JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia), 12(1), 2023, 1–16
DOI: http://dx.doi.org/10.15408/jp3i.v10i1.20417
http://journal.uinjkt.ac.id/index.php/jp3i

Multidimensional Rasch Analysis of Gender Differences in Tes Intelegensi Kolektif Indonesia–Tinggi (TIKI-T)

Whisnu Yudiana, Airin Triwahyuni, Hery Susanto

Departement of Psychology, Faculty of Psychology, Universitas Padjadjaran, Indonesia

whisnu.yudiana@unpad.ac.id

Abstract

The empirical evidence on gender differences in the g-factor or general intelligence and various cognitive abilities remains contradictory. Some studies have found that there are no gender differences in general intelligence, while others have found differences between genders in verbal, spatial, and numerical abilities as measured by standardized cognitive tests. This study aims to examine the presence of differential item functioning (DIF) on standardized tests that measure verbal, numerical, and spatial/nonverbal abilities, as well as gender differences in item level. The multidimensional Rasch model was used to identify DIF based on four cognitive domains in the Tes Intelegensi Kolektif Indonesia-Tinggi (TIKI-T) test. A total of 1,443 undergraduate students were tested. The results of the study showed that while there were several unbiased items, some items were clearly biased against males or females. The DIF was higher in the numerical and verbal subtests for female-male differences, while the DIF on male tests corresponded to spatial/nonverbal subtest performance. The theoretical and practical implications of the results are discussed.

Keywords: gender differences, intelligence, multidimensional rasch analysis, differential item functioning

Abstrak

Bukti empiris mengenai perbedaan gender dalam g-factor *atau kecerdasan umum dan berbagai kemampuan kognitif masih kontradiktif. Beberapa penelitian menemukan bahwa tidak ada perbedaan gender dalam kecerdasan umum, sementara penelitian lain menemukan bahwa terdapat perbedaan antar gender pada kemampuan verbal, spasial, dan numerik yang diukur dengan tes kognitif terstandar. Penelitian ini bertujuan untuk menguji* differential item functioning *(DIF) pada tes terstandarisasi yang mengukur kemampuan verbal, numerik, dan spasial/nonverbal, serta perbedaan gender pada level butir soal. Model Rasch multidimensi digunakan untuk mengidentifikasi DIF berdasarkan empat domain kognitif dalam tes Tes Intelegensi Kolektif Indonesia-Tinggi (TIKI-T). Sebanyak 1.443 mahasiswa sarjana menjadi sampel. Hasil penelitian menunjukkan bahwa meskipun ada beberapa item yang tidak bias, beberapa item jelas-jelas bias terhadap laki-laki atau perempuan. DIF lebih tinggi pada subtes numerik dan verbal untuk perbedaan perempuan dan laki-laki, sedangkan DIF pada tes laki-laki berhubungan dengan kinerja subtes spasial/nonverbal. Implikasi teoritis dan praktis dari hasil penelitian ini akan dibahas.*

Kata kunci: *perbedaan antargender, intelegensi, model rasch multidimensi,* differential item functioning

Introduction

Studies on gender differences in cognitive or intellectual abilities have been widely conducted. The results from many studies are inconsistent. Several research were conducted to determine gender difference in general intelligence or spearman "g". The result showed that males obtained higher score than women (Lynn & Irwing, 2004; Lynn & Kanazawa, 2011; Steinmayr et al., 2010). However, others found that there were no significant differences in general intelligence scores between males and females children (Deary et al., 2003; Savage-McGlynn, 2012).

Research on cognitive differences between gender was not only based on "g", but also on specific factors such as: verbal (Hyde & Linn, 1988; Lim, 1994), numerical (Lim, 1994) and spatial abilities (Lim, 1994), processing speed abilities (Camarata & Woodcock, 2006), fluid intelligence (Colom, 2002), and latent traits of general intelligence (Reynolds et al., 2008). The outcomes of those research indicated that there were a small advantage on female on overall verbal scale, with no mean difference found in nonverbal/figural abilities, and a small advantage on male was found in numerical ability. In a study conducted by Strand et al (2006) involving a representative sample of UK pupils, it was found that females outperformed males in verbal reasoning, exhibiting a mean score of 2.2 points higher than their male counterparts. However, the results showed that the mean score difference between females and males for non-verbal reasoning (NVR) was only 0.3 standard points, while for quantitative reasoning (QR), males scored 0.7 points higher than females.

Several methods have been used to analyze difference cognitive ability between gender, such as: t-test or anova (Camarata & Woodcock, 2006; Colom et al., 2002), exploratory factor analysis (Colom et al., 2000), confirmatory factor analysis (Abad et al., 2004; Deary et al., 2007; Lim, 1994; Savage-McGlynn, 2012), and difference item functioning (DIF), which is also called item bias (Abad et al., 2004). Confirmatory factor analysis procedure is also used to large samples of subjects that have been administered a variety of intellectual or cognitive factors. This method is able to lend an exploration of gender differences in cognitive abilities (Camarata & Woodcock, 2006). However, before making conclusion about gender differences in cognitive ability, it is important to validate that there is no bias in the instrument. Occasionally, the items in the test have been known to be biased against particular subgroup. Thus, it has become a matter of considerable concern to users of test results (Hambleton & Swaminathan, 1985). In addition, it is crucial to detect biased items in order to confirm that the instrument measures the same trait in all the subgroup of the population to which the test is administrated (Lord & Stocking, 1988).

The purpose of this study is to validate the instrument for measuring intelligence or cognitive ability across gender in Indonesian population. Tes Intelegensi Kolektif Indonesia–Tinggi (TIKI-T) (Drenth et al., 1977) was used in this study. The Tes Inteligensi Kolektif Indonesia (TIKI) was developed in 1976 as a tool to measure the intelligence levels of the Indonesian population (Drenth et al., 1977). The TIKI-T variant was designed for individuals in the highest grade of SMA and the beginning of higher education. Its primary application is to facilitate decisions regarding admission to tertiary education or selection in organizational contexts. The TIKI-T has been widely used in educational settings, such as predicting academic success (Permatasari, 2016), assessing and placing gifted students, and evaluating international standard schools (Nisa, 2013). Additionally, the TIKI-T has been frequently employed in organizational contexts for selection and placement decisions. Despite its extensive use, there has been a lack of efforts to evaluate and maintain the quality of the test. Since its publication, the TIKI-T has remained unchanged, including its paper-and-pencil test format, the order of the items, the sequence of subtests, and norm. This reseach administered the a TIKI-T short form as part of the complete TIKI-T (Drenth et al., 1977). The TIKI-T short form provided several scores that represent several cognitive abilities,

namely: Arithmetic (numerical ability), Component (spatial/nonverbal ability), Word Relations (verbal ability), and Figure Classification (spatial/nonverbal ability).

All subtests in the TIKI-T short form also measured one construct known as "g" factor. All the subtests were then correlated (Drenth et al., 1977). To consider the correlations between latent traits into calculation, one needs a multidimensional model that simultaneously calibrates all the tests and thus utilizes the correlations to increase measurement precision. In Item Responses Theory (IRT) terminology, the model is termed as multidimensional item response model (Wu et al., 2007, 2016). Specifically, in evaluating scores on the TIKI-T, it is important to consider both the subtest level and the test level. Initially, the IQ score was calculated by simply adding the number of correct answers across all subtests without taking into account the correlation between subtests. If the subtests are found to be uncorrelated, it is possible to use a unidimensional model to scale each ability separately. if the subtests are found to be correlated, a multidimensional model would be a more appropriate analysis method compared to the unidimensional model. This is because a multidimensional model takes into account the correlations between subtests and allows for the examination of abilities across multiple dimensions (Wu et al., 2007, 2016). It is important to note that the selection of the appropriate analysis method should be based on the characteristics of the test items and the goals of the analysis. Therefore, when examining gender differences across dimensions on the TIKI-T, it is crucial to consider both the subtest scores and the test-level scores to obtain a comprehensive understanding of performance differences between genders. a multidimensional item response modeling was used to examine the four domains of Arithmetic (AR), Component (CO), Word Relations (WR), and Figure Classification (FC).

This paper aims to serve two primary objectives: (1) To investigate the empirical evidences supporting the four subtests using a higher order of four-dimensional model, and (2) To explore gender performance within and across the four subtests of TIKI-T. This multidimensional model may yield useful information on gender differences in several cognitive abilities.

Methods

Participants

The study involved 1443 undergraduate students from Universitas Padjadjaran, Bandung, Indonesia, who were randomly selected from nine study programs out of more than 50. The students belonged to various programs such as Animal Husbandry, Psychology, Sundanese Literature, Biology, Pharmacy, Geological Engineering, International Relations, Library Sciences, and Sociology. The percentage of students in each program varied, with Animal Husbandry having the highest percentage (23.4%), followed by Psychology (18.4%), and Sundanese Literature (14.3%). The majority of the participants were first (40%) and second (48%) year students, while only a small percentage (2%) were third year students. The reason for the majority of females (66.9%) in the sample was that these study programs were chosen by more females than males. On average, the participants were between 17–22 years old (M=19.34, SD=0.91).

The study was conducted on a voluntary basis, where students chose to participate if they wanted to. The data collection process involved several steps. Firstly, the researchers explained the purpose of the study and the data collection process to the targeted students. Secondly, the researchers collected informed consent from the students, which indicated their willingness to participate in the study. Thirdly, the data was collected in classrooms with 40 to 60 students per room. The testing was conducted by an experienced instructor with 2 to 3 assistants under the supervision of a psychologist.

Instrument

Each participant completed the TIKI-T (Drenth et al., 1977). The test consits of four subtest: (1) Arithmetic (40 items with 7 minutes time limit), Component (26 items with 7 minutes time limit), Word Relations (40 items with 5 minutes time limit), and Figure Classification (30 items with 7 minutes time limit). Following general instructions and practice problems, the TIKI-T was administered with a 40-minute time limit. Before each test was started, instructions and practice items were presented to the subjects to make sure they understood how to complete the tests.

Multidimensional Rasch Model

The basic assumption in the Rasch model is that a set of items measures only one trait, known as unidimensionality (Bond & Fox, 2015; Hambleton & Swaminathan, 1985). However, for a standardized test, sometimes one instrument measures several traits, which violates the basic assumption of the Rasch model (Ansley & Forsyth, 1985). Furthermore, the correlations between traits are often highly correlated. In such cases, a multidimensional Rasch model is appropriate to analyze this problem.

The multidimensional Rasch model is an extension of the Rasch Model, in which the population distribution is a multivariate distribution rather than univariate (Wu et al., 2016). The model is able to cover simultaneous analyses for all domains, rather than partial analysis. As a result, the latent correlation between constructs could be acquired without measurement errors. To examine the dimensionality of the TIKI-T and to make comparison across gender, a Multidimensional Rasch Models for the dichotomous variables was selected for item calibration and ability estimation in this study. Based on the assumption that every item measures only one latent variable, the multidimensional between-item model was employed as an approach in this study (Adams et al., 1997). The Conquest 2.0 with Monte Carlo method to estimate the item parameters and WLE method were used to produce students' ability estimate (Wu et al., 2007). Weighted likelihood estimation (WLE) has been developed and shown to have lower bias than maximum likelihood estimation (MLE) while having equivalent asymptotic variance and normal distribution (Warm, 1989). To get the well estimated latent variance-covariance matrix the analysis used 2000 nodes (Volodin & Adams, 1995).

A deviance value and number of parameter between unidimensional and multidimensional models were examined to decide the best model for the test. Next, a Wright Map was analyzed to examine the relation between item difficulty and person ability. Lastly, to determine the quality of the item, Weighted Fit Mean Squared (WFMS) and Unweighted Fit Mean Squared (UWFMS) were examined. Linacre, 2012) proposed that misfit statistics can be considered moderately high if they range between 1.5 and 2.0, and severe if they are higher than 2.0. Another approach, the WFMS and UWFMS values between 0.75 and 1.33 are the general consideration for item fit (Adams & Khoo, 1996; Bond & Fox, 2015). WFMS lower than 0.75 is known as overfit, meanwhile, WFMS higher than 1.33 is identified as underfit.

Differential Item Functioning

The term "differential item functioning" (DIF) is closely related to the concept of bias in an item. Osterlind (1983) defined bias as a systematic error that occurs in the measurement process. However, the term bias is not commonly used to describe item bias. Instead, researchers use the term DIF to explain the bias in an item. DIF occurs when individuals from different groups with the same ability level have different probabilities of answering a particular item correctly (Osterlind, 1983; Temel et al., 2022; Wu et al., 2016). This indicates that the item has different levels of difficulty for different groups of individuals (e.g male and female). At the test level which explains the different between score, the term known as Differential Test Fuctioning (DTF)(Temel et al., 2022). To identify DIF between groups, Bond & Fox

(2015) and Wu et al. (2016) proposed a substantial difference between groups in item estimate values to be a difference of ±0.50 logits. To perform this analysis, the data from the TIKI-T set were subjected to Rasch analysis using Conquest 2.0 software (Wu et al., 2007).

Data Analysis Strategy

An initial analysis was conducted in this research. Curtis & Boman (2007) recommend that before conducting primary test analyses, it is crucial to examine person fit statistics to determine the degree to which a person's response pattern aligns with the model pattern. Person fit statistics provide a measure of the fit between an individual's response pattern and the expected response pattern based on the model (Freitas et al., 2014). If a person's fit value is greater than 1.3, their response pattern may be unreliable, leading to unpredictable data patterns (Bond & Fox, 2015). Therefore, it is advisable to exclude individuals with poor person fit values from subsequent analyses. This ensures that only high-quality data is used for analysis, leading to more accurate and reliable results.

Earlier, it was mentioned that the analysis of the TIKI-T involved Multidimensional Rasch Models for four subtests with scale alligment, as well as a Differential Item Functioning (DIF) analysis. These analyses were conducted using ConQuest 2.0 software, as described by Wu et al. (2007). Furthermore, item discrimination analysis was conducted to evaluate the quality of the test. The item discrimination index values were classified according to the following criteria: values less than or equal to 0.20 were considered poor, values between 0.21 and 0.24 were considered acceptable, values between 0.25 and 0.34 were considered good, and values greater than or equal to 0.35 were considered excellent (Date et al., 2019). The analysis was conducted using JASP 0.17.1 (JASP Team, 2023)

Data Analysis Strategy

An initial analysis was conducted in this research. Curtis & Boman (2007) recommend that before conducting primary test analyses, it is crucial to examine person fit statistics to determine the degree to which a person's response pattern aligns with the model pattern. Person fit statistics provide a measure of the fit between an individual's response pattern and the expected response pattern based on the model (Freitas et al., 2014). If a person's fit value is greater than 1.3, their response pattern may be unreliable, leading to unpredictable data patterns (Bond & Fox, 2015). Therefore, it is advisable to exclude individuals with poor person fit values from subsequent analyses. This ensures that only high-quality data is used for analysis, leading to more accurate and reliable results.

Earlier, it was mentioned that the analysis of the TIKI-T involved Multidimensional Rasch Models for four subtests with scale alligment, as well as a Differential Item Functioning (DIF) analysis. These analyses were conducted using ConQuest 2.0 software, as described by Wu et al. (2007). Furthermore, item discrimination analysis was conducted to evaluate the quality of the test. The item discrimination index values were classified according to the following criteria: values less than or equal to 0.20 were considered poor, values between 0.21 and 0.24 were considered acceptable, values between 0.25 and 0.34 were considered good, and values greater than or equal to 0.35 were considered excellent (Date et al., 2019). The analysis was conducted using JASP 0.17.1 (JASP Team, 2023).

Results and Discussion

Results

Person Fit Analysis

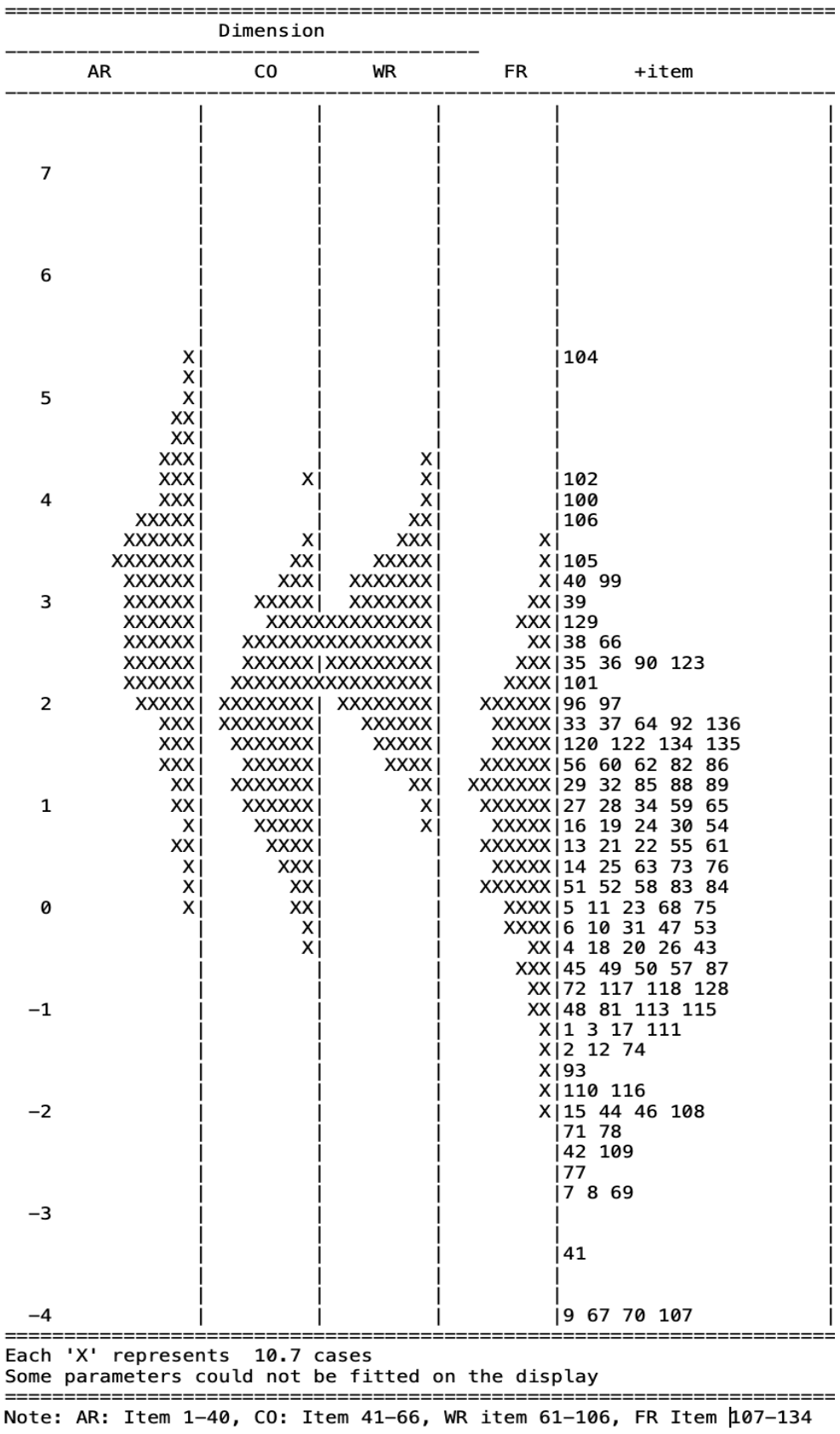
This study examined 1,433 undergraduate students from a single university in Bandung, Indonesia, representing eight faculties. Initially, a person fit index analysis was conducted using the weighted mean square (infit) as the indicator. Table 2 summarizes the person fit statistics for four subtests of the

test, with mean weighted MNSQ values ranging from 0.01 to 16.71, indicating a wide range of infit values. As a result, 412 subjects were excluded due to their person fit indexes exceeding 1.3 classified as misfit or underfit persons. Since underfitting can reduce the quality of subsequent measurements (Bond & Fox, 2015), the analysis was based on a total of 1,031 subjects.

Multidimensional Rasch Model

Two analyses were conducted to choose the best fitting model between unidimensional and a higher order of four-dimensional models for comparison. This study used deviance statistics as a consideration for choosing the best model for a data set (Wu et al., 2007) . Deviance for unidimensional model of the TIKI-T was 71,221.16 with 134 number of parameters, while deviance for the higher order of four-dimensional models model was 69,300.52 with 279 number of parameters. The difference in deviance statistics of these two models is 1,921 with 145 degrees of freedom, where degrees of freedom are the difference in number of parameters estimated in the unidimensional and multidimensional models. The difference of deviance was statistically significant at alpha = .001 level. This provides statistical support for the use of a Multidimensional Rasch Model, along with the theoretical support on the basis of the assessment design for the four content topics.

In Multidimensional Rasch Model, item parameters and person abilities estimates were calibrated to be on the same logit metric. Thus, within a single dimension all model parameter estimates can be compared on the same scale. Wright Maps is a visualization of the items which are indicated by item number and each individual person's performance which are represented by an 'X". Basically, Wright Maps only report the relationship between item difficulty and person ability. Meanwhile, error estimate and fit indices for the item and the person are reported separately (Bond & Fox, 2015). For item targeting, it is important that item difficulty distribution covers the span of person ability distribution to provide accurate measures of person's proficiency over the whole scale. A lack of items in difficulty range will lead to large errors in ability estimation and low item reliability (Bond & Fox, 2015).



Sources: Personal data (2023)

Figure 1. Wright Map for a Higher Order of Four-Dimensional Model of *TIKI-T.*

In Figure 1, in general, items difficulty in the TIKI-T did not cover the distribution of the person's abilities. Specifically, in Dimension 1 (AR), Dimension 2 (CO), and Dimension 3 (WR), the items seemed too easy. The items did not cover for the top students. Approximately, more than 15 out of 40 items in AR subtest had item difficulties estimate bellow 0 logit score. Meanwhile, for CO subtest, 11 out of 26 items had item difficulties estimate bellow 0 logit score. For WR subtest, even more items were relatively easy for the students. The FC subtest was the best dimension compared with the other subtests. Almost all items were able to cover the spread of person's abilities. These results affected the reliability of item difficulty estimates for each subtest. Receptively, the reliability of item difficulty estimates for AR, CO, WR and FC were 0.65, 0.77, 0.63 and 0.87. Item reliability can be interpreted the same way as Cronbach's alpha is interpreted (Bond & Fox, 2015). With a cut point of 0.7 subtest, AR and WR were unreliable (Loewenthal, 2001).

Further analysis was conducted to determine the qualities of the items. The items were classified into four classifications: poor item, acceptable item, good items and exelent item (Date et al., 2019). The focus of the analysis is to items in the poor classification. There are six item (15%) on the arithmetic subtest, eight (20%) item on the Word relation subtest, and one item (3.3 %) in the poor classifications. In the literature these kind of items need to be deleted or revision. The results is presented in Table 1.

Table 1. Item Discrimination Classification.

Dimension	Poor Items (D < 0.2)	Acceptable items (0.21 < D < 0.24)	Good Items (0.25 < D < 0.34)	Excellent Items (D > 0.34)
AR	6 (15%)	5 (12.5%)	5 (12.5%)	24 (60%)
CO	0 (0 %)	2 (7.7%)	6 (23.1%)	18 (69.2%)
WR	8 (20%)	12 (30%)	8 (20%)	12 (30%)
FC	1 (3.3%)	3 (10%)	5 (16.7%)	21 (70%)

Sources: Personal data (2023)

In addition, Table 2 shows the correlation between dimensions reported from ConQuest 2.0. The correlation ranges from 0.412 to 0.637. Initialy measurement error in person measures, can cause computed correlations to underestimate the true correlations between latent traits. This underestimation is commonly referred to as attenuation due to measurement error. However, the multidimensional approach addresses this issue by directly estimating the correlation while taking into account the impact of measurement error, resulting in an unbiased estimate without any attenuation because the correlation was directly computed using an estimated variance matrix (Andersen & Madsen, 1977). Thus, the correlation reflects the latent correlation(Wu et al., 2016). The CO was highly correlated with FC. These variables measured the same construct of spatial/nonverbal ability. FC also moderately correlated with AR and WR. This indicated that those variables measure some common cognitive abilities.

Table 2. Correlation Coefficients between Variables.

Dimension	1	2	3	4
1. AR		**0.526**	**0.454**	**0.807**
2. CO	0.412		**0.408**	**0.827**
3. WR	0.473	0.502		**0.547**
4. FC	0.525	0.637	0.560	
Variance	1.511	1.079	0.610	1.561

Sources: Personal data (2023)

Note: Upper triangle covariance matrix, lower triangle correlation matrix

Commonly, infit and outfit statistics are two frequently used item fit statistics that are used to compare the observed response patterns with the expected response patterns (Bond & Fox, 2015). ConQuest provides a weighted fit mean square (WFMS) statistics known as infit statistics and an unweighted fit means square (UWFMS) known as outfit statistics (Wu et al., 2007). The WFMS or infit values between 0.75 and 1.33 are the general consideration for item fit (Adams & Khoo, 1996; Bond & Fox, 2015). WFMS lower than 0.75 is known as overfit, meanwhile, WFMS higher than 1.33 is identified as underfit.

Table 3. Summary of WFMS and UWFMS from the *TIKI-T.*

	WFMS Item Fit Type			UWFMS Item Fit Type		
Dimension	Overfit (< 0.75)	Fit (0.75 - 1.33)	Underfit (>1.33)	Overfit (< 0.75)	Fit (0.75 - 1.33)	Underfit (>1.33)
AR	0 (0 %)	39 (98 %)	1 (3 %)	8 (20 %)	26 (65 %)	6 (15 %)
CO	0 (0 %)	26 (100%)	0 (0 %)	4 (15 %)	22 (85 %)	0 (0 %)
WR	0 (0 %)	38 (95 %)	2 (5 %)	8 (20 %)	30 (75 %)	2 (5 %)
FC	0 (0 %)	30 (100%)	0 (0 %)	5 (17 %)	24 (80 %)	1 (3 %)

Sources: Personal data (2023)

Table 3 illustrates the summary of WFMS and UWFMS types for each dimension. Based on the WFMS, almost all dimensions showed reasonably good fit with the model. Only one item on the AR subtest (item number 9) and two items on the WR subtest (items number 1 and 2) were classified as underfit items. These occurred because all participants answered those items correctly. For the UWFMS indices, approximately 15–20% items for each dimension were overfit, meanwhile, around 3–15% items were indicated as underfit. The wide discrepancy between the fit statistics could result from the fact that the items are too easy. Hence, the items failed to distinguish among students with different ability level.

Differential Item Functioning Based on Gender

The analyses were conducted by examining fit indices of the items. After investigating the WFMS, evidences in Table 4 show that most items in the TIKI-T were within acceptable range (0.75-1.33), with only some fell out of range. It was found that the WFMS value of 5 items fell out of the predetermined range in male group, which indicates that the items did not fit. Items that were on the AR subtest are items number 9 and 36, and items that were on the WR subtests are items number 1, 2, and 10. Meanwhile, for the female group, four items were identified as underfit. The underfit item that was on the AR subtest is item number 9, whereas items that were on WR subtest are items number 1 and 2, and item that was on FC subtest is item number 16. As these items were classified as underfit in both male and female models, these items were identified as bad items, including those on the test that should be considered. Thus, based on item WFMS criterion, it is evident that gender bias was a problem in TIKI-T.

Table 4. Summary of WFMS across Gender from the *TIKI-T*

	Male (WFMS)			Female (WFMS)		
Dimension	Overfit (< 0.75)	fit (0.75 - 1.33)	Underfit (>1.33)	Overfit (< 0.75)	fit (0.75 - 1.33)	Underfit (>1.33)
AR	0 (0 %)	38 (95 %)	2 (5 %)	0 (0 %)	39 (97 %)	1 (3 %)
CO	0 (0 %)	26 (100%)	0 (0 %)	0 (0 %)	26 (100%)	0 (0 %)
WR	0 (0 %)	37 (93 %)	3 (7 %)	0 (0 %)	38 (95 %)	2 (5 %)
FC	0 (0 %)	30 (100%)	0 (0 %)	0 (0 %)	29 (97 %)	1 (3 %)

Sources: Personal data (2023)

Differential Item Functioning (DIF) was presented when examinees from male and female groups had differing probabilities of success on an item, after they have been equaled on the ability of cognitive ability (Wu et al., 2016; Zumbo, 2007). DIF can be identified by checking the difference in item difficulty for people of two groups with the same level of ability. The initial analysis determined the mean location for each gender. The mean location of the female group was -0.020, while the male group was 0.020. Overall, this result implied that males had more difficulties to answer the questions in TIKI-T rather than females. The chi-square test of the parameter equality was 3.88 with 1df (degree of freedom). Probability value of 0.049 showed that the difference was statistically significant. This result was an indication of DIF on the items of TIKI-T.

Table 5. Summary of DIF between Gender from the Short Form of the TIKI-T.

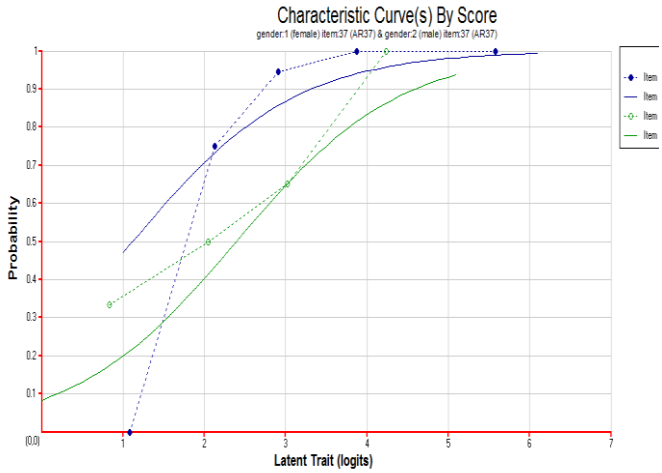
Dimension	Easier Item for Female $D_1 - D_2 > 0.50$	Non DIF item	Easier item for Male $D_1 - D_2 < -0.50$
AR	16 (40 %)	21 (53 %)	3 (8%)
CO	2 (8 %)	20 (77%)	4 (15 %)
WR	9 (23 %)	28 (70 %)	3 (8%)
FC	3 (10 %)	24 (80 %)	3 (10 %)

Sources: Personal data (2023)

In order to determine the extent of DIF (Differential Item Functioning), the disparity in difficulty between items for males and females was measured. This was done by subtracting D2, the item difficulty for females, from D1, the item difficulty for males, resulting in the value D1-D2. If the resulting value is negative, it means that the item too easy for males than females, while a positive value indicates the opposite. Using this criterion, it is obvious that a vast majority of the TIKI-T was apparently in favor of one gender or the other. However, the difference between estimate of an item for male and female groups may not be a sufficient indication as a bias for particular group. As mentioned earlier, a difference in item estimate ± 0.50 range is large enough to have both statistical and substantive meanings (Bond & Fox, 2015; Hungi, 2005). Meanwhile, a difference in standardized difference in item estimate over ± 2.00 can also be used as a criterion (Hungi, 2005; Wu et al., 2016). In addition, some correction formula has to be used for large samples (Hungi, 2005). Therefore, with 1,032 samples, the cut off point for standardized difference is ± 3.21.

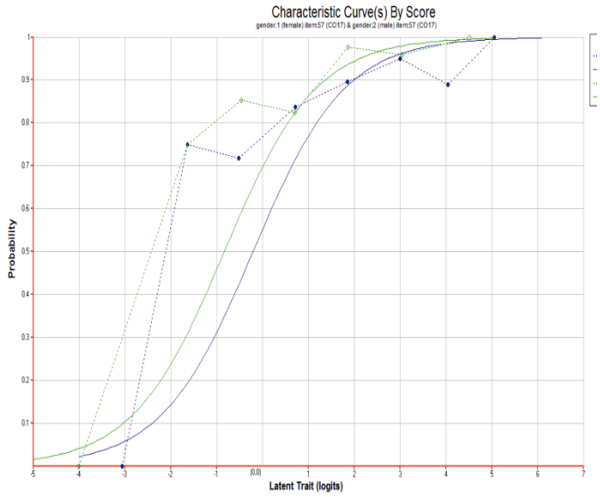
Table 5 shows the summary of DIF from the TIKI-T. Based on the criteria, overall, around 53– 80% items were not indicated as DIF items. Marked as DIF items, the items in the AR and WR subtests were markedly easier to answer by female compared to male group. While on the CO subtest, the proportion of easier items for male group was slightly higher than female group. For the FC subtest, the proportion of DIF items on female and male groups was exactly similar.

In a more detailed manner, almost 50% of items in the AR subtest were identified as DIF items. It was found that 16 items (items number 2, 3, 4, 5, 7, 13, 14, 15, 23, 27, 29, 30, 32, 35, 37, and 40) were evidently easier for female compared to male group. Meanwhile, only three items (items number 9, 18, and 38) were markedly easier for male rather than female group. In the CO subtest, two items (items number 1 and 26) were identified as easy for female group and four items (items number 12, 17, 20, and 23) were noticeably easier for male rather than female group. Next, nine items ((items number 4, 5, 10, 11, 21, 22, 27, 37, and 38) were evidently easy items for female group in the WR, and only three items (items number 1, 8, and 32) were markedly easier items for male than female group. Lastly, equally, three items were identified as easy items for both male group (items number 3, 14, and 20) and female groups (item number 1, 4, and 27). These DIF items look somewhat problematic because there was significant variance found in the items.



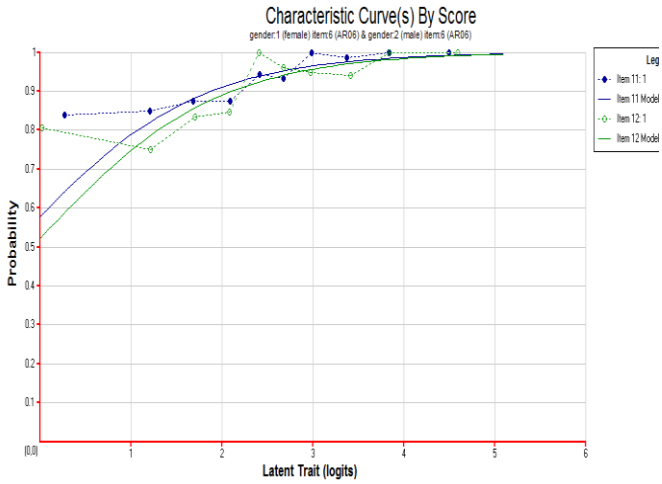
Sources: Personal data (2023)

Figure 2. ICC for Item 37 in AR Subtest (Biased in Favor of Female, D1 - D2 = -1.24 and st(D_1-D_2) =-3.47)



Sources: Personal data (2023)

Figure 3. ICC for Item 17 in CO Subtest (Biased in Favor of Male, D1 - D2 = 0.66 and st(D_1-D_2) =3.13)



Sources: Personal data (2023)

Figure 4. ICC for Item 06 in AR Subtest (Non-bias, D1 - D2 = 0.19 and st(D_1-D_2) =0.18)

Figures 2 to 3 show the Item Characteristic Curves (ICC) of the *TIKI-T,* identified as problematic items as explained in the preceding paragraphs (that is, item number 37 in AR subtest and item number 17 in CO subtest). Figure 4 is an example of non-DIF item (in this case item number 6 in AR subtest). The ICCs of Figures 3 to 5 were obtained from ConQuest 2.0 software (Wu et al., 2007).

In Figure 4, the ICC of female group was mostly higher than that of male for medium to high ability person, meaning that, for medium to high ability, this item is biased in favor of female. On the other hand, for item 17 in CO subtest (Figure 4), the ICC for male group was clearly higher than female, meaning that males stand greater chances than females in getting these items correctly at the same ability level. By comparing the value of D_1-D_2 between Figure 2 and Figure 3, it can be concluded that the bigger the difference value, the larger the distance between groups in ICC. Meanwhile, in Figure 4, the ICC for female and male groups was nearly on the same line. This item represents non-bias item between male and female groups.

Discussion

The present study aimed at investigating gender differences in cognitive abilities and differential item functioning using TIKI-T. This research explored the DIF between male and female groups under the a higher order of Multidimensional Rasch Model. Before discussing these results, it is important to acknowledge some limitations that should be considered in interpreting the findings. Specifically, this investigation was based on data that only represents the sample from one university in Bandung. Additionally, the majority of the students included in the study were female. This can be attributed to the fact that the chosen major of study in this university tends to attract more female students. Although the Rasch model results demonstrated sample invariance, it is important to note the gender imbalance in the respondent population.

This study takes a multidimensional Rasch approach to evaluate gender differences across the four domains of TIKI-T. The multidimensional Rasch model was shown to have better fit model than the unidimensional Rasch model. The four domains were correlated to support the higher four-factor models. Nevertheless, based on Wright Maps, some items need to be revised. Items in AR, CO, and WR subtests were too easy to answer and were not able to cover all range of person's abilities, especially high ability students. This condition also affected item reliability which is lower than standard (Bond & Fox, 2015).

The findings showed that although there were several unbiased items, some are clearly biased against male. The DIF was higher in female–male differences in numerical (AR) and verbal (WR) subtests. Meanwhile, DIF on male test corresponded to spatial/nonverbal subtest (CO). The AR test measures individuals' ability to solve simple numerical problems that require arithmetic operations, such as addition, subtraction, multiplication, and division (Drenth et al., 1977). The data revealed the presence of differential item functioning (DIF), indicating that females outperformed males on almost half arithmetic items. The findings on numerical ability were encountered in previous research. Some literatures describe that DIF occurs across gender in large samples such as PISA or TIMMS studies. In these researches, female students showed higher performance in numeracy subject compared with male students (Fitriati, 2014; Liu et al., 2008). However, In the contrary the reseach in Armed Services Vocational Aptitude Battery (ASVAB) showed that white-male outperformed in arithmatical subtest rather than female (Gibson & Harvey, 2003). Therefore, the results from this study add another information about gender difrerence in numerical ability.

In the verbal (WR) subtest, participants are required to identify two words with either identical or contrasting meanings (Drenth et al., 1977). Content analysis revealed a gender difference in performance, which may be attributed to differences in familiarity with the context of the items. For instance, males performed better on the word pair 'miser-philanthropist,' while females showed better performance on 'uninspired and crazed.' It has been well documented that female's superior performance in the test is related with verbal subtest (Hyde & Linn, 1988; Kimura & Hampson, 1994), while male performed well in spatial/nonverbal subtest (Colom, 2002; Geary et al., 2000). Normally, research found that males are higher in numerical and spatial problems (Geary et al., 2000).

In this study, the nonverbal subtest (CO and FR) was examined to assess the participant's ability to manipulate and transform figural material, as well as classify figural objects. The majority of items (77-80%) were found to be free from DIF. Moreover, after examining the test content, no specific characteristics of the items were indicative of gender effects on item response. These findings are consistent with previous research, which indicated that the analysis of gender DIF supports the equivalence of APM-SF items across male and female respondents indicating that no item appeared to be easier or harder for males compared to females (Chiesi et al., 2012).

Some researcher suggested at least three methods to deal with DIF items, such as: removing DIF items from the test, splitting DIF items into two new items, or retaining DIF items in the data (Wu et al., 2016). Clearly, items with very large DIF are nominees for deletion . However, in terms of deleting the items, the process should be done carefully, in relation to theoretical framework and test specification of the test. The process can be started by deleting the largest DIF magnitude and progressively working back for item decreasing magnitude of DIF. Sometimes, information about DIF items is important to control items composition. At least the composition of DIF items in favor to male is equal to female. In Indonesia, the norm group for intelligence tests is typically based on age. However, for the TIKI-T test, DIF can be addressed by using different norm groups based on gender, which is uncommon. The development of gender-based norms for TIKI-T was achieved using the Rasch model, which estimates person ability based on item parameters related to gender.

Conclusion

Multidimensional Rasch Analysis is a powerful method for evaluating the psychometric properties of instruments, particularly in the field of psychology. By analyzing multiple dimensions for each construct, MRA can provide detailed information on the relationship between the construct and the performance of each item, with a focus on identifying items that exhibit differential item functioning (DIF) across genders. The identification of DIF items is essential for evaluating potential sources of bias in psychological assessments. The findings of this study suggest that when comparing intelligence across genders using available TIKI-T norms, test users must consider the possibility of differential item functioning (DIF) and take steps to ensure unbiased scoring. In addition, the removal of DIF items from the TIKI-T and the development of a new normative scoring method such as item response theory, without these items is one approach to mitigating the impact of such bias and ensuring that the test scores are unbiased. Future research could use Multidimensional Rasch Analysis to investigate potential sources of bias in employee selection settings and across different regions of Indonesia. Such research would help to increase the generalizability of the TIKI test results and inform the development of more inclusive and equitable psychological assessments.

Acknowledgment

We would like to thank you to head of nine study programs at Universitas Padjadjaran for providing support in data collection.

Conflict of Interest

All authors have no conflict of interest to disclose.

Authors Contribution

WY contributed to conception, design of the study and stastical analyses. AT organized the database and provided literature review. HS contributed to design study and the statistical analysis . All authors wrote sections of the manuscript, contributed to manuscript revision, read, and approved the submitted version.

References

Abad, F. J., Colom, R., Rebollo, I., & Escorial, S. (2004). Sex differential item functioning in the Raven's Advanced Progressive Matrices: Evidence for bias. *Personality and Individual Differences*, *36*(6), 1459–1470. https://doi.org/10.1016/S0191-8869(03)00241-1

Adams, R. J., & Khoo, S. T. (1996). Quest. *Melbourne, Australia: ACER*.

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1–23. https://doi.org/10.1177/0146621697211001

Andersen, E., & Madsen, M. (1977). Estimating the parameters of the latent population distribution. *Psychometrika*, *42*, 357–374.

Ansley, T. N., & Forsyth, R. A. (1985). An examination of the characteristics of unidimensional IRT parameter estimates derived from two-dimensional data. *Applied Psychological Measurement*, *9*(1), 37–48.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (Third Edition). Routledge.

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

Chiesi, F., Ciancaleoni, M., Galli, S., Morsanyi, K., & Primi, C. (2012). Item response theory analysis and differential item functioning across age, gender and country of a short form of the advanced progressive matrices. *Learning and Individual Differences*, *22*(3), 390–396.

Colom, R. (2002). *Sex differences in fluid intelligence among high school graduates*. *32*(2016), 445–451. https://doi.org/https://doi.org/10.1016/S0191-8869(01)00040-X

Colom, R., Juan-Espinosa, M., Abad, F., & García, L. F. (2000). Negligible sex differences in general intelligence. *Intelligence*, *28*(1), 57–68. https://doi.org/10.1016/S0160-2896(99)00035-5

Curtis, D. D., & Boman, P. (2007). X-ray your data with Rasch. *International Education Journal*, *8*(2), 249–259.

Date, A. P., Borkar, A. S., Badwaik, R. T., Siddiqui, R. A., Shende, T. R., & Dashputra, A. V. (2019). Item analysis as tool to validate multiple choice question bank in pharmacology. *International Journal of Basic & Clinical Pharmacology*, *8*(9), 1999–2003. https://doi.org/http://dx.doi.org/10.18203/2319-2003.ijbcp20194106

Deary, I. J., Irwing, P., Der, G., & Bates, T. C. (2007). Brother-sister differences in the g factor in intelligence: Analysis of full, opposite-sex siblings from the NLSY1979. *Intelligence*, *35*(5), 451–456. https://doi.org/10.1016/j.intell.2006.09.003

Deary, I. J., Thorpe, G., Wilson, V., Starr, J. M., & Whalley, L. J. (2003). Population sex differences in IQ at age 11: The Scottish mental survey 1932. *Intelligence*, *31*(6), 533–542. https://doi.org/10.1016/S0160-2896(03)00053-9

Drenth, P. J. D., Dengah, B., Bleichrodt, N., Soemarto, & Poespadibrata, S. (1977). *Test intelligensi kolektip Indonesia*. Swets & Zeitlinger.

Fitriati. (2014). Differential item functioning: Analysis of TIMMS mathematics test items using Australian and Indonesian database. *Makara Hubs-Asia*, *18*(2), 127–139. https://doi.org/10.7454/mssh.v18i2.3467

Freitas, S., Prieto, G., Simões, M. R., & Santana, I. (2014). Psychometric properties of the Montreal Cognitive Assessment (MoCA): an analysis using the Rasch model. *The Clinical Neuropsychologist*, *28*(1), 65–83. https://doi.org/10.1080/13854046.2013.870231

Geary, D. C., Saults, S. J., Liu, F., & Hoard, M. K. (2000). Sex differences in spatial cognition, computational fluency, and arithmetical reasoning. *Journal of Experimental Child Psychology*, *77*(4), 337–353. https://doi.org/10.1006/jecp.2000.2594

Gibson, S. G., & Harvey, R. J. (2003). Gender and ethnicity based differential item functioning on the Armed Services Vocational Aptitude Battery. *Equal Opportunities International*, *22*(4), 1–15.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Kluwer.

Hungi, N. (2005). Employing the Rasch model to detect biased items. In *Applied Rasch measurement: A book of exemplars* (pp. 139–157). Springer.

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, *104*(1), 53–69. https://doi.org/10.1037/0033-2909.104.1.53

JASP Team. (2023). *JASP (Version 0.17.1)*. [Computer Software].

Kimura, D., & Hampson, E. (1994). Cognitive patten in men and women is influenced by fluctuations in sex hormones. *Current Directions in Psychological Science*, *3*(2), 57–61.

Lim, T. K. (1994). Gender-related differences in intelligence: Application of confirmatory factor analysis. *Intelligence*, *19*(2), 179–192. https://doi.org/10.1016/0160-2896(94)90012-4

Linacre, J. M. (2012). *A user's guide to WINSTEPS MINISTEP Rasch-model computer programs.: program manual 3.73. 0. 2011*.

Liu, O. L., Wilson, M., & Paek, I. (2008). A multidimensional Rasch analysis of gender differences in PISA mathematics. *Journal of Applied Measurement*, *9*(1), 18–35.

Loewenthal, K. M. (2001). *An introduction to psychological tests and scales*. Psychology Press.

Lord, F. M., & Stocking, M. L. (1988). Item response theory. In J. P. Keeves (Ed.), *Educational Research, Methodology, and Measurement: An International Handbook* (pp. 269–272). Pergamon Press.

Lynn, R., & Irwing, P. (2004). Sex differences on the advanced progressive matrices in college students. *Personality and Individual Differences*, *37*(1), 219–223. https://doi.org/10.1016/j.paid.2003.08.028

Lynn, R., & Kanazawa, S. (2011). A longitudinal study of sex differences in intelligence at ages 7, 11, and 16 years. *Personality and Individual Differences*, *51*(3), 321–324.

Nisa, D. H. M. (2013). Implementasi kelas akselerasi dalam pembelajaran pendidikan agama Islam (PAI) di SMA Negeri 1 Kediri. *Didaktika Religia*, *1*(1).

Osterlind, S. J. (1983). *Test item bias*. Sage Publication.

Permatasari, T. O. (2016). Faktor kognitif dan non-kognitif pada seleksi mahasiswa baru sebagai prediktor terhadap prestasi akademik. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *20*(1), 80–89.

Reynolds, M. R., Keith, T. Z., Ridley, K. P., & Patel, P. G. (2008). Sex differences in latent general and broad cognitive abilities for children and youth: Evidence from higher-order MG-MACS and MIMIC models. *Intelligence*, *36*(3), 236–260. https://doi.org/10.1016/j.intell.2007.06.003

Savage-McGlynn, E. (2012). Sex differences in intelligence in younger and older participants of the Raven's standard progressive matrices plus. *Personality and Individual Differences*, *53*(2), 137–141. https://doi.org/10.1016/j.paid.2011.06.013

Steinmayr, R., Beauducel, A., & Spinath, B. (2010). Do sex differences in a faceted model of fluid and crystallized intelligence depend on the method applied? *Intelligence*, *38*(1), 101–110. https://doi.org/10.1016/j.intell.2009.08.001

Strand, S., Deary, I. J., & Smith, P. (2006). Sex differences in cognitive abilities test scores: a UK national picture. *The British Journal of Educational Psychology*, *76*(Pt 3), 463–480. https://doi.org/10.1348/000709905X50906

Temel, G. Y., Rietz, C., Machunsky, M., & Bedersdorfer, R. (2022). Examining and improving the gender and language DIF in the VERA 8 Tests. *Psych*, *4*(3), 357–374. https://doi.org/https://doi.org/10.3390/psych4030030

Volodin, N. A., & Adams, R. J. (1995). Identifying and estimating a D-dimensional item response model. *International Objective Measurement Workshop, University of California, Berkeley, California*.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450.

Wu, M., Adams, R. J., Wilson, M. R., & Haldane, S. a. (2007). ACER ConQuest version 2.0 : generalised item response modelling software. In *Educational Research*. ACER Press.

Wu, M., Tam, H. P., & Jen, T. (2016). *Educational measurement for applied researchers: Theory into practice*. Springer Nature. https://doi.org/10.1007/978-981-10-3302-5

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*(2), 223–233. https://doi.org/https://doi.org/10.1080/15434300701375832