REID (Research and Evaluation in Education), 8(1), 2022, 55-66
Available online at: http://journal.uny.ac.id/index.php/reid

Gender differential item functioning on the Kentucky Inventory of Mindfulness Skills instrument using logistic regression

Sumin[1*]; Fitri Sukmawati[1]; Nurdin[2]
[1]Institut Agama Islam Negeri Pontianak, Indonesia
[2]Dinas Pendidikan Provinsi Aceh, Indonesia
[*]Corresponding Author. E-mail: amien.ptk@gmail.com

ARTICLE INFO

Article History
Submitted:
15 June 2022
Revised:
29 June 2022
Accepted:
30 June 2022

Keywords
item biases; differential item functioning; logistic regression

Scan Me:

ABSTRACT

The item differential function (DIF) describes a situation in which testees of similar ability but from different demographic groups have varying chances of achieving the same result. This study aims to identify the function of uniform and non-uniform differential items on the Kentucky Inventory of Mindfulness Skills Instrument using logistic regression techniques and determine the impact of DHF on construct validity. This study uses a survey method with a quantitative approach. The study involved 602 people, divided into two groups based on gender: 301 women and 301 men. The Kentucky Inventory of Mindfulness Skills (KIMS) is a 39-item online questionnaire that measures mindfulness. KIMS has been proven to meet content, construct, and factor validity and has good test-retest reliability and internal consistency estimators. This study uses Regression Logistics to detect DIF, analyzed with R Studio 4.1.3 software. Research results found 17 DIF items detected using logistic regression, 13 uniform DIF items, and four non-uniform DIF. Through CFA, we have succeeded in proving that DIF-free items are proven to have better construct validity. The implications of this study are expected to inspire counseling psychologists to be more careful in using rating scales or instruments. The validity and reliability of the measures are not strong enough to justify that all measuring instruments are correct. However, it is also necessary to check for item bias or functional differential items to ensure that each item on the scale or instrument is understandable to all demographic groups and does not benefit only certain demographic groups.

How to cite:
Sumin, S., Sukmawati, F., & Nurdin, N. (2022). Gender differential item functioning on the Kentucky Inventory of Mindfulness Skills instrument using logistic regression. *REID (Research and Evaluation in Education), 8*(1), 55-66. doi:https://doi.org/10.21831/reid.v8i1.50809

INTRODUCTION

A psychologist is often faced with clients with diverse backgrounds, such as gender, education, religion, region, and age. Client demographic diversity can affect their response to test questions. Therefore, psychologists must ensure that the instruments used are free of demographic group bias. This statement is empirical evidence from the study of Saygin and Atwater (2021) that gender differences in self-assessment and the tendency to abstain or not fill out certain items in all test domains can be dominated by either male or female, depending on the level of difficulty and variety of questions presented. In line with the results of research conducted by Cuevas and Cervantes (2012), each test parameter that is different between two or more subpopulation groups, such as the level of item difficulty, can be a sign of a threat to the validity of the test. This is because test results will require different interpretations for each group. Therefore, it is very important to pay attention to the function of the differential items in the scale or instrument development stage.

Differential item functioning (DIF) refers to the conditions experienced by individuals with the same ability from different demographic groups but have different opportunities to achieve scores on the same item. DIF problems can be a potential problem in test instruments and non-test instruments. The existence of DIF poses a serious threat to construct validity. Thus, the DIF assessment is an important first step in this process (Abedalaziz et al., 2018).

Item Response Theory (IRT) provides a powerful method for detecting item bias from a set of measuring instruments developed as a differential item functioning. Item bias detection using the IRT approach is a development of bias detection through the classical test theory (CTT) approach, which has several limitations, the CTT approach cannot distinguish between situations where (a) the subgroups have different methods, and the test is biased , versus (b) different ways, but the test is not biased (Abedalaziz, 2010).

The item differential function (DIF) is generally assessed to test the test's fairness prerequisites (Stark et al., 2006), It has evolved into a standard procedure that is used in all-encompassing educational examinations such as the Trends in Mathematics and Science Study (TIMSS) and the Program for International Student Assessment (PISA). DIF indicates a balance in the chances that two groups will actually respond to or support an item, even when participants in both groups have the same ability level. DIF indicates a balance in the chances that two groups will actually respond to or support an item, even when participants in both groups have the same level of ability (Chen & Jin, 2018).

Recent efforts to establish alternative measurement techniques, such as performance appraisals, original assessments, portfolio assessments, and even non-cognitive personality assessments, have inspired interest in examining DIF in various item types, particularly polytomous assessment items. According to Wiberg (2007), many methods can be used to detect DIF, and the most popular in item response theory (IRT) are the Manel-Haenszel method, Logistic Regression, and Likelihood Ratio Test. However, Swaminathan and Rogers (1990) revealed that the logistic regression procedure was stronger than the Mantel-Haenszel procedure for detecting uniform and non-uniform DIF. Besides detecting uniform and non-uniform DIF, the logistic regression method can also detect DIF in polytomous data with ordinal measurement degrees. This is in line with the results of the Camilli and Congdon (1999) study said that the use of cox regression and logistic regression is suitable for detecting DIF in polytomous responses. Logistic regression (LR) was established as a realistic approach for finding the differential item function (DIF). It was first developed in the 1960s. The fact that this method allows for some leeway in the way that the regression equation is specified is one of the reasons why it should be used (Mazor et al., 1995).

Based on those empirical study, this study aims to detect uniform and non-uniform differential items functioning using a logistic regression approach on the Kentucky Inventory of Mindfulness Skills Instrument and to determine the impact of DIF on construct validity. DIF occurs when the manifest group has "different probabilities of correct answers, even though group members have the same ability on the test" (Bandalos, 2018). Manifest groups are broken down into focus groups and reference groups. Focus groups are generally identified as minority or disadvantaged groups, whereas reference groups are usually the majority normative group (Martinková et al., 2017). For instance, if gender linked DIF research focuses on bias against women, referral groups are mandatory for men, and focus groups are compulsory for women.

There are two types of DIF, namely uniform DIF and non-uniform DIF. Uniform DIF reports when psychometric traits are measured consistently, whereas non-uniform DIF reports when psychometric traits are measured inconsistently. Discovering differential item functions has inspired the development of many DIF approaches. This approach may be broken down into two categories: parametric and non-parametric, depending on whether it is used to analyze observable or latent variables. It can identify a uniform or non-uniform DIF that is either suitable for polytomous or dichotomous score data. Based on these features, Wiberg (2007) has classified the DIF method, and Table 1 presents several DIF methods.

Table 1. Types of DIF Methods

No.	DIF Methods	Dichotomous	Polytomous	Uniform	Non-Uniform
1.	Manel-Haenszel	Yes	Yes	Yes	No
2.	Standardization	Yes	No	Yes	No
3.	Chi-Square Techniques	Yes	No	Yes	No
4.	Sibtest	Yes	Yes	Yes	Yes
5.	Logistic Regression	Yes	Yes	Yes	Yes
6.	Likelihood Ratio Test	Yes	Yes	Yes	Yes
7.	General IRT-LR	Yes	Yes	Yes	Yes
8.	IRT LRT	Yes	Yes	Yes	Yes
9.	IRT Methods	Yes	Yes	Yes	Yes
10.	Lor's Chi-Squared Test	Yes	No	Yes	Yes
11.	Log-Linear Models	Yes	Yes	Yes	Yes
12.	Mixed Effect model	Yes	Yes	Yes	Yes

Logistic Regression Approach to Detect DIF

The logistic regression approach to detect DIF was introduced by Swaminathan and Rogers (1990) as a result of its ability to do straightforward calculations while concurrently detecting both uniform and non-uniform DIF, this strategy has emerged as one of the most widely used DIF detection approaches. In order to simulate the chance of obtaining an item right using conditioning factors (such as the total observed test score), group membership, and interactions between the conditioning variables and group membership, logistic regression techniques were utilized. If the regression coefficients associated with group membership or group conditioning interactions for an item are statistically different from zero, that item is said to demonstrate differential item functioning (DIF). Maximum likelihood (ML) estimates are often used to estimate regression coefficients, and asymptotic distributions are typically relied upon for statistical testing hypotheses on the degree of independence (Lee, 2017).

According to Desjardins and Bulut (2018), the DIF detection method using logistic regression is based on the comparison of a series of logistic regression models in which the probability of correctly answering a dichotomous item (or favouring a particular response option) is predicted by the estimated examinee trait, group membership, and the interaction of the estimated examinee trait with group membership. The logistic regression model can predict the probability of a correct answer on an item, and it can be expressed by a mathematical equation as follows (Swaminathan & Rogers, 1990), where u is response to item, θ is the observed ability of a test taker, and β_1 is the slope parameter.

$$P(u = 1|\theta) = \frac{ee^{(\beta_0+\beta_1\theta)}}{[1 + e^{(\beta_0+\beta_1\theta)}]},$$

However, this model is a standard logistic regression model used to predict the dichotomous dependent variable. The aforementioned logistic regression model can determine DIF by determining separate equations for the two groups of interest (focal group and reference group) through the following mathematical equation as seen in Formula (1), in which u_{ij} is the response of the testee i in group j for a particular item, β_{0j} is the slope parameter, β_{1j} is the slope parameter for group j, and θ_{ij} is the ability of testee i in group j.

$$P\left(u_{ij} = 1|\theta_{ij}\right) = \frac{e^{\left(\beta_{0j}+\beta_{1j}+\theta_{1j}\right)}}{\left[1+e^{\left(\beta_{0j}+\beta_{1j}+\theta_{1j}\right)}\right]}, \quad i = 1,2,\ldots,n_j, j = 1,2. \ldots\ldots\ldots\ldots (1)$$

The accepted definition of DIF is that an item exhibits DIF if individuals with the same ability but from different groups do not have the same chance of successfully working on an item. Therefore, no DIF appears if the LR curves for the two groups are the same, i.e., if $\beta_{01} = \beta_{02}$, and $\beta_{11} = \beta_{12}$. If $\beta_{11} = \beta_{12} \neq \beta_{02}$, the curves are parallel but do not coincide, and therefore a uniform DIF is identified.

If $\beta_{01} = \beta_{02}$, and $\beta_{11} \neq \beta_{12}$, the curves are not parallel, and it shows that DIF is non-uniform. An alternative but equivalent to model (1) is (Desjardins & Bulut, 2018; Swaminathan & Rogers, 1990) as seen in Formula (2). In Formula (3), the variable g represents group membership which is defined as in Formula (4).

$$P(u = 1) = \frac{e^z}{[1+e^z]}, \text{.............} (2)$$

where

$$z = \tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta g) \text{ ....} (3)$$

$$g\begin{cases} 1 \text{ if the participant is a member of group 1} \\ 0 \text{ if the participant is a member of group 2} \end{cases} \text{ ....} (4)$$

The term θg is the correlation between two predictor variables, g, and θ. Using the coding given above, the parameter τ_2 relates to group differences on the items, and θ_3 relates to the interaction between group membership and ability. Judging from the model parameters in equation (1) are:

$\tau_2 = \beta_{01} - \beta_{02}$ and $\tau_3 = \beta_{11} - \beta_{12}$

An item shows a uniform DIF if $\tau_2 \neq 0$, and $\tau_3 = 0$, shows non-uniform DIF if $\tau_3 \neq 0$ (although $\tau_2 = 0$).

According to Desjardins and Bulut (2018), in addition to individual comparisons for uniform DIF and non-uniform DIF, an omnibus test can also compare Model 0 with Model 2 when uniform and non-uniform DIF are considered simultaneous. The significant R-squared difference between Model 0 and Model 2 indicates that the investigated item is DIF, and a follow-up analysis is required to determine the type of DIF.

METHOD

We used to survey or non-experimental methods with a quantitative approach. Survey methods are used to explore a population's trend, behaviour, or opinion by examining a sample of the population described quantitatively. From this sample, the researcher generalizes or makes claims about the population (Creswell & Poth, 2016).

Participant

This study did not conduct a direct survey to obtain data, but we used survey data released by http://openpsychometrics.org/_rawdata/KIMS.zip. The data on this website is open access to be downloaded freely. Used by the public. This survey involved 602 respondents. Respondents were grouped into two based on gender, namely, 301 women and 301 men. Respondents were randomly selected for undergraduate students and outpatients with personality disorders in Kentucky USA. We did not get much information about where the survey was conducted because our data were obtained from a survey conducted by Baer et al. (2004).

Instrument

We used the online Kentucky Inventory of Mindfulness Skills (KIMS) instrument developed by Baer et al. (2004). The instrument consists of 39 items (Q1-Q39), construction items using a Likert scale with five responses. Each response is given a score indicating the frequency or rank of the respondent's responses. Score 1 = Never or very rarely true, 2 = Rarely true, 3 = Sometimes true, 4 = Often true, 5 = Very often or always true (0 = none selected). The results of the exploratory factor analysis and confirmatory factor analysis conducted by Baer et al. (2004) shows that the items make up four factors: Observing, Describing, Acting Consciously, and Receiving. Without Judgment is multidimensional. The items in the KIMS instrument meet factor validity, reliability with the test-retest method, and internal consistency.

Tools of Data Analysis

This study uses Regression Logistics as a tool to detect DIF. We use logistic regression to offer flexibility in both ordinal and nominal data with dichotomous and polytomous scores. The data analysis process was conducted using the Open-Source R Studio software version 4.1.3.

We did prove the validity of the constructs using Confirmatory Factor Analysis (CFA) to prove the impact of DIF on the validity of the constructs, as stated by Abedalaziz et al. (2018). The construct validity proven was convergent validity and discriminant validity (Campbell & Fiske, 1959; Jöreskog, 1969). The validity of our construct value of Average Variance Extracted (AVE). A construct was proven to meet convergent validity if AVE > 0.5 and was also proven to meet discriminant validity if $\sqrt{AVE} > 0{,}7$ (Fornell & Larcker, 1981; Ghazali, 2014).

FINDINGS AND DISCUSSION

Findings

Table 2. Logistic Regression DIF Statistic

Items	DIF Uniform		DIF Non-Uniform	
	Stat	P-value	Stat	P-value
Q1	0.000	1.000	0.000	1.000
Q2	0.000	1.000	0.000	1.000
Q3	0.413	0.521	0.221	0.639
Q4	0.023	0.880	0.506	0.477
Q5	0.276	0.600	0.014	0.905
Q6	0.142	0.706	0.171	0.679
Q7	0.386	0.535	0.081	0.776
Q8	4.096	***0.043***	0.000	1.000
Q9	1.464	0.226	0.000	1.000
Q10	2.935	0.087	0.000	1.000
Q11	4.052	***0.044***	0.000	1.000
Q12	0.000	1.000	0.000	1.000
Q13	0.921	0.337	0.012	0.914
Q14	0.001	0.971	0.094	0.760
Q15	0.002	0.961	0.229	0.632
Q16	2.634	0.105	0.000	1.000
Q17	1.464	0.226	0.000	1.000
Q18	0.226	0.635	0.663	0.416
Q19	0.554	0.457	1.223	0.269
Q20	0.659	0.417	8.960	**0.003**
Q21	2.955	0.086	0.000	1.000
Q22	5.239	***0.022***	0.000	1.000
Q23	4.151	***0.042***	0.000	0.999
Q24	0.004	0.950	0.027	0.869
Q25	1.347	0.246	0.002	0.967
Q26	2.700	0.100	0.224	0.636
Q27	0.454	0.500	0.344	0.558
Q28	1.004	0.316	0.000	1.000
Q29	0.023	0.881	0.124	0.725
Q30	0.617	0.432	3.664	0.056
Q31	1.964	0.161	2.686	0.101
Q32	0.164	0.686	0.312	0.576
Q33	0.000	1.000	0.000	1.000
Q34	3.766	0.052	0.000	0.987
Q35	0.007	0.931	0.089	0.765
Q36	4.078	***0.043***	0.000	0.997
Q37	2.296	0.130	2.137	0.144
Q38	5.207	***0.023***	0.000	1.000
Q39	5.239	***0.022***	0.000	1.000

Detection of uniform and non-uniform Differential Item Functions using Logistics regression method using R Study software shows the syntax. Matched items are those with P-Value > 0.5. This study used a non-cognitive instrument to measure the mindfulness construct, so no anchor items were provided.

Based on the Logistic regression DIF statistics in Table 2, there are six items, namely, items Q8, Q11, Q22, Q23, Q36, Q38 and Q39, that were detected as not fit on uniform DIF because they had P-Value > 0.05. In contrast, item 20 indicated not fit on non-uniform DIF. Visually, the items identified as unfit in the Uniform DIF and non-uniform identified by Logistic Regression can be seen in Figure 1 and Figure 2.

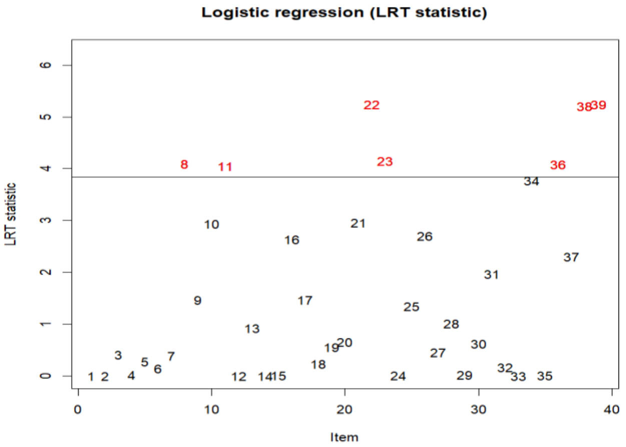

Figure 1. Item Fit on DIF Uniform Identified

Figure 1 shows the items above the DIF detection threshold (horizontal line) of 3,842 at a significance level of 5%. This gives us confidence that Q8, Q11, Q22, Q23, Q36, Q38 and Q39 are above the threshold, which means that the detected items are not fit.

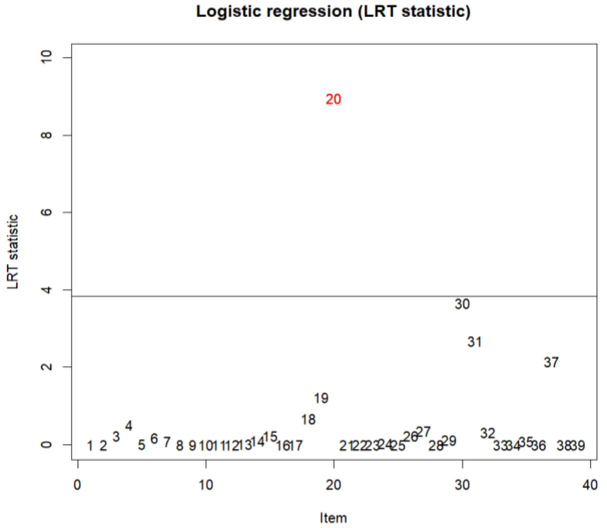

Figure 2. Item Fit on DIF Non-Uniform Identified

Figure 2 shows one above the DIF detection threshold (horizontal line) of 3.845 at a significance level of 5%. This gives us confidence that the Q20 items that are above the threshold are items that are detected as not fitting in the non-uniform DIF. Furthermore, the items identified as unfit for the uniform or non-uniform DIF are presented in a curve that shows the probability of the focal group (female respondents) being compared with the reference group (male respondents) to obtain a certain score on the same abilities.

The DIF Item Information Curve in Figure 3 shows Q8 and Q9 items, the probability of the focal group getting a score of 0 to 40 has a chance of 1, significantly different from the reference group, which has a probability of between 0.4 to 1.0 for Q8 items and about 0.27 to 1.00 for Q11 items. Furthermore, in items numbered Q22, Q38 and Q39, there is a clear difference in the ability of the focal group and the reference group. The focal group has a higher chance of getting a score of 10 to 40, while the reference group has a lower chance of getting a score between 10 to 30 and has a higher chance of getting a score of 10 to 30. high scores to get a score between 30 to 40. On the other hand, items Q23 and Q36 in the focal group have a small chance of getting a low score but a very high chance of getting a high score (30 to 40), while the reference group has a higher chance of getting a score between 20 to 30. Q20 items are items that do not fit in the non-uniform DIF. The focal group has the same opportunity to get low and high scores at all ability levels. However, the reference group only has the opportunity to get a score between 30 to 40 but has no chance of getting a score below 30.

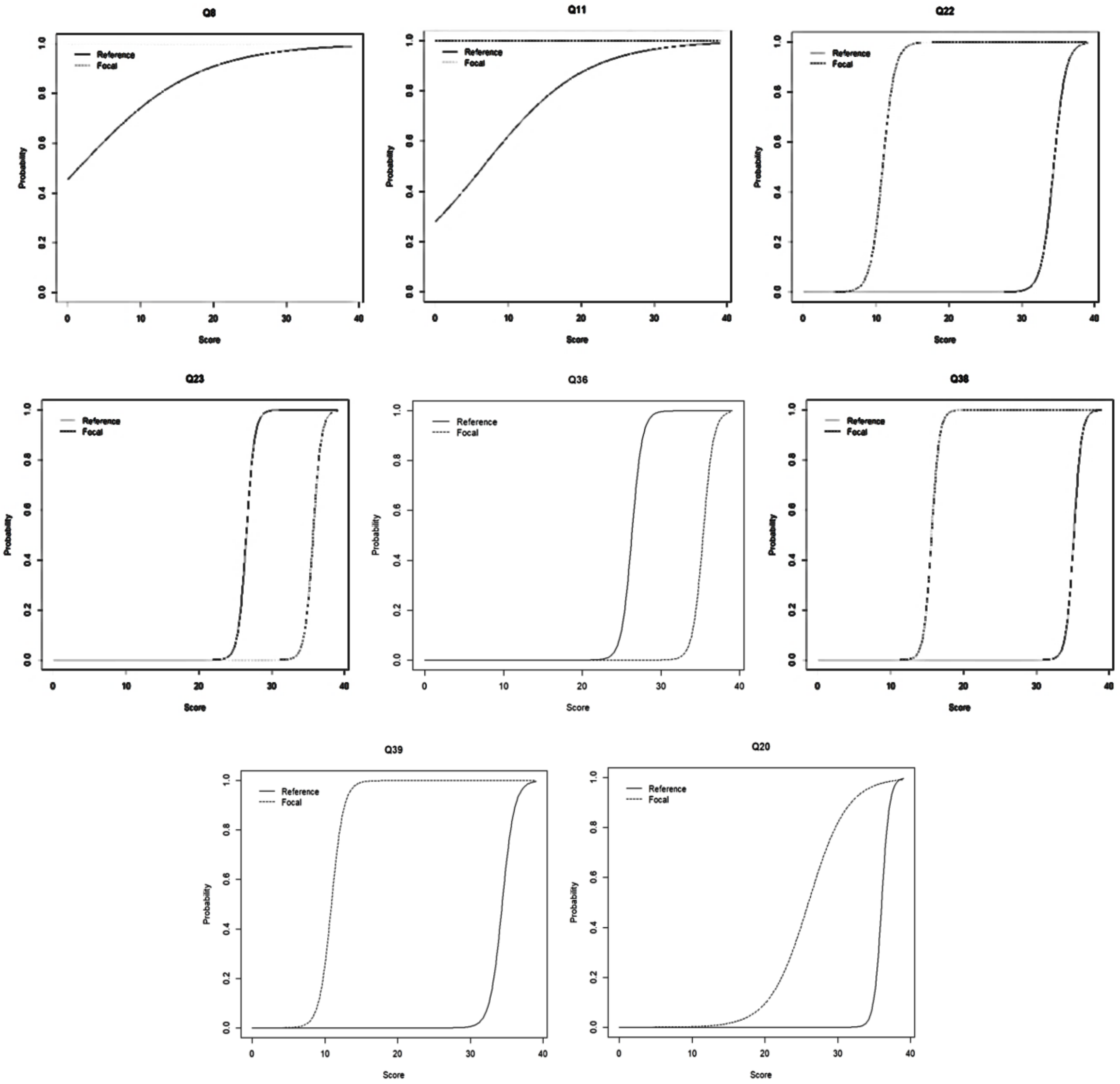

Figure 3. Information Item DIF

After the significance test for DIF was performed, the effect size, R2, was calculated using the extent by Zumbo (1999), with limitations for the categorization by Jodoin and Gierl (2001). Threshold value for effect size criteria Zumbo and Thomas (ZT); 0.000 – 0.130 = "A" (negligible effect), 0.131 – 0.260 = "B" (moderate effect), 0.261 – 1.000 = "C' (large effect), while the threshold value for the criteria of Jodoin and Gierl; 0.000-0.035 = "A" (negligible effect), 0.036 – 0.070 = "B" (moderate effect), and 0.071-1.000 = "C" (large effect). Through the R study program, we get the Zumbo and Thomas (ZT) and Jodoin and Gierl (JG) categories.

Several types of R^2 are available to calculate ΔR^2. In this study, Nagelkerke R^2 was used to measure the uniform DIF effect size, while the non-uniform DIF is calculated by subtracting R^2 (Zumbo, 1999). Items are inferred to display DIF if the DIF effect size is categorized as medium or large. An item is inferred to have no DIF when the DIF effect size is categorized as negligible.

Table 3. Effect size (Nagelkerke's R^2)

Items	**DIF Uniform**			**DIF Non-Uniform**		
	R^2	**ZT**	**JG**	**R^2**	**ZT**	**JG**
Q1	NaN**	-	-	NaN**	-	-
Q2	NaN**	-	-	NaN**	-	-
Q3	0.011	A	A	0.006	A	A
Q4	0.001	A	A	0.019	A	A
Q5	0.008	A	A	0.000	A	A
Q6	0.003	A	A	0.003	A	A
Q7	0.011	A	A	0.002	A	A
Q8	0.111	A	**C**	0.000	A	A
Q9	0.100	A	**C**	0.000	A	A
Q10	0.112	A	**C**	0.000	A	A
Q11	0.110	A	**C**	0.000	A	A
Q12	NaN	-	-	NaN	-	-
Q13	0.020	A	A	0.000	A	A
Q14	0.000	A	A	0.002	A	A
Q15	0.000	A	A	0.009	A	A
Q16	0.100	A	**C**	0.000	A	A
Q17	0.100	A	**C**	0.000	A	A
Q18	0.006	A	A	0.018	A	A
Q19	0.010	A	A	0.022	A	A
Q20	0.008	A	A	0.109	A	**C**
Q21	0.110	A	**C**	0.000	A	A
Q22	0.140	B	**C**	0.000	A	A
Q23	0.071	A	**C**	0.000	A	A
Q24	0.000	A	A	0.000	A	A
Q25	0.035	A	**B**	0.000	A	A
Q26	0.035	A	**B**	0.003	A	A
Q27	0.008	A	A	0.006	A	A
Q28	0.026	A	A	0.000	A	A
Q29	0.001	A	A	0.003	A	A
Q30	0.011	A	A	0.063	A	**B**
Q31	0.026	A	A	0.035	A	**B**
Q32	0.002	A	A	0.004	A	A
Q33	0.000	A	A	0.000	A	A
Q34	0.077	A	**C**	0.000	A	A
Q35	0.000	A	A	0.001	A	A
Q36	0.070	A	**C**	0.000	A	A
Q37	0.049	A	**B**	0.045	A	**B**
Q38	0.194	B	**C**	0.000	A	A
Q39	0.140	B	**C**	0.000	A	A

**NaN = Not A Number

Based on effect size (Nagelkerke's R2), there are thirteen Uniform DIF items with large effect sizes according to Jodoin and Gierl (JG) criteria: items Q8, Q9, Q10, Q11, Q16, Q17, Q21, Q22, Q23, Q34, Q36, Q38, and Q39. Items Q25, Q26, and Q37 have small effect sizes. There are four non-uniform DIF items based on the effect size criteria in Table 3; Q20 items have a large effect size, and items Q30, Q31, and Q37 have a medium effect size. Specifically, the statement of each item that contains DIF is shown in Table 4.

Table 4. Description of DIF Identified Items

Items Number	Item Statements	DIF Type
Q8	“I tend to evaluate whether my perception is right or wrong.”	Uniform
Q9	“When I walk, I notice the sensation of my body moving.”	Uniform
Q10	“I am good at coming up with words to express my perceptions, such as how something tastes, smells, or sounds.”	Uniform
Q11	“I was driving on autopilot without paying attention to what I was doing.”	Uniform
Q16	"I believe some of my thoughts are abnormal or bad and I shouldn't think that way."	Uniform
Q17	“I pay attention to how food and drink affect my thoughts, body sensations, and emotions.”	Uniform
Q20	“I make judgments about whether my thoughts are good or bad.”	Non-Uniform
Q21	“I notice sensations, like the wind in my hair or the sun on my face.”	Uniform
Q22	"When I have a sensation in my body, it's hard for me to describe it because I can't find the right words."	Uniform
Q23	"I don't pay attention to what I'm doing because I'm daydreaming, worried, or distracted."	Uniform
Q25	“I pay attention to sounds, like the ticking of a clock, birds chirping, or cars passing by.”	Uniform
Q26	“Even when I feel really upset, I can find a way to put it into words.”	Uniform
Q30	“I purposely stayed aware of my feelings.”	Non-Uniform
Q31	“I tend to do several things at once rather than focusing on one thing at a time.”	Non-Uniform
Q34	“My natural tendency is to put my experiences into words.”	Uniform
Q36	“I disagree with myself when I have irrational ideas.”	Uniform
Q37	“I pay attention to how my emotions affect my thoughts and behaviour.”	Non-Uniform
Q38	"I'm really engrossed in what I'm doing, so all my attention is focused on it."	Uniform
Q39	“I noticed when my mood started to change.”	Uniform

The existence of a uniform DIF indicates that the difference in the probability that women get a specific score better or worse than men is the same at all levels of ability and is influenced by the ability of each gender to understand the sentences in the statements of each items. Meanwhile, non-uniform DIF means that the difference in the chances of women getting a specific score better or worse than men is not the same at all ability levels, which is caused by statements or items that only benefit certain genders. This can be seen from the statement of instrument items identified by DIF in Table 4. The items identified in the non-uniform DIF type tend to be liked by women because they tend to measure feelings and emotions, while the items identified in the uniform DIF type tend to measure the behaviour of both sexes (female and male), the difference in the chances of the two sexes giving a positive or negative response to this instrument tends to be the same.

Comparison of Evidence of Construct Validity Before and After DIF Detection

The results of the study of Baer et al. (2004) have identified through the EFA that the KIMS instrument constitutes four multidimensional factors. We performed repeated CFA to prove the validity of the constructs before and after amputating the 18 items identified as uniform and non-uniform DIF. The results of analysis using R Studio obtained the value of Average Variance Extracted (AVE) as follows.

Table 5. Proof of Contract Validity

Factor	Before Item DIF Deleted		After Item DIF Deleted	
	AVE	$\sqrt{AVE}$	AVE	$\sqrt{AVE}$
Observe	0.344	0.587	0.344	0.587
Describes	0.495	0.704	0.570	0.755
Acting with Consciousness	0.303	0.551	0.357	0.597
Accept without Judgment	0.573	0.757	0.592	0.769

AVE values less than the threshold of 0.5 obtained for the four constructs of mindfulness-based on the results of the second-order CFA using R Studio through the maximum likelihood estimation that we summarized in Table 5, while "Accepting without assessment" has an AVE greater than 0.50, indicating that the "Accept" factor without consideration is proven to meet convergent validity. This is done to get the AVE value before the DIF detection stage. The square root value of the AVE factors "Describes" and "Accept without Judgment" before the DIF item was deleted had a value above the threshold of 0.7, which indicates that these two factors have also been proven to meet discriminant validity. After the deletion of items containing DIF, there was an increase in both convergent and discriminant validity, although not significant. Before deleting the DIF item, only one factor met convergent validity. After deleting the "Describe" factor, it also meets convergent validity.

Discussion

Based on the logistic regression DIF statistics in Table 2 and Table 3, seven items are not significant or do not fit uniform DIF; Q8, Q11, Q22, Q23, Q36, Q38, and Q39 indicated uniform DIF, while item 20 indicated no fit for non-uniform DIF. After evaluating the effect size (Nagelkerke's R^2), there are thirteen Uniform DIF items with large effect sizes according to the criteria of Jodoin and Gierl (2001), namely, items Q8, Q9, Q10, Q11, Q16, Q17, Q21, Q22, Q23, Q34, Q36, Q38, and Q39. Items Q25, Q26, and Q37 have small effect sizes. There are four non-uniform DIF items based on effect size criteria; Q20 items have a large effect size, and items Q30, Q31, and Q37 have a medium effect size. According to Huang et al. (2022), items detected by DIF must be amputated (excluded) from the instrument because items containing DIF can give misleading conclusions. This is in line with Özdemir (2015) that:

> *Purification of items based on repeated deletion of DIF items minimizes inflation I and improve the accuracy of the results. The DIF element has been shown to amplify Type-I errors, resulting in many non- DIF items being misclassified as DIF.*

On the other hand, this study verify studies conducted by Saygin and Atwater (2021), Ozgümüs et al. (2020), and Cuevas and Cervantes (2012) that differences in education and gender can affect the results of the assessment. The items identified by the DIF that we present in Table 4 are items that can be perceived differently by male and female respondents or are considered more dominant in exploring certain gender feelings. These findings inspire psychometricians to pay attention to items that have the potential for gender or other demographic bias.

The use of logistic regression in detecting DIF on the Kentucky Inventory Mindfulness instrument succeeded in simultaneously detecting uniform and non-uniform DIF. The identified DIF uniforms show that the differences in the chances of women and men obtaining a particular score are the same at all levels of ability, influenced by the ability of each gender to understand the sentence in the statement of each item. Meanwhile, non-uniform DIF means that the difference in the chances of women and men getting a particular score is not the same at all levels of ability; it can be caused by statements or items that only benefit a specific gender.

Through confirmatory analysis, we have proved that DIF-free items have better construct validity; it is in line with Abedalaziz et al. (2018) that the existence of DIF poses a severe threat to the validity of the construct. However, in this study, the increase is not significant since many factors affect the validity of a measuring instrument, one of which is that based on our review of the study by Baer et al. (2004) as the initial developer of the KIMS instrument, some items with a negative loading factor are not discarded. They also set a loading factor threshold too low (0.4). However, we also did not reverify the validity factor. We did not amputate items with a negative factor loading or <0.5 since the aim of this study was only to detect DIF and prove the effect on construct validity (convergent and discriminant validity). The same study in the future needs to calibrate the detected items containing DIF and prove the validity of the factors, to ensure all items have a factor loading >0.5 and meet all the model accuracy criteria required in the CFA.

CONCLUSION

This study found that 19 items were detected using logistic regression, including 15 items detected as uniform DIF and four as non-uniform DIF. We have also proved that the constructs' validity increased, although not significantly, after the removal of DIF.

The study implications are that the findings of this study are expected to inspire counseling psychologists to be more careful in using rating scales or instruments. The validity and reliability measures are not strong enough to justify that all measuring instruments are correct. However, checking for item bias or differential items is also necessary. Functional ensures that each item of the scale or instrument can be understood by all demographic groups and does not only benefit certain demographic groups.

ACKNOWLEDGMENT

We express our sincere gratitude to all parties who have contributed to the writing of this paper, especially to the lecturers and friends of the research and educational evaluation doctoral program at Yogyakarta State University and other parties whom we cannot mention one by one here.

REFERENCES

Abedalaziz, N. (2010). Detecting gender related DIF using logistic regression and Mantel-Haenszel approaches. *Procedia-Social and Behavioral Sciences*, *7*, 406–413. https://doi.org/10.1016/j.sbspro.2010.10.055.

Abedalaziz, N., Leng, Â. C. H., & Alahmadi, Â. A. (2018). Detecting a gender-related differential item functioning using transformed item difficulty. *MOJES: Malaysian Online Journal of Educational Sciences*, *2*(1), 16–22. https://ejournal.um.edu.my/index.php/MOJES/article/view/12820.

Baer, R. A., Smith, G. T., & Allen, K. B. (2004). Assessment of mindfulness by self-report: The Kentucky inventory of mindfulness skills. *Assessment*, *11*(3), 191–206. https://doi.org/10.1177/1073191104268029.

Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford Publications.

Camilli, G., & Congdon, P. (1999). Application of a method of estimating DIF for polytomous test items. *Journal of Educational and Behavioral Statistics*, *24*(4), 323–341. https://doi.org/10.3102/10769986024004323.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81-105. https://doi.org/10.1037/h0046016.

Chen, H.-F., & Jin, K.-Y. (2018). Applying logistic regression to detect Differential Item Functioning in multidimensional data. *Frontiers in Psychology*, *9*, 1302. https://doi.org/10.3389/fpsyg.2018.01302.

Creswell, J. W., & Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications.

Cuevas, M., & Cervantes, V. H. (2012). Differential item functioning detection with logistic regression. *Psychologie et Mathématiques, 3*, 45-59. https://doi.org/10.4000/msh.12274.

Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using* R. CRC Press.

Fornell, C., & Larcker, D. F. (1981). *Structural equation models with unobservable variables and measurement error: Algebra and statistics*. Sage Publications.

Ghazali, I. (2014). *SEM metode alternatif dengan menggunakan Partial Least Squares (PLS)*. Universitas Diponegoro Semarang.

Huang, T. -W., Wu, P. -C., & Mok, M. M. C. (2022). Examination of gender-related differential item functioning through Poly-BW indices. *Frontiers in Psychology, 13*, 821459. https://doi.org/10.3389/fpsyg.2022.821459.

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

Jöreskog, K. G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *34*(2), 183–202. https://doi.org/10.1007/BF02289343.

Lee, S. (2017). Detecting differential item functioning using the logistic regression procedure in small samples. *Applied Psychological Measurement*, *41*(1), 30–43. https://doi.org/10.1177/0146621616668015.

Martinková, P., Drabinová, A., Liaw, Y.-L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE—Life Sciences Education*, *16*(2), rm2. https://doi.org/10.1187/cbe.16-10-0307.

Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement*, *32*(2), 131–144. https://doi.org/10.1111/j.1745-3984.1995.tb00459.x.

Özdemir, B. (2015). A comparison of IRT-based methods for examining differential item functioning in TIMSS 2011 mathematics subtest. *Procedia-Social and Behavioral Sciences*, *174*, 2075–2083. https://doi.org/10.1016/j.sbspro.2015.02.004.

Özgümüs, A., Rau, H. A., Trautmann, S. T., & König-Kersting, C. (2020). Gender Bias in the Evaluation of Teaching Materials. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.01074.

Saygin, P. O., & Atwater, A. (2021). Gender differences in leaving questions blank on high-stakes standardized tests. *Economics of Education Review*, *84*, 102162. https://doi.org/10.1016/j.econedurev.2021.102162.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, *91*(6), 1292-1306. https://doi.org/10.1037/0021-9010.91.6.1292.

Swaminathan, H., & Rogers, H. J. (1990). Detecting Differential Item Functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361–370. http://www.jstor.org/stable/1434855.

Wiberg, M. (2007). *Measuring and detecting differential item functioning in criterion-referenced licensing test: A theoretic comparison of methods*. Institutionen för beteendevetenskapliga mätningar, Umeå.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF)*. National Defense Headquarters.