0.205	0.090	0.073	0.003	0.008	0.005
P9	0.863	0.673	0.382	0.000	0.001	0.001
P10	0.363	0.650	0.852	0.002	0.002	0.000
P11	0.326	0.589	0.757	0.001	0.002	0.000
NI5	0.166	0.225	0.301	0.003	0.005	0.002
NI8	0.996	0.864	0.589	0.000	0.000	0.000
NI9	0.716	0.437	0.217	0.000	0.002	0.002
NI10	0.350	0.264	0.180	0.002	0.006	0.004
NI17	0.122	0.165	0.270	0.004	0.006	0.002

Note. M1-M2 = Model 1 vs. Model 2 (uniform DIF), M1-M3 = Model 1 vs. Model 3 (total DIF), M2-M3 = Model 2 vs. Model 3 (non-uniform DIF).

As per Table 39, no notable differences were observed on any items between English and Zulu participants ($R^2 < 0.035$).

Table 40. *Differential Item Functioning Across Language Groups*

	p-values for χ2 difference tests			Change in Nagelkerke's R^2		
Item	M1-M2	M1-M3	M2-M3	M1-M2	M1-M3	M2-M3
			Language (English vs. Afrikaans)			
NP4	0.992	0.762	0.461	0.000	0.001	0.001
NP6	0.727	0.933	0.900	0.000	0.000	0.000
NP7	0.001	0.003	0.425	0.017	0.018	0.001
NP12	0.031	0.095	0.851	0.010	0.010	0.000
NP13	0.574	0.040	0.013	0.000	0.009	0.009
P5	0.099	0.054	0.078	0.004	0.008	0.004
P7	0.022	0.074	0.953	0.007	0.007	0.000
P8	0.006	0.024	0.980	0.011	0.011	0.000
P9	0.530	0.518	0.338	0.000	0.002	0.001
P10	0.342	0.563	0.619	0.002	0.002	0.000
P11	0.263	0.515	0.783	0.002	0.002	0.000
NI5	0.734	0.757	0.506	0.000	0.001	0.001
NI8	0.707	0.640	0.386	0.000	0.001	0.001
NI9	0.032	0.080	0.494	0.006	0.006	0.001
NI10	0.359	0.574	0.605	0.002	0.003	0.001
NI17	0.549	0.790	0.738	0.000	0.001	0.000

Note. M1-M2 = Model 1 vs. Model 2 (uniform DIF), M1-M3 = Model 1 vs. Model 3 (total DIF), M2-M3 = Model 2 vs. Model 3 (non-uniform DIF).

As per Table 40, no notable differences were observed on any items between English and Afrikaans participants. In addition to English and Zulu, and English and Afrikaans, differences between English and Pedi were also evaluated (see Appendix A). Similarly, no notable differences were observed ($R^2 < 0.035$).

Table 41. *Differential Item Functioning Across Educational Groups*

	p-values for χ2 difference tests			Change in Nagelkerke's R^2		
Item	M1-M2	M1-M3	M2-M3	M1-M2	M1-M3	M2-M3
Education (Grade 12 vs. Diploma/Certificate)						
NP4	0.992	0.944	0.735	0.000	0.000	0.000
NP6	0.865	0.423	0.193	0.000	0.006	0.006
NP7	0.419	0.374	0.251	0.002	0.006	0.004
NP12	0.461	0.567	0.442	0.003	0.005	0.003
NP13	**0.000**	0.001	0.948	**0.051**	**0.051**	0.000
P5	0.773	0.685	0.412	0.000	0.002	0.002
P7	0.688	0.886	0.775	0.000	0.001	0.000
P8	0.928	0.912	0.674	0.000	0.001	0.001
P9	0.628	0.862	0.805	0.001	0.001	0.000
P10	0.732	0.929	0.863	0.000	0.000	0.000
P11	0.629	0.302	0.141	0.001	0.008	0.007
NI5	0.999	0.853	0.573	0.000	0.001	0.001
NI8	0.319	0.570	0.715	0.004	0.004	0.000
NI9	0.094	0.112	0.211	0.010	0.016	0.006
NI10	0.011	0.027	0.371	0.022	0.024	0.003
NI17	0.518	0.170	0.077	0.001	0.012	0.010

Note. M1-M2 = Model 1 vs. Model 2 (uniform DIF), M1-M3 = Model 1 vs. Model 3 (total DIF), M2-M3 = Model 2 vs. Model 3 (non-uniform DIF).

As per **Table 41**, a moderate effect size of 0.051 was established for item NP13 between Grade 12 and Diploma/Certificate participants. The effect size leaned closer to the negligible cut-off ($R^2 > 0.35$) point than the large cut-off ($R^2 > 0.70$) point. Otherwise, no notable differences were observed ($R^2 < 0.035$).

Table 42. *Differential Item Functioning Across Educational Groups*

	p-values for χ2 difference tests			Change in Nagelkerke's R^2		
Item	M1-M2	M1-M3	M2-M3	M1-M2	M1-M3	M2-M3
Education (Bachelor's degree vs. Honours degree)						
NP4	0.798	0.967	0.974	0.000	0.000	0.000
NP6	0.489	0.491	0.332	0.002	0.006	0.004
NP7	0.458	0.669	0.615	0.005	0.007	0.002
NP12	0.983	0.876	0.607	0.000	0.002	0.002
NP13	0.086	0.215	0.724	0.013	0.014	0.001
P5	0.987	0.999	0.961	0.000	0.000	0.000
P7	0.045	0.127	0.766	0.014	0.014	0.000
P8	0.601	0.286	0.136	0.001	0.010	0.009
P9	0.602	0.793	0.662	0.001	0.002	0.001
P10	0.336	0.420	0.368	0.007	0.012	0.006
P11	0.654	0.606	0.371	0.001	0.003	0.003
NI5	0.098	0.086	0.142	0.015	0.027	0.012
NI8	0.883	0.558	0.285	0.000	0.004	0.004
NI9	0.990	0.283	0.112	0.000	0.010	0.010
NI10	0.184	0.330	0.502	0.011	0.013	0.003
NI17	0.730	0.918	0.819	0.000	0.001	0.000

Note. M1-M2 = Model 1 vs. Model 2 (uniform DIF), M1-M3 = Model 1 vs. Model 3 (total DIF), M2-M3 = Model 2 vs. Model 3 (non-uniform DIF).

As per Table 42, no notable differences were observed between participants with Bachelor's degrees and those with Honours degrees ($R^2 < 0.035$).

In summary, from a DIF perspective, no notable concerns were observed in terms of how the items functioned across different gender, and specific ethnic and language groups. This offers preliminary support to refrain from developing gender-, ethnic-, and language-based norms.

Measurement invariance

Putnick and Bornstein (2016) propose that before any meaningful comparisons between groups can be made, measurement invariance should be established. "Measurement invariance assesses the

(psychometric) equivalence of a construct across groups …" (Putnick & Bornstein, 2016, p. 72). Widaman and Reise's (1997) steps were followed for measurement invariance testing. Hence, configural, metric (or weak factorial), scalar (or strong factorial), and strict (or residual/invariant uniqueness) models were specified and tested sequentially in Mplus with the WLSMV estimator. In each model, an additional constraint was added. De Beer and Morin's (2022) (B) ESEM invariance syntax generator for Mplus was used but had to be slightly modified to accommodate the dichotomous nature of the data. To draw parallels between the models, delta changes (Δ) in CFI, TLI, and RMSEA were assessed. Changes of-0.01 (CFI and TLI) and 0.015 (RMSEA) from one model to the next signalled noninvariance (Chen, 2007; Cheung & Rensvold, 2002). Table 43 reports the measurement invariance results for gender, and specific ethnic, language, and educational groups.

Table 43. *Bifactor ESEM Measurement Invariance Testing for Gender, Ethnicity, Language, and Education*

Model	χ^2	*df*	CFI	TLI	RMSEA	SRMR	CM	ΔCFI	ΔTLI	ΔRMSEA
Gender (Women, *n* = 639; Men, *n* = 675)										
M1: Configural	125.285*	124	1.000	1.000	0.004 [0.000, 0.020]	0.030	-	-	-	-
M2: Metric	183.467*	168	0.998	0.997	0.012 [0.000, 0.022]	0.039	M1	-0.002	-0.003	0.008
M3: Scalar	198.343*	172	0.997	0.995	0.015 [0.000, 0.024]	0.042	M2	-0.001	-0.002	0.003
M4: Strict	214.965*	184	0.996	0.995	0.016 [0.000, 0.024]	0.042	M3	-0.001	0.000	0.001
Ethnicity (Black African, *n* = 588; White, *n* = 188)										
M1: Configural	126.417*	124	0.999	0.999	0.007 [0.000, 0.026]	0.039	-	-	-	-
M2: Metric	182.012*	168	0.997	0.995	0.015 [0.000, 0.028]	0.053	M1	-0.002	-0.004	0.008
M3: Scalar	186.500*	172	0.997	0.995	0.015 [0.000, 0.028]	0.053	M2	0.000	0.000	0.000
M4: Strict	209.556*	184	0.994	0.992	0.019 [0.000, 0.030]	0.055	M3	-0.003	-0.003	0.004
Language (English, *n* = 588; Zulu, *n* = 142; Afrikaans, *n* = 224)										
M1: Configural	177.077*	186	1.000	1.000	0.000 [0.000, 0.022]	0.044	-	-	-	-
M2: Metric	281.071*	282	1.000	1.000	0.000 [0.000, 0.023]	0.064	M1	0.000	0.000	0.000
M3: Scalar	300.502*	290	0.998	0.997	0.011 [0.000, 0.026]	0.063	M2	-0.002	-0.003	0.011
M4: Strict	303.270*	306	1.000	1.000	0.000 [0.000, 0.022]	0.065	M3	0.002	0.003	-0.011
Education (Grade 12 + Diploma/Certificate, *n* = 301; Bachelor's/Honours degree, *n* = 273)										
M1: Configural	103.978*	124	1.000	1.000	0.000 [0.000, 0.013]	0.048	-	-	-	-
M2: Metric	163.896*	168	1.000	1.000	0.000 [0.000, 0.025]	0.062	M1	0.000	0.000	0.000
M3: Scalar	165.294*	172	1.000	1.000	0.000 [0.000, 0.023]	0.066	M2	0.000	0.000	0.000
M4: Strict	183.183*	184	1.000	1.000	0.000 [0.000, 0.026]	0.068	M3	0.000	0.000	0.000

Note. χ^2 = Chi-square, *df* = Degrees of Freedom, CFI = Comparative Fit Index, TLI = Tucker-Lewis Index, SRMR = Standardised Root Mean Square Residual, RMSEA = Root Mean Square Error of Approximation with 90% Confidence Intervals, CM = Comparison Model, ΔCFI = Change in CFI, ΔTLI = Change in TLI, ΔRMSEA = Change in RMSEA.

As per Table 43, all models displayed good fit statistics (CFI, TLI > 0.95; RMSEA < 0.06; SRMR < 0.08; Hu & Bentler, 1999). Furthermore, no problematic delta changes were observed for CFI (Δ-0.01), TLI (Δ-0.01), or RMSEA (Δ 0.015) from one model to the next (Chen, 2007; Cheung & Rensvold, 2002). Hence, strict measurement invariance for the Numeratum was established across gender, and specific ethnic, language, and educational groups. This meant that meaningful group comparisons could be made at the total score level.

Mean differences across groups

Before mean group comparisons were made, the homogeneity of variance assumption was tested with the *leveneTest* function in the *car* (Fox & Weisberg, 2019) package. Results from Levene's (1960) test showed that the homogeneity of variance assumption was violated across almost all groups. Therefore, depending on the number of groups (two or more), either Welch two-sample t-tests or Welch ANOVAs were performed to assess mean group differences. These were calculated with the *t.test* function (two groups) or the *oneway.test* function (> two groups) in the *stats* (R Core Team, 2023) package. Post hoc tests were carried out with the *posthoc_anova* function in the *biostat* (Gegzna, 2020) package. Table 44 reports the results of the Welch two-sample t-test for the gender groups.

Table 44. *Mean Differences Between Gender Groups*

	Men		Women		x̄ diff.	*t*	*df*	Effect size [a]
	M	*SD*	*M*	*SD*				
Total	11.50	3.71	10.21	3.88	1.29	6.187	1299.1	0.34*

Note. x̄ diff. = Mean difference between comparison groups, *t* = Welch statistic, *df* = Degrees of freedom, Effect size = Cohen's *d*. *$p < 0.05$.

[a] Values of 0.20, 0.50, and 0.80 correspond to small-, medium-, and large effect sizes (Cohen, 1988).

As per Table 44, statistically significant mean differences were found between male and female participants. A Cohen's *d* value of 0.34 indicated a small effect size (Cohen, 1988). Other descriptive statistics showed that obtaining correct answers on all items of the test were accomplishable for both gender groups. In the following analyses, more than two groups were involved, hence Welch ANOVAs were performed.

Results from the Welch ANOVA showed statistically significant mean differences between ethnic groups [$F(2, 302.4) = 38.64$, $p < 0.001$], language groups [$F(3, 298.39) = 24.194$, $p < 0.001$], and educational groups [$F(3, 308.34) = 68.318$, $p < 0.001$]. Post hoc Games-Howell tests showed that statistically

significant mean differences were found between different ethnic, language, and educational groups that had at least 100 observations. For practical purposes, only statistically significant results were displayed. Table 45 reports these results.

Table 45. *Mean Differences Between Ethnic, Language, and Educational Groups*

Mean differences based on Ethnicity												
Group	Ethnicity	*n*	Mean	Variance	*SD*	Comp.	x̄ diff.	LCI	UCI	*t*	*df*	*p*
1	Black	588	10.015	15.191	3.898	2 vs. 1	1.293	0.561	2.026	4.16	333.76	<0.001
2	White	188	11.309	13.348	3.654	3 vs. 1	2.769	2.006	3.532	8.57	213.96	<0.001
3	Indian	125	12.784	9.832	3.136	3 vs. 2	1.475	0.564	2.387	3.81	291.44	<0.001
Mean differences based on Home Language												
Group	Language	*n*	Mean	Variance	*SD*	Comp.	x̄ diff.	LCI	UCI	*t*	*df*	*p*
1	English	501	12.162	11.780	3.432	2 vs. 1	-2.612	-3.493	-1.731	7.68	218.04	<0.001
2	Zulu	142	9.549	13.100	3.619	3 vs. 1	-1.077	-1.830	-0.324	3.69	399.62	0.001
3	Afrikaans	224	11.085	13.809	3.716	4 vs. 1	-1.932	-2.997	-0.867	4.72	133.26	<0.001
4	Pedi	100	10.230	14.401	3.795	3 vs. 2	1.536	0.522	2.549	3.91	306.00	<0.001
Mean differences based on Education												
Group	Education	*n*	Mean	Variance	*SD*	Comp.	x̄ diff.	LCI	UCI	*t*	*df*	*p*
1	Grade 12	166	8.133	12.879	3.589	3 vs. 1	4.098	3.142	5.054	11.07	316.72	<0.001
2	Diploma	135	8.756	13.708	3.702	4 vs. 1	4.363	3.361	5.366	11.25	275.18	<0.001
3	Degree	156	12.231	9.263	3.043	3 vs. 2	3.475	2.438	4.512	8.66	259.75	<0.001
4	Honours	117	12.496	8.528	2.920	4 vs. 2	3.740	2.660	4.820	8.96	247.88	<0.001

Note. *n* = Group sample size, *SD* = Standard Deviation, Comp. = Groups being compared, x̄ diff. = Mean difference between comparison groups, LCI = Lower Confidence Interval of mean difference, UCI = Upper Confidence Interval of mean difference, *t* = t-statistic as derived from Games-Howell test, *df* = Degrees of Freedom, Black = Black African.

At this stage, we only knew that statistically significant mean differences existed and between which groups. To assess the magnitude of these mean differences, effect sizes were calculated in a pairwise manner with the *cohens_d* function in the *rstatix* (Kassambara, 2023) package. Cohen's (1988) guidelines were followed in this regard to establish small- (0.20), medium- (0.50), and large (0.80) effect sizes.

Regarding ethnicity, small effects were found between Black African and White (d = 0.34), and White and Indian groups (d = 0.43). A medium effect was found between Black African and Indian groups (d = 0.78). Other descriptive statistics showed that obtaining correct answers on all items of the test were accomplishable for all the preceding ethnic groups. The Black African and White groups had higher variability in their scores compared to the Indian group. Furthermore, although many ethnic groups did not disclose their highest level of education, the Indian group had the largest percentage (56.90%) of participants that fell within the Bachelor's or Honours degree group. Although medium effect sizes were found, we ascribe this mainly to variability in scores across the different groups, as well as educational differences. Hence, at this stage, we believe that ethnic-based norms are not justified.

Regarding home language groups, small effect sizes were observed between Afrikaans and English (d = 0.30), and Afrikaans and Zulu groups (d = 0.42). Medium effect sizes were observed between English and Zulu (d = 0.74), and between English and Pedi groups (d = 0.53). No statistically significant differences were observed between Afrikaans and Pedi, or between the Zulu and Pedi groups ($p > 0.05$). Other descriptive statistics showed that obtaining correct answers on all items of the test were accomplishable for all the preceding language groups. The Zulu, Afrikaans, and Pedi groups had higher variability in their scores compared to the English group. Furthermore, although many language groups did not disclose their highest level of education, the English group had the largest percentage (51.48%) of participants that fell within the Bachelor's or Honours degree group. Although medium effect sizes were found, we ascribe this mainly to variability in scores across the different groups, as well as educational differences. Hence, at this stage, we believe that language-based norms are not justified.

Regarding educational level, large effect sizes were observed between Grade 12 participants and those with a Bachelor's (d = 1.23) or Honours degree (d = 1.33). Large effect sizes were also observed between participants with a Diploma or Certificate and those with a Bachelor's (d = 1.03) or Honours degree (d = 1.12). No statistically significant differences were observed between the Grade 12 and Diploma or Certificate groups, or between the Bachelor's degree or Honours degree groups ($p > 0.05$). Implications regarding the effect sizes are discussed in the norms section.

CHAPTER 8: CORRELATION BETWEEN THE VERBATIM AND NUMERATUM

To enable comparisons between the Verbatim and the Numeratum, the data of participants who completed all items of both the shortened questionnaires were merged (*n* = 1263). Next, bifactor ESEM models were separately analysed in Mplus for the Verbatim and Numeratum. As part of these analyses, factor scores were saved. Gorsuch (1983) advises to only use factor score estimates when factor determinacy (FD) values exceed 0.90. The FD values were 97.1 (Verbatim) and 95.5 (Numeratum), respectively. A correlation of 0.76 was established based on the factor scores obtained for the general factors of the respective questionnaires. Hence, regarding discriminant validity, the preceding correlation seemed adequate as per the < 0.80 threshold (Rönkkö & Cho, 2022).

As bifactor ESEM models constrain inter-factor correlations to zero, an additional analysis was conducted in *R* to assess Pearson and Spearman's Rank correlations between the Verbatim- and Numeratum's total and scale scores using the *rcorr* function in the *Hmisc* (Harrell, 2023) package. **Table 46** reports these results. The correlations predominantly had medium to large effect sizes (Cohen, 1988). This confirmed the relatedness, yet uniqueness of the scales.

Table 46. *Pearson and Spearman's Rank Correlations between the Verbatim and Numeratum Scales*

	S	O	A	R	VI	V Tot	NP	P	NI	N Tot
S	-	-	-	-	-	-	0.47	0.49	0.47	0.56
O	-	-	-	-	-	-	0.47	0.54	0.54	0.61
A	-	-	-	-	-	-	0.51	0.58	0.56	0.66
R	-	-	-	-	-	-	0.48	0.58	0.57	0.65
VI	-	-	-	-	-	-	0.45	0.50	0.55	0.59
V Tot	-	-	-	-	-	-	0.59	0.69	0.68	0.78
NP	0.47	0.46	0.49	0.47	0.45	0.59	-	-	-	-
P	0.52	0.55	0.58	0.57	0.50	0.69	-	-	-	-
NI	0.49	0.53	0.55	0.56	0.55	0.68	-	-	-	-
N Tot	0.59	0.61	0.64	0.63	0.59	0.78	-	-	-	-

Note. Pearson correlations are below the diagonal, Spearman's rank correlations are above the diagonal. S = Synonyms, O = Opposites, A = Analogies, R = Reasoning, VI = Verbal Interpretation, V Total = Verbatim Total, NP = Number Problems, P = Patterns, NI = Numerical Interpretation, N Tot = Numeratum Total. Values of 0.10, 0.30, and 0.50 correspond to small-, medium-, and large effects (Cohen, 1988). All correlations significant at $p < 0.001$ level.

CHAPTER 9: NORMS

As observed in the previous section, the mean differences between the Grade 12 vs. the Diploma/Certificate group were not statistically significant. Similarly, there were no statistically significant mean differences between the Bachelor's degree and Honour's degree group. There were however mostly large mean differences between Grade 12 vs. Bachelor's degree or Honour's degree groups. The same applied to the Diploma/Certificate group in comparison to the latter. This, accompanied by no serious concerns regarding differential item functioning and measurement invariance led to the development of two norm groups, one where Grade 12 and those with a Diploma/Certificate are grouped together, and one where those with a Bachelor's degree or Honours degree are grouped together. In the development of these norms, all available data from those who fell into the preceding categories were used (see **Table 47** for a sociodemographic composition of the norm groups). All norms were developed with Stanscore 5 (Barrett, 2018) and were hand smoothed afterwards.

Table 47. *Sociodemographic Composition of the Norm Groups*

Variable	Grade 12/Diploma (*N* = 1076)		Bachelor's/Honours (*N* = 1127)	
	Verbatim			
Gender	*n*	%	*n*	%
Women	565	52.5%	622	55.2%
Men	511	47.5%	505	44.8%
Ethnicity	*n*	%	*n*	%
Black African	649	62.0%	695	62.7%
White	191	18.3%	175	15.8%
Coloured	75	7.2%	66	6.0%
Indian/Asian	75	7.2%	123	11.1%
Other	56	5.4%	50	4.5%
Language	*n*	%	*n*	%
English	279	25.9%	350	31.1%
Zulu	164	15.2%	144	12.8%
Afrikaans	152	14.1%	107	9.5%
Xhosa	83	7.7%	102	9.1%
Sotho	75	7.0%	79	7.0%
Venda	50	4.6%	49	4.3%
Pedi	90	8.4%	112	9.9%
Tsonga	60	5.6%	56	5.0%
Tswana	91	8.5%	93	8.3%
Ndebele	11	1.0%	4	0.4%
Swati/Swazi	20	1.9%	29	2.6%
Other	1	0.1%	2	0.2%
	Numeratum			
Variable	**Grade 12/Diploma (*N* = 617)**		**Bachelor's/Honours (*N* = 560)**	
Gender	*n*	%	*n*	%
Women	299	48.5%	301	53.8%
Men	318	51.5%	259	46.2%
Ethnicity	*n*	%	*n*	%
Black African	348	58.0%	340	62.8%
White	125	20.8%	95	17.6%
Coloured	45	7.5%	26	4.8%

Indian/Asian	50	8.3%	62	11.5%
Other	32	5.3%	18	3.3%
Language	*n*	%	*n*	%
English	174	28.2%	176	31.4%
Zulu	93	15.1%	62	11.1%
Afrikaans	85	13.8%	68	12.1%
Xhosa	40	6.5%	54	9.6%
Sotho	39	6.3%	30	5.4%
Venda	26	4.2%	32	5.7%
Pedi	55	8.9%	55	9.8%
Tsonga	38	6.2%	26	4.6%
Tswana	55	8.9%	41	7.3%
Ndebele	2	0.3%	4	0.7%
Swati/Swazi	10	1.6%	12	2.1%

Note. Missing data are not reported and were not considered in the calculation of percentages.

CHAPTER 10: CONCLUDING COMMENTS

Overall, the psychometric properties of the Verbatim and Numeratum were satisfactory, and the assessments appear to be appropriate for use in South African samples. Some of the key psychometric findings are listed below:

- To ascertain whether scale scores, a total score, or both should be interpreted, results as gathered from reliability indicators, Haberman's test, Rasch analysis, and bifactor analysis were examined and suggest that a total Verbatim and Numeratum score should be interpreted.
- The Verbatim and Numeratum showed acceptable reliability at the total score level. Acceptable reliability coefficients were also established across specific gender, ethnic, language, and educational groups.
- Person reliability and person separation indices indicated that the Verbatim and Numeratum can adequately distinguish between high and low scorers.
- Good model fit was established for the Bifactor ESEM models of the Verbatim and Numeratum, with a well-defined general factor for each measure. The specific factors were fairly weakly defined compared to the general factor. This supports previous notions to interpret total rather than scale scores.
- Strict measurement invariance was established for the Verbatim and Numeratum across specific gender, ethnic, language, and educational groups.
- Pairwise differential item functioning comparisons between specific gender, ethnic, language, and educational groups produced no serious concerns for the Verbatim or Numeratum.
- Small mean differences were found for gender on the Verbatim and Numeratum.
- Medium mean differences were found between specific ethnic groups, as well as specific language groups. We ascribed these differences mainly to variability in scores across the different groups, as well as educational differences. Hence, at this stage, we believed that ethnic- or language-based norms were not justified.

- The medium to large effect mean differences between specific educational groups however led to the development of two norms groups: Norm group 1 – Grade 12/Diploma or Certificate, Norm group 2 – Bachelor's degree/Honours degree.

REFERENCES

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing, 38*(1), 6–40. https://doi.org/10.1177/0265532220927487

Ball, T. M., Squeglia, L. M., Tapert, S. F., & Paulus, M. P. (2020). Double dipping in machine learning: Problems and solutions. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *5*(3), 261–263. https://doi.org/10.1016/j.bpsc.2019.09.003

Barrett, P. (2018). *Stanscore 5: Normalized – standardized scores norm table generation.* https://www.pbarrett.net/software/Stanscore.pdf

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental measurement in the human sciences* (3rd ed.). Routledge. https://doi.org/10.4324/9781315814698

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* Cambridge University Press. https://doi.org/10.1017/CBO9780511571312

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*(1), 1–22. https://doi.org/10.1037/h0046743

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R Environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling, 14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Choi, S. W., Gibbons, L. E., & Crane, P. K. (2016). *lordif: Logistic ordinal regression differential item functioning using IRT. R package version 0.3-3*. https://CRAN.R-project.org/package=lordif

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

Combrinck, C. (2020). Is this a useful instrument? An introduction to Rasch measurement models. In S. Kramer, S. Laher, A. Fynn, & H. H. Janse van Vuuren (Eds.), *Online readings in research methods*. Psychological Society of South Africa. https://doi.org/10.17605/OSF.IO/BNPFS

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297–334. https://doi.org/10.1007/BF02310555

De Beer, L. T., & Morin, A. J. S. (2022). *(B)ESEM invariance syntax generator for Mplus*. https://statstools.app/b_esem/

Dorigo, M., & Stützle, T. (2010). Ant colony optimization: Overview and recent advances. In M. Gendreau & J.-Y. Potvin (Eds.), *Handbook of metaheuristics* (Vol. 146, pp. 227–263). Springer. https://doi.org/10.1007/978-1-4419-1665-5_8

Dueber, D. M. (2017). *Bifactor Indices Calculator: A Microsoft Excel-based tool to calculate various indices relevant to bifactor CFA models.* https://doi.org/10.13023/edp.tool.01

Eid, M., Krumm, S., Koch, T., & Schulze, J. (2018). Bifactor models for predicting criteria by general and specific factors: Problems of nonidentifiability and alternative solutions. *Journal of Intelligence*, *6*(3), 42. https://doi.org/10.3390/jintelligence6030042

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *The American Psychologist*, *63*(7), 591–601. https://doi.org/10.1037/0003-066X.63.7.591

Fisher, W. (1992). Reliability, separation, strata statistics. *Rasch Measurement Transactions, 6*(3), 238. https://www.rasch.org/rmt/rmt63i.htm

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Gao, J., Han, X., Pan, G., & Yang, Y. (2017). High dimensional correlation matrices: The central limit theorem and its applications. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79*(3), 677–693. https://doi-org.uplib.idm.oclc.org/10.1111/rssb.12189

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

Gegzna, V. (2020). *biostat: Routines for basic (Bio)statistics*. https://gegznav.github.io/biostat/

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.). Erlbaum.

Grubbs, F. E. (1950). Sample criteria for testing outlying observations. *Annals of Mathematical Statistics, 21,* 27–58. https://doi.org/10.1214/aoms/1177729885

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

Harding, B., Tremblay, C., & Cousineau, D. (2014). Standard errors: A review and evaluation of standard error estimators using Monte Carlo simulations. *The Quantitative Methods for Psychology*, *10*, 107–123.

Harrell, Jr. F. (2023). *Hmisc: Harrell Miscellaneous. R package version 5.1-0*. https://CRAN.R-project.org/package=Hmisc

Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But…. *Communication Methods and Measures*, *14*(1), 1–24. https://doi.org/10.1080/19312458.2020.1718629

Hoekstra, R., Kiers, H., & Johnson, A. (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00137

Howard, J. L., Gagné, M., Morin, A. J. S., & Forest, J. (2018). Using bifactor exploratory structural equation modeling to test for a continuum structure of motivation. *Journal of Management, 44*(7), 2638–2664. https://doi.org/10.1177/0149206316645653

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Praeger Publishers/Greenwood Publishing Group.

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education, 14*(4), 329–349. https://doi.org/10.1207/S15324818AME1404_2

Kassambara, A. (2023). *rstatix: Pipe-friendly framework for basic statistical tests. R package version 0.7.2.* https://CRAN.R-project.org/package=rstatix

Kell, H. J., & Lang, J. W. B. (2017). Specific abilities in the workplace: More important than *g*? *Journal of Intelligence*, *5*(2), 13. https://doi.org/10.3390/jintelligence5020013

Kelley, K. (2022). *MBESS: The MBESS R package. R package version 4.9.2*. https://CRAN.R-project.org/package=MBESS

Kerlinger, F. N., & Lee, H. B. (2000). *Foundations of Behavioral Research* (4th ed.) Wadsworth Publishing.

Koh, K. (2014). Univariate normal distribution. In A. C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 6817–6819). Springer. https://doi.org/10.1007/978-94-007-0753-5_3109

Komsta, L. (2022). *outliers: Tests for outliers. R package version 0.15*. https://CRAN.R-project.org/package=outliers

Korkmaz, S., Goksuluk, D., & Zararsiz, G. (2014). MVN: An R package for assessing multivariate normality. *The R Journal, 6*(2), 151–162.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), *Contributions to probability and statistics* (pp. 278–292). Stanford University Press.

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

Linacre, J. M. (2020a). *Winsteps® (Version 4.6.1)* [Computer software]. Winsteps.com. https://www.winsteps.com/

Linacre, J. M. (2020b). *Winsteps® Rasch measurement computer program User's Guide. Version 4.6.1.* Winsteps.com.

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

Lorenzo-Seva, U., & Ferrando, P. J. (2021). MSA: The forgotten index for identifying inappropriate items before computing exploratory item factor analysis. *Methodology*, 17(4), 296–306. https://doi.org/10.5964/meth.7185

Lorenzo-Seva, U., & Ferrando P. J. (2023). *FACTOR (Version 12.04.01)* [Computer software]. Universitat Rovira i Virgili. https://psico.fcep.urv.cat/utilitats/factor/index.html

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*. Lawrence Erlbaum.

Meijer, R. R., Boevé, A. J., Tendeiro, J. N., Bosker, R. J., & Albers, C. J. (2017). The use of subscores in higher education: When is this useful? *Frontiers in Psychology, 8,* 305. https://doi.org/10.3389/fpsyg.2017.00305

Morin, A. J. S., Myers, N. D., & Lee, S. (2020). Modern factor analytic techniques: Bifactor models, exploratory structural equation modeling (ESEM), and Bifactor-ESEM. In G. Tenenbaum, R. C. Eklund, & N. Boiangin (Eds.), *Handbook of sport psychology: Exercise, methodologies, and special topics* (pp. 1044–1073). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119568124.ch51

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide* (8th ed.). Muthén & Muthén.

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

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Olaru, G., & Danner, D. (2021). Developing cross-cultural short scales using Ant Colony Optimization. *Assessment*, *28*(1), 199–210. https://doi.org/10.1177/1073191120918026

Olaru, G., & Jankowsky, K. (2022). The HEX-ACO-18: Developing an age-invariant HEXACO short scale using Ant Colony Optimization. *Journal of Personality Assessment*, *104*(4), 435–446. https://doi.org/10.1080/00223891.2021.1934480

Osborne, J. W., & Overbay, A. (2019). The power of outliers (and why researchers should ALWAYS check for them). *Practical Assessment, Research, and Evaluation, 9,* Art. 6. https://doi.org/10.7275/qf69-7k43

Postlethwaite, B. E. (2011). *Fluid ability, crystallized ability, and performance across multiple domains: A meta-analysis* (Doctoral dissertation). The University of Iowa, Iowa. https://doi.org/10.17077/etd.zopi8wvs

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

R Core Team (2023). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. https://www.R-project.org/

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

Raîche, G. (2005). Critical eigenvalue sizes in standardized residual principal components analysis. *Rasch Measurement Transactions, 19*, 1012. https://doi.org/10.1016/B978-012471352-9/50004-3.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory.* Routledge Taylor and Francis Group. https://doi.org/10.4324/9780203841624

Ree, M. J., & Carretta, T. R. (2022). Thirty years of research on general and specific abilities: Still not much more than g. *Intelligence*, *91*, 101617. https://doi.org/10.1016/j.intell.2021.101617

Republic of South Africa. *Health Professions Act*, *No. 56 of 1974*. Government Printers.

Revelle, W. (2023). *psych: Procedures for psychological, psychometric, and personality research. Northwestern University., R package version 2.3.3*. https://CRAN.R-project.org/package=psych

Robitzsch, A. (2022). *sirt: Supplementary item response theory models. R package version 3.12-66*. https://CRAN.R-project.org/package=sirt

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016a). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, *98*(3), 223–237. https://doi.org/10.1080/00223891.2015.1089249

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016b). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*(2), 137–150. https://doi.org/10.1037/met0000045

Rönkkö, M., & Cho, E. (2022). An updated guideline for assessing discriminant validity. *Organizational Research Methods*, *25*(1), 6–14. https://doi.org/10.1177/1094428120968614

Rosenbusch, H., Soldner, F., Evans, A. M., & Zeelenberg, M. (2021). Supervised machine learning methods in psychology: A practical introduction with annotated R code. *Social and Personality Psychology Compass, 15*(2), Article e12579. https://doi.org/10.1111/spc3.12579

Spearman, C. (1904). 'General intelligence,' objectively determined and measured. *The American Journal of Psychology, 15*(2), 201–293. https://doi.org/10.2307/1412107

Spector, P. E. (1992). *Summated rating scale construction: An introduction.* Sage Publications, Inc. https://doi.org/10.4135/9781412986038

Teo, T., & Fan, X. (2013). Coefficient alpha and beyond: Issues and alternatives for educational research. *The Asia-Pacific Education Researcher*, *22*(2), 209–213. https://doi.org/10.1007/s40299-013-0075-z

Tesio, L., Caronni, A., Simone, A., Kumbhare, D., & Scarano, S. (2024). Interpreting results from Rasch analysis 2. Advanced model applications and the data-model fit assessment. *Disability and Rehabilitation*, *46*(3), 1–14. https://doi.org/10.1080/09638288.2023.2169772

The jamovi project (2023). *jamovi* (Version 2.3) [Computer Software]. https://www.jamovi.org

Tsagris, M., & Frangos, C. (2020). *Cronbach: Cronbach's alpha. R package version 0.1*. https://CRAN.R-project.org/package=Cronbach

Tuszynski, J. (2021). *caTools: Tools: Moving Window Statistics, GIF, Base64, ROC AUC, etc. R package version 1.18.2*. https://CRAN.R-project.org/package=caTools

Van Iddekinge, C. H., Aguinis, H., Mackey, J. D., & DeOrtentiis, P. S. (2018). A meta-analysis of the interactive, additive, and relative effects of cognitive ability and motivation on performance. *Journal of Management, 44*(1), 249–279. https://doi.org/10.1177/0149206317702220

van Zyl, Casper J.J., & Taylor, Nicola. (2015). *Verbatim & Numeratum Technical Manual.* Johannesburg: JVR Psychometrics (Pty) Ltd.

van Zyl, L. E., & ten Klooster, P. M. (2022). Exploratory structural equation modeling: Practical guidelines and tutorial with a convenient online tool for Mplus. *Frontiers in Psychiatry*, *12*, 795672. https://doi.org/10.3389/fpsyt.2021.795672

Waller, N. G. (2023). *fungible: Psychometric functions from the Waller lab. University of Minnesota. R package 2.3*. https://CRAN.R-project.org/package=fungible

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281–324). American Psychological Association. https://doi.org/10.1037/10222-009

Zijlmans, E. A. O., Tijmstra, J., van der Ark, L. A., & Sijtsma, K. (2018). Item-score reliability in empirical-data sets and its relationship with other item indices. *Educational and Psychological Measurement*, *78*(6), 998–1020. https://doi.org/10.1177/0013164417728358

APPENDIX A: SUPPLEMENTARY TABLES

Supplementary Table 1. *Differential Item Functioning Across Ethnic Groups - Verbatim*

	p-values for χ2 difference tests			Change in Nagelkerke's R^2		
Item	M1-M2	M1-M3	M2-M3	M1-M2	M1-M3	M2-M3
			Ethnicity (Black African vs. Indian)			
S1	0.409	0.704	0.884	0.002	0.002	0.000
S3	0.064	0.013	0.022	0.005	0.013	0.008
S5	0.004	0.009	0.320	0.006	0.007	0.001
S10	0.923	0.921	0.693	0.000	0.000	0.000
S11	0.002	0.002	0.113	0.007	0.009	0.002
O4	0.005	0.020	0.779	0.007	0.008	0.000
O6	0.080	0.212	0.852	0.002	0.002	0.000
O9	0.073	0.137	0.382	0.003	0.003	0.001
O10	0.001	0.003	0.666	0.009	0.009	0.000
O11	0.648	0.725	0.510	0.000	0.000	0.000
O12	0.002	**0.000**	0.009	0.008	0.013	0.005
A5	0.009	0.010	0.120	0.006	0.008	0.002
A6	0.080	0.002	0.002	0.002	0.009	0.007
A7	0.321	0.505	0.537	0.001	0.001	0.000
A8	0.158	0.003	0.002	0.002	0.010	0.009
A9	**0.000**	**0.000**	0.237	0.016	0.017	0.001
A10	0.767	0.023	0.006	0.000	0.006	0.006
R3	0.162	0.132	0.148	0.001	0.003	0.002
R5	0.124	0.102	0.138	0.002	0.003	0.002
R6	0.146	0.001	**0.000**	0.002	0.011	0.009
R7	0.452	0.694	0.683	0.000	0.001	0.000
R8	0.097	0.155	0.325	0.003	0.004	0.001
R9	0.082	0.211	0.754	0.002	0.003	0.000
VI2	0.011	0.038	0.794	0.005	0.005	0.000
VI4	0.622	0.724	0.525	0.000	0.000	0.000
VI6	0.507	0.802	0.972	0.000	0.000	0.000
VI13	0.120	0.298	0.987	0.002	0.002	0.000
VI14	0.662	0.130	0.049	0.000	0.003	0.003

Note. M1-M2 = Model 1 vs. Model 2 (uniform DIF), M1-M3 = Model 1 vs. Model 3 (total DIF), M2-M3 = Model 2 vs. Model 3 (non-uniform DIF).

Supplementary Table 2. *Differential Item Functioning Across Ethnic Groups - Verbatim*

	p-values for χ2 difference tests			Change in Nagelkerke's R^2		
Item	M1-M2	M1-M3	M2-M3	M1-M2	M1-M3	M2-M3

Ethnicity (White vs. Indian)						
S1	0.833	0.391	0.176	0.000	0.016	0.016
S3	0.525	0.209	0.099	0.002	0.013	0.011
S5	0.026	0.068	0.518	0.010	0.011	0.001
S10	0.006	0.016	0.394	0.024	0.026	0.002
S11	0.716	0.927	0.891	0.000	0.000	0.000
O4	0.125	0.103	0.138	0.009	0.017	0.008
O6	0.006	0.009	0.174	0.016	0.020	0.004
O9	0.630	0.890	0.991	0.001	0.001	0.000
O10	0.219	0.399	0.568	0.005	0.006	0.001
O11	0.166	0.383	0.942	0.005	0.005	0.000
O12	0.862	0.728	0.437	0.000	0.002	0.002
A5	0.165	0.198	0.252	0.005	0.008	0.003
A6	0.374	0.621	0.685	0.002	0.002	0.000
A7	0.001	0.004	0.368	0.018	0.020	0.001
A8	0.412	0.215	0.121	0.001	0.006	0.005
A9	0.717	0.226	0.092	0.000	0.005	0.005
A10	0.977	0.976	0.828	0.000	0.000	0.000
R3	0.329	0.087	0.048	0.002	0.011	0.009
R5	0.843	0.921	0.724	0.000	0.000	0.000
R6	0.621	0.223	0.097	0.000	0.005	0.005
R7	0.227	0.313	0.353	0.003	0.004	0.002
R8	0.331	0.473	0.457	0.003	0.004	0.002
R9	0.943	0.989	0.897	0.000	0.000	0.000
VI2	0.750	0.201	0.078	0.000	0.007	0.006
VI4	0.401	0.415	0.305	0.001	0.003	0.002
VI6	0.177	0.370	0.680	0.004	0.004	0.000
VI13	0.037	0.109	0.781	0.009	0.009	0.000
VI14	0.930	0.391	0.171	0.000	0.004	0.004

Note. M1-M2 = Model 1 vs. Model 2 (uniform DIF), M1-M3 = Model 1 vs. Model 3 (total DIF), M2-M3 = Model 2 vs. Model 3 (non-uniform DIF).

Supplementary Table 3. *Differential Item Functioning Across Language Groups - Verbatim*

	p-values for χ2 difference tests			Change in Nagelkerke's R^2		
Item	M1-M2	M1-M3	M2-M3	M1-M2	M1-M3	M2-M3
			Language (English vs. Xhosa)			
S1	0.078	0.209	0.910	0.014	0.014	0.000
S3	0.290	0.523	0.676	0.002	0.003	0.000
S5	0.240	0.178	0.150	0.002	0.004	0.002
S10	0.102	0.131	0.240	0.004	0.007	0.002
S11	0.038	0.115	0.960	0.004	0.004	0.000
O4	**0.000**	0.002	0.771	0.019	0.020	0.000
O6	0.584	0.805	0.714	0.000	0.000	0.000
O9	0.060	0.009	0.016	0.004	0.012	0.007
O10	0.797	0.892	0.688	0.000	0.000	0.000
O11	0.839	0.386	0.173	0.000	0.002	0.002
O12	0.043	0.006	0.014	0.005	0.011	0.007
A5	0.601	0.855	0.842	0.000	0.000	0.000
A6	0.504	0.729	0.665	0.000	0.001	0.000
A7	0.458	0.485	0.344	0.000	0.001	0.001
A8	0.040	0.007	0.017	0.004	0.011	0.006
A9	**0.000**	**0.000**	0.164	0.011	0.012	0.002
A10	0.422	0.306	0.190	0.001	0.002	0.002
R3	0.134	0.273	0.552	0.002	0.003	0.000
R5	0.723	0.373	0.174	0.000	0.002	0.002
R6	0.945	0.167	0.059	0.000	0.004	0.004
R7	0.396	0.609	0.603	0.001	0.001	0.000
R8	0.467	0.684	0.631	0.001	0.001	0.000
R9	0.698	0.914	0.864	0.000	0.000	0.000
VI2	0.024	0.061	0.489	0.005	0.006	0.000
VI4	0.718	0.731	0.481	0.000	0.001	0.000
VI6	0.638	0.565	0.337	0.000	0.001	0.001
VI13	0.476	0.627	0.515	0.000	0.001	0.000
VI14	0.759	0.954	0.993	0.000	0.000	0.000

Note. M1-M2 = Model 1 vs. Model 2 (uniform DIF), M1-M3 = Model 1 vs. Model 3 (total DIF), M2-M3 = Model 2 vs. Model 3 (non-uniform DIF).

Supplementary Table 4. *Differential Item Functioning Across Language Groups - Verbatim*

	p-values for χ2 difference tests			Change in Nagelkerke's R^2		
Item	M1-M2	M1-M3	M2-M3	M1-M2	M1-M3	M2-M3
			Language (English vs. Pedi)			
S1	0.769	0.955	0.940	0.000	0.000	0.000
S3	0.041	0.117	0.720	0.010	0.010	0.000
S5	0.149	0.337	0.767	0.002	0.002	0.000
S10	0.183	0.412	0.977	0.003	0.003	0.000
S11	**0.000**	0.002	0.834	0.011	0.011	0.000
O4	0.020	0.054	0.527	0.009	0.009	0.001
O6	0.165	0.350	0.681	0.002	0.002	0.000
O9	0.217	0.436	0.709	0.002	0.002	0.000
O10	0.044	0.124	0.718	0.005	0.005	0.000
O11	0.288	0.568	0.977	0.002	0.002	0.000
O12	**0.000**	**0.000**	0.136	0.021	0.023	0.002
A5	0.310	0.597	0.965	0.001	0.001	0.000
A6	0.562	0.845	0.974	0.000	0.000	0.000
A7	0.062	0.146	0.543	0.003	0.003	0.000
A8	0.372	0.460	0.386	0.001	0.002	0.001
A9	**0.000**	**0.000**	0.098	0.021	0.023	0.002
A10	0.161	**0.000**	**0.000**	0.002	0.024	0.022
R3	0.831	0.628	0.347	0.000	0.001	0.001
R5	0.840	0.958	0.832	0.000	0.000	0.000
R6	0.005	0.018	0.913	0.008	0.008	0.000
R7	0.297	0.486	0.552	0.001	0.001	0.000
R8	0.004	0.007	0.205	0.012	0.014	0.002
R9	0.932	0.608	0.320	0.000	0.001	0.001
VI2	0.127	0.092	0.118	0.002	0.005	0.002
VI4	0.712	0.631	0.376	0.000	0.001	0.001
VI6	0.397	0.643	0.684	0.001	0.001	0.000
VI13	1.000	0.093	0.029	0.000	0.005	0.005
VI14	0.776	0.588	0.322	0.000	0.001	0.001

Note. M1-M2 = Model 1 vs. Model 2 (uniform DIF), M1-M3 = Model 1 vs. Model 3 (total DIF), M2-M3 = Model 2 vs. Model 3 (non-uniform DIF).

Supplementary Table 5. *Differential Item Functioning Across Ethnic Groups - Numeratum*

	p-values for χ2 difference tests			Change in Nagelkerke's R^2		
Item	M1-M2	M1-M3	M2-M3	M1-M2	M1-M3	M2-M3
			Ethnicity (Black African vs. Indian)			
NP4	0.403	0.115	0.057	0.001	0.009	0.007
NP6	0.100	0.129	0.238	0.004	0.006	0.002
NP7	0.740	0.717	0.456	0.000	0.001	0.001
NP12	0.933	0.729	0.429	0.000	0.001	0.001
NP13	0.022	0.072	0.995	0.008	0.008	0.000
P5	0.313	0.557	0.699	0.001	0.001	0.000
P7	0.451	0.694	0.686	0.001	0.001	0.000
P8	0.955	0.965	0.795	0.000	0.000	0.000
P9	0.086	0.145	0.338	0.004	0.005	0.001
P10	0.166	0.283	0.437	0.003	0.004	0.001
P11	0.608	0.678	0.474	0.000	0.001	0.001
NI5	0.331	0.400	0.346	0.001	0.002	0.001
NI8	0.479	0.743	0.761	0.001	0.001	0.000
NI9	0.741	0.942	0.920	0.000	0.000	0.000
NI10	0.399	0.575	0.530	0.001	0.002	0.001
NI17	0.867	0.180	0.065	0.000	0.005	0.005

Note. M1-M2 = Model 1 vs. Model 2 (uniform DIF), M1-M3 = Model 1 vs. Model 3 (total DIF), M2-M3 = Model 2 vs. Model 3 (non-uniform DIF).

Supplementary Table 6. *Differential Item Functioning Across Ethnic Groups - Numeratum*

	p-values for χ2 difference tests			Change in Nagelkerke's R^2		
Item	M1-M2	M1-M3	M2-M3	M1-M2	M1-M3	M2-M3
			Ethnicity (White vs. Indian)			
NP4	0.023	0.053	0.409	0.022	0.025	0.003
NP6	0.240	0.060	0.039	0.005	0.020	0.015
NP7	0.001	0.003	0.989	0.037	0.037	0.000
NP12	0.216	0.183	0.173	0.007	0.015	0.008
NP13	0.976	0.928	0.699	0.000	0.000	0.000
P5	0.037	0.109	0.784	0.015	0.016	0.000
P7	0.008	0.030	0.937	0.022	0.022	0.000
P8	0.106	0.142	0.256	0.009	0.014	0.004
P9	0.058	0.144	0.589	0.010	0.011	0.001
P10	0.493	0.790	0.964	0.003	0.003	0.000
P11	0.893	0.390	0.172	0.000	0.006	0.006
NI5	0.887	0.906	0.673	0.000	0.001	0.001
NI8	0.273	0.382	0.394	0.004	0.006	0.002
NI9	0.541	0.484	0.299	0.001	0.004	0.003
NI10	0.905	0.517	0.253	0.000	0.008	0.008
NI17	0.292	0.367	0.345	0.004	0.006	0.003

Note. M1-M2 = Model 1 vs. Model 2 (uniform DIF), M1-M3 = Model 1 vs. Model 3 (total DIF), M2-M3 = Model 2 vs. Model 3 (non-uniform DIF).

Supplementary Table 7. *Differential Item Functioning Across Language Groups - Numeratum*

	p-values for χ2 difference tests			Change in Nagelkerke's R^2		
Item	M1-M2	M1-M3	M2-M3	M1-M2	M1-M3	M2-M3
			Language (English vs. Pedi)			
NP4	0.894	0.969	0.832	0.000	0.000	0.000
NP6	0.513	0.477	0.305	0.001	0.003	0.002
NP7	0.017	0.007	0.038	0.014	0.024	0.010
NP12	0.645	0.396	0.200	0.001	0.005	0.004
NP13	0.600	0.652	0.446	0.000	0.001	0.001
P5	0.471	0.714	0.695	0.001	0.001	0.000
P7	0.149	0.322	0.668	0.003	0.003	0.000
P8	0.097	0.208	0.540	0.005	0.005	0.001
P9	0.541	0.026	0.009	0.001	0.012	0.011
P10	0.536	0.360	0.197	0.001	0.005	0.004
P11	0.900	0.825	0.543	0.000	0.001	0.001
NI5	0.059	0.097	0.295	0.007	0.009	0.002
NI8	0.819	0.792	0.520	0.000	0.001	0.001
NI9	0.232	0.360	0.434	0.002	0.003	0.001
NI10	0.877	0.430	0.197	0.000	0.004	0.004
NI17	0.635	0.363	0.180	0.000	0.004	0.003

Note. M1-M2 = Model 1 vs. Model 2 (uniform DIF), M1-M3 = Model 1 vs. Model 3 (total DIF), M2-M3 = Model 2 vs. Model 3 (non-uniform DIF).

APPENDIX B: COMPLETION TIME CALCULATIONS

In this section, we outline the steps for calculating the time limits imposed on each subtest of the latest versions of the Verbatim and Numeratum, respectively.

Step 1. First, we calculated the mean and trimmed mean (10%) completion time for each section. Below, the time allocations as listed in the booklets of the previous versions of the Verbatim and Numeratum are presented first (i.e., how much time was allowed), followed by the trimmed mean (10%) and the mean for completed cases only (i.e., how long participants actually took to complete the questionnaires).

Verbatim - Average Time per Section (Completed cases only)					
10 min	10 min	10 min	20 min	20 min	Original time allocation
Synonyms	Opposites	Analogies	Reasoning	Interpretation	
03:18:37	04:27:02	04:27:43	16:17:54	08:00:10	Trimmed mean 10%
03:26:36	04:33:21	04:34:07	15:55:23	08:12:03	Mean

Numeratum - Average Time per Section (Completed cases only)			
20 min	20 min	20 min	Original time allocation
Number Problems	Patterns	Interpretation	
09:41:24	11:51:15	17:53:49	Trimmed mean 10%
09:49:55	11:48:04	17:27:49	Mean

Step 2. To calculate time limits, the preceding mean scores were first converted to seconds, then divided by the number of original V&N items in each section, and then multiplied by the number of revised items in each section of the new V&N. For example, if we take the first row of Synonyms, 03:18:37 was converted to seconds: ~ 199 seconds. 199 was divided by the number of original V&N items which is 12 for Synonyms, and then multiplied by the number of items in the new V&N, which is 5 for Synonyms. Hence, the following calculation: (199/12*5) = ~ 83 seconds. It will then be approximately 1 and a half minutes when converting back to minutes. These times were then rounded up to the next minute (e.g., 1 and a half minutes becomes 2 minutes) or in certain instances, more time was added for reading (e.g., Interpretation section). Based on the preceding calculations, the following time limits apply to the different V&N sections:

Verbatim 2.0 - Total +/- 25 min				
2 min	3 min	3 min	10 min	7 min
Synonyms	**Opposites**	**Analogies**	**Reasoning**	**Interpretation**
82,92 sec	133,5 sec	146,18 sec	533,45 sec	240,50 sec
86,25 sec	137 sec	150 sec	521,45 sec	246,50 sec

Numeratum 2.0 - Total +/- 20 min		
4 min	7 min	9 min
Number Problems	**Patterns**	**Interpretation**
207,86 sec	388,36 sec	452,21 sec
210,71 sec	386,73 sec	441,26 sec