Matrigma®

by ASSESSIO

South African Manual Supplement

Compiled by:

Nicola Taylor

Cobi Hayes

Anders Sjöberg

Sofia Sjöberg

November 2017

Printed in South Africa by JvR Psychometrics (Pty) Ltd under licence.

For more information, contact research@jvrafrica.co.za or any of the offices below.

JvR Psychometrics forms part of the JvR Africa Group

The JvR Africa Group

SOUTH AFRICA | NAMIBIA | BOTSWANA

JOHANNESBURG
15 hunter avenue | ferndale | randburg | johannesburg | 2194
p.o. box 2560 | pinegowrie | 2123 | johannesburg | south africa
+27 11 781 3705/6/7 (tel) | +27 11 781 3703 (fax) | info@jvrafrica.co.za (email)

CAPE TOWN
oak leaf terrace | 1 edmar street | old oak office park | bellville | 7530
p.o. box 5534 | tygervalley | 7536 | cape town | south africa
+27 861 333 916 (tel) | +27 861 004 225 (fax) | capeinfo@jvrafrica.co.za (email)

FREE STATE
8 hugenote street | jordania | bethlehem | 9700 | south africa
+27 58 303 3843 (tel) | freestate-info@jvrafrica.co.za (email)

NAMIBIA
9A | lenie street | ludwigsdorf | windhoek | namibia
+264 (0) 81 151 6220 (tel) | namibiainfo@jvrafrica.co.na (email)

BOTSWANA
unit 10 | hillside view | kgale | gaborone | botswana
+267 7211 7343 (tel) | botswanainfo@jvrafrica.co.bw (email)

www.jvrpsychometrics.co.za
www.psychologyafrica.com
www.jvrafrica.co.na
www.jvrafrica.co.bw

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The Matrigma is a non-verbal assessment of general mental ability that requires individuals to complete a series of non-verbal problem-solving tasks. More specifically, it taps into fluid intelligence which is the ability to reason out and solve problems that are of a new and different nature, without relying on previously gained knowledge (Horn & Cattell, 1966).

The 2013 itinerant norms lacked in representativeness of the larger South African population and JvR Psychometrics in partnership with Assessio embarked on a research project to collect a more representative norm sample. This supplement will outline the findings from the statistical analyses conducted with the new norm sample and will suggest new norms that are representative of the larger South African population.

User qualifications

The Matrigma is a measure of general mental ability, which is defined as a psychological construct. In line with the Health Professions Act (No. 56 of 1974), only registered psychology professionals are allowed to use the Matrigma.

Appropriate use

The Matrigma is intended to assess individuals older than 18 years and is intended for a normal population, not clinical, psychiatric, or psychopathological samples. The Matrigma can be used in occupational contexts for personnel selection and professional development.

The development of the Matrigma

For more information on the development of the Matrigma, please refer to the Matrigma Technical Manual (Mabon & Sjöberg, 2013).

Administration

Web-based administration

The Matrigma is an online assessment that can be accessed through the JvR Online portal. It is important that assessment administrators understand how participants complete an online assessment, are able to answer participants' questions or concerns, and can use the online administrative platform. The JvR Client Services can set up individual user platforms for clients upon request. The user will receive detailed instructions on how to use the system and the JvR Client Services is also available for additional technical support on weekdays from 8am to 5pm (call 011 781 3705 or email clientservices@jvrafrica.co.za).

The platform allows users to have full control over their account and offers the following benefits:

- Accessible 24 hours a day, 7 days a week from any PC with internet capability
- Assessment results are available in 90 seconds or less after assessment completion
- Ordering of credits can be done online
- Permissions can be set for confidentiality (allowing only particular users access to view and order reports)

Training

Training on the Matrigma is not a requirement as only registered psychology professionals are allowed to use the assessment, but an optional e-learning course is available through JvR Academy (call 011 781 3705 or e-mail training@jvrafrica.co.za). Delegates registered for the eLearning training will be taken through a web-based course. The course contains six sections with the last section being a quiz (70% pass mark) and takes approximately 60 minutes to complete.

SAMPLE DESCRIPTION

Table 1 gives an overview of the current norm sample. JvR Psychometrics invited current and new clients to partake in the data collection phase. Free Matrigma reports were offered to clients who used the Matrigma with participants who matched the stratified criteria set out by JvR Psychometrics. The aim was to collect data from a sample that is representative of the South African working adult population. The inclusion criteria to form a part of the norm sample was that individuals should be South African citizens, over 18 years of age, and either had been or were currently employed. The final sample consisted of 472 individuals who completed the Matrigma assessment via the JvR Online portal.

Table 1. Sample overview

Population Group	*N*	%
Gender		
Women	202	42.80
Men	270	57.20
	472	**100**
Age		
Age 19-29	162	34.32
Age 30-40	163	34.53
Age 41-50	111	23.52
Age 51-60	32	6.78
Age 61-68	4	.85
	472	**100**
Ethnicity		
Asian	3	0.72
Black	164	39.42
Coloured	34	8.17
Indian	37	8.89
White	168	40.38
Not provided	10	2.40
	416	**100**
Employment Status		
Retired	2	0.50
Unemployed	77	19.35
Unknown	13	3.27
Working full-time	259	65.08
Working on contract	35	8.79
Working part-time	12	3.02
	398	**100**
Job level		
Business Owner	10	2.98
Employee	139	41.37
Executive	18	5.36
Manager	121	36.01
Supervisor	48	14.29
	336	**100**

According to Statistics South Africa (StatsSA, 2016), South Africa consists of around 55.6 million people and Table 2 gives a comparative overview of the current sample against the 2011 population census conducted through Statistics South Africa (2012). The results of the most recent Labour Force Survey (StatsSA, 2017) are also reported in Table 2. The percentages indicate South Africans currently listed as employed. The age groups in the Labour Force Survey do not correspond directly to those in the sample or census, so the various age groups are listed separately.

Table 2. Sample comparison to South African population

Sample	Sample %	SA population %		Labour force %	
Gender					
Women	42.80	51.04		43.96	
Men	57.20	48.96		56.04	
Age*					
Age 19-29	34.32	Age 20-29	33.08	Age 15-24	7.61
Age 30-40	34.53	Age 30-39	25.95	Age 25-34	30.48
Age 41-50	23.52	Age 40-49	18.89	Age 35-44	31.30
Age 51-60	6.78	Age 50-59	13.47	Age 45-54	21.24
Age 61-68	.85	Age 60-69	8.60	Age 55-64	9.35
Ethnicity					
Asian/Indian	9.61	2.47		3.27	
Black	39.42	80.66		74.69	
Coloured	8.17	8.76		10.13	
White	40.38	8.12		11.91	

Note. * = The StatsSA age groups between 20 – 69 make up 57.48% of the overall South African population and the percentages in Table 2 are a summary of that age group's representation within the 57.48%, not against the total population.

From Table 2, it is apparent that women are slightly under-represented when compared to the population, but when compared to working adults, the percentage appears correct. Asian/Indian participants were over-represented in the sample, but the size of the group is still relatively small overall. The Black respondent group is under-represented. However, in order to be able to make meaningful comparisons across population groups, having similarly sized samples for Black and White respondents is useful.

RELIABILITY

The Matrigma can be completed in 5 different forms, ranging from Form A to E. In each one of these forms the order of the questions is different. The internal consistency reliability coefficients for each of the Matrigma forms are presented in Table 3.

Table 3. Reliability coefficients for the different Matrigma forms

Form	*N*	Mean	S.D	α*	λ2**
A	93	17.19	5.36	.86	.87
B	85	16.23	5.37	.85	.86
C	97	16.38	6.06	.88	.89
D	94	16.34	5.96	.87	.88
E	103	16.78	4.80	.81	.82
	472	16.59	5.51	.85	.86

Note. * = Cronbach's estimate of reliability; ** = Guttmann's Lambda 2.

For the purpose of this report, two measures of internal consistency were reported: Cronbach's Alpha and Guttmann's Lambda 2. Guttman's Lambda 2 is a more robust measure of internal consistency (Osburn, 2000). All of the Matrigma forms had reliability estimates above .80, suggesting good internal consistency. In other words, the items in the Matrigma all seem to be measuring a similar construct.

In order to determine if the reliability of the Matrigma was consistent across different population groups in the sample, we ran the reliabilities for various subgroups for each form of the Matrigma. Table 4 gives an indication of the different subgroup sizes and

Table 5 provides the results from the reliability analyses.

Table 4. Subgroup population sizes across Forms.

Group	Form A	Form B	Form C	Form D	Form E
Gender					
Women	40	33	43	40	46
Men	53	52	54	54	57
Ethnicity					
Black/African	39	28	34	31	35
White/Caucasian	29	32	37	34	36
Language					
English 1st Language	38	29	38	26	37
English 2nd Language	55	56	59	68	66

Table 5. Reliability for subgroups across Forms.

	Form A		Form B		Form C		Form D		Form E	
	α	λ2	α	λ2	α	λ2	α	λ2	α	λ2
Gender										
Women	.733	.767	.865	.877	.895	.905	.830	.848	.819	.841
Men	.900	.908	.847	.862	.866	.876	.898	.905	.801	.819
Ethnicity										
Black/African	.885	.895	.851	.868	.901	.911	.878	.894	.794	.820
White/Caucasian	.862	.876	.815	.834	.831	.848	.882	.893	.799	.822
Language										
English 1st Language	.819	.837	.830	.852	.821	.842	.870	.884	.834	.851
English 2nd Language	.878	.888	.864	.874	.897	.905	.876	.886	.788	.810

Subgroups with small sample sizes, like the Asian/Indian and Coloured ethnicity groups, were excluded from analyses. Language groups were grouped together in English first and second language participants due to smaller language groups.

All of the subgroups for the different Matrigma forms yielded good reliability coefficients ($r > .730$), indicating that the Matrigma measures the same construct in each of the subgroups across all five forms.

GROUP DIFFERENCES

Matrigma forms

A one-way analysis of variance with post-hoc tests was run to determine if any statistically significant differences existed between participants who completed different forms of the Matrigma. There were no statistically significant differences between how participants scored on the different Matrigma forms (C-score: F (4,467) = .480, p = .751; raw score: F (4, 467) = .509, p = .729).

These results support the use of multiple forms of the Matrigma without the form that the participant received having a potential influence on their overall performance.

Gender

A *t*-test was conducted in order to investigate whether there were any statistically significant differences between how men and women scored on the Matrigma. C-scores and raw scores were used for the analysis. Table 6 gives an overview of how men and women performed on the Matrigma.

Table 6. Gender differences

		C-score		Raw score	
Gender	*N*	*M*	S.D.	*M*	S.D.
Women	202	4.71	2.165	16.42	5.343
Men	270	4.84	2.324	16.73	5.645

Although men scored slightly higher than women, the difference was not statistically significant (C-score: t (470) = .635; p = .526; raw score: t (470) = .572, p= .552) and yielded a small effect size (C-score: d = .060; raw score: d = 0.059). In other words, men and women are expected to score similarly on the Matrigma and therefore there is no need to include gender-specific norm groups.

Age

Correlation between age and Matrigma score

According to Staff, Hogan, and Whalley (2014), fluid ability declines with age. In order to investigate this effect in the current sample, a Spearman rho correlation was run between individuals' age and their Matrigma raw scores. A negative correlation (r = -.251, p = .000) was obtained, suggesting that older respondents scored lower on the Matrigma than younger respondents. This is also confirmed by most of the results from the age group comparisons where on average, younger groups scored .58 C-score units higher than their older counterparts.

The classic ageing pattern suggests that a decline in fluid intelligence is specifically in relation to visual-spatial information processing (Schretlen, et al., 2000). The Matrigma consists of a series of visual non-verbal reasoning questions that the participants need to solve. These declines could, however, be due to combinations of ageing with processing speed or working memory (Salthouse, 1991). The Matrigma is a timed assessment and declines in processing speed could influence the scores of older people on timed assessments.

Age group comparison

The sample was grouped into 5 distinct age categories that are presented in Table 7.

Table 7. Age group mean scores

Age category	*N*	C-score		Raw score	
		M	S.D.	*M*	S.D.
19-29	162	5.33	2.102	17.88	5.055
30-40	163	4.88	2.205	16.82	5.440
41-50	111	4.18	2.426	15.16	6.043
51-60	32	3.57	1.867	13.97	4.561
61-68	4	4.51	1.260	16.25	2.754

In order to determine whether any statistically significant differences existed between how different age groups performed on the Matrigma a one-way analysis of variance was run. The oldest age group (61-68) was excluded from these analyses due to the small sample size. Table 8 and Table 9 give an overview of the post hoc test results for the different age groups on C-scores and raw scores, respectively.

Table 8. Age group differences (C-scores)

Age category		Mean difference	S.E.	*p*	Cohen's *d*
19-29	30-40	.44	.24	.352	.21
	41-50	1.14	.27	.000*	.51
	51-60	1.75	.425	.000*	.88
30-40	41-50	.69	.27	.075	.30
	51-60	1.31	.425	.018*	.64
41-50	51-60	.61	.44	.637	.28

*Note: * = Statistically significant at the p < .05 level.*

Table 9. Age group differences (raw scores)

Age category		Mean difference	S.E.	*p*	Cohen's *d*
19-29	30-40	1.07	.60	.385	.20
	41-50	2.72	.67	.000*	.49
	51-60	3.91	1.04	.002*	.81
30-40	41-50	1.65	.66	.094	.20
	51-60	2.85	1.04	.051	.57
41-50	51-60	1.19	1.08	.805	.22

*Note: * = Statistically significant at the p < .05 level.*

From Table 8 it can be seen that there were four groups with statistically significant differences in their C-scores. The age group from 19-29 years scored higher than all of the other groups and statistically significantly higher than the 41-50 (C-scores: F (4, 467) = 7.109 , p =.000, d = .21; raw score: F (4,467) = 6.232, p = .000, d =.20) and 51-60 year olds (C-score: F (4, 467) = 7.109, p = .000; raw score: F (4,467) = 6.232, p = .002, d = .81). The 30-40 year old group scored higher than all of the older age groups and significantly higher than the 51 – 60 year old group on their C-scores (F (4, 467) = 7.109, p = .018, d = .64), but not statistically significant on their raw scores (F (4,467) = 6.232, p = .051, d = .57). All of the significant differences resulted in moderate to large effect sizes (.49 - .88), suggesting that

the practical implications of these differences might warrant age specific norms. The overall trend is that younger participants in the sample scored higher than their older counterparts.

Ethnicity

The sample consisted of 4 different ethnic groups: Asian/Indian, Black, Coloured and White. Table 10 gives an overview of the mean scores for each one of the population groups in the sample.

Table 10. Mean differences in ethnic groups

	N	C-score		Raw score	
Ethnicity		*M*	*S.D.*	*M*	*S.D.*
Black	164	4.16	2.297	15.10	5.835
Coloured	34	4.52	2.492	16.15	5.668
Asian/Indian	40	5.53	1.909	18.35	4.544
White	168	5.20	2.117	17.60	5.061

A one-way analysis of variance was conducted to investigate differences in C-scores obtained between different ethnic groups within the sample. Table 11 shows that there were statistically significant differences between different ethnic groups.

Table 11. Post hoc results for ethnic group differences (C-scores)

Ethnicity		Mean difference	S.E.	*P*	Cohen's *d*
Asian/Indian	Black	1.366	0.389	0.003*	0.65
	Coloured	1.010	0.514	0.204	0.46
	White	0.326	0.388	0.835	0.16
Black	Coloured	-0.357	0.416	0.827	0.15
	White	-1.040	0.242	0.000*	0.47
Coloured	White	-0.684	0.415	0.353	0.29

Note: * = Statistically significant at the p < .05 level.

The results indicated statistically significant differences between the mean C-scores for Black and Asian/Indian participants. Asian/Indian people obtained higher scores compared to the Black participants (F (3,402) = 8.02, p=.003). This difference yielded a moderate to large effect size (d = .65), which could suggest that it might be necessary to have different norms for these two population groups. There were also statistically significant differences between the mean C-scores of Black and White participants (F (3,402) = 8.02, p<.000, d = .47). These results suggest that the possibility of having separate norm groups for Black and White participants should be considered. There were no statistically significant differences found between the other groups.

We also ran a one-way analysis of variance on the raw scores for participants from the various ethnic groups. These results are presented in Table 12.

Table 12. Post hoc results for ethnic group differences (raw scores)

Ethnicity		Mean difference	S.E.	*P*	Cohen's D
Asian/Indian	Black	3.252	0.951	0.004*	0.62
	Coloured	2.203	1.258	0.299	0.43
	White	0.749	0.949	0.859	0.16
Black	Coloured	-1.049	1.016	0.730	0.18
	White	-2.504	0.592	0.000*	0.46
Coloured	White	-1.454	1.014	0.479	0.27

Note: * = Statistically significant at the p < .05 level.

The same pattern was found with raw scores as with the C-scores. White participants scored higher than Black participants (F (3,402) = 7.64, p<.000*) with a moderate effect size (d = .46). This would suggest that it might be necessary to have separate norm groups for White and Black participants. There were also statistically significant differences between Black participants' and Asian/Indian participants' mean scores (F (3,402) = 7.64, p=.004) with a moderate to large effect size (d = .62), suggesting that it might be valuable to have separate norm groups for Asian/Indian and Black ethnic groups. There were no statistically significant differences between other ethnic groups.

Education level

Participants were grouped into different levels of education in order to establish if there were any statistically significant differences between how participants from different educational backgrounds score on the Matrigma. Table 13 gives an overview of the mean scores (both on C-scores and raw scores) for the different education levels.

Table 13. Mean differences between different educational levels

	N	C-score		Raw score	
Education level		*M*	*S.D.*	*M*	*S.D.*
Grade 10/Standard 8 & Grade 12/Standard 10	81	3.89	2.307	14.47	5.867
Some university	11	4.86	2.087	17.00	4.561
Certificate/Diploma/Degree	202	4.73	2.231	16.44	5.559
Honour's degree	83	5.27	2.132	17.80	4.973
Master's degree	37	5.22	2.409	17.76	5.351
PhD	2	4.63	1.618	16.50	3.536

From Table 13 it is clear that participants who completed an education beyond high school scored higher on the Matrigma than those who did not complete their high school education. One of the groups (PhD) had too few participants to run subsequent analyses and was excluded from the post hoc tests.

Table 14 provides an overview of the post hoc results for differences between educational levels on their C-scores, while

Table 15 gives an overview of the differences in raw scores between different education levels.

Table 14. Post hoc results for educational level differences (C-scores)

Education Level		Mean differences	S.E.	*p*	*d*
Grade 10/Standard 8 & Grade 12/Standard 10	Some university	-0.969	0.719	0.758	.425
	Certificate/Diploma/Degree	-0.840	0.294	0.052	.373
	Honours	-1.385	0.350	0.001*	.622
	Masters	-1.333	0.444	0.034*	.569
Some university	Certificate/Diploma/Degree	0.130	0.693	1.000	-.058
	Honours	-0.415	0.718	0.992	.193
	Masters	-0.363	0.769	0.997	.154
Certificate/Diploma/ Degree	Honours	-0.545	0.292	0.424	.245
	Masters	-0.493	0.400	0.821	.217
Honours	Masters	0.052	0.442	1.000	-.023

Note: * = Statistically significant at the p < .05 level.

Table 15. Post hoc results for education level differences (raw scores)

Education Level		Mean differences	S.E.	*p*	*d*
Grade 10/Standard 8 & Grade 12/Standard 10	Some university	-2.531	1.756	0.702	.441
	Certificate/Diploma/Degree	-1.971	0.719	0.069	.349
	Honours	-3.326	0.854	0.002*	.613
	Masters	-3.288	1.084	0.031*	.576
Some university	Certificate/Diploma/Degree	0.559	1.692	0.999	-.102
	Honours	-0.795	1.754	0.998	.162
	Masters	-0.757	1.877	0.999	.146
Certificate/Diploma/ Degree	Honours	-1.355	0.713	0.403	.252
	Masters	-1.316	0.977	0.759	.239
Honours	Masters	0.038	1.080	1.000	-.008

Note: * = Statistically significant at the p < .05 level

There were no statistically significant differences between how participants with a high school education (Grade 10 or higher) scored in comparison to participants with some tertiary education (some university or certificate/diploma/degree). There were also no statistically significant differences between how participants with different levels of tertiary education scored on the Matrigma. The only statistically significant results were between how post-graduate (Honours and Masters level) participants scored from participants who only had a secondary education.

Participants with a secondary education scored lower on the Matrigma than participants with Honours Degrees (C-score: F (5,410) = 3.630, p = .001, d = .622; raw score: F (5,410) = 3.580, p = 002, d = .613). The participants with a Master's degree also scored significantly higher on the Matrigma than those who only had some level of secondary education (C-score: F (5, 410) = 3.630, p = .0034, d = .569; raw score: F (5,410) = 3.580, p = .031, d = .576).

Although there were significant differences between how participants with a post-graduate education scored from those with only secondary education, there were no statistically significant differences between how participants with secondary and those with some tertiary education performed. Further, there were no differences between how participants with varying levels of tertiary education

performed on the Matrigma. It is therefore argued that, at this stage, it is not possible to establish whether the results necessitate separate norm groups for varied education levels. More data is required in order to confirm the above findings.

Language

The sample was divided into two groups, namely those who indicated English as their first language, and those who indicated a different first language, who were subsequently grouped together as the English second language group. It was investigated whether any statistically significant differences existed between how first language English speakers performed compared to the rest of the sample. Although the Matrigma is a non-verbal assessment, the instructions on how to complete the assessment were given to the participants in English. The breakdown of the sample is presented in Table 16, and the results from an independent t-test to determine differences in performance is presented in Table 17.

Table 16. Language group mean scores

	N	C-score		Raw score	
Language group		*M*	S.D.	*M*	S.D.
English 1st Lang	168	5.17	2.145	17.57	5.009
English 2nd Lang	304	4.57	2.290	16.06	5.712

Table 17. Language group differences in Matrigma performance

	Levene's Test for Equality of Variances		t-test for Equality of Means					
Score	F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	S.E.	Effect size
C-score	0.881	0.348	2.785	470	0.006	0.600	0.215	.270
Raw score	2.070	0.151	2.863	470	0.004	1.506	0.526	.281

There were statistically significant differences between how first language English and second language English participants scored on the Matrigma. Those participants who indicated that they were first language English speakers scored higher than the other participants (C-score: t (470) = 2.785, p = .006, d = .270; raw score: t (470) = 2.863, p = .004; d = .281). However, the effect sizes were small and therefore it was concluded that there is insufficient data at this stage to necessitate the calculation of separate norm groups for first and second language English speakers.

Summary

1. There were no statistically significant differences between how men and women performed on the Matrigma.
2. Younger participants scored significantly higher than the older population groups. However, further research is required to confirm these findings with larger samples of people above the age of 60. There was also a negative correlation between age and Matrigma scores, which is in line with research on cognitive ability.
3. Participants from different ethnic groups performed significantly differently on the Matrigma, particularly participants from the Black population group whose scores were significantly lower than those of the Asian/Indian and the White sample groups. The effect sizes ranged from small to medium, and practitioners are advised to bear these differences in mind when interpreting results. The largest difference was one raw score point.
4. Education level comparisons indicated that participants with a post-graduate level of education scored significantly higher than those with only a secondary level of education. There were no other significant differences between levels of education.
5. First language English speakers scored statistically significantly higher than other participants. However, the magnitude of these differences was small.
6. The differences found between different population groups were further assessed during the Rasch Analysis section to determine if these differences are due to item bias.

RASCH ANALYSIS

The Rasch model (Rasch, 1960) is known as a fundamental measurement model, and is based on the assumption that the probability of achieving higher scores on a test increases as the ability of the individual increases, and decreases as the ability of the individual decreases (Green & Frantom, 2002). In other words, the probability of correctly answering an item on a test is a function of the difficulty of the item and the ability of the person. The unit of measurement in Rasch analysis is the logit (or log-odds unit). The mean logit score is set at 0, with higher scores indicating greater difficulty and negative scores indicating lesser difficulty (Bond & Fox, 2001). In this section, item fit and differential item functioning were investigated for each Form of the Matrigma.

Item Fit

Fit is an indication of the degree to which responses conform to a logical pattern (Green & Frantom, 2002). Items can "overfit" (be too predictable) or "underfit" (be too unpredictable) the model. Items can be removed from the model on the basis of these fit indices to allow better fit to the model.

For each Form of the Matrigma, the item statistics are presented in table format, the 'misfitting' items are discussed and graphic representations of how the item difficulty compares to the participants' ability to respond correctly to them (item map) are presented.

Three specific statistics will be important to understand this section, namely:

1. **Measure** – this statistic refers to how easy it is for a participant to get an item correct. Items with a negative value in the measurement column are easier; in other words, more participants are likely to get these items correct. Items with positive values are more difficult for the participants and they are likely to struggle more with these items.
2. **Mean-square statistic (MNSQ)** – for both infit and outfit, MNSQ gives an indication of how well each item fits against the predictions of the Rasch model for the specific item. These values are expected to be close to 1.0. Items with good fit will generally have scores ranging between .70 and 1.35 (Linacre, 2015). Items with a score smaller than .70 might be redundant, i.e., they are measuring the same thing as other items and do not add any additional information. Items with a score greater than 1.35 might be measuring a different construct than what the test was intended to measure.
3. **Standardised fit statistics (ZSTD)** – for both infit and outfit, ZSTD is a z-score output of a t-test to determine how well the data fit the Rasch model. Scores should central around 0.0. Scores higher than 0.0 indicate a lack of predictability – the item did not function as the Rasch model predicted, while scores below 0.0 indicate too much predictability – i.e., not enough variance in response patterns. Items with ZSTD statistics above 2.0 and below -2.0 are flagged for further investigation.

Infit

Infit statistics refer to a weighted fit that is not influenced by specific outliers in the data and is more sensitive to the pattern of responses for a specific sample on the test items.

Outfit

Outfit statistics are sensitive to outlier data points. This statistic is influenced by data points that fall outside the expected response pattern, i.e., extremely low scores on specific questions due to time constraints where many participants could not answer the question. It is a less robust measure of item fit, but still gives valuable information about the outlier data points that warrant further investigation.

Differential Item Functioning

Differential item function (DIF) refers to the probability that equally able participants might not have the same response patterns (i.e., correct or wrong) for specific items, based on one or more of their population specifications, i.e., gender, ethnicity, etc. (Westers & Kelderman, 1991). A psychological measure that contains items with significant DIF might be unfair towards specific population groups and it is crucial that these items are identified and investigated for future inclusion, adaptation or removal from the measure (De Beer, 2004; Strobl, Kopf, & Zeileis, 2011).

In order to investigate whether there were any items that suggest bias across different population groups, differential item functioning analyses for three specific population variations were run:

- **Ethnicity** – specifically between the White and Black population groups as they had sufficient sample sizes.
- **Gender** – between men and women.
- **Language** – between first and second language English speakers. All participants who did not indicate English as their first language were grouped together into the English Second Language grouping in order to identify potential differences between how the items functioned in the two groups.

Three specific statistics in the investigation of DIF are reported on:

- **DIF contrast** – Gives an indication of the difference between item difficulties for the two groups being compared. A negative DIF contrast value suggests that the item was easier for the first group. In other words, participants from the second group were less likely to get the item correct. A positive DIF value indicates that the item was easier for the second group, and that they were more likely to get the item correct. Items with a DIF contrast greater than an absolute value of .50 were identified for further investigation. The significance of the DIF was considered by exploring the Rasch-Welch and Mantel-Haenszel probabilities.
- **Rasch-Welch** – The Rasch-Welch test is a *t*-test that estimates a Rasch difficulty for the item for each group in the DIF comparison. The Rasch-Welch test allows for missing data in the dataset.
- **Mantel-Haenszel** – The Mantel-Haenszel test is a chi-square estimate of item difficulty differences and is used with dichotomous data. The Mantel-Haenszel statistic is the industry standard for

reporting DIF in psychometric instruments, but at times cannot be estimated due to small sample sizes.

Sample Breakdown for Population Groups per Form

Table 18. Rasch sample breakdown for population groups in each Form

Form	Ethnicity	*N*	Gender	*N*	Language	*N*
A	White	29	Women	40	English 1st Language	38
	Black	39	Men	53	English 2nd Language	55
B	White	32	Women	33	English 1st Language	29
	Black	28	Men	52	English 2nd Language	56
C	White	37	Women	43	English 1st Language	38
	Black	34	Men	54	English 2nd Language	59
D	White	34	Women	40	English 1st Language	26
	Black	31	Men	54	English 2nd Language	68
E	White	36	Women	46	English 1st Language	37
	Black	35	Men	57	English 2nd Language	66

Form A

Item Fit

Table 19 provides an overview of how well the items of the Matrigma Form A fit the Rasch model. From the measure column it can be seen that the items are relatively equally spread out between difficult (positive scores) and easy (negative scores) items. The item difficulty will be plotted against the participants' ability in the following section.

Only one item (item 11) was identified for misfit based on the infit statistics (MNSQ = .69, ZSTD = -1.6). The data suggests that this item might be redundant and does not add any additional information to the scale. Items with a MNSQ statistic below .70 do not influence the structure of the test, but could inflate the test's reliability due to the fact that there are sufficient other items in the test to cover the construct being measured.

Six further items were identified for further investigation based on their outfit statistics:

- Potentially redundant items – 7, 8 and 12.
- Items potentially measuring a different construct – 4, 29 and 30.

The infit statistics for all of these items fell within the standard parameters and when investigating the standardised fit statistics (ZSTD) for these items only one item (item 29) fell outside the parameters (ZSTD = 4.3). This could indicate that the response pattern for this item differs significantly from what the Rasch model expects. It is important to note here that only 13 participants got this item correct. The Matrigma is a timed assessment and it could be possible that many participants did not attempt item 29 due to time limitations. Due to the position of item 29 in the assessment, and the time limit on the test, the potential misfit of this item was deemed to be low risk.

Based on the item fit statistics, all of the items for Form A of the Matrigma performed as expected in terms of the Rasch model. No major risks due to misfit were identified in this Form.

Table 19. Item statistics: Form A

Item	Measure	Model S.E.	Infit		Outfit	
			MNSQ	ZSTD	MNSQ	ZSTD
1	-2.38	0.43	1.01	0.1	1.78	1.2
2	-2.80	0.49	0.91	-0.1	0.87	0.1
3	-3.41	0.62	1.23	0.6	1.53	0.8
4	-3.86	0.74	1.16	0.5	**2.84**	1.4
5	-2.04	0.39	1.23	0.9	1.37	0.8
6	-1.90	0.37	0.96	-0.1	1.04	0.2
7	-3.07	0.54	0.88	-0.2	**0.45**	-0.5
8	-0.85	0.29	0.81	-1.2	**0.60**	-1.6
9	-1.02	0.30	1.18	1.0	1.43	1.4
10	-1.52	0.34	1.22	1.0	1.81	1.8
11	-1.52	0.34	**0.69**	-1.6	**0.49**	-1.5
12	-2.57	0.46	0.92	-0.1	**0.44**	-0.8
13	-0.69	0.28	0.82	-1.2	0.73	-1.1
14	-0.85	0.29	1.04	0.3	1.10	0.5
15	0.37	0.24	1.11	1.0	1.16	1.0
16	0.06	0.25	0.80	-1.8	0.70	-2.0
17	0.42	0.24	0.96	-0.3	0.93	-0.4
18	0.60	0.24	0.96	-0.4	0.96	-0.2
19	1.11	0.24	0.86	-1.6	0.79	-1.2
20	0.60	0.24	0.88	-1.2	0.83	-1.2
21	1.22	0.24	0.99	-0.1	0.95	-0.2
22	2.52	0.29	0.94	-0.4	1.06	0.3
23	2.01	0.26	1.02	0.2	1.08	0.4
24	1.51	0.24	0.89	-1.2	0.95	-0.2
25	1.57	0.24	1.03	0.4	1.02	0.1
26	3.34	0.36	1.00	0.1	0.89	0.0
27	2.52	0.29	1.04	0.3	0.97	0.1
28	2.79	0.31	1.02	0.2	0.91	-0.1
29	2.88	0.32	1.25	1.3	**4.44**	**4.3**
30	4.99	0.72	1.05	0.3	**2.18**	1.2
M	**0.00**	**0.35**	**0.99**	**-0.1**	**1.21**	**0.2**
P.SD	**2.23**	**0.14**	**0.14**	**0.8**	**0.79**	**1.2**

Differential Item Functioning

Ethnicity

Figure 1 gives a graphic representation of how each item performed based on item difficulty for the White and Black participants.

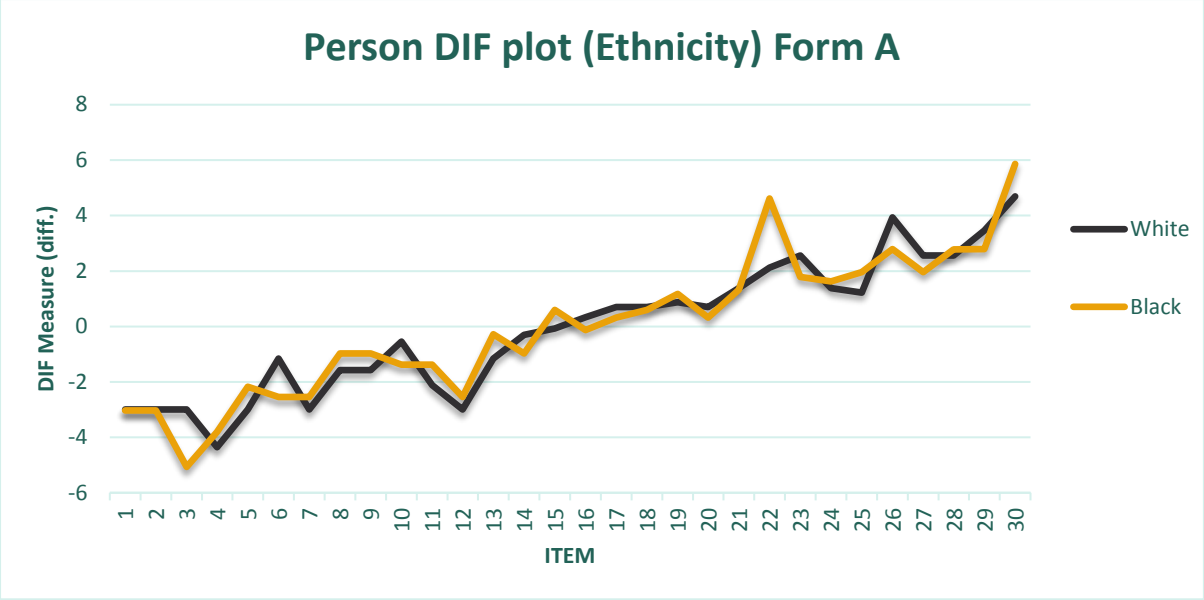

Figure 1. Person DIF plot (Ethnicity) Form A.

From the graph, several items can be identified with large differences between the difficulty levels of the item for the two population groups. These differences are not all in the same direction, however. In other words, one population group is not consistently finding the items more difficult than the other group. In order to investigate these items further, the Rasch-Welch and Mantel-Haenszel probabilities for the items are presented in Table 20. In this case, the White participants were the reference group, while Black participants formed the focal group.

Although the majority of items had DIF contrast values greater than .50 logits, only item 22 showed a statistically significant difference between groups. In other words, the probability that these other items will cause bias based on ethnic differences between the population groups is not likely. Based on the Rasch-Welch test, item 22 shows significant DIF at the $p<.05$ level, but the Mantel-Haenszel probability is not significant. Due to the small sample sizes these results need to be interpreted with care and it was concluded that there is not sufficient evidence indicating that this item will indeed be biased against White participants in other samples. In addition, new research suggests that statistical significance should only be interpreted at the $p<.005$ level (Benjamin et al., 2017), which would render this finding non-significant.

Table 20. DIF between White and Black participants on Form A

Item	DIF Contrasts	Joint S.E.	Rasch-Welch			Mantel-Haenszel	
			t	d.f.	Prob.	Chi-squ	Prob.
1	0.04	1.34	0.03	50	0.979	0.067	0.796
2	0.04	1.34	0.03	50	0.979	0.125	0.724
3	**2.07**	2.15	0.96	56	0.339	0.000	1.000
4	**-0.55**	2.16	-0.26	42	0.800	-*	-*
5	**-0.82**	1.24	-0.66	41	0.513	0.143	0.706
6	**1.38**	0.88	1.57	62	0.121	0.450	0.502
7	-0.45	1.27	-0.35	44	0.725	-*	-*
8	**-0.60**	0.81	-0.74	47	0.462	0.071	0.790
9	**-0.60**	0.81	-0.74	47	0.462	0.625	0.429
10	**0.83**	0.70	1.19	59	0.237	0.040	0.841
11	**-0.74**	0.94	-0.79	44	0.435	0.080	0.778
12	-0.45	1.27	-0.35	44	0.725	0.017	0.896
13	**-0.88**	0.72	-1.22	48	0.230	0.264	0.607
14	**0.67**	0.65	1.03	59	0.309	0.023	0.879
15	**-0.68**	0.60	-1.13	56	0.263	1.084	0.298
16	0.46	0.59	0.78	59	0.439	0.342	0.559
17	0.38	0.57	0.67	59	0.507	0.230	0.632
18	0.10	0.56	0.17	59	0.862	0.192	0.661
19	-0.30	0.56	-0.53	59	0.595	0.050	0.822
20	0.38	0.57	0.67	59	0.507	1.706	0.192
21	0.07	0.57	0.12	60	0.906	0.017	0.895
22	**-2.50**	1.12	-2.23	48	**0.031**	3.534	0.060
23	**0.77**	0.64	1.21	56	0.230	2.120	0.145
24	-0.24	0.58	-0.41	60	0.681	0.191	0.663
25	**-0.74**	0.59	-1.26	61	0.212	0.991	0.320
26	**1.15**	0.91	1.26	49	0.213	0.028	0.868
27	**0.60**	0.64	0.94	57	0.353	0.268	0.605
28	-0.23	0.70	-0.32	62	0.749	0.183	0.669
29	**0.67**	0.81	0.83	55	0.410	0.306	0.580
30	**-1.17**	2.11	-0.55	54	0.581	0.042	0.838

Note: * - Mantel-Haenszel test statistics are not estimable due to small sample sizes.

Gender

A graphical representation of how the two gender groups performed on the 30 items is given in Figure 2.

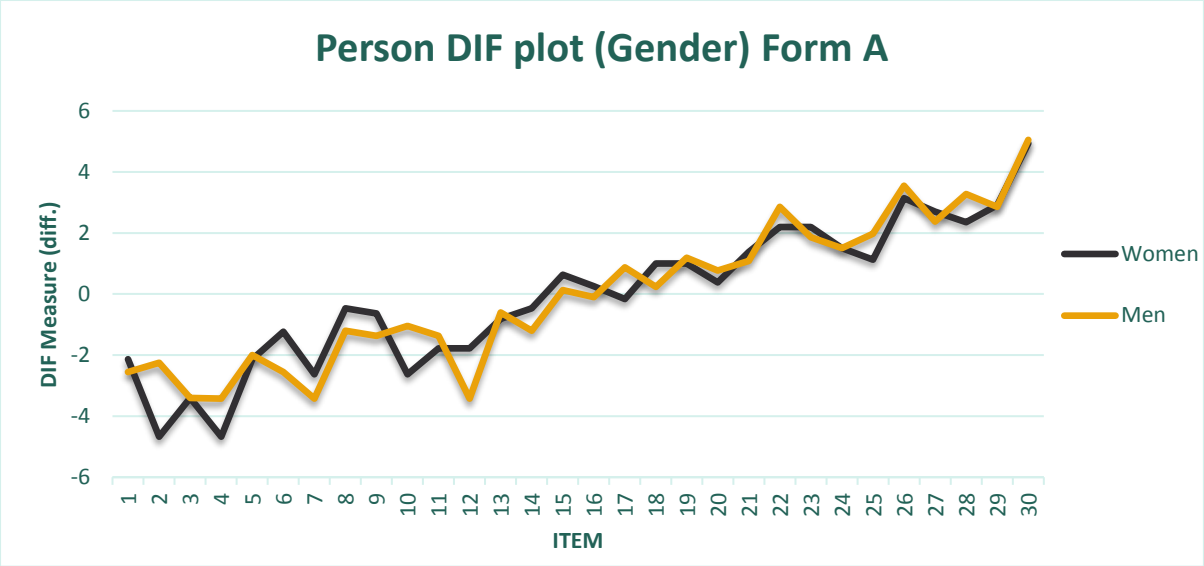

Figure 2. Person DIF plot (Gender) Form A.

It is clear from Figure 2 that several items were identified with large DIF contrasts. These differences were however not unidirectional. Table 21 gives an overview of the DIF statistics between men and women.

Half of the items in the Matrigma Form A were identified for potential DIF between men and women, but only one of these items had a statistically significant probability (Rasch-Welch) of actually causing DIF due to gender differences (Item 17). The item's Mantel-Haenszel statistic did however not yield a statistically significant probability, suggesting that the DIF contrast is caused by factors other than gender differences. The probability that the same item will yield DIF contrasts in a different sample is not likely and therefore it is suggested that none of the current items need to be further investigated for DIF.

Table 21. DIF between women and men participants on Form A

Item	DIF Contrasts	Joint S.E.	Rasch-Welch			Mantel-Haenszel	
			t	d.f.	Prob.	Chi-squ	Prob.
1	0.42	0.86	0.49	83	0.628	0.277	0.599
2	**-2.42**	1.92	-1.26	45	0.215	0.235	0.628
3	0.00	1.30	0.00	75	1.000	0.023	0.879
4	**-1.25**	2.00	-0.62	52	0.536	0.008	0.927
5	-0.14	0.81	-0.18	77	0.859	0.399	0.528
6	**1.32**	0.75	1.76	87	0.082	1.714	0.191
7	**0.80**	1.08	0.74	87	0.462	-*	-*
8	**0.73**	0.57	1.28	87	0.205	2.336	0.126
9	**0.74**	0.59	1.24	87	0.218	0.012	0.912
10	**-1.58**	0.86	-1.84	59	0.071	0.714	0.398
11	-0.40	0.71	-0.57	75	0.571	0.000	1.000
12	**1.65**	0.96	1.72	85	0.088	1.250	0.264
13	-0.21	0.57	-0.37	81	0.709	0.003	0.954
14	**0.73**	0.57	1.28	87	0.205	0.956	0.328
15	**0.51**	0.49	1.04	85	0.299	1.322	0.250
16	0.36	0.50	0.71	85	0.477	0.282	0.595
17	**-1.04**	0.50	-2.07	81	**0.041**	1.788	0.181
18	**0.77**	0.48	1.59	85	0.116	2.746	0.098
19	-0.19	0.48	-0.39	84	0.695	0.037	0.849
20	-0.39	0.48	-0.81	84	0.422	1.078	0.299
21	0.29	0.48	0.60	84	0.551	0.025	0.875
22	**-0.67**	0.58	-1.15	87	0.252	1.076	0.300
23	0.33	0.52	0.64	83	0.524	0.007	0.933
24	0.00	0.49	0.00	84	1.000	0.010	0.920
25	**-0.85**	0.50	-1.72	86	0.090	3.290	0.070
26	-0.41	0.73	-0.55	87	0.582	0.229	0.633
27	0.33	0.58	0.56	82	0.575	0.116	0.734
28	**-0.93**	0.64	-1.46	87	0.147	0.107	0.743
29	0.05	0.63	0.07	84	0.942	0.223	0.637
30	-0.11	1.44	-0.08	86	0.937	0.563	0.453

Note: * - Mantel-Haenszel test statistics are not estimable.

Language

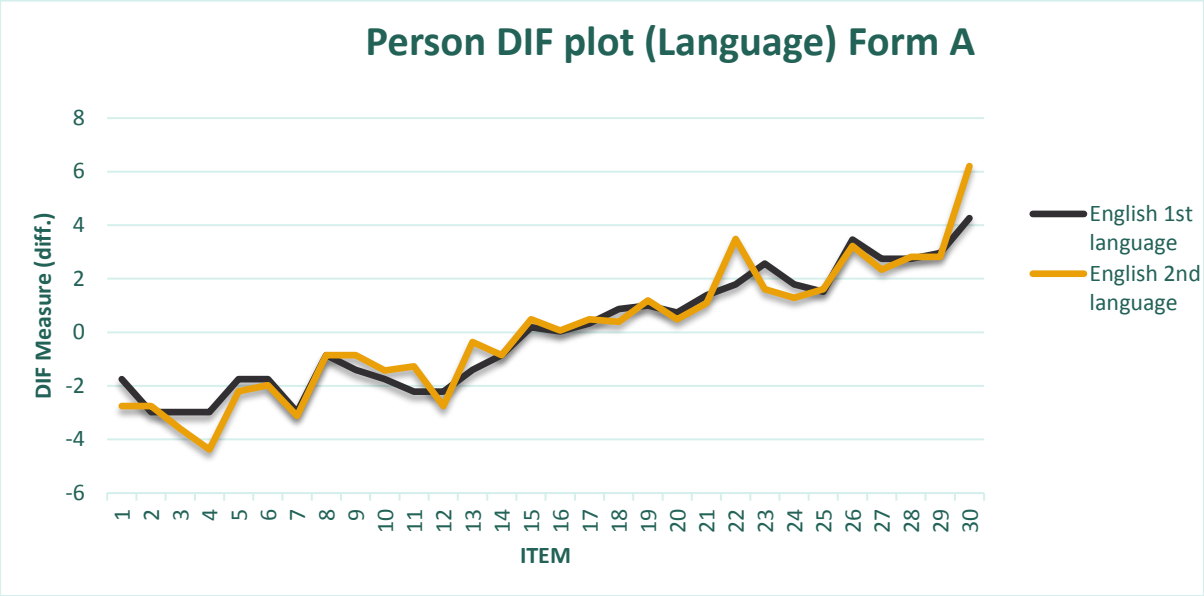

Figure 3. Person DIF plot (Language) Form A

Nearly one third of the items displayed large differences in item difficulty between the two language groupings. When examining Figure 3, it is clear that these differences are not unidirectional, but rather that there is a relatively equal split between the number of items in each direction that had large DIF contrasts for the two groups.

Nine of the items in Form A displayed large DIF contrasts between the two language groups, but only item 22 yielded statistically significant probability (at the $p < .05$ level) for DIF with both the Rasch-Welch and Mantel-Haenszel tests. This could indicate that item 22 will potentially also be flagged for DIF in other samples (although with a stricter cutoff of $p<.005$, this would fall away). This item was more difficult to answer for the English second language participants and should be investigated further to understand why the item performed in this manner.

Table 22. DIF between English 1st language and 2nd language participants on Form A.

Item	DIF Contrasts	Joint S.E.	Rasch-Welch			Mantel-Haenszel	
			t	d.f.	Prob.	Chi-squ	Prob.
1	1.00	0.85	1.18	81	0.241	0.315	0.575
2	-0.23	1.18	-0.19	57	0.849	0.263	0.608
3	**0.62**	1.28	0.49	71	0.629	0.003	0.961
4	**1.40**	1.47	0.95	85	0.343	0.080	0.777
5	0.45	0.80	0.57	73	0.571	0.022	0.883
6	0.23	0.78	0.30	70	0.769	0.019	0.891
7	0.14	1.22	0.11	62	0.910	0.235	0.628
8	0.00	0.60	0.00	72	1.000	0.182	0.670
9	**-0.55**	0.67	-0.83	64	0.409	1.358	0.244
10	-0.32	0.75	-0.43	64	0.668	0.086	0.769
11	**-0.95**	0.85	-1.12	55	0.268	0.007	0.934
12	**0.53**	0.94	0.57	72	0.573	0.500	0.480
13	**-1.04**	0.65	-1.59	61	0.116	0.450	0.502
14	0.00	0.60	0.00	72	1.000	0.023	0.879
15	-0.30	0.50	-0.61	77	0.545	0.106	0.744
16	-0.03	0.51	-0.05	77	0.959	0.044	0.834
17	-0.16	0.49	-0.32	78	0.748	0.001	0.971
18	0.48	0.48	1.00	80	0.322	0.741	0.389
19	-0.19	0.48	-0.40	80	0.687	0.003	0.960
20	0.25	0.48	0.51	79	0.609	0.050	0.823
21	0.29	0.48	0.61	80	0.543	0.000	0.988
22	**-1.71**	0.65	-2.62	85	**0.011**	6.639	**0.010**
23	**0.95**	0.54	1.77	74	0.081	1.126	0.289
24	0.49	0.49	1.00	79	0.320	0.296	0.587
25	-0.10	0.49	-0.20	81	0.845	0.185	0.667
26	0.24	0.73	0.32	80	0.746	0.050	0.823
27	0.41	0.58	0.70	78	0.486	1.193	0.275
28	-0.07	0.62	-0.11	83	0.912	0.069	0.793
29	0.14	0.63	0.22	81	0.826	0.069	0.794
30	**-1.94**	1.98	-0.98	67	0.329	0.500	0.480

Form B

Item Fit

The item fit statistics for Form B of the Matrigma are presented in Table 23. Several items had outfit statistics outside of the parameters of 'good fit', but none of these items were identified as problematic based on their infit statistics. Due to the average MNSQ for the outfit statistics still being relatively close to 1 (*M = 1.1*), the ZSTD mean score being equal to zero, and the fact that outfit statistics are influenced by outliers, it was determined that none of the items for Form B need to be identified for misfit.

Table 23. Item statistics: Form B

Item	Measure	Model	Infit		Outfit	
		S.E.	MNSQ	ZSTD	MNSQ	ZSTD
1	-2.78	0.44	1.33	1.0	0.89	0.0
2	-2.78	0.44	1.11	0.5	1.13	0.4
3	-2.27	0.38	0.88	-0.4	0.87	-0.1
4	-2.42	0.40	1.11	0.5	1.17	0.5
5	-2.27	0.38	0.73	-1.0	0.39	-1.4
6	-1.64	0.33	0.94	-0.2	0.89	-0.1
7	-2.13	0.37	0.71	-1.2	**0.49**	-1.1
8	-1.34	0.31	1.02	0.2	0.88	-0.3
9	-1.16	0.30	1.05	0.4	1.05	0.3
10	-1.64	0.33	0.75	-1.3	**0.51**	-1.4
11	-1.08	0.29	0.85	-0.9	0.71	-0.9
12	-1.87	0.35	0.96	-0.1	**0.68**	-0.7
13	-0.40	0.26	1.01	0.1	0.93	-0.2
14	-0.61	0.27	0.96	-0.2	0.83	-0.6
15	-0.34	0.26	0.89	-0.9	0.84	-0.6
16	-0.01	0.25	0.91	-0.9	0.80	-0.8
17	0.30	0.25	1.25	2.5	1.20	0.8
18	1.04	0.25	0.93	-0.7	0.82	-0.4
19	0.30	0.25	0.98	-0.2	0.93	-0.2
20	0.73	0.25	1.18	1.9	**1.51**	1.6
21	1.30	0.26	1.17	1.6	**2.06**	2.3
22	1.88	0.28	0.83	-1.2	**0.58**	-1.1
23	1.73	0.28	0.97	-0.2	0.83	-0.3
24	1.58	0.27	0.93	-0.5	0.85	-0.3
25	1.65	0.27	0.89	-0.9	0.78	-0.5
26	2.54	0.34	1.04	0.2	1.13	0.4
27	2.05	0.30	1.08	0.5	**1.61**	1.4
28	2.92	0.38	1.06	0.3	**1.94**	1.5
29	3.07	0.40	1.17	0.7	1.27	0.6
30	3.64	0.49	1.04	0.2	**4.28**	2.9
M	**0.00**	**0.32**	**0.99**	**0.0**	**1.10**	**0.0**
P.SD	**1.89**	**0.07**	**0.15**	**0.9**	**0.70**	**1.0**

Differential Item Functioning

Ethnicity

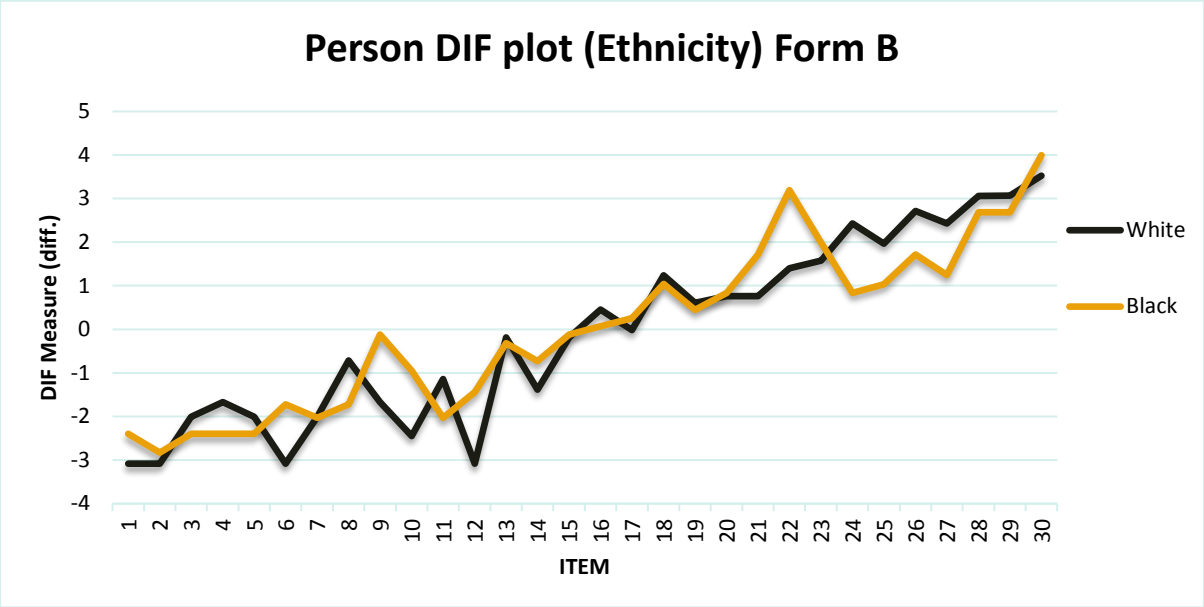

Figure 4. Person DIF plot (Ethnicity) Form B

The differences in item difficulty between Black and White participants are plotted in Figure 4. From the figure, it is clear that several items displayed large DIF contrasts between the two groups. These differences are not unidirectional, with some items being more difficult for one group, while other items were easier for the same group. This gives an indication that the assessment as a whole is not bias to a specific ethnic group in the sample. Table 24 provides an overview of the probability of these items showing DIF due to ethnic differences in the sample.

Three items were highlighted that show significant DIF (at the $p < .50$ level) between the two ethnic groups: Items 9, 22 and 24. All three items had statistically significant probabilities of causing DIF in other samples based on their Rasch-Welch statistics, but none of them were statistically significant based on the Mantel-Haenszel test. It was therefore concluded that none of these items are expected to cause DIF based on ethnic differences in future samples.

Table 24. DIF between White and Black participants on Form B.

Item	DIF Contrasts	Joint S.E.	Rasch-Welch			Mantel-Haenszel	
			t	d.f.	Prob.	Chi-squ	Prob.
1	**-0.69**	1.09	-0.63	53	0.531	0.006	0.937
2	-0.25	1.13	-0.22	56	0.824	0.105	0.746
3	0.39	0.88	0.44	57	0.663	0.100	0.752
4	**0.73**	0.84	0.87	55	0.389	0.272	0.602
5	0.39	0.88	0.44	57	0.663	0.500	0.480
6	**-1.36**	1.05	-1.30	50	0.198	0.112	0.739
7	0.02	0.85	0.03	57	0.978	0.000	1.000
8	1.00	0.70	1.43	53	0.158	0.190	0.663
9	**-1.55**	0.71	-2.19	56	**0.033**	3.760	0.053
10	**-1.50**	0.86	-1.74	52	0.087	0.701	0.403
11	**0.89**	0.75	1.19	54	0.240	2.278	0.131
12	**-1.64**	1.03	-1.59	48	0.118	1.361	0.243
13	0.14	0.60	0.22	56	0.824	0.011	0.917
14	**-0.66**	0.69	-0.95	57	0.345	0.245	0.621
15	-0.06	0.60	-0.10	56	0.921	0.065	0.799
16	0.38	0.59	0.65	56	0.517	0.037	0.848
17	-0.28	0.59	-0.47	56	0.642	0.033	0.856
18	0.20	0.61	0.32	56	0.749	0.003	0.954
19	0.16	0.59	0.27	56	0.788	0.024	0.876
20	-0.07	0.59	-0.12	55	0.902	0.276	0.599
21	**-0.96**	0.64	-1.49	52	0.143	1.866	0.172
22	**-1.79**	0.88	-2.03	41	**0.049**	0.601	0.438
23	-0.41	0.69	-0.59	53	0.555	0.009	0.927
24	**1.60**	0.68	2.35	57	**0.022**	3.603	0.058
25	**0.93**	0.64	1.45	57	0.153	1.786	0.181
26	**1.00**	0.75	1.33	57	0.190	2.014	0.156
27	**1.18**	0.69	1.71	57	0.093	0.736	0.391
28	0.38	0.91	0.41	57	0.681	0.269	0.604
29	0.39	0.91	0.43	57	0.672	0.000	1.000
30	-0.47	1.29	-0.36	49	0.718	0.042	0.838

Gender

Although more than half of the items for Form B showed item difficulty differences larger than .50 for the two gender groups, these differences were not all in the same direction. For both genders, there were items that were easier or more difficult than for the other gender. It does not appear that the assessment is bias against a specific gender group, but to further investigate these items, the probability of them portraying DIF in other samples are presented in Table 25.

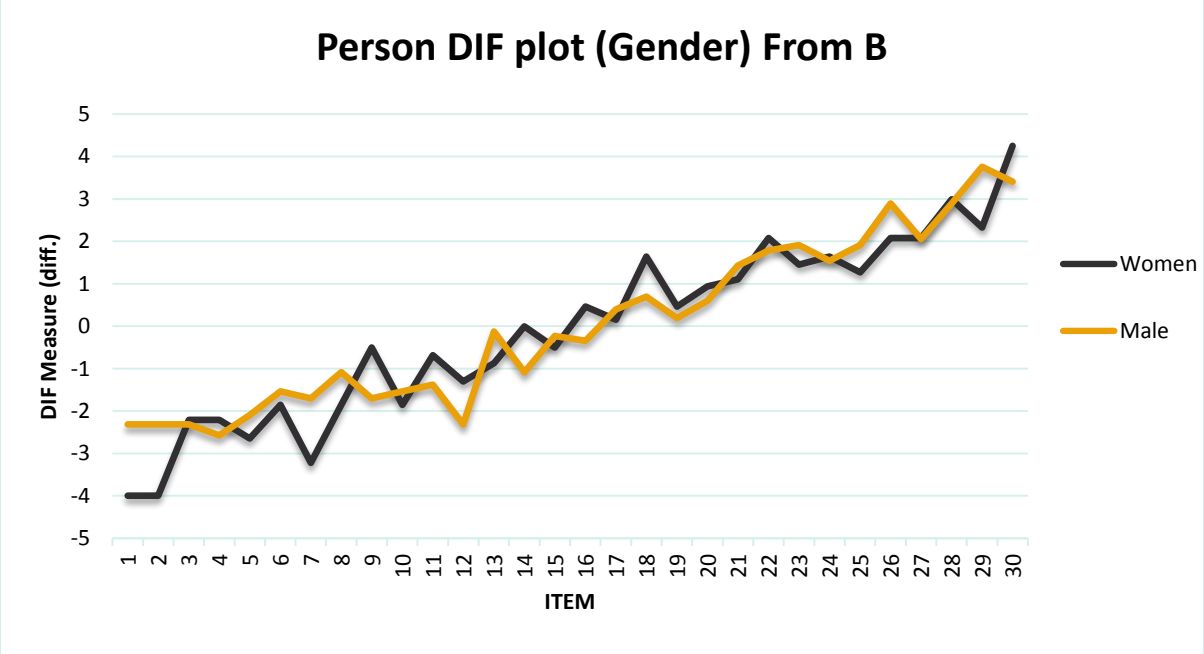

Figure 5. Person DIF plot (Gender) Form B

Of the 16 items that were flagged due to their large DIF contrast values, only item 9 displayed a statistically significant probability that the item will also cause DIF in other samples. This probability is based only on the Rasch-Welch test, and the Mantel-Haenszel test did not pick up any significant probability. It is therefore concluded that there was insufficient evidence for this item to be flagged as causing DIF-based gender differences.

Table 25. DIF between women and men participants on Form B.

Item	DIF Contrasts	Joint S.E.	Rasch-Welch			Mantel-Haenszel	
			t	d.f.	Prob.	Chi-squ	Prob.
1	**-1.68**	1.07	-1.57	48	0.123	0.071	0.790
2	**-1.68**	1.07	-1.57	48	0.123	3.821	0.051
3	0.11	0.79	0.14	67	0.889	0.052	0.819
4	0.37	***0.81***	0.45	70	0.654	0.022	0.883
5	**-0.55**	0.84	-0.66	58	0.514	0.235	0.628
6	-0.32	0.70	-0.46	63	0.647	0.000	0.991
7	**-1.51**	0.92	-1.65	49	0.105	0.033	0.855
8	**-0.77**	0.68	-1.14	58	0.259	1.252	0.263
9	**1.20**	0.59	2.03	78	**0.046**	3.663	0.056
10	-0.32	0.70	-0.46	63	0.647	0.013	0.911
11	**0.69**	0.58	1.19	74	0.237	1.192	0.275
12	**1.01**	0.69	1.46	78	0.148	0.001	0.978
13	**-0.75**	0.55	-1.36	64	0.180	0.460	0.498
14	**1.08**	0.55	1.98	75	0.051	0.704	0.401
15	-0.27	0.53	-0.50	67	0.616	0.021	0.885
16	**0.81**	0.52	1.56	71	0.123	1.555	0.212
17	-0.24	0.51	-0.48	68	0.635	0.018	0.893
18	**0.94**	0.55	1.73	62	0.089	1.858	0.173
19	0.27	0.51	0.54	69	0.594	0.044	0.835
20	0.34	0.51	0.66	67	0.510	0.140	0.709
21	-0.33	0.53	-0.62	69	0.536	0.705	0.401
22	0.29	0.60	0.48	63	0.629	0.013	0.911
23	-0.46	0.56	-0.82	70	0.418	0.482	0.488
24	0.10	0.56	0.18	66	0.860	0.044	0.834
25	**-0.64**	0.55	-1.15	72	0.253	0.982	0.322
26	**-0.82**	0.68	-1.21	76	0.231	0.083	0.774
27	0.02	0.61	0.04	65	0.971	0.118	0.731
28	0.10	0.79	0.12	63	0.902	0.106	0.745
29	**-1.43**	0.82	-1.74	82	0.085	0.289	0.591
30	**0.84**	1.18	0.71	50	0.478	0.015	0.903

Language

Several items displayed large DIF contrast for the two language groups, but they did not all influence one specific group in a similar direction. Each group experienced some of the items with large DIF contrasts as either easier or more difficult than the other group. In other words, it does not appear that the assessment as a whole is negatively or positively influencing one specific language group.

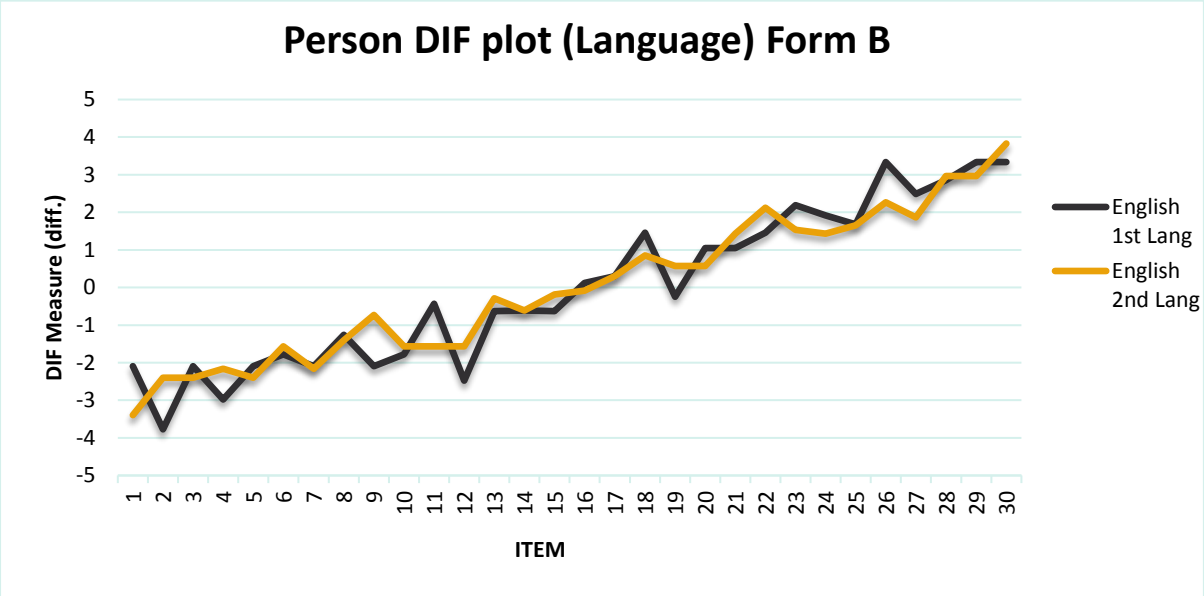

Figure 6. Person DIF plot (Language) Form B

None of the items flagged for having large DIF contrasts showed a statistically significant probability of displaying DIF in other samples based on language differences. None of the items are therefore flagged for further investigation and the research concluded that those items with large DIF contrasts was based on a sample specific characteristic, rather than language differences.

Table 26. DIF between English 1st language and 2nd language participants on Form B.

Item	DIF Contrasts	Joint S.E.	Rasch-Welch			Mantel-Haenszel	
			t	d.f.	Prob.	Chi-squ	Prob.
1	**1.30**	0.88	1.48	78	0.143	0.121	0.728
2	**-1.37**	1.17	-1.17	41	0.248	1.663	0.197
3	0.31	0.78	0.40	66	0.694	0.854	0.355
4	**-0.82**	0.91	-0.90	49	0.372	0.030	0.863
5	0.31	0.78	0.40	66	0.694	0.160	0.689
6	-0.21	0.68	-0.31	59	0.757	0.003	0.958
7	0.06	0.76	0.08	63	0.933	0.225	0.635
8	0.14	0.63	0.23	63	0.820	0.189	0.664
9	**-1.37**	0.68	-2.02	47	0.050	2.335	0.127
10	-0.21	0.68	-0.31	59	0.757	0.021	0.885
11	**1.14**	0.60	1.89	71	0.062	3.904	0.048
12	**-0.91**	0.77	-1.17	50	0.247	0.633	0.426
13	-0.34	0.55	-0.62	57	0.538	0.000	0.986
14	0.00	0.56	0.00	59	1.000	0.001	0.972
15	-0.44	0.55	-0.81	56	0.422	0.617	0.432
16	0.21	0.53	0.39	57	0.698	0.040	0.841
17	0.00	0.52	0.00	56	1.000	0.091	0.763
18	**0.61**	0.55	1.10	52	0.277	1.174	0.279
19	**-0.82**	0.53	-1.55	55	0.126	2.398	0.122
20	0.48	0.54	0.89	54	0.378	2.400	0.121
21	-0.38	0.55	-0.69	57	0.491	0.074	0.786
22	**-0.67**	0.59	-1.13	61	0.265	0.089	0.765
23	**0.65**	0.62	1.05	49	0.300	0.422	0.516
24	0.49	0.59	0.83	51	0.412	0.155	0.694
25	0.03	0.58	0.05	54	0.960	0.031	0.860
26	**1.07**	0.85	1.27	42	0.212	0.929	0.335
27	**0.62**	0.67	0.92	48	0.360	0.088	0.767
28	-0.11	0.79	-0.13	57	0.894	0.123	0.725
29	0.37	0.89	0.42	49	0.676	0.014	0.905
30	-0.49	0.99	-0.50	64	0.620	0.033	0.855

Form C

Item Fit

The item statistics for Form C are displayed in

Table 27. There were ten items flagged for potential outfit, all falling outside the parameters for the

Item	Measure	Model	Infit		Outfit	
		S.E.	MNSQ	ZSTD	MNSQ	ZSTD
1	-2.98	0.45	1.37	1.1	0.70	-0.2
2	-1.35	0.29	0.79	-1.2	**0.66**	-1.1
3	-2.06	0.35	0.82	-0.7	0.73	-0.5
4	-2.32	0.37	0.88	-0.4	**0.65**	-0.6
5	-1.94	0.34	0.76	-1.1	0.80	-0.3
6	-1.94	0.34	0.72	-1.3	**0.51**	-1.3
7	-1.83	0.33	0.70	-1.5	**0.46**	-1.6
8	-1.18	0.28	1.06	0.4	1.19	0.7
9	-1.18	0.28	0.91	-0.5	0.86	-0.4
10	-1.44	0.30	1.14	0.8	1.29	0.9
11	-0.95	0.27	0.86	-0.9	0.77	-0.9
12	-1.18	0.28	0.94	-0.3	0.80	-0.6
13	-0.29	0.25	0.93	-0.6	0.98	0.0
14	-0.47	0.25	0.89	-0.9	0.85	-0.6
15	-0.29	0.25	1.03	0.3	1.14	0.7
16	0.01	0.24	1.00	0.0	0.97	-0.1
17	0.40	0.24	0.97	-0.2	0.90	-0.3
18	-0.54	0.26	1.23	1.7	**1.63**	2.4
19	0.68	0.23	1.04	0.5	1.13	0.6
20	0.35	0.24	1.02	0.2	1.17	0.7
21	1.12	0.24	1.10	1.0	**1.50**	1.4
22	1.87	0.26	0.86	-1.1	0.95	0.0
23	1.36	0.24	1.03	0.3	1.09	0.4
24	2.08	0.27	1.07	0.5	0.93	0.0
25	1.94	0.27	0.93	-0.5	0.95	0.0
26	2.01	0.27	0.93	-0.5	1.01	0.2
27	2.49	0.30	1.04	0.3	**1.68**	1.4
28	2.49	0.30	1.30	1.6	**4.14**	4.2
29	2.16	0.28	0.86	-0.9	**0.61**	-1.0
30	3.02	0.35	1.26	1.1	**1.99**	1.6
M	**0.00**	**0.29**	**0.98**	**-0.1**	**1.10**	**0.2**
P.SD	**1.68**	**0.05**	**0.16**	**0.9**	**0.66**	**1.2**

outfit statistics. All of these items, however, still fell within the parameters of the infit statistic. The mean MNSQ value for both the infit and outfit were still around 1.00, and the standardised fit statistic mean value for both fit indices were close to 0.0, leading to the conclusion that the data did not warrant the exclusion of any of these items from the assessment. Therefore, none of the items are identified for misfit.

Table 27. Item statistics: Form C

Item	Measure	Model	Infit		Outfit	
		S.E.	MNSQ	ZSTD	MNSQ	ZSTD
1	-2.98	0.45	1.37	1.1	0.70	-0.2
2	-1.35	0.29	0.79	-1.2	**0.66**	-1.1
3	-2.06	0.35	0.82	-0.7	0.73	-0.5
4	-2.32	0.37	0.88	-0.4	**0.65**	-0.6
5	-1.94	0.34	0.76	-1.1	0.80	-0.3
6	-1.94	0.34	0.72	-1.3	**0.51**	-1.3
7	-1.83	0.33	0.70	-1.5	**0.46**	-1.6
8	-1.18	0.28	1.06	0.4	1.19	0.7
9	-1.18	0.28	0.91	-0.5	0.86	-0.4
10	-1.44	0.30	1.14	0.8	1.29	0.9
11	-0.95	0.27	0.86	-0.9	0.77	-0.9
12	-1.18	0.28	0.94	-0.3	0.80	-0.6
13	-0.29	0.25	0.93	-0.6	0.98	0.0
14	-0.47	0.25	0.89	-0.9	0.85	-0.6
15	-0.29	0.25	1.03	0.3	1.14	0.7
16	0.01	0.24	1.00	0.0	0.97	-0.1
17	0.40	0.24	0.97	-0.2	0.90	-0.3
18	-0.54	0.26	1.23	1.7	**1.63**	2.4
19	0.68	0.23	1.04	0.5	1.13	0.6
20	0.35	0.24	1.02	0.2	1.17	0.7
21	1.12	0.24	1.10	1.0	**1.50**	1.4
22	1.87	0.26	0.86	-1.1	0.95	0.0
23	1.36	0.24	1.03	0.3	1.09	0.4
24	2.08	0.27	1.07	0.5	0.93	0.0
25	1.94	0.27	0.93	-0.5	0.95	0.0
26	2.01	0.27	0.93	-0.5	1.01	0.2
27	2.49	0.30	1.04	0.3	**1.68**	1.4
28	2.49	0.30	1.30	1.6	**4.14**	4.2
29	2.16	0.28	0.86	-0.9	**0.61**	-1.0
30	3.02	0.35	1.26	1.1	**1.99**	1.6
M	**0.00**	**0.29**	**0.98**	**-0.1**	**1.10**	**0.2**
P.SD	**1.68**	**0.05**	**0.16**	**0.9**	**0.66**	**1.2**

Differential Item Functioning

Ethnicity

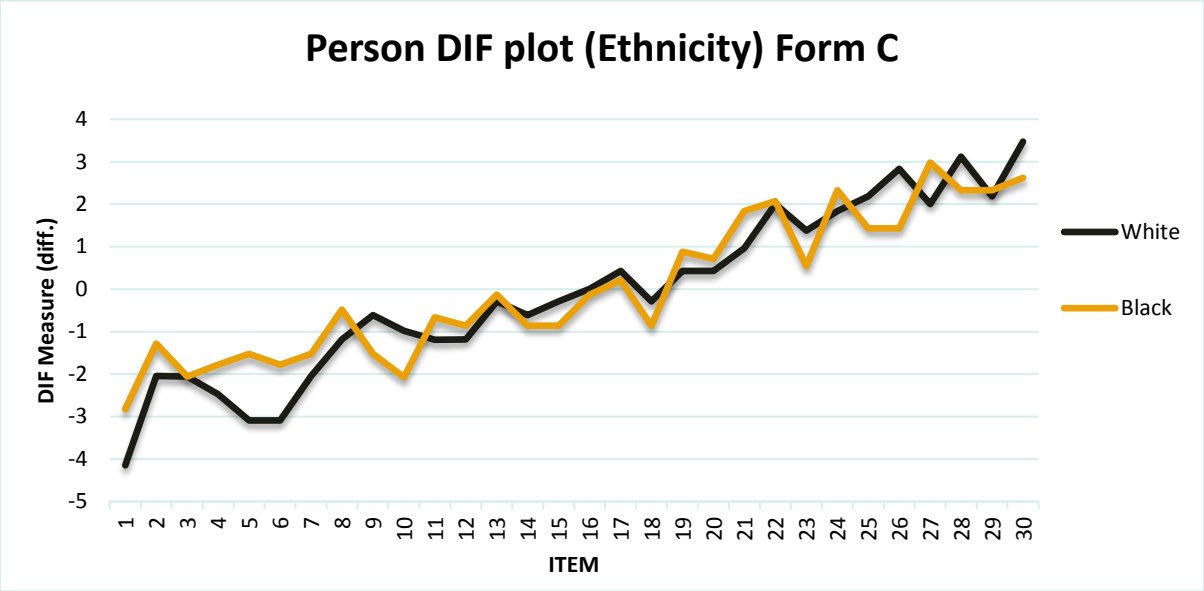

Figure 7. Person DIF plot (Ethnicity) Form C

The DIF plot in Figure 7 indicates numerous items with large differences between how difficult or easy the two ethnic population groups experienced the items. These differences are not unidirectional; in other words, both ethnic groups appeared to have experienced certain items as more difficult than the other population group. Therefore, the assessment as a whole does not unfairly bias one ethnic group over the other. To explore whether specific items were flagged for potential bias, Rasch-Welch and Mantel-Haenszel tests were run and the results are displayed in Table 28.

Item 26 was the only item with large DIF contrast that also showed a statistically significant (Mantel-Haneszel test) probability of causing DIF between ethnic groups in other samples. It is recommended that this item is explored further to determine why it appears to be biased against the White population group.

Table 28. DIF between White and Black participants on Form C

Item	DIF Contrasts	Joint S.E.	Rasch-Welch			Mantel-Haenszel	
			t	d.f.	Prob.	Chi-squ	Prob.
1	**-1.32**	1.39	-0.95	55	0.330	0.047	0.829
2	**-0.76**	0.77	-0.98	65	1.000	0.169	0.681
3	0.00	0.83	0.00	67	0.435	0.002	0.968
4	**-0.69**	0.88	-0.79	64	0.125	0.007	0.934
5	**-1.57**	1.01	-1.56	56	0.206	0.007	0.934
6	**-1.31**	1.02	-1.28	58	0.509	0.185	0.667
7	**-0.52**	0.79	-0.66	66	0.275	0.096	0.757
8	**-0.70**	0.64	-1.10	67	0.165	2.478	0.115
9	**0.91**	0.65	1.40	64	0.133	0.448	0.503
10	**1.09**	0.72	1.52	63	0.415	0.234	0.628
11	**-0.53**	0.64	-0.82	67	0.621	0.121	0.728
12	-0.32	0.65	-0.50	67	0.785	0.003	0.955
13	-0.16	0.57	-0.27	67	0.681	0.078	0.780
14	0.25	0.61	0.41	66	0.337	0.005	0.944
15	**0.57**	0.59	0.97	65	0.805	0.002	0.967
16	0.14	0.56	0.25	66	0.690	0.067	0.796
17	0.22	0.55	0.40	66	0.335	0.679	0.410
18	0.58	0.59	0.97	65	0.416	0.287	0.592
19	-0.45	0.56	-0.82	65	0.610	0.008	0.931
20	-0.28	0.55	-0.51	66	0.148	0.808	0.369
21	-0.87	0.60	-1.47	62	0.919	0.080	0.778
22	-0.07	0.64	-0.10	64	0.138	0.761	0.383
23	0.84	0.56	1.50	66	0.461	0.002	0.962
24	-0.49	0.66	-0.74	61	0.225	0.821	0.365
25	**0.75**	0.62	1.23	67	0.041	2.849	0.091
26	**1.41**	0.67	2.09	67	0.204	4.116	0.043
27	**-0.98**	0.76	-1.29	56	0.302	0.009	0.925
28	**0.80**	0.77	1.04	67	0.834	0.010	0.920
29	-0.14	0.68	-0.21	64	0.319	0.001	0.971
30	**0.85**	0.85	1.00	67	0.330	0.047	0.829

Gender

Figure 8 represents the differences in item difficulty between women and men participants on Form C of the Matrigma. From the figure, it can be seen that the item difficulty differences are not unidirectional and that different items appeared to be more difficult for the different gender groupings. This suggests that the measure as a whole does not influence the performance of a specific gender group in any particular direction.

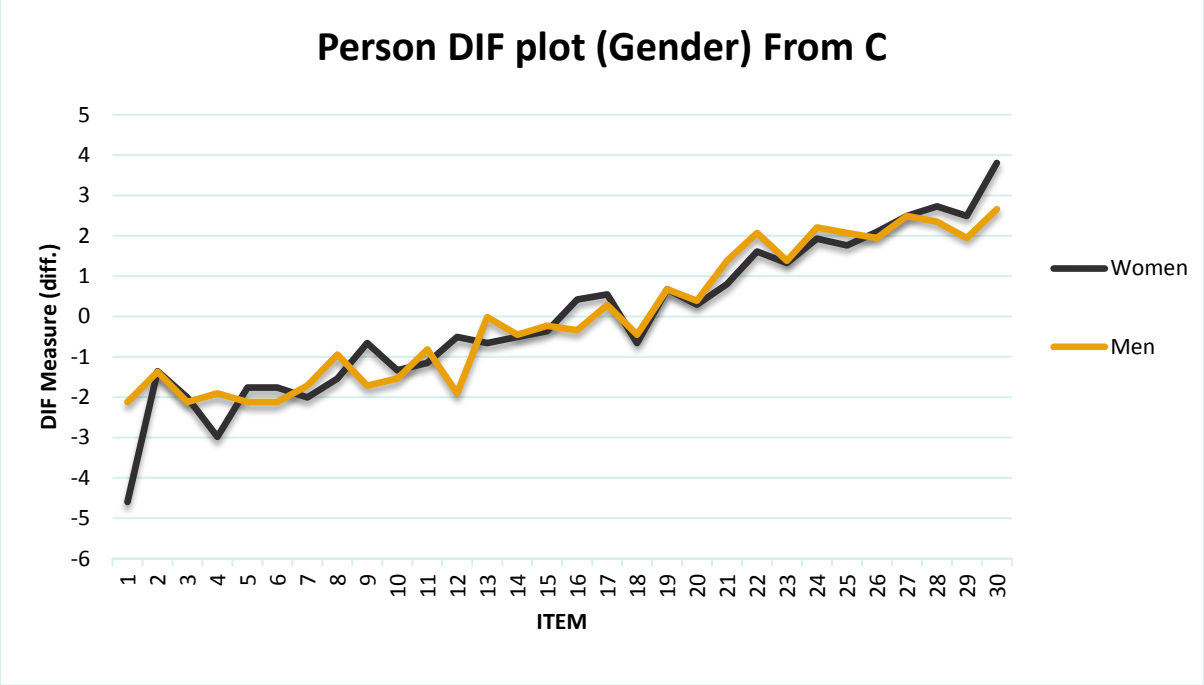

Figure 8. Person DIF plot (Gender) Form C

Table 29 gives an overview of the results of the Rasch-Welch and Mantel-Haneszel tests. Ten items with large DIF contrasts (> .50, < -.50) were flagged in the table. Only one of these items (number 12) had a statistically significant probability of causing DIF in other samples. Item 12 yielded a significant result on the Rasch-Welch test. Due to the sample size and the non-significant result from the Mantel-Haenszel test for item 12, it was concluded that there is insufficient evidence that the item is indeed causing DIF due to gender differences.

Table 29. Gender DIF Form C

Item	DIF Contrasts	Joint S.E.	Rasch-Welch			Mantel-Haenszel	
			t	d.f.	Prob.	Chi-squ	Prob.
1	**-2.48**	0.98	-2.52	66	0.014	0.567	0.452
2	0.02	0.59	0.04	89	0.969	0.015	0.902
3	0.11	0.70	0.16	91	0.873	0.050	0.823
4	**-1.08**	0.78	-1.38	78	0.171	2.036	0.154
5	0.36	0.68	0.53	92	0.597	0.155	0.694
6	0.36	0.68	0.53	92	0.597	0.040	0.842
7	-0.29	0.67	-0.44	86	0.661	0.000	0.995
8	**-0.59**	0.59	-1.01	84	0.314	0.000	0.990
9	**1.06**	0.58	1.83	93	0.071	3.896	0.048
10	0.20	0.60	0.34	91	0.737	0.917	0.338
11	-0.34	0.55	-0.60	87	0.547	0.158	0.691
12	**1.40**	0.59	2.37	93	**0.020**	3.024	0.082
13	**-0.64**	0.51	-1.26	86	0.210	0.038	0.846
14	-0.06	0.51	-0.11	89	0.914	0.334	0.563
15	-0.14	0.50	-0.27	89	0.787	0.078	0.780
16	**0.76**	0.49	1.56	91	0.122	2.355	0.125
17	0.26	0.47	0.54	89	0.588	0.004	0.953
18	-0.20	0.52	-0.39	88	0.694	0.020	0.888
19	0.00	0.47	0.00	89	1.000	0.016	0.900
20	-0.09	0.48	-0.20	89	0.845	0.287	0.592
21	**-0.59**	0.48	-1.22	90	0.227	0.103	0.748
22	-0.46	0.53	-0.88	91	0.381	0.809	0.368
23	-0.06	0.49	-0.12	88	0.902	0.946	0.331
24	-0.28	0.55	-0.51	89	0.611	0.022	0.883
25	-0.31	0.53	-0.58	90	0.563	0.456	0.500
26	0.15	0.55	0.27	86	0.786	0.001	0.982
27	0.00	0.61	0.00	87	1.000	0.014	0.908
28	0.38	0.63	0.61	82	0.543	0.652	0.420
29	**0.55**	0.59	0.93	81	0.353	0.090	0.765
30	**1.15**	0.85	1.35	66	0.181	1.967	0.161

Language

Figure 9 presents the item difficulty curves for both first language and second language English speakers from the current sample. It is clear that several items appeared more difficult for each one of these groups, with several large differences in the item difficulty for more than a third of the items. These differences are not all specific to one group and therefore it was concluded that Form C of the Matrigma as a whole entity is not biased against a specific language grouping.

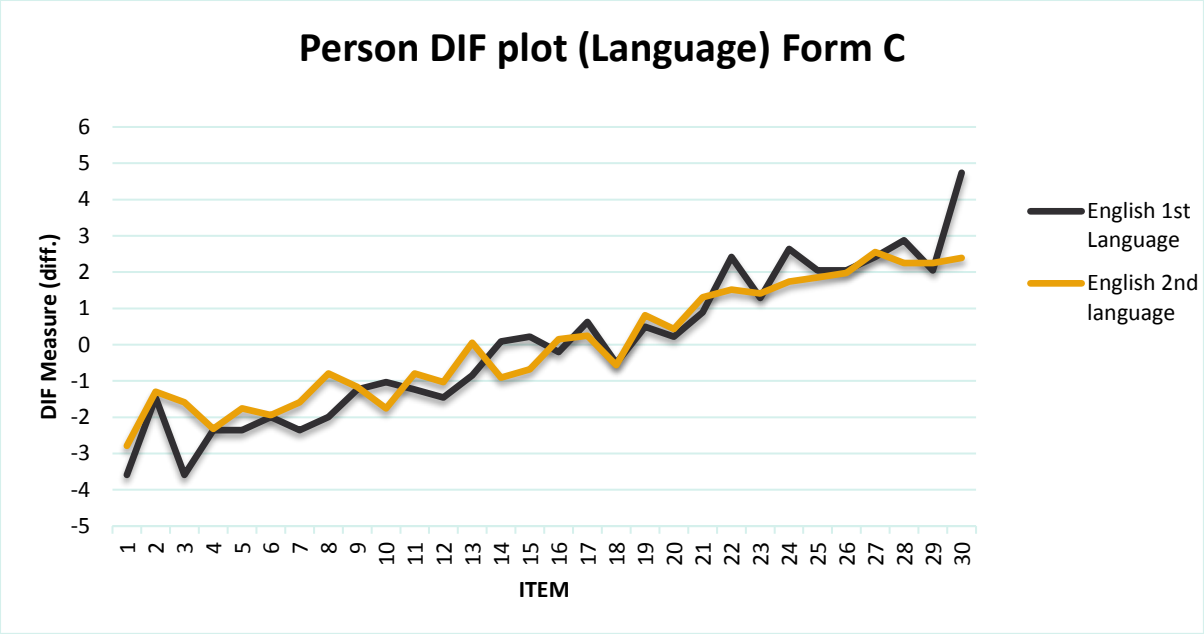

Figure 9. Person DIF plot (Language) Form C

From

Table 30, it is clear that only one of the flagged items showed significant probability that the items cause different language groups to perform differently on the item. It is suggested that this item is further explored to determine why the data suggests that the item shows bias against language groups.

Table 30. Language DIF Form C

Item	DIF Contrasts	Joint S.E.	Rasch-Welch			Mantel-Haenszel	
			t	d.f.	Prob.	Chi-squ	Prob.
1	**-0.80**	1.15	-0.69	55	0.491	0.381	0.537
2	-0.16	0.61	-0.26	76	0.797	0.057	0.811
3	**-2.00**	1.11	-1.80	47	0.078	0.898	0.343
4	-0.03	0.78	-0.04	73	0.967	0.000	0.992
5	-0.60	0.75	-0.80	66	0.429	0.009	0.923
6	-0.05	0.71	-0.07	74	0.941	0.067	0.796
7	**-0.76**	0.75	-1.02	64	0.312	0.000	0.987
8	**-1.21**	0.66	-1.83	63	0.072	2.121	0.145
9	-0.07	0.59	-0.12	78	0.903	0.005	0.941
10	**0.72**	0.60	1.20	86	0.233	0.233	0.629
11	-0.44	0.57	-0.77	74	0.441	0.499	0.480
12	-0.42	0.60	-0.70	73	0.484	0.002	0.965
13	**-0.90**	0.53	-1.70	73	0.093	1.695	0.193
14	**0.99**	0.51	1.95	86	0.055	2.454	0.117
15	**0.90**	0.50	1.80	85	0.076	1.592	0.207
16	-0.35	0.49	-0.71	79	0.482	0.155	0.694
17	0.38	0.48	0.80	82	0.428	0.029	0.865
18	0.06	0.52	0.11	80	0.914	0.060	0.807
19	-0.32	0.48	-0.67	81	0.503	0.019	0.891
20	-0.21	0.48	-0.44	80	0.662	0.326	0.568
21	-0.42	0.48	-0.86	83	0.392	0.371	0.543
22	**0.91**	0.56	1.63	73	0.108	1.821	0.177
23	-0.12	0.49	-0.24	83	0.813	0.062	0.804
24	**0.90**	0.58	1.53	72	0.129	2.151	0.143
25	0.19	0.54	0.36	81	0.721	0.026	0.872
26	0.07	0.55	0.13	82	0.897	0.000	0.992
27	-0.13	0.61	-0.22	85	0.830	0.002	0.969
28	**0.63**	0.63	1.00	74	0.323	0.000	0.991
29	-0.20	0.56	-0.35	85	0.727	0.001	0.974
30	**2.34**	1.10	2.12	47	**0.039**	3.965	**0.047**

Form D

Item Fit

The fit statistics for Form D appear in Table 31. Eleven items were identified for potential misfit based on their outfit statistics. Of these 11 items, only two items had infit statistics falling beyond the parameters of good fit (items 1 and 2). Item 1 had an infit MNSQ value above 1.30, indicating that this item might potentially be measuring a different construct than the rest of the items in the assessment. Item 2 yielded an infit MNSQ value below .70, indicating that this item might be redundant and not be adding any additional information about candidates' performance.

Table 31. Item-fit statistics: Form D

Item	Measure	Model	Infit		Outfit	
		S.E.	MNSQ	ZSTD	MNSQ	ZSTD
1	-2.75	0.43	**1.37**	1.2	0.73	-0.3
2	-2.43	0.39	**0.63**	-1.4	**0.30**	-1.6
3	-2.28	0.37	0.78	-0.8	0.56	-0.9
4	-1.60	0.31	0.81	-1.0	**0.60**	-1.2
5	-1.80	0.33	1.02	0.2	0.93	0.0
6	-1.41	0.30	1.03	0.3	0.94	-0.1
7	-1.24	0.29	0.99	0.0	0.81	-0.6
8	-1.80	0.33	0.84	-0.7	**0.50**	-1.4
9	-1.50	0.30	0.89	-0.5	**0.66**	-1.1
10	-1.41	0.30	1.08	0.5	**1.37**	1.2
11	-1.01	0.27	0.84	-1.1	0.73	-1.0
12	-1.01	0.27	0.86	-0.9	0.65	-1.4
13	-0.05	0.24	0.95	-0.5	0.83	-0.7
14	-0.59	0.26	0.93	-0.6	0.86	-0.5
15	0.07	0.24	0.96	-0.4	0.84	-0.6
16	-0.53	0.25	0.88	-1.0	0.84	-0.7
17	0.46	0.24	1.03	0.3	0.97	0.0
18	0.13	0.24	1.26	2.6	**1.42**	1.7
19	-0.10	0.24	1.14	1.4	1.14	0.7
20	0.97	0.24	0.98	-0.1	0.99	0.1
21	0.46	0.24	1.01	0.1	0.98	0.0
22	1.27	0.25	1.04	0.4	1.01	0.2
23	1.09	0.24	0.89	-1.1	0.94	-0.1
24	1.21	0.25	1.16	1.5	**2.31**	2.8
25	1.80	0.27	1.04	0.3	1.10	0.4
26	2.03	0.28	1.16	1.1	**2.74**	3.0
27	2.37	0.30	0.97	-0.1	0.99	0.1
28	2.03	0.28	0.87	-0.9	0.74	-0.5
29	3.34	0.41	1.20	0.7	**3.10**	2.5
30	4.29	0.60	1.06	0.3	**8.88**	4.0
M	**0.00**	**0.30**	**0.99**	**0.0**	**1.32**	**0.1**
P.SD	**1.74**	**0.08**	**0.15**	**0.9**	**1.53**	**1.4**

When investigating the mean MNSQ values for the infit statistic, it is close to 1.00. Neither item 1 nor 2 had ZSTD statistics beyond the acceptable parameters, and the mean ZSTD for infit statistics was 0.00. These results lead to the conclusion that the evidence is insufficient to support misfit of items 1 and 2.

Differential Item Functioning

Ethnicity

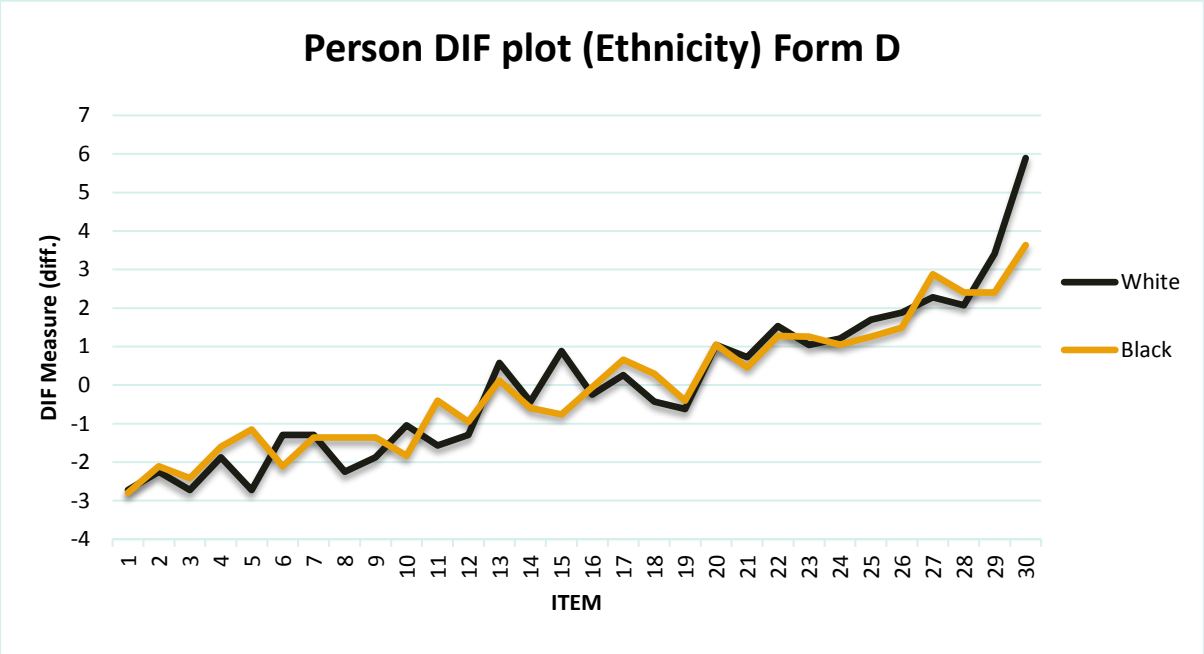

Figure 10. Person DIF plot (Ethnicity) Form D

Figure 10 gives an indication of how two ethnic groups (White and Black) performed on the Matrigma in terms of item difficulty measures. From the graph, it is evident that several items had large item difficulty differences (>.50, <-.50) for the two population groups. It is however also notable that these differences are not unidirectional – i.e., the test is not consistently easier for one of the specific ethnic groups. This led to the conclusion that Form D of the Matrigma is not biased against a specific ethnic group, but the individual item statistics are also explored in Table 32.

One third of the items for Form D were flagged for potential DIF based on the size of their DIF contrast values. Of these 10 items, only 1 item (item 15) yielded statistically significant probabilities on both the Rasch-Welch and Mantel-Haenszel tests. This could indicate that the item will also cause different ethnic groups to perform differently on the item in other samples. Further research with a larger sample size is needed to establish if this item is indeed biased against White participants (i.e., causing them to score lower).

Table 32. DIF between White and Black participants on Form D.

Item	DIF Contrasts	Joint S.E.	Rasch-Welch			Mantel-Haenszel	
			t	d.f.	Prob.	Chi-squ	Prob.
1	0.07	0.97	0.07	61	0.943	0.111	0.739
2	-0.15	0.84	-0.18	61	0.861	0.000	1.000
3	-0.30	0.94	-0.32	60	0.747	0.143	0.706
4	-0.28	0.76	-0.37	61	0.712	0.014	0.906
5	**-1.57**	0.86	-1.83	53	0.073	1.141	0.286
6	**0.81**	0.74	1.10	60	0.277	0.005	0.947
7	0.07	0.69	0.10	61	0.922	0.433	0.511
8	**-0.89**	0.79	-1.12	58	0.265	0.091	0.763
9	**-0.52**	0.75	-0.70	60	0.489	0.033	0.855
10	**0.78**	0.70	1.12	61	0.268	0.083	0.774
11	**-1.16**	0.68	-1.70	60	0.095	2.358	0.125
12	-0.34	0.67	-0.51	61	0.615	0.000	0.984
13	0.45	0.57	0.79	61	0.433	0.003	0.958
14	0.17	0.61	0.27	61	0.784	0.005	0.943
15	**1.65**	0.58	2.82	60	**0.007**	4.226	**0.040**
16	-0.19	0.59	-0.32	61	0.747	0.125	0.724
17	-0.40	0.59	-0.68	60	0.501	0.107	0.743
18	**-0.72**	0.60	-1.20	61	0.235	1.454	0.228
19	-0.22	0.61	-0.35	61	0.724	0.147	0.702
20	0.00	0.60	-0.01	59	0.993	0.357	0.550
21	0.27	0.58	0.46	60	0.647	0.201	0.654
22	0.26	0.62	0.41	59	0.683	0.074	0.785
23	-0.22	0.61	-0.35	58	0.725	0.014	0.907
24	0.16	0.60	0.27	59	0.787	0.283	0.595
25	0.44	0.63	0.70	60	0.486	0.019	0.891
26	0.39	0.65	0.60	59	0.552	0.001	0.983
27	**-0.59**	0.88	-0.67	48	0.505	0.186	0.666
28	-0.33	0.78	-0.42	53	0.675	0.191	0.662
29	**1.01**	0.90	1.12	61	0.266	0.009	0.925
30	**2.26**	2.11	1.07	51	0.290	-*	-*

Note: * = Mantel-Haenszel statistics could not be estimated.

Gender

In Figure 11, the various item difficulties are plotted for both women and men candidates. From the graph, it is clear that many of the items performed differently for the two gender groups. These differences are not unidirectional, but rather random and the item level statistics are further explored in Table 33.

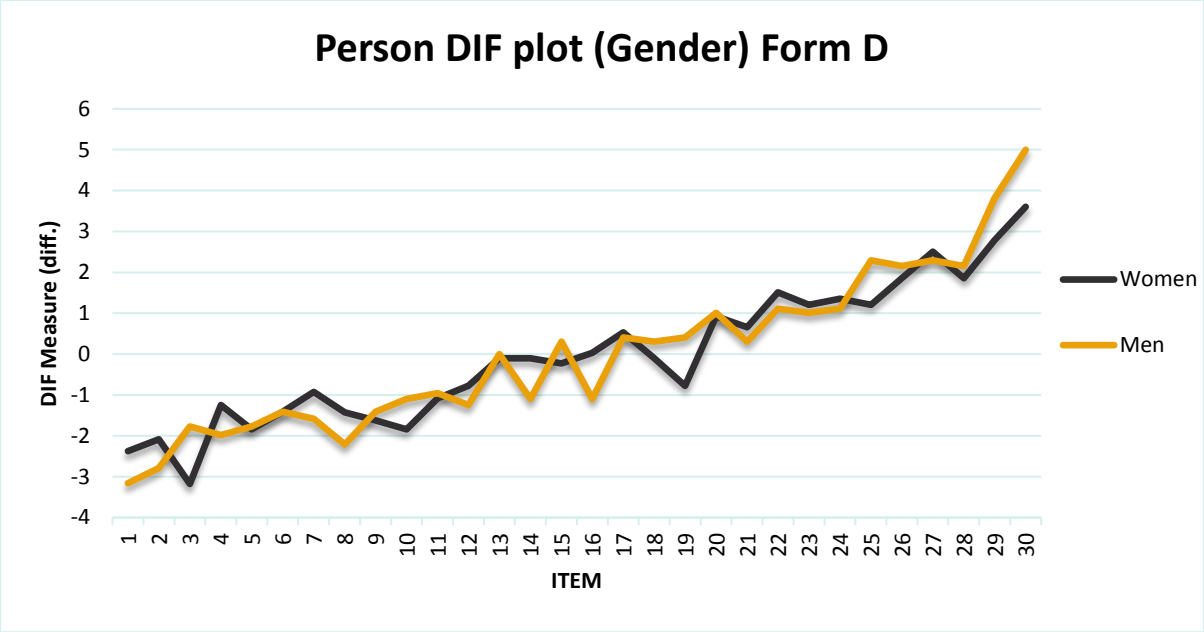

Figure 11. Person DIF plot (Gender) Form D

Only two items (19 and 25) showed statistically significant probability on the Rasch-Welch test, indicating that these items might cause differences between how women and men participants perform in other samples. These items were, however, not significant based on the Mantel-Haenszel results and due to the small population size there is insufficient evidence to support that these items will cause lower performance for men in subsequent groups.

Table 33. DIF between women and men participants on Form D.

Item	DIF Contrasts	Joint S.E.	Rasch-Welch			Mantel-Haenszel	
			t	d.f.	Prob.	Chi-squ	Prob.
1	**0.78**	0.85	0.92	90	0.361	0.089	0.765
2	**0.70**	0.78	0.90	90	0.369	1.511	0.219
3	**-1.40**	0.84	-1.66	67	0.101	0.695	0.405
4	**0.73**	0.63	1.17	90	0.245	0.743	0.389
5	-0.07	0.65	-0.10	86	0.918	0.083	0.773
6	0.00	0.60	0.00	87	1.000	0.055	0.815
7	**0.66**	0.58	1.14	90	0.256	0.068	0.795
8	**0.79**	0.66	1.19	90	0.236	0.061	0.805
9	-0.21	0.61	-0.35	85	0.728	0.104	0.747
10	**-0.74**	0.61	-1.21	80	0.230	0.819	0.365
11	-0.13	0.55	-0.23	86	0.820	0.079	0.779
12	0.47	0.55	0.86	89	0.390	0.324	0.569
13	-0.10	0.48	-0.20	86	0.843	0.017	0.897
14	**1.00**	0.52	1.90	90	0.060	1.607	0.205
15	**-0.54**	0.48	-1.11	84	0.269	0.664	0.415
16	**1.12**	0.52	2.15	90	0.034	1.723	0.189
17	0.12	0.48	0.26	85	0.796	0.022	0.882
18	-0.41	0.48	-0.85	85	0.398	0.455	0.500
19	**-1.18**	0.50	-2.38	82	**0.020**	2.060	0.151
20	-0.09	0.49	-0.18	84	0.860	0.025	0.875
21	0.35	0.48	0.74	84	0.464	0.056	0.813
22	0.40	0.51	0.77	80	0.444	0.336	0.562
23	0.19	0.50	0.39	82	0.698	0.092	0.762
24	0.24	0.51	0.47	81	0.637	1.425	0.233
25	**-1.09**	0.54	-2.03	89	**0.046**	0.288	0.591
26	-0.30	0.57	-0.54	83	0.594	0.004	0.950
27	0.20	0.64	0.32	76	0.749	0.215	0.643
28	-0.30	0.57	-0.54	83	0.594	0.037	0.847
29	**-1.03**	0.83	-1.24	90	0.219	0.020	0.887
30	**-1.39**	1.27	-1.10	88	0.275	1.049	0.306

Language

In Figure 12 the various item difficulties are plotted for first and second language English speakers. From the graph, it is clear that some of the items performed differently for the two language groups. These differences are not unidirectional, but rather random and the item level statistics are further explored in Table 34.

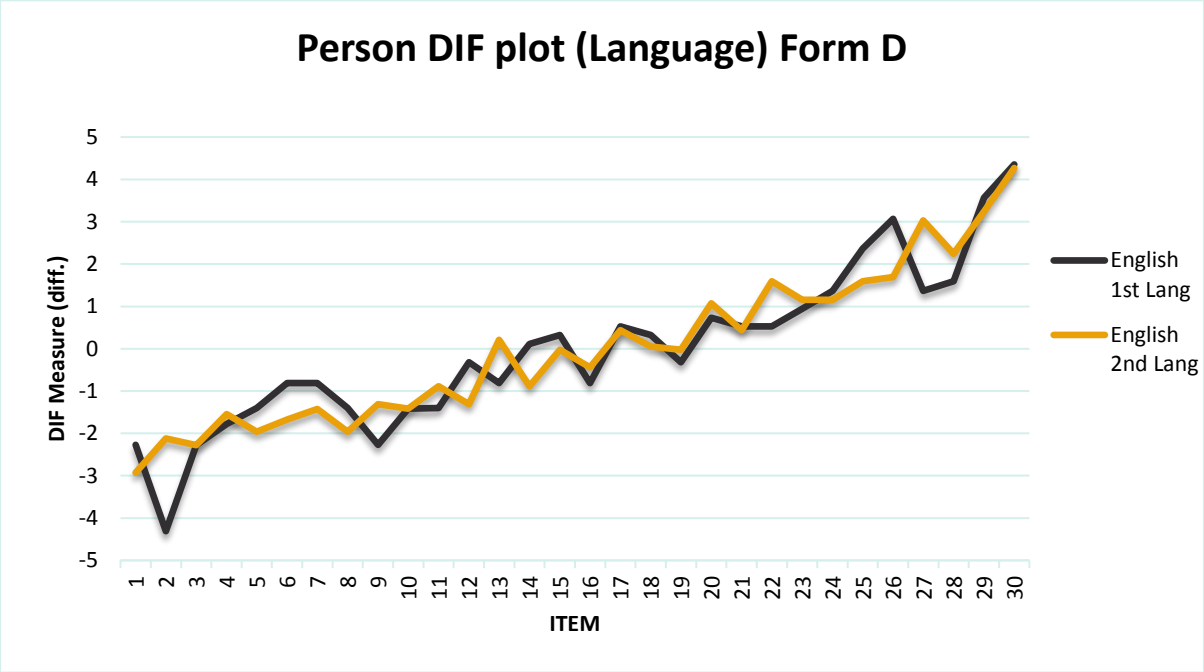

Figure 12. Person DIF plot (Language) Form D

Only item 27 showed statistically significant probability on the Rasch-Welch test, indicating that it might cause differences between how English first language speakers and English second language speakers perform in other samples. The item was not significant based on their Mantel-Haenszel results and further might not have been answered by all of the participants due to the timed nature of the assessment, which leads to the conclusion that there is insufficient evidence to support that the item will cause lower performance for men in subsequent groups.

Table 34. DIF between English 1st language and 2nd language participants on Form D.

Item	DIF Contrasts	Joint S.E.	Rasch-Welch			Mantel-Haenszel	
			T	d.f.	Prob.	Chi-squ	Prob.
1	**0.66**	0.92	0.72	45	0.476	0.032	0.859
2	**-2.19**	1.89	-1.16	26	0.257	0.624	0.430
3	0.00	0.88	0.00	39	1.000	0.036	0.850
4	-0.23	0.74	-0.31	39	0.756	0.000	0.985
5	**0.56**	0.70	0.80	46	0.430	0.020	0.889
6	**0.87**	0.63	1.38	49	0.175	1.010	0.315
7	**0.61**	0.62	0.99	47	0.325	0.244	0.622
8	**0.56**	0.70	0.80	46	0.430	0.015	0.904
9	**-0.96**	0.84	-1.15	33	0.258	1.828	0.176
10	0.00	0.68	0.00	41	1.000	0.012	0.914
11	**-0.51**	0.66	-0.77	38	0.445	0.181	0.671
12	**0.98**	0.58	1.69	49	0.098	1.553	0.213
13	**-1.02**	0.58	-1.74	39	0.089	1.150	0.284
14	1.00	0.55	1.81	47	0.077	2.883	0.090
15	0.35	0.54	0.65	43	0.517	0.404	0.525
16	-0.37	0.59	-0.63	40	0.532	0.120	0.729
17	0.09	0.53	0.17	43	0.867	0.000	0.989
18	0.27	0.53	0.51	43	0.615	0.052	0.821
19	-0.30	0.55	-0.53	41	0.597	0.064	0.800
20	-0.34	0.54	-0.63	44	0.535	0.001	0.970
21	0.09	0.53	0.17	43	0.867	0.047	0.828
22	**-1.07**	0.55	-1.95	47	0.057	3.250	0.071
23	-0.21	0.54	-0.39	44	0.697	0.002	0.962
24	0.21	0.55	0.39	43	0.699	0.004	0.953
25	**0.77**	0.63	1.23	40	0.227	1.737	0.188
26	**1.38**	0.73	1.90	35	0.066	2.600	0.107
27	**-1.66**	0.65	-2.56	67	**0.013**	1.302	0.254
28	**-0.65**	0.60	-1.09	51	0.279	0.447	0.504
29	0.32	0.91	0.36	44	0.724	0.091	0.762
30	0.09	1.28	0.07	49	0.947	0.406	0.524

Form E

Item Fit

Table 35 gives an overview of the item fit statistics for Form E of the Matrigma. Based on the outfit statistics, eleven items were identified for potential misfit. Upon further investigation, however, none of these items fell outside the parameters for infit statistics. This, along with the mean infit MNSQ and the mean infit ZSTD scores respectively being close to 1.00 and 0.00, it was concluded that none of the Matrigma Form E items should be flagged for misfit.

Table 35. Item fit statistics: Form E

	Measure	Model	Infit		Outfit	
		S.E.	MNSQ	ZSTD	MNSQ	ZSTD
1	-2.55	0.40	0.91	-0.2	**1.34**	0.8
2	-2.72	0.42	0.94	-0.1	0.78	-0.2
3	-2.55	0.40	0.96	0.0	1.06	0.3
4	-1.47	0.28	0.95	-0.3	0.81	-0.5
5	-1.55	0.29	0.78	-1.2	**0.56**	-1.5
6	-1.17	0.26	0.83	-1.2	**0.64**	-1.4
7	-1.55	0.29	0.81	-1.1	**0.69**	-0.9
8	-0.97	0.25	0.98	-0.1	0.86	-0.5
9	-2.40	0.38	1.29	1.1	**1.38**	0.9
10	-1.10	0.26	0.85	-1.1	0.80	-0.8
11	-0.85	0.25	0.86	-1.2	0.73	-1.3
12	-0.91	0.25	0.94	-0.4	0.90	-0.4
13	-0.56	0.24	0.98	-0.2	1.01	0.1
14	-0.40	0.23	0.78	-2.4	0.73	-1.7
15	0.06	0.22	0.96	-0.5	0.96	-0.2
16	-1.17	0.26	0.75	-1.8	**0.56**	-1.9
17	0.49	0.22	1.02	0.2	1.15	1.0
18	-0.79	0.24	1.25	2.0	**1.47**	2.0
19	0.40	0.22	0.94	-0.8	1.01	0.1
20	0.06	0.22	1.08	1.0	1.07	0.5
21	0.88	0.22	0.94	-0.7	0.95	-0.2
22	1.29	0.23	1.14	1.4	**2.51**	5.4
23	2.02	0.26	1.08	0.6	**1.25**	0.9
24	1.82	0.25	0.91	-0.7	0.87	-0.4
25	2.32	0.29	0.95	-0.2	0.88	-0.3
26	2.24	0.28	1.12	0.8	2.00	2.5
27	2.02	0.26	0.88	-0.8	0.73	-0.9
28	2.68	0.32	1.20	1.0	**3.22**	3.7
29	3.14	0.37	1.28	1.0	**2.12**	1.9
30	3.29	0.39	1.26	0.9	**2.48**	2.2
M	**0.00**	**0.28**	**0.99**	**-0.2**	**1.18**	**0.3**
P.SD	**1.76**	**0.06**	**0.15**	**1.0**	**0.64**	**1.6**

Differential Item Functioning

Ethnicity

From Figure 13, it is clear that several items appear to have large DIF contrast, i.e., there are large differences between how difficult the item was for the two ethnic groups. These differences are not unidirectional, however, indicating that the assessment does not unfairly advantage one ethnic over the other. To explore this further the item analysis is presented in

Table 36.

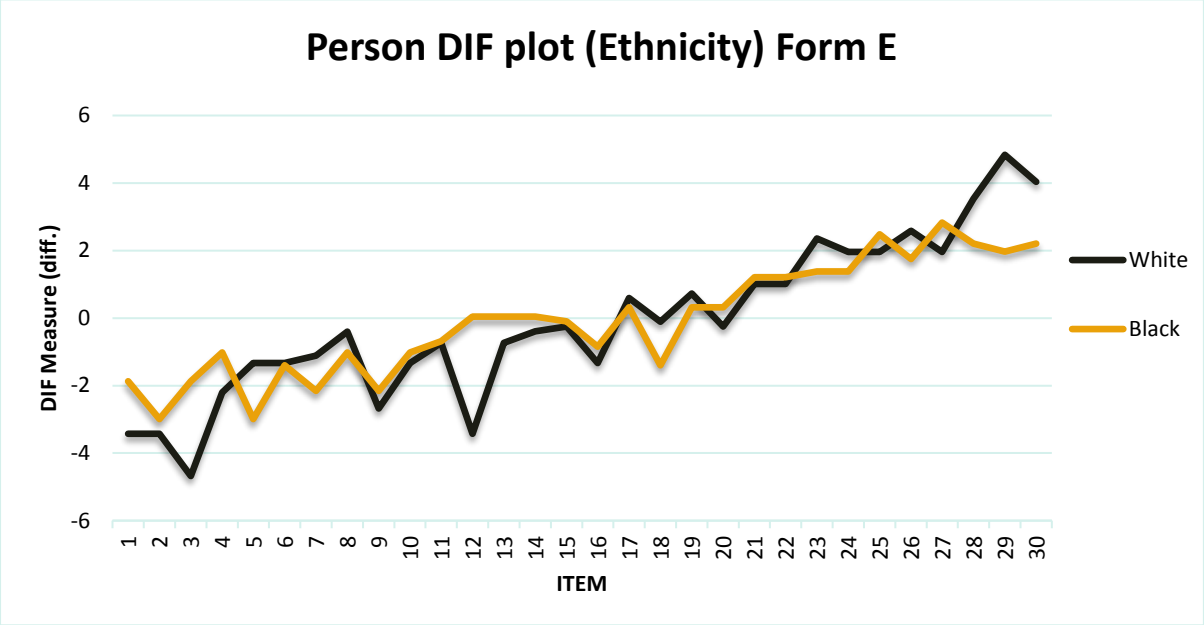

Figure 13. Person DIF plot (Ethnicity) Form E

Only two of the items that were flagged for potential DIF (items 12 and 29) had statistically significant results on one or both the Rasch-Welch and Mantel-Haenszel tests. Item 29 only showed a statistical probability on the Rasch-Welsch test and due to the timed nature of the Matrigma, it could be influenced by many of the participants in the sample not responding to the item. Item 12, however, also yielded statistically significant results on the Mantel-Haenszel test, suggesting that this item will also function differently for different ethnic groups in other samples.

Table 36. DIF between White and Black participants on Form E

Item	DIF Contrasts	Joint S.E.	Rasch-Welch			Mantel-Haenszel	
			t	d.f.	Prob.	Chi-squ	Prob.
1	**-1.56**	1.15	-1.36	51	0.181	0.150	0.699
2	-0.44	1.28	-0.34	63	0.732	0.385	0.535
3	**-2.81**	1.91	-1.47	40	0.149	0.804	0.370
4	**-1.19**	0.76	-1.57	60	0.122	1.029	0.310
5	**1.66**	0.90	1.86	56	0.069	3.820	0.051
6	0.07	0.66	0.11	67	0.911	0.431	0.512
7	**1.05**	0.72	1.45	64	0.153	2.885	0.089
8	**0.62**	0.58	1.07	67	0.289	0.331	0.565
9	**-0.52**	0.94	-0.55	64	0.583	0.366	0.545
10	-0.31	0.64	-0.48	67	0.632	0.002	0.968
11	-0.05	0.58	-0.08	67	0.937	0.002	0.963
12	**-3.47**	1.10	-3.17	44	**0.003**	10.544	**0.001**
13	**-0.78**	0.56	-1.39	67	0.170	1.470	0.225
14	-0.44	0.54	-0.81	67	0.422	0.018	0.894
15	-0.16	0.54	-0.29	67	0.774	0.062	0.803
16	-0.48	0.63	-0.76	66	0.450	0.221	0.639
17	0.27	0.53	0.51	67	0.612	0.096	0.757
18	1.30	0.59	2.18	65	0.033	1.397	0.237
19	0.41	0.53	0.77	67	0.444	0.024	0.876
20	**-0.58**	0.54	-1.07	67	0.289	0.048	0.826
21	-0.20	0.55	-0.37	67	0.713	0.007	0.934
22	-0.20	0.55	-0.37	67	0.713	1.394	0.238
23	**0.97**	0.63	1.56	67	0.124	0.097	0.756
24	**0.58**	0.60	0.96	68	0.339	1.937	0.164
25	**-0.53**	0.70	-0.76	63	0.452	0.239	0.625
26	**0.83**	0.67	1.24	67	0.218	0.519	0.471
27	**-0.87**	0.76	-1.15	59	0.253	0.012	0.913
28	**1.32**	0.83	1.60	64	0.114	0.258	0.612
29	**2.87**	1.16	2.48	48	**0.017**	0.987	0.320
30	**1.83**	0.93	1.97	59	0.054	1.000	0.317

Gender

The item difficulty for the two gender groups are plotted for each item in Figure 14. The graph suggests that many of the Matrigma Form E items perform differently for women and men. The DIF contrasts are however not unidirectional, suggesting that the assessment as a whole is not potentially bias toward a specific gender group. In order to investigate if any items are likely to perform differently for women and men candidates in other samples, Rasch-Welch and Mantel-Haenszel tests were run. These results are presented in Table 37.

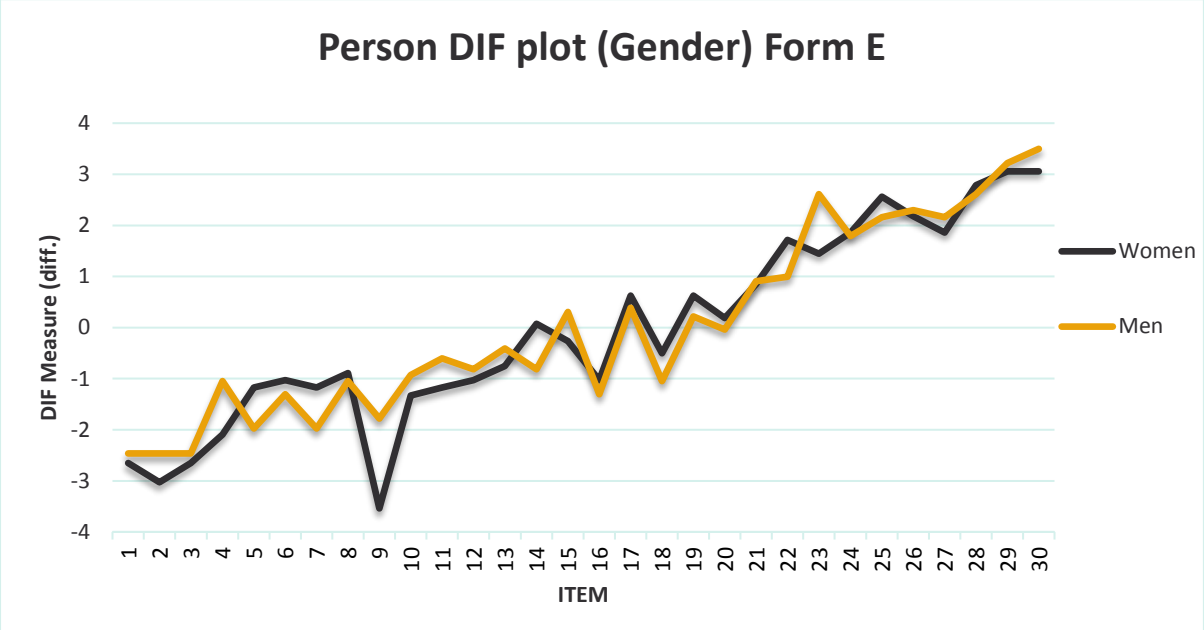

Figure 14. Person DIF plot (Gender) Form E

Ten of the items from Form E showed large DIF contrasts (>.50, <-.50), but only one of these items yielded a statistically significant probability of causing DIF in other samples. Item 23 was flagged for yielding a statistically significant result on the Rasch-Welch test, but did not have a statistically significant probability on the Mantel-Haenszel test. This led to the conclusion that it is unlikely that the item will be flagged for DIF in future samples, when looking at gender differences.

Table 37. DIF between women and men participants on Form E.

Item	DIF Contrasts	Joint S.E.	Rasch-Welch			Mantel-Haenszel	
			t	d.f.	Prob.	Chi-squ	Prob.
1	-0.19	0.79	-0.24	97	0.813	0.057	0.812
2	**-0.56**	0.85	-0.66	92	0.508	0.118	0.731
3	-0.19	0.79	-0.24	97	0.813	0.025	0.876
4	**-1.05**	0.60	-1.75	84	0.083	1.206	0.272
5	**0.80**	0.60	1.34	99	0.185	0.714	0.398
6	0.28	0.53	0.52	98	0.602	0.114	0.736
7	**0.80**	0.60	1.34	99	0.185	0.123	0.726
8	0.16	0.51	0.31	97	0.756	0.015	0.902
9	**-1.76**	0.89	-1.97	70	0.053	1.101	0.294
10	-0.40	0.53	-0.76	93	0.447	0.001	0.982
11	**-0.57**	0.50	-1.13	91	0.260	0.214	0.643
12	-0.21	0.50	-0.42	94	0.674	0.018	0.895
13	-0.35	0.48	-0.73	93	0.466	1.498	0.221
14	**0.89**	0.47	1.90	98	0.061	0.618	0.432
15	**-0.57**	0.45	-1.27	93	0.208	1.962	0.161
16	0.28	0.53	0.52	98	0.602	0.009	0.924
17	0.24	0.44	0.54	94	0.593	0.494	0.482
18	**0.54**	0.49	1.10	98	0.274	0.520	0.471
19	0.41	0.44	0.92	95	0.360	0.116	0.734
20	0.23	0.45	0.51	95	0.610	0.722	0.395
21	-0.06	0.45	-0.13	95	0.900	0.067	0.797
22	**0.72**	0.48	1.50	90	0.137	1.582	0.209
23	**-1.16**	0.55	-2.12	99	**0.036**	2.572	0.109
24	0.07	0.51	0.14	94	0.892	0.074	0.786
25	0.40	0.59	0.68	90	0.496	0.182	0.670
26	-0.12	0.56	-0.21	96	0.835	0.017	0.897
27	-0.30	0.53	-0.56	97	0.576	0.117	0.732
28	0.18	0.64	0.28	92	0.780	0.108	0.743
29	-0.16	0.74	-0.22	96	0.825	0.146	0.703
30	-0.44	0.77	-0.57	99	0.571	0.017	0.895

Language

Figure 15 provides a plot of the differences that English first language and second language participants experienced in terms of item difficulties. It is clear from the plot that some items were perceived as more or less difficult by the two groups, but these differences or not unidirectional. In other words, the plot would suggest that Form E of the Matrigma is not biased against a specific language group.

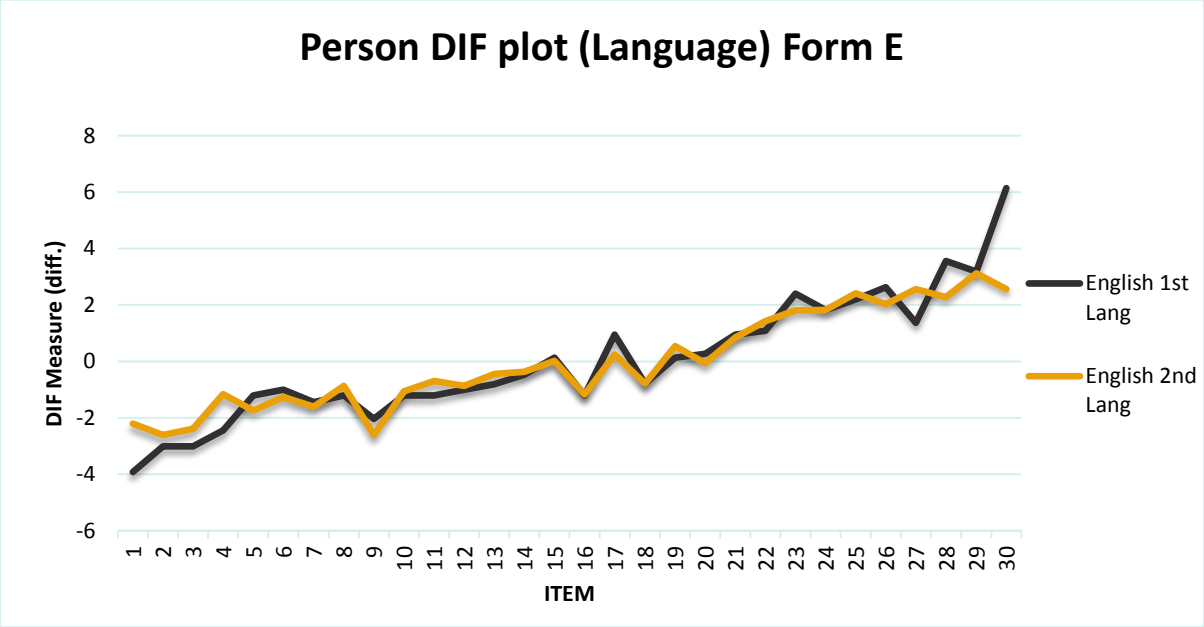

Figure 15. Person DIF plot (Language) Form E

Table 38 provides an overview of the probability tests run to estimate if the items showing large DIF contrasts will also be flagged for DIF in future samples. Of the ten items that displayed large differences in item difficulty for the two groups, only one (item 27) was flagged for the probability that the item will also cause DIF in other samples. The Rasch-Welch statistic for item 27 indicates that this item could potentially perform differently for English first and second language users in other samples. The Mantel-Haenszel test was however not significant and, taking the timed nature of the assessment into account, there is not sufficient evidence to conclude that this item will indeed cause DIF in other samples.

Table 38. DIF between English 1st language and 2nd language participants on Form E.

Item	DIF Contrasts	Joint S.E.	Rasch-Welch			Mantel-Haenszel	
			t	d.f.	Prob.	Chi-squ	Prob.
1	**-1.72**	1.20	-1.43	46	0.159	0.696	0.404
2	-0.40	0.96	-0.42	61	0.679	0.002	0.961
3	**-0.62**	0.94	-0.66	57	0.514	0.024	0.876
4	**-1.29**	0.75	-1.71	51	0.094	0.309	0.578
5	**0.53**	0.60	0.89	77	0.377	0.189	0.664
6	0.26	0.55	0.48	72	0.634	0.547	0.459
7	0.16	0.61	0.27	71	0.790	0.043	0.836
8	-0.34	0.56	-0.60	65	0.549	0.293	0.588
9	**0.57**	0.77	0.74	79	0.462	0.430	0.512
10	-0.15	0.56	-0.26	67	0.793	0.034	0.855
11	**-0.51**	0.55	-0.93	63	0.358	0.058	0.810
12	-0.13	0.54	-0.24	68	0.815	0.210	0.647
13	-0.36	0.51	-0.70	67	0.484	0.307	0.580
14	-0.10	0.49	-0.20	70	0.841	0.045	0.832
15	0.11	0.46	0.24	73	0.813	0.102	0.750
16	-0.04	0.57	-0.06	68	0.950	0.039	0.843
17	0.70	0.46	1.52	73	0.132	0.490	0.484
18	0.00	0.52	0.00	69	1.000	0.121	0.728
19	-0.41	0.46	-0.89	73	0.377	1.352	0.245
20	0.32	0.46	0.70	74	0.485	1.001	0.317
21	0.10	0.46	0.21	74	0.834	0.009	0.923
22	-0.34	0.48	-0.72	78	0.474	2.294	0.130
23	**0.59**	0.56	1.05	70	0.299	0.018	0.892
24	0.00	0.52	0.00	77	1.000	0.417	0.518
25	-0.21	0.58	-0.37	82	0.712	0.027	0.868
26	**0.60**	0.60	1.00	69	0.320	0.001	0.976
27	**-1.19**	0.55	-2.17	93	**0.033**	2.459	0.117
28	**1.28**	0.74	1.73	58	0.090	0.251	0.616
29	0.06	0.75	0.09	81	0.931	0.394	0.530
30	**3.58**	1.90	1.89	39	0.067	2.484	0.115

Summary

1. After conducting item fit analyses for all five Forms of the Matrigma, it was concluded that the items fit the Rasch model for each of the five Forms. No items were specifically highlighted for misfit.
2. Differential item functioning tests were then run for three subgroups within the sample; namely ethnicity, gender and language groupings. The following items were flagged for potential differential item functioning in these groups:

Form	Ethnicity	Gender	Language
A	22	17	22*
B	9, 22, 24	9	-
C	26*	12	30*
D	15*	19, 25	5, 27*
E	12*, 29	23	27

Note: * = items with statistically significant Mantel-Haenszel probabilities.

3. Due to the nature of the items being diagrammatic, it is more likely that these singular instances of flagged items are due to sample differences, and not the nature of the item. If bias was due to the diagrammatic nature of the assessment, we would expect consistent directional differences to be found in all items of the test. Due to the small sample sizes, the fact that the items were not the same for ethnicity and language in most cases, and that the direction of DIF was not consistent, the likelihood that these items demonstrate bias is low.

OVERALL SUMMARY

South African norms for the Matrigma were generated based on data collected from 472 working South African adults. Overall, the psychometric properties were good, and the assessment appears to be appropriate for use in South African samples. The following points are a summary of the results of the psychometric analysis:

1. The Matrigma showed good reliability across all five forms and for specific subgroups within each sample.
2. No statistically significant differences existed between how participants scored on the different forms of the Matrigma.
3. There were statistically significant differences between how people from different age categories scored, with younger participants scoring higher than older participant groups. There was also a negative correlation between age and Matrigma scores. These findings are in line with previous research on cognitive ability.
4. There were no statistically significant differences between how men and women scored on the Matrigma.
5. The results indicated that there were statistically significant differences between how different ethnic groups scored on the Matrigma. Black participants scored significantly lower than both Asian/Indian and White participants. The Asian/Indian sample was small, so it is recommended that more research be done with a larger sample. The effect sizes ranged from small to medium, and practitioners are advised to bear these differences in mind when interpreting results. The largest difference was one raw score point.
6. When comparing different levels of education there were significant differences between the performance of participants with only secondary level education and those with postgraduate education. There were no other significant differences between different levels of tertiary education.
7. Due to the sizes of the different language group samples, they were combined into only two language groups – English first and second language speakers. There were statistically significant differences between how these groups scored, with small effect sizes.
8. Rasch analysis indicated that the Matrigma items fit the Rasch model well for all five of the forms. None of the items were identified for misfit.
9. Items flagged for potential DIF were seemingly reflections of sample performance. Due to the small sample sizes, the fact that the items flagged were not the same for ethnicity and language in most cases, and that the direction of DIF was not consistent, the likelihood that these items demonstrate bias was deemed low.

REFERENCES

Benjamin, D.J. et al. (January, 2018). Redefine statistical significance. *Nature Human Behaviour, 2*, 6-10. doi:10.1038/s41562-017-0189-z.

De Beer, M. (2004). Use of differential item functioning (dif) analysis for bias analysis in test construction. *South African Journal of Industrial Psychology, 30*(4), 52-58.

Mabon, H., & Sjöberg, A. (2011). *Matrigma technical manual*. Stockholm, Sweden: Assessio International AB.

Salthouse, T. A. (1991). Mediation of adult age differences in cognition by reductions in working memory and speed of processing. *Psychological Science, 2*, 179-183.

Schretlen, D., Pearlson, G. D., Anthony, J. C., Aylward, E. H., Augustine, A. M., Davis, A., & Barta, P. (2000). Elucidating the contribution of processing speed, executive ability, and frontal lobe volume to normal age-related differences in fluid intelligence. *Journal of the International Neuropsychological Society, 6*, 52–61.

Staff, R. T., Hogan, M. J., & Whalley, L. J. (2014). Aging trajectories of fluid intelligence in late life: The influence of age, practice and childhood IQ on Raven's Progressive Matrices. *Intelligence, 47*, 194-201.

Statistics South Africa. (2012). *Census 2011* [Statistical release P-0301.4]. Pretoria: South Africa.

Statistics South Africa. (2017). *Quarterly Labour Force Survey: Quarter 2, 2017* [Statistical Release P0211]. Pretoria, South Africa: Author.

Strobl, C., Kopf, J., & Zeileis, A. (2011). A new method for detecting differential item functioning in the Rasch model. *Working Papers in Economics and Statistics, No. 2011-01*. Innsbruck, Austria: University of Innsbruck. Retrieved from http://hdl.handle.net/10419/73503.

Taylor, N. (2008). The Raven's Standard and Advanced Progressive Matrices among adults in South Africa. In J. Raven and J. Raven (Eds.), *Uses and abuses of intelligence: Studies advancing Spearman and Raven's quest for non-arbitrary metrics* (pp. 371-391), New York, NY: Royal Fireworks Press.

Westers, P., & Kelderman, H. (1991). Examining Differential Item Functioning due to Item Difficulty and Alternative Attractiveness. *Psychometrika, 57*(1), 107-118.