

Addressing gender discrimination in cognitive assessment using the English Comprehension Test

**Author:**Danille E. Arendse¹ **Affiliation:**

¹Department of Psychology,
Faculty of Humanities,
University of Pretoria,
Pretoria, South Africa

Corresponding author:

Danille Arendse,
arendse.danielle@gmail.com

Dates:

Received: 14 Jan. 2020

Accepted: 04 Dec. 2020

Published: 11 Mar. 2021

How to cite this article:

Arendse, D.E. (2021).
Addressing gender
discrimination in cognitive
assessment using the English
Comprehension Test.
*SA Journal of Industrial
Psychology/SA Tydskrif vir
Bedryfsielkunde*, 47(0), a1776.
<https://doi.org/10.4102/sajip.v47i0.1776>

Copyright:

© 2021. The Authors.
Licensee: AOSIS. This work
is licensed under the
Creative Commons
Attribution License.

Orientation: The empirically designed English Comprehension Test (ECT) is theorised to measure verbal reasoning and is currently undergoing validation. The test development produced two versions of the ECT, namely, ECT version 1.2 and ECT version 1.3. This study focuses on the latest test version, ECT version 1.3.

Research purpose: The purpose of this study was to statistically explore the performance of men and women who were assessed by the empirically designed ECT.

Motivation for the study: Cognitive assessment has often been used as a discriminatory tool against gender, race and/or languages. The discrimination against race and gender were the consequences of a patriarchal system and Apartheid in South Africa, as black men and women were deemed to be subordinate to white men. With the demise of Apartheid, measures have been put in place to guard against unfair assessment practices. In addition, legislation was developed to ensure that test developers and test users employed assessments that did not unfairly prejudice individuals based on their race, gender and language. These measures are imperative to ensure fairness and equal opportunities for men and women across race and language groups.

Research design, approach and method: This study used a quantitative cross-sectional design. The ECT was administered to a non-probability convenience sample of 881 individuals. The data were analysed by differential test functioning (DTF) in Winsteps and analysis of variance (ANOVA) in the Statistical Product and Service Solutions (SPSS) package.

Main findings: The results indicated that the majority of the test items did not present any bias, but five possibly biased items were identified across gender groups in the test. These five items that were possibly biased appear to be affected by language and not gendered knowledge, and this, however, necessitates further investigation. The ANOVA results only indicated statistically significant differences across the different language groups, thereby confirming the DTF results.

Practical/managerial implications: A major limitation of this study is the restriction of range and lack of generalisability.

Contribution/value-add: This study promotes the use of DTF and ANOVA as a means of ensuring fairness in assessment practices across gender groups. Moreover, it contributes to cross-cultural test development and validation research in South Africa.

Keywords: gender; language; psychometric testing; differential test functioning; English Comprehension Test.

Introduction

Orientation

There is a need to develop tests locally for the diverse population of South Africa, which presents a unique combination of multilingualism and multiculturalism that innately affect performance in internationally created tests (Bekwa, 2016; Foxcroft, 2004; Foxcroft, Roodt, & Abrahams, 2013; Laher & Cockcroft, 2017; Arendse, 2018). The culturally complex context of South Africa makes this a formidable, yet imperative task. The researcher undertook this task and empirically created the English Comprehension Test (ECT). The ECT, a South African empirically developed test, was identified as a measure for verbal reasoning (Arendse & Maree, 2019; Arendse, 2018, 2020) but it is still in the validation process.

Read online:

Scan this QR
code with your
smart phone or
mobile device
to read online.

Research purposes and objectives

South African organisational contexts continuously require tests that are applicable to the South African workforce and serve as a valid and reliable measurement instrument. As testing forms a large part of organisational selections for specific positions, it is essential that the tests used in these contexts are valid and reliable because they are used for decision-making purposes. More importantly, the assessments in the organisational context are often used to measure cognitive performance and functioning (Foxcroft, 2004; Foxcroft & Aston, 2006; Foxcroft et al., 2013; Laher & Cockcroft, 2017; Muleya, Fourie, & Schlebusch, 2017). As the ECT is considered a cognitive assessment, it is crucial to assess whether there is differential performance in the test because of gender. As language is a factor that has previously been found to affect performance (Foxcroft & Aston, 2006; Laher & Cockcroft, 2017), the interaction between the gender and language may provide useful insights. The objectives of the study were as follows:

- Objective 1: To examine whether there are any differences in men and women with regard to performance in the ECT (cognitive assessment).
- Objective 2: To examine the interaction between gender and the different language groups in the ECT total score.

Literature review

The context of psychometric testing is crucial to understanding the existence of bias in testing, and thus, it is important to delve into the issues promoting this discrimination (Foxcroft, 2004; Foxcroft & Aston, 2006; Foxcroft et al., 2013; Muleya et al., 2017). Gender discrimination, particularly prejudice towards women, has persisted because of the dominant patriarchal influence in the global context. These patriarchal acts of injustice caused the rise of feminist movements as a means of challenging the hegemonic systems of thought (Phakeng, 2015; Stone & Coetzee, 2005). The responsiveness of feminists to patriarchal mechanisms of exclusion was vital in the fight for gender equality in the workplace and in university spaces (Phakeng, 2015; Stone & Coetzee, 2005). This awareness led to the criticism of Perry's college-stage theory (1970) for only focusing on males' cognitive development at university, and consequently, Belenky, Clinchy, Goldberg and Tarule (1986) developed a female perspective. Belenky et al. (1986) decided to replicate the study in order to develop an understanding of cognitive development of female college students. Belenky and Staunton (1998) and Belenky et al. (1986) identified seven positions through which women progress on their journey to acquiring new information. These positions describe the process by which a woman becomes actively involved in debates, knowledge production and confidence in her cognitive and moral development (Garrison, 2009).

One of the methods used for the exclusion of women in the workplace and university was psychometric assessment, which created the segregation of women because of theorised intelligence deficits. Globally, women were discriminated

from men through the application of psychometric instruments, which indicated that they were cognitively inferior to men and were thus not capable of occupying certain positions (Camarata & Woodcock, 2006; Hur, te Nijenhuis, & Jeong, 2017; Hyde, 1981; Miller & Halpern, 2014; Palejwala & Fine, 2015; Toivainen, Papageorgiou, Tosto, & Kovas, 2017). This discrimination was very powerful as it appeared to be scientifically proven that women were less intelligent than men in male-dominated spaces. It was found in several studies that men tended to score better in verbal analogies and spatial relations tasks (Camarata & Woodcock, 2006; Hur et al., 2017; Hyde, 1981; Miller & Halpern, 2014; Palejwala & Fine, 2015; Toivainen et al., 2017), whilst women were found to be cognitively stronger in certain domains than men, such as verbal ability (reasoning) and other verbal-related cognitive assessments, such as word memory, anagrams, reading, writing, general and mixed verbal ability assessments (Griskevica & Rascevska, 2009; Hur et al., 2017; Hyde, 1981; Miller & Halpern, 2014; Palejwala & Fine, 2015; Strand, Deary, & Smith, 2006; Toivainen et al., 2017; Wai, Hodges, & Makel, 2018; Wilsenach & Makaure, 2018). These gender differences were explained as biological differences between the male and female sexes because of hormones or genetic differences between men and women (Hur et al., 2017; Miller & Halpern, 2014; Toivainen et al., 2017; Wilsenach & Makaure, 2018). Other reasons provided were the different socialisation of men and women, particularly through school and cultural influences, which, therefore, led to differences in the observation of intelligence across the genders (Miller & Halpern, 2014; Wilsenach & Makaure, 2018). Another aspect included the psychosocial influence that led to differential performance in intelligence measures for men and women, which was also informed by the gender stereotypes associated with men and women (Hur et al., 2017; Miller & Halpern, 2014). However, relative evidence existed, which indicated that there were no gender differences in general intelligence (Camarata & Woodcock, 2006; Griskevica & Rascevska, 2009; Hur et al., 2017; Palejwala & Fine, 2015; Strand et al., 2006; Toivainen et al., 2017).

Despite these diverse views on gender differences in cognitive assessment across numerous studies in American, European and Asian countries, in Africa very few studies have been conducted on differences in cognitive abilities between men and women (Hur et al., 2017). A recent study on differences in cognitive assessment in South Africa, however, noted a disconcerting trend amongst Grade 3 boys in this country, because they were consistently achieving much lower scores compared with women in Grade 3 (Wilsenach & Makaure, 2018). Another study conducted by Bakhiet and Lynn (2015) using the Ravens Coloured Progressive Matrices (a non-verbal intelligence assessment) on Xhosa South African schoolchildren found that the intelligence scores of these schoolchildren were similar to Zulu South African schoolchildren who had performed the same test some years before. The intelligence scores obtained by the Xhosa and Zulu South African schoolchildren allowed the researchers to conclude that the education these

schoolchildren were receiving was not increasing their cognitive level (Bakhiet & Lynn, 2015). This conclusion, however, raises issues regarding the appropriateness of tests that originated from different geographical locations as they do not always consider the multicultural and multilingual context of South Africa (Foxcroft, 2004; Laher & Cockcroft, 2017).

In South Africa, psychometric instruments previously served to perpetuate the ideology of the Apartheid government (Laher & Cockcroft, 2013). The combination of race and gender amplified the discrimination, which was not different in the American context, where African American and Mexican individuals were discriminated in psychometric assessments (Kennedy, Allaire, Gamaldo, & Whitfield, 2012; Sireci & Parker, 2006). Moreover, a further implication of the diverse languages in South Africa was the intensified degree of discrimination in the use of psychometric assessments (Laher & Cockcroft, 2013). The level of bias, therefore, intersected on three levels: race, gender and language. The recognition of these injustices prompted the introduction of laws to prohibit discrimination and enforced fair testing practices for employment and educational purposes, such as the *Employment Equity Act* (Act 55 of 1998) and the *Health Professions Act* (Act 56 of 1974) of South Africa. Psychologists have thus made concerted efforts to limit bias in testing, and have adapted international tests for use in South Africa and created norms for the South African population as corrective procedures in testing (He & Van de Vijver, 2012; Laher & Cockcroft, 2013, 2017; Malda, Van de Vijver, & Temane, 2010; Muleya et al., 2017; Van de Vijver & Tanzer, 2004). There have also been substantial efforts to address the gender, racial and language discrimination associated with psychometric testing, which include the comprehensive validation of testing instruments to limit all bias against any racial and gender group as far as possible (Foxcroft & Aston, 2006; He & Van de Vijver, 2012; Malda et al., 2010; Van de Vijver & Tanzer, 2004). Language has been discriminated against English additional language individuals and has been identified as one of the most important factors affecting performance in tests (Foxcroft & Aston, 2006). In terms of cognitive assessment, language has also been found to affect performance in tests related to verbal comprehension (Foxcroft & Aston, 2006). A recent study by Reilly, Neumann and Andrews (2019) found gender differences in reading and writing achievement scores and attributed some of these gender differences to language and culture. In light of these issues, gender cannot be considered in isolation but should instead be regarded as part of other factors that affect men and women in completing assessments.

The relevance of addressing racist and sexist research was recently emphasised in the retracted article by Nieuwoudt, Dickie, Coetsee, Engelbrecht and Terblanche (2019), in which they claimed that coloured women had an increased risk of low cognitive functioning and were presented with low education levels. The term 'coloured' refers to the legal

classification of racially mixed individuals in South Africa. This legal classification was created during Apartheid in South Africa and has remained a legal classification in South Africa until the present (Adhikari, 2006; Isaacs-Martin, 2018). The article by Nieuwoudt et al. (2019) was petitioned and later retracted by the publishers as it was heavily criticised for perpetuating racist and sexist ideologies as well as colonial stereotypes, with Apartheid underpinnings, of coloured women. The uncritical use of the term coloured to homogenise a racially diverse group of women was found in the article (Boswell, Erasmus, Johannes, Mahomed, & Ratele, 2019). Moreover, the authors were criticised for applying the flawed methodology in addition to using an international test that was not culturally adapted to the South African population, which would, therefore, have provided potentially biased results. The consequences of such results are far-reaching and cause psychological damage as the conclusions generated by this study were generalised to all coloured women (Boswell et al., 2019).

When reviewing the literature on the use of psychometric instruments in South Africa, our history reminds us that we need to be cautious of the manner that assessments have been used to promote racist science. The uncritical use of measurement instruments had led to unfair assessment and biased conclusions. These conclusions, based on the literature, are not singular, but can represent multiple factors. In this manner, a feminist theory such as intersectionality (Crenshaw, 1988) becomes an important way of assessing the different aspects that affect performance in assessments. Intersectionality argues that individuals can be oppressed on multiple grounds simultaneously (Crenshaw, 1988). With reference to the literature, some of the aspects, such as gender, race, language and culture, can simultaneously oppress individuals completing cognitive assessments when the assessment had not been subjected to sufficient validation for their population and context. In this article, the use of intersectionality (Crenshaw, 1988) allows the author to focus predominantly not only on gender but also on language as another factor that can affect performance in assessment. Although there have been important developments and interventions to guide test developers and test users on the fairness, validity and reliability of assessments, the retracted research study by Nieuwoudt et al. (2019) reminds us of the importance of continually validating assessments in the South African context. For this reason, this study was guided by two research questions, namely: are there any differences in the performance of men and women in the ECT? and what is the interaction effects of gender and language on the ECT total score?

Research design

Research approach

This study used a quantitative cross-sectional design. The ECT is theorised to measure verbal reasoning and is currently undergoing validation. The ECT was administered

to a non-probability convenience sample of 882 individuals. The data were analysed by differential test functioning (DTF) in Winsteps and a two-way ANOVA in Statistical Product and Service Solutions (SPSS). Differential test functioning analysis was used to assess differences across gender groups in the ECT. This statistical analysis allows the researcher to assess whether the different genders, man and woman, performed similarly in the test. It is worth noting that DTF is of critical relevance in cross-cultural and multilingual research (Sireci & Berberoglu, 2000). A two-way ANOVA was conducted to assess the interaction effects of gender and language on the ECT total score (Lee & Lee, 2018). This was anticipated to provide additional information on gender differences and to assess whether language had an interactional effect on gender.

Research method

Research participants

The sample size consisted of 881 individuals, with the age of the female sample ($N = 213$) ranging from 18 to 42 years and with most participants aged 18 years. The female racial distribution consisted of black African ($N = 165$), white ($N = 24$), coloured ($N = 20$) and Indian ($N = 3$) groups. The age of the male sample ($N = 666$) ranged from 18 to 41 years, with most participants aged 19. The male racial distribution consisted of black African people ($N = 517$), white people ($N = 111$), coloured people ($N = 30$) and Indian people ($N = 8$). All 11 languages (English, Afrikaans, IsiXhosa, IsiZulu, Sepedi, SiSwati, Tshivenda, IsiNdebele, Sotho, Setswana and Xitsonga) and all nine provinces of South Africa were present in the sample. For the two-way ANOVA, the languages were grouped into the following: (1) Afrikaans, (2) English and (3) African languages, as shown in Table 1. All participants had completed Grade 12 (highest school grade). It should be noted that the majority of the sample population were relatively young and included considerably more men than women.

Measuring instrument

The ECT is an empirically created test that is theorised to measure verbal reasoning (Arendse & Maree, 2019; Arendse, 2018, 2020). It is comprised of comprehension and language sections, which include multiple-choice questions that are dichotomously scored. The language section also includes a written answer section with four sentence construction items. This test has been used on individuals from different linguistic and cultural backgrounds, as well as on different age groups in South Africa. The ECT is currently still in development and is, at present, undergoing validation.

TABLE 1: Sample characteristics for English Comprehension Test 1.3.

Sample	Number
Women	213
Men	666
Afrikaans (1)	131
English (2)	67
African languages (3)	676

The development of the ECT, thus far, has led to the piloting of two test versions, namely, ECT 1.2 (39 items) with a time limit of 45 min and ECT 1.3 (42 items) with no time limit (Arendse & Maree, 2019; Arendse, 2018). As the ECT is still in development, it has only been used for research purposes. The ECT has been used as a screening tool for verbal reasoning in educational and organisational settings. It may assist organisational practitioners in screening for verbal reasoning, which is often required in organisational positions. This study is part of the validation and further development of the ECT. However, this article focuses only on ECT 1.3, the latest test version.

Research procedure and ethical considerations

A convenience sampling method was used to collect data as the participants were attending selections and were available after assessments had been completed. After individuals were done with the selection process, they had a lunch break. After the break, the participants were informed of the ECT for research purposes and their consent to participate in the research requested, after which they completed the test. The intention behind carrying out the research after the selection process was to avoid the research having an impact on the performance of participants in the selection. It should be noted that fatigue must be considered because of the time when the research took place (Arendse & Maree, 2019; Arendse, 2018, 2020). As the sample comprised of people seeking employment, the participants in the study can be considered to be job seekers from various backgrounds and ages and, therefore, regarded as relevant to organisational psychologists.

The ethical considerations for this study were anonymity, because no identifying information was required for the study, and confidentiality, because demographic variables were treated with confidentiality. All participants gave their written consent to participate in the study. Safeguarding information is important, and thus, only relevant project members are able to access the research.

Statistical analysis

Measurement invariance was explored by conducting a DTF analysis within a Rasch framework using Winsteps (Linacre, 2009). As the sample size of the men and women differ significantly, DTF is able to evaluate differential performance in the test without this being affected by the sample size (Bond & Fox, 2007). Differential test functioning requires the use of item difficulties, referred to as item measures in Rasch, for gender comparison (Linacre, 2012). The analysis, therefore, is used to assess the performance of men and women to establish whether test items caused differential performance in the case of either gender. The two-way ANOVA assessed the interactional effects of gender and the different language groups on the ECT total score. The two-way ANOVA was conducted in SPSS, and the Tukey post hoc test was run for significant variables (Lee & Lee, 2018).

Ethical considerations

Ethical clearance was obtained from the Faculty of Humanities Research Ethics Committee at the University of Pretoria for the PhD study (No. GW20150407HS) from which these results were obtained.

Results

The normality of ECT version 1.3 was assessed by the skewness coefficient -0.256 and the kurtosis coefficient -0.082 . The skewness and kurtosis coefficients were within the commonly established -1.000 to $+1.000$ ranges. This indicates that the data are normally distributed, and thus, the analyses of the study can be run (Arendse, 2020).

Differential test functioning results

Comparison of the average performance of women and men is shown in Table 2. It can be observed that there are some items in which both genders have the same or similar means, whilst it also shows that men and women have higher means in respect of different items.

According to the DTF statistics (Table 1), the items that have the highest t statistical values (as they are greater than the 1.96 cut-off for 95% confidence interval) include items 4 (-3.07), 6 (-3.15), 32 (2.84), 36 (2.69) and 40 (-2.21). These items are statistically different for the two genders and can be considered as possibly biased.

In Table 3, the person and item infit and outfit Mean-Square Statistics (MNSQ) values for men and women were both acceptable as they were close or equal to 1 (Linacre, 2002). These results indicated that both men and women presented a good model fit according to each person's ability and the item difficulty. The person separation values for the men and women are below 2 (Baghaei & Amrahi, 2011), which suggests that there is limited variation in the abilities of men and women. The item separation value is much higher than 2 (Baghaei & Amrahi, 2011), which suggests that there is a relative range of item difficulties across the genders in the test. It is, however, apparent that the item difficulties for the male group are more varied compared with the female group. The item reliability across gender groups is considered excellent reliability values.

In Table 4, the empirical slope of 0.942 is considered acceptable and suggests that the items are relatively similar for both genders, with only a few item differences. The correlation of the male and female intercepts is 0.986, which suggests that the items are measuring the same construct across gender groups. The reliability for both genders is well over 0.90, which indicates that high internal consistency is present (Erguven, 2014; Nunnally & Bernstein, 1994; Suhr & Shay, 2009).

The test items across gender groups were acceptable in terms of their variation of difficulties. Although the majority of the test items did not show any bias across genders, five

TABLE 2: Female and male means for English Comprehension Test 1.3.

Test items	Mean		t statistic
	Women	Men	
Item 1	0.28	0.32	-1.09725
Item 2	0.72	0.68	1.081687
Item 3	0.72	0.74	-0.66146
Item 4	0.78	0.67	3.073706
Item 5	0.42	0.42	-0.0084
Item 6	0.65	0.75	-3.15231
Item 7	0.72	0.69	0.754847
Item 8	0.36	0.38	-0.75133
Item 9	0.85	0.87	-0.90649
Item 10	0.76	0.7	1.656174
Item 11	0.62	0.59	0.815126
Item 12	0.50	0.56	-1.59664
Item 13	0.70	0.71	-0.44357
Item 14	0.74	0.73	0.304498
Item 15	0.65	0.62	0.677828
Item 16	0.88	0.83	1.742391
Item 17	0.88	0.88	0.193821
Item 18	0.79	0.76	0.770091
Item 19	0.83	0.79	1.064563
Item 20	0.96	0.94	0.818736
Item 21	0.93	0.92	0.57756
Item 22	0.95	0.94	0.895919
Item 23	0.08	0.05	1.45013
Item 24	0.52	0.55	-0.89076
Item 25	0.16	0.17	-0.48818
Item 26	0.30	0.37	-1.96883
Item 27	0.22	0.27	-1.64688
Item 28	0.51	0.56	-1.30252
Item 29	0.92	0.94	-1.0991
Item 30	0.97	0.96	0.985776
Item 31	0.93	0.91	1.145414
Item 32	0.82	0.72	2.847114
Item 33	0.57	0.57	-0.30252
Item 34	0.71	0.70	0.264585
Item 35	0.82	0.80	0.645391
Item 36	0.64	0.54	2.697479
Item 37	0.81	0.79	0.552242
Item 38	0.60	0.65	-1.43732
Item 39	0.15	0.13	-0.83341
Item 40	0.10	0.13	-2.21625
Item 41	0.38	0.37	-1.48219
Item 42	0.17	0.18	-1.73874

Source: Arendse, D.E. (2018). *Exploring the construct validity and reliability of the English comprehension test*. Unpublished Doctoral thesis, University of Pretoria.

TABLE 3: Average fit statistics for men and women for English Comprehension Test version 1.3.

Rasch statistics	Men	Women
Person: infit MNSQ	1.00	1.00
Person: outfit MNSQ	1.02	1.02
Person: separation	1.81	1.80
Person: reliability	0.77	0.76
Items: infit MNSQ	0.99	0.98
Items: outfit MNSQ	1.02	1.02
Items: separation	13.60	7.77
Items: reliability	0.99	0.98
Total	665	212

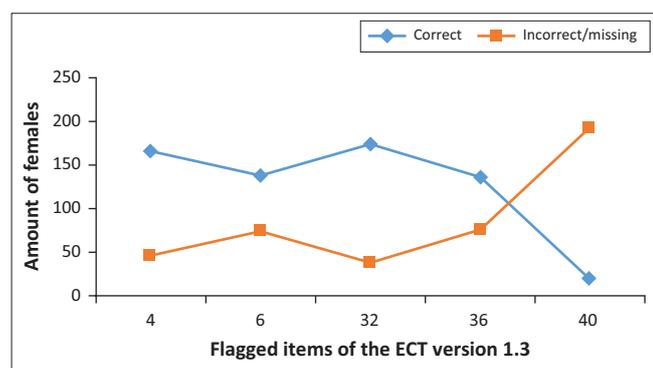
Source: Arendse, D.E. (2018). *Exploring the construct validity and reliability of the English comprehension test*. Unpublished Doctoral thesis, University of Pretoria.

test items were identified (items 4, 6, 32, 36, and 40) as statistically different across the two genders.

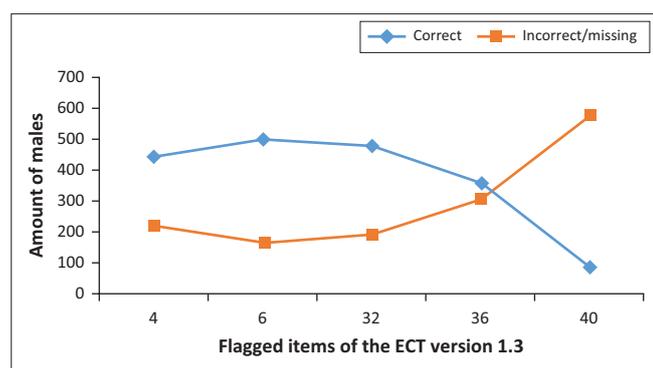
TABLE 4: Differential test functioning statistics for the male and female groups for English Comprehension Test version 1.3.

Statistics	Men	Women
Mean	-0.00095	0.000476
Empirical slope	0.941751	-
Correlation	0.986065	-
Reliability	0.994904151	0.984493321
Disattenuated correlation	0.996342012	-

Source: Arendse, D.E. (2018). *Exploring the construct validity and reliability of the English comprehension test*. Unpublished Doctoral thesis, University of Pretoria.



ECT, English Comprehension Test.

FIGURE 1: Female performance on flagged items of English Comprehension Test, version 1.3.

ECT, English Comprehension Test.

FIGURE 2: Male performance on flagged items of English Comprehension Test, version 1.3.

Both Figures 1 and 2 show the performance of women and men in the flagged (statistically different) items to be relatively similar.

In the assessment of the item content of these statistically different items, as shown in Table 5, it was observed that no recognisable gender discrimination is apparent in the item content. The respective means displayed in Table 5 indicated that women performed higher in three of the five items.

Two-way ANOVA results

In Table 6, the descriptive statistics for gender and the respective language groups are shown. The descriptive statistics for gender indicates the number of men and women in each language group. As expected, the largest numbers belong to the male group. It is, however, worth noting that the means across the different genders are similar, with a small difference across these average scores.

TABLE 5: Statistically different items for English Comprehension Test version 1.3.

Items	Mean	
	Women	Men
4. Which statement is <i>true</i> according to the information given in the passage?	0.78	0.67
6. Which statement is <i>false</i> according to the information given in the passage?	0.65	0.75
32. Choose the <i>correct form of the word</i> in the following sentences: Whilst the medics were on duty for the hospital strikes, they saw seriously injured (A. children, B. child,) in the ward.	0.82	0.72
36. Choose the best word from the list of words provided below which has an <i>opposite meaning (antonym)</i> to the following words: Desperate	0.64	0.54
40. Rearrange the words below in order to make a sentence. <i>All the words must be used in the sentence:</i> Time calculated involves a job-making decisions about pilot's and space.	0.10	0.13

TABLE 6: Descriptive statistics for gender and language groups.

Variable	Language	Mean	Standard deviation	Number
Women	Afrikaans	28.97	4.533	32
	English	30.25	7.407	16
	African languages	25.08	5.109	165
	Total	26.05	5.516	213
Men	Afrikaans	29.10	4.075	99
	English	32.10	3.976	51
	African languages	24.60	5.254	511
	Total	25.85	5.550	661

TABLE 7: ANOVA statistics for gender and language groups.

Source	Type III sum of squares	df	Mean square	F	Significance
Gender	17.210	1	17.210	0.672	0.413
Language groups	2863.901	2	1431.951	55.893	0.000†
Gender * Language group	63.246	2	31.623	1.234	0.292

Note: Dependent variable: ECT total; $R^2 = 0.170$; Adjusted $R^2 = 0.165$.

ECT, English Comprehension Test.

†, The only statistically significant value.

From Table 7, it can be observed that there was no statistically significant difference in the mean ECT total score between men and women, $F(2, 54) = 0.672$, $p = 0.413$. There was, however, a statistically significant difference in the mean ECT total score for the different language groups, $F(2, 54) = 55.893$, $p = 0.000$. The interaction between the gender and language had non-significant effects on the dependent variable, ECT total score, $F(2, 54) = 1.234$, $p = 0.292$.

In Figure 3, the estimated marginal means plot provides a graphical image of the study results. Based on the graph, the lines for men and women appear in a relatively parallel form, thus suggesting that there might not be an interaction effect in the data.

As the interaction results for gender and language were statistically insignificant, the Tukey post hoc test was run on the variable language groups, which was previously reported as statistically significant. Table 8 indicates the results of the Turkey post hoc test for the language groups. Based on the post hoc test results, the differences between Afrikaans and English ($p = 0.002$), Afrikaans and African languages ($p = 0.000$) and English and African languages ($p = 0.000$) are all statistically significant.

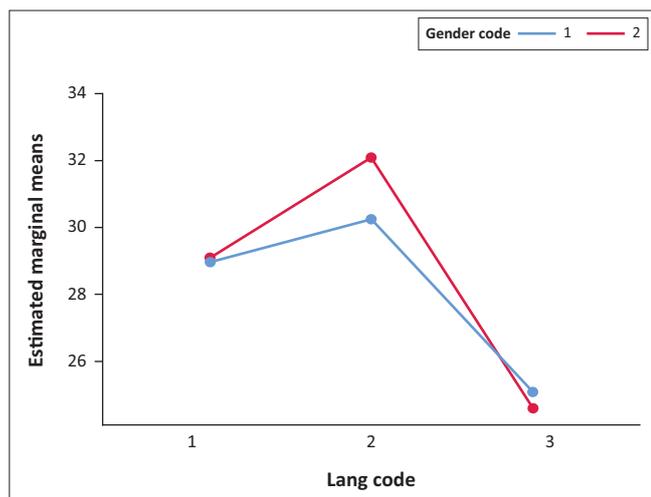


FIGURE 3: Estimated marginal means plot for gender and language groups.

TABLE 8: Multiple comparisons for the language groups (Turkey post hoc test).

(I) Lang code	(J) Lang code	Mean difference (I-J)	Standard error	Significance	95% Confidence interval	
					Lower bound	Upper bound
Afrikaans	English	-2.59*	0.760	0.002	-4.37	-0.80
	African languages	4.35*	0.483	0.000	3.22	5.48
English	Afrikaans	2.59*	0.760	0.002	0.80	4.37
	African languages	6.94*	0.648	0.000	5.42	8.46
African languages	Afrikaans	-4.35*	0.483	0.000	-5.48	-3.22
	English	-6.94*	0.648	0.000	-8.46	-5.42

Note: The error term is mean square (error) = 25.620.

*, The mean difference is significant at the 0.05 level.

Discussion

Outline of the results

This study was aimed at exploring gender differences in the ECT by investigating the test through DTF analysis. The findings of the study indicated that the fit statistics across gender groups revealed high reliability values and good average infit and outfit MNSQ values, which gives some certainty that the gender performance in the test was not necessarily biased and that both genders performed similarly in the test items. The results also showed that the performance of the participants in the sample was problematic in terms of their limited variation of abilities across gender groups. The persons' limited ability levels imply that they were unable to perform better, but the items were not the cause of any specific bias linked to their ability. When observing the content of the items identified as statistically different (Table 5), there appears to be no obvious gender discrimination. The literature on gender differences relating to cognitive assessment suggests that women are more skilled at verbal tasks than men (Griskevica & Rascevska, 2009; Hur et al., 2017; Hyde, 1981; Miller & Halpern, 2014; Palejwala & Fine, 2015; Strand et al., 2006; Toivainen et al., 2017; Wai et al., 2018; Wilsenach & Makaure, 2018), but this finding cannot be concluded in respect of the ECT on the basis of three statistically different items.

An additional analysis was conducted to assess the interaction between gender and language groups. The reasoning for this was the fact that language is a factor that also affects performance in tests (Bekwa, 2016; Foxcroft & Aston, 2006; Reilly et al., 2019; Arendse, 2018). The two-way ANOVA was conducted to examine the effect of gender and the different language groups on the ECT total score. The ANOVA results indicated that there was no statistically significant interaction between the effects of gender and the different language groups on the ECT total score, $F(2, 54) = 1.234, p = 0.292$. The small mean differences were statistically non-significant, and the estimated marginal means plot was also indicative of negligible differences across gender groups. This relates to the findings of the DTF, in which gender differences were only found to affect five items in the test. This is a positive finding, as assessments should not be found to discriminate across genders. Furthermore, it indicates that the test is not biased towards individuals on the basis of their gender. Although gender was found not to discriminate in the ECT total score, there were statistically significant differences in the mean ECT total score for different language groups. When examining these language differences, mean differences were found across three language groups (Afrikaans, English and African language groups). This confirms findings in the literature that language may cause differential performance in assessments.

Practical implications

One way in which the five statistically different items from the DTF can be interpreted is that these differences observed across gender groups are indicative of language-related differences. This was confirmed by the two-way ANOVA, in which gender and the interaction between the gender and the different language groups were found to be statistically non-significant. The findings, however, confirmed that the different language groups had statistically significant differences across their means. One cannot escape the presence of language inhibiting performance in a test, when the majority of the participants are black Africans and are predominately English second- or third-language speakers. The official languages associated with black African members include IsiXhosa, IsiZulu, Sepedi, Setswana, Tshivenda, Sesotho, SiSwati, IsiNdebele and Xitsonga. These languages comprised of the African languages group. The remaining two official languages, English and Afrikaans, formed the other two language groups. When attempting to make sense of the language differences observed in the two-way ANOVA, the African languages vary substantially from the English language, in that there are instances where there is no African equivalent for an English word (Schaap, 2011). This may have had an impact on how African language-speaking individuals interpreted the items in the test. Moreover, these items could have different social meanings attached to certain words across the individuals based on their first language (Radden, 2008). This may also be true for both the African and Afrikaans language groups. In addition, the language differences may be indicative of how individuals differ in their thinking because of language differences and semantic

structures (Boroditsky, 2011; Gentner & Goldin-Meadow, 2003). These differences may also connect to Vygotsky's emphasis on the influence of culture and language on cognitive development (Ormrod, 2008; Vygotsky, 1978) as the environment and learning opportunities can influence this (Van der Pool & Catano, 2008). The finding that gender discrimination was limited to five statistically significant items, which on further investigation indicated no statistically significant mean differences, is a positive one. It also points to the necessity of exploring gender differences as this affects performance in a test. The empirical ECT, which is still in development, can be considered gender neutral in terms of item content, as the five items were not prejudicing individuals because of associated gender knowledge. It is, however, of concern that language was found to affect performance, but this is also a common bias that most tests fall prey to in multilingual and multi-cultural contexts, such as South Africa (Bekwa, 2016; Foxcroft, 2004; Foxcroft et al., 2013; Laher & Cockcroft, 2017). This is a significant finding that will assist in further developing and validating the ECT.

This study relates to other cross-cultural studies that have found background factors to have an effect on the performance of individuals in cognitive assessments (He & Van de Vijver, 2012; Van de Vijver & Tanzer, 2004). Thus, when interpreting the results of this study in light of the language differences, it should be noted that in cross-cultural contexts, the background and culture of different persons completing the test need to be considered (Van de Vijver & Rothmann, 2004). In this study, culture and language are interrelated as words can be regarded as the overlap of race and class (Cooper, 2018). The intersection between the race and class includes culture, which may explain the differences observed across gender groups and, more specifically, the differences between languages. This notion of language differences was also found by research conducted on the verbal scale of the Wechsler Intelligence Scale for Children, which showed higher loadings on linguistics and culture for aboriginal children (Flanagan & Ortiz, 2001).

African individuals in different contexts across the globe were discriminated in cognitive assessment because of their perceived poor performance in cognitive assessments. Their poor performance was linked to historic and educational inequalities (Kennedy et al., 2012; Laher & Cockcroft, 2013, 2017). In light of numerous cross-cultural findings (Flanagan & Ortiz, 2001; Foxcroft, 2004; He & Van de Vijver, 2012; Van de Vijver & Rothmann, 2004), and this study in particular, these conclusions appear to be inadequate for explaining the findings across gender groups but may explain the differences across language groups. In South Africa, it can be deduced that language is influenced by race and culture (Cooper, 2018; Foxcroft & Aston, 2006; Ormrod, 2008; Vygotsky, 1978). Thus, the implication of language, which cannot be separated from race and culture, provides a more substantial reasoning for the observed gender differences in this study. This is because of language often being influenced by a person's race and

culture, which forms part of the contextual aspects impacting on language (Cooper, 2018; Foxcroft & Aston, 2006; Ormrod, 2008; Vygotsky, 1978). The intersection of race and culture in language needs to be considered when examining the results of non-native English individuals in English cognitive assessments. This consideration will limit bias and prevent discrimination against individuals from different cultures in cognitive assessments (Flanagan & Ortiz, 2001; Laher & Cockcroft, 2013, 2017; Van de Vijver & Rothmann, 2004). This consideration also allows the use of intersectionality (Crenshaw, 1988) as a lens through which quantitative findings can be understood.

Limitations and recommendations

A limitation of this study is that the results are not generalisable as convenience sampling was carried out. Although the sample consisted of a relative range of ages, the majority of the sample consisted of young adults. Having predominantly male participants in the sample is another limitation of the study, as the genders were not equally represented. In terms of the language group distribution, the majority of the sample comprised of African language speakers, and thus, the language groups were not equal. There is also the possibility that fatigue may have impacted the participants' performance in the test and should be considered.

The recommendations for the ECT are that the associated language issues with the possibly biased items identified in the ECT need to be examined further and either removed or rephrased for better interpretation. A factor analysis and reliability analysis of the male and female samples should be performed to confirm whether these factor structures correspond with the overarching factor structure and reliability of the ECT.

Conclusion

Psychometric assessments have been known to discriminate and were previously implicated in emphasising gender and racial differences on the basis of cognitive assessments. As the ECT can be considered a cognitive assessment because it measures verbal reasoning (Arendse, 2018), the identification of five biased items from the DTF analysis was the cause of concern. It was, however, argued that the item content did not suggest gendered knowledge but rather tapped into language differences across individuals. This finding, therefore, contradicts numerous international studies that had found women outperformed men in verbal cognitive assessment. The two-way ANOVA found that no statistically significant differences were observed between men and women in the ECT total score. In addition, the interaction between gender and language had non-significant effects on the dependent variable, ECT total mean score. This confirmed that no gender discrimination was observed in the ECT. The findings nevertheless indicated that there were statistically significant differences in the ECT total mean score for the different language groups. It was observed in the post hoc analysis that

there were statistically significant mean differences amongst all the different language groups, Afrikaans, English and African languages. The language issue has been plaguing psychometric testing and test development in South Africa for years, and it remains an immense task to ensure that the tests produced or used are not prejudicing any individuals or limiting their opportunities unfairly. The findings of this study are, therefore, a step towards rectifying the discrimination of the past, in terms of both gender and language. The identification of biased items is imperative to the further development and validation of the ECT and will require further investigation. This study promotes the use of DTF and ANOVA as a means of ensuring fairness in assessment practices across gender groups. Consequently, this study contributes to cross-cultural test development. This study also highlighted the importance of incorporating intersectionality into quantitative studies to ensure that bias in cognitive assessment is addressed.

Acknowledgements

Competing interests

The author declares that they have no financial or personal relationships that may have inappropriately influenced them in writing this research article.

Author's contribution

D.E.A. is the sole author of this research article.

Funding information

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Data availability

The data that support the findings of this study are available from the author upon reasonable request.

Disclaimer

The views and opinions expressed in this article are those of the author and do not necessarily reflect the official policy or position of any affiliated agency of the author.

References

- Adhikari, M. (2006). Hope, fear, shame, frustration: Continuity and change in the expression of coloured identity in white supremacist South Africa, 1910–1994. *Journal of Southern African Studies*, 32(3), 468–487. <https://doi.org/10.1080/03057070600829542>
- Arendse, D.E. (2020). The impact of different time limits and test versions on reliability in South Africa. *African Journal of Psychological Assessment*, 2(0), a14. <https://doi.org/10.4102/ajopa.v2i0.14>
- Arendse, D.E. (2018). *Exploring the construct validity and reliability of the English comprehension test*. Unpublished Doctoral thesis, University of Pretoria.
- Arendse, D.E., & Maree, D. (2019). Exploring the factor structure of the English comprehension test. *South African Journal of Psychology*, 49(3), 376–390. <https://doi.org/10.1177/0081246318805268>
- Baghaei, P., & Amrahi, N. (2011). Validation of a multiple choice English vocabulary test with the Rasch Model. *Journal of Language Teaching and Research*, 2(5), 1052–1060. <https://doi.org/10.4304/jltr.2.5.1052-1060>
- Bakhiet, S.F.A., & Lynn, R. (2015). A study of the intelligence of Xhosa children in South Africa. *Mankind Quarterly*, 55(4), 335–339. <https://doi.org/10.46469/mq.2015.55.4.4>
- Bekwa, N.N. (2016). *The development and evaluation of Africanised items for multicultural cognitive assessment*. Unpublished doctoral dissertation, University of South Africa.
- Belenky, M.F., & Staunton, A. (1998). Women and transformative learning: Connected ways of knowing and developing public voice. In J. Mezirow, V. Marsick, C.A. Smyth, & C. Wiessner (Eds.), *Changing adult frames of reference: Proceedings of First National Transformative Learning Conference* (pp. 9–15). New York, NY: Teachers College.
- Belenky, M.F., Clinchy, B.M., Goldberger, N.R., Nancy, R., & Tarule, J.M. (1986). *Women's ways of knowing: The development of self, voice and mind*. New York, NY: Basic Books.
- Bond, T.G., & Fox, C.M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Boroditsky, L. (2011). How language shapes thought. *Scientific American*. Retrieved from www.ScientificAmerican.com
- Boswell, B., Erasmus, Z., Johannes, S., Mahomed, S., & Ratele, K. (2019). Racist science: The burden of Black bodies and minds. *The Thinker: A Pan-African Quarterly for Thought Leaders*, 2019(81), 4–8.
- Camarata, S., & Woodcock, R. (2006). Sex differences in processing speed: Developmental effects in males and females. *Intelligence*, 34(3), 231–252. <https://doi.org/10.1016/j.intell.2005.12.001>
- Cooper, A. (2018). You can't write in Kaapse Afrikaans in your question paper...the terms must be right: Race- and class-infused language ideologies in educational places on the Cape flats. *Educational Research for Social Change*, 7(1), 30–45. <https://doi.org/10.17159/2221-4070/2018/v7i1a3>
- Crenshaw, K. (1988). Race, reform and retrenchment: Transformation and legitimation in anti discrimination law. *Harvard Law Review*, 101(7), 1331–1387. <https://doi.org/10.2307/1341398>
- Erguven, M. (2014). Two approaches to psychometric process: Classical test theory and item response theory. *Journal of Education*, 2(2), 23–30.
- Flanagan, D.P., & Ortiz, S.O. (2001). *Essentials of cross-battery assessment*. New York, NY: John Wiley and Sons, Inc.
- Foxcroft, C.D. (2004). Planning a psychological test in the multicultural South African context. *SA Journal of Industrial Psychology*, 30(4), 8–15. <https://doi.org/10.4102/sajip.v30i4.171>
- Foxcroft, C.D., & Aston, S. (2006). Critically examining language bias in the South African adaptation of the WAIS-III. *SA Journal of Industrial Psychology*, 32(4), 97–102. <https://doi.org/10.4102/sajip.v32i4.243>
- Foxcroft, C.D., Roodt, G., & Abrahams, F. (2013). Psychological assessments. A brief retrospective overview. In C. Foxcroft & G. Roodt (Eds.), *Introduction of psychological assessments in the South Africa context* (4th ed., pp. 9–27). Cape Town: Oxford University Press.
- Garrison, M.S. (2009). The cognitive development of collegiate students: A brief literature review. *The Campbellville Review*, 87–100. Retrieved from http://www.campbellville.edu/websites/cu/images/library/campbellville_review/vol_4/the_cognitive_development_of_collegiate_students_Garrison.pdf
- Gentner, D., & Goldin-Meadow, S. (2003). Whither Whorf. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind. Advances in the study of language and thought* (pp. 3–14). Cambridge, MA: MIT Press.
- Griskevica, I., & Rascevska, M. (2009). The relationship among cognitive abilities and demographics factors in Latvia. *Baltic Journal of Psychology*, 10(1–2), 55–72.
- He, J., & Van de Vijver, F. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture*, 2(2), 1–19. <https://doi.org/10.9707/2307-0919.1111>
- Health Professions Council of South Africa (HPCSA). (1974). *Health Professions Act, Act 56 of 1974*. Pretoria: South African Press.
- Hur, Y., Te Nijenhuis, J., & Jeong, H. (2017). Testing Lynn's theory of sex differences in intelligence in a large sample of Nigerian school-aged children and adolescents (N > 11 000) using Raven's standard progressive matrices plus. *Mankind Quarterly*, 57(3), 428–437. <https://doi.org/10.46469/mq.2017.57.3.11>
- Hyde, J.S. (1981). How large are cognitive gender differences? *American Psychologist*, 36(8), 892–901. <https://doi.org/10.1037/0003-066X.36.8.892>
- Isaacs-Martin, W. (2018). Minority identities and negative attitudes toward immigrants: Prejudice and spatial difference amongst the coloured population in south Africa. *African Review*, 10(1), 41–57. <https://doi.org/10.1080/09744053.2017.1399562>
- Kennedy, S.W., Allaire, J.C., Gamaldo, A.A., & Whitfield, K.E. (2012). Race differences in intellectual control beliefs and cognitive functioning. *Experiential Aging Research*, 38(3), 247–264. <https://doi.org/10.1080/0361073X.2012.672122>
- Laher, S., & Cockcroft, K. (2013). *Psychological assessment in South Africa: Research applications*. Johannesburg: Wits University Press.
- Laher, S., & Cockcroft, K. (2017). Moving from culturally biased to culturally responsive assessment practises in low-resource, multicultural settings. *Professional Psychology: Research and Practise*, 48(2), 115–121. <https://doi.org/10.1037/pro0000102>
- Lee, S., & Lee, D.K. (2018). What is the proper way to apply the multiple comparison test? *Korean Journal of Anaesthesiology*, 71(5), 353–360. <https://doi.org/10.4097/kja.d.18.00242>
- Linacre, J.M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, 16(2), 878.
- Linacre, J.M. (2009). *Winsteps version 3.68 computer software*. Retrieved from www.winsteps.com

- Linacre, J.M. (2012). *Winsteps tutorials 4*. Retrieved from www.winsteps.com/forum
- Malda, M., Van de Vijver, F.J.R., & Temane, Q.M. (2010). Rugby versus soccer in South Africa: Content familiarity contributes to cross-cultural differences in cognitive test scores. *Intelligence*, 38(6), 582–595. <https://doi.org/10.1016/j.intell.2010.07.004>
- Miller, D.I., & Halpern, D.F. (2014). The new science of cognitive sex differences. *Trends in Cognitive Sciences*, 18(1), 37–45. <https://doi.org/10.1016/j.tics.2013.10.011>
- Muleya, V.R., Fourie, L., & Schlebusch, S. (2017). Ethical challenges in assessment centres in South Africa. *South African Journal of Industrial Psychology*, 43(2), 1–20. <https://doi.org/10.4102/sajip.v43i0.1324>
- Nieuwoudt, S., Dickie, K.E., Coetsee, C., Engelbrecht, L., & Terblanche, E. (2019). Age- and education-related effects on cognitive functioning in coloured South African women. *Aging, Neuropsychology, and Cognition*, 27(6), 1–17. <https://doi.org/10.1080/13825585.2019.1598538>
- Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw Hill.
- Ormrod, J.E. (2008). *Cognitive development. Educational psychology, developing learners* (6th ed.). Upper Saddle River, NJ: Pearson Education Inc.
- Palejwala, M.H., & Fine, J.G. (2015). Gender differences in latent cognitive abilities in children aged 2 to 7. *Intelligence*, 48, 96–108. <https://doi.org/10.1016/j.intell.2014.11.004>
- Perry, W.G. (1970). *Forms of intellectual and ethical development in the college years: A scheme*. New York, NY: Holt, Rinehart and Winston.
- Phakeng, M. (2015). Leadership: The invisibility of African women and the masculinity of power. *South African Journal of Science*, 111(11/12), 1–2. <https://doi.org/10.17159/sajs.2015/a0126>
- Radden, G. (2008). The cognitive approach to language. In J. Andor, B. Hollosy, T. Laczko, & P. Pelyvas (Eds.), *When grammar minds language and literature, Festschrift for Prof Bela Korponay on the occasion of his 80th birthday* (pp. 387–412). Debrecen: Institute of English and American Studies.
- Reilly, D., Neumann, D.L., & Andrews, G. (2019). Gender differences in reading and writing achievement: Evidence from the National Assessment of Educational Progress (NAEP). *American Psychologist*, 74(4), 445–458. <https://doi.org/10.1037/amp0000356>
- Republic of South Africa. (1998). *Employment Equity Act 55 of 1998*. Cape Town: South African Government.
- Schaap, P. (2011). The differential item functioning and structural equivalence of a non-verbal cognitive ability test for five language groups. *South African Journal of Industrial Psychology*, 37(1), 137–152. <https://doi.org/10.4102/sajip.v37i1.881>
- Sireci, S.G., & Berberoglu, G. (2000). Using bilingual respondents to evaluate translated adapted items. *Applied Measurement in Education*, 13(3), 229–248. https://doi.org/10.1207/S15324818AME1303_1
- Sireci, S.G., & Parker, P. (2006). Validity on trial: Psychometric and legal conceptualizations of validity. *Educational Measurement Issues and Practices*, 25(3), 27–34. <https://doi.org/10.1111/j.1745-3992.2006.00065.x>
- Stone, K., & Coetzee, M. (2005). Levelling the playing field: Reducing barriers to mentoring for women protégés in the South African organisational context. *South African Journal of Human Resource Management*, 3(3), 33–39. <https://doi.org/10.4102/sajhrm.v3i3.76>
- Strand, A., Deary, I.J., & Smith, P. (2006). Sex differences in cognitive abilities test scores: A UK national picture. *British Journal of Educational Psychology*, 76(Pt 3), 463–480. <https://doi.org/10.1348/000709905X50906>
- Suhr, D.D., & Shay, M. (2009). *Guidelines for reliability, confirmatory and exploratory analysis*. Paper presented at the SAS Global Forum, Washington, DC.
- Toivainen, T., Papageorgiou, K.A., Tosto, M.G., & Kovas, Y. (2017). Sex differences in non-verbal and verbal abilities in childhood and adolescence. *Intelligence*, 64, 81–88. <https://doi.org/10.1016/j.intell.2017.07.007>
- Van de Vijver, F.J.R., & Rothmann, S. (2004). Assessment in multicultural groups. The South African case. *South African Journal of Industrial Psychology*, 30(4), 1–7. <https://doi.org/10.4102/sajip.v30i4.169>
- Van de Vijver, F.J.R., & Tanzer, N.K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue Europeenne de psychologie appliquee*, 54(2), 119–135. <https://doi.org/10.1016/j.erap.2003.12.004>
- Van der Pool, M., & Catano, V.M. (2008). Comparing the performance of native north Americans and predominantly White military recruits on verbal and non-verbal measures of cognitive ability. *International Journal of Selection and Assessment*, 16(3), 239–248. <https://doi.org/10.1111/j.1468-2389.2008.00430.x>
- Vygotsky, L.S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.
- Wai, J., Hodges, J., & Makel, M.C. (2018). Sex differences in ability tilt in the right tail of cognitive abilities: 35-year examination. *Intelligence*, 67, 76–83. <https://doi.org/10.1016/j.intell.2018.02.003>
- Wilsenach, C., & Makaure, P. (2018). Gender effects on phonological processing and reading development in Northern Sotho children learning to read in English: A case study of Grade 3 learners. *South African Journal of Childhood Education*, 8(1), 1–12. <https://doi.org/10.4102/sajce.v8i1.546>