International Journal of Assessment Tools in Education
2023, Vol. 10, No. 3, 434–453
https://doi.org/10.21449/ijate.1279152
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

Investigating the sources of differential item functioning: A sample critical thinking motivation scale

Fatma Betül Kurnaz[1,*], **Hüseyin Yıldız**[2]

[1]Karabuk University, Faculty of Letter, Department of Educational Sciences, Karabük, Türkiye
[2]Australian Council for Educational Research, Melbourne, Australia

ARTICLE HISTORY
Received: Apr. 07, 2023
Revised: July 17, 2023
Accepted: July 27, 2023

Keywords:
Differential item functioning,
Critical thinking,
Critical thinking motivation.

Abstract: Investigating the existence of items with differential item functioning (DIF) may provide more accurate comparisons of group differences in studies that aim to compare scores obtained in a test by groups with different characteristics. In the present study, a scale measuring critical thinking motivation that was adapted to the Turkish culture was applied to 817 participants, who were high school graduates, university students, and university graduates. The aim of the study was to examine whether the data collected from these participants had DIF or not. Hence, DIF analysis of the collected data was performed via the "lordif" function in the R "lordif" package. DIF was found to occur in twelve items, three of which were related to gender and nine to level of education. While it was revealed that the content of the items was the source of gender related DIF, the source of DIF related to level of education was found to be the language and expression of the items.

1. INTRODUCTION

The investigation of the psychometric properties of measurement tools measuring motivational factors related to such cognitive factors as success, intelligence, and critical thinking can facilitate understanding of the construct that is of cultural interest. In the development and adaptation of scales that measure affective factors related to cognitive features such as learning motivation, critical thinking dispositions, and beliefs about learning and knowing these cultural differences can provide important information to researchers developing or adapting these measurement tools. It is stated in the related literature that researchers need not only be well-informed, but also to provide explanations about the psychometric properties of developed or adapted measurement tools (Crocker & Algina, 1986). Adaptation of measurement tools developed in different cultures creates discussions on the problem that the construct expected to be measured via a measurement tool can show variations across cultures (Cole et al., 1993; Ferne & Rupp, 2007). Hence, it is stated in the relevant literature that the construct validity, item and test bias as well as cultural norms of measurement tools adapted to different cultures particularly need to be investigated when the aim is to make inter-cultural comparisons (Byrne et al., 2009).

*CONTACT: Fatma Betül Kurnaz betulkurnaz@gmail.com Karabuk University, Faculty of Letter, Department of Educational Sciences, Karabuk, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2023

As a result of the adaptation of a scale developed in one culture to another culture, experts may seek evidence that the original and adapted measurement forms ensure the equivalence of the construct measured, that the scale can reveal the difference between groups in a culture-independent manner, and that the effect of culture and language on the construct measured is reduced. For this reason, studies that provide evidence on how the results obtained from the application of the adapted scale represent the construct in the target culture gain importance. Such studies may require in-depth qualitative analyses of cultural characteristics as well as statistical evidence.

When a comparison needs to be made among the scores obtained from groups that have different characteristics, investigating the presence of differential item functioning (DIF) can enable more accurate comparisons regarding group differences (Galic et al., 2014). The present study examines the DIF and its sources in a scale measuring critical thinking motivation, which was developed in one culture and then adapted to the Turkish culture.

1.1. Critical Thinking and Motivation

Critical thinking as defined by Ennis (1996) is "reasonable, reflective thinking that is focused on deciding what to believe or do" (p. 166). French et al. (2014) also define critical thinking as "the conscious process a person does when s/he explores a situation or a problem from different perspectives" (p. 275). Critical thinking, therefore, enables an individual to solve a problem more effectively (Ennis, 1993) and also to produce more effective strategies when solving problems (Glevey, 2006), and thus facilitates lifelong learning skills (Halpern, 1998).

Ennis (1996) considers critical thinking dispositions as a component of critical thinking skills and emphasizes that critical thinking dispositions, such as "being open to alternatives", should be accepted as part of the critical thinking skill. There are views in the literature supporting that critical thinking dispositions are essential for the use of critical thinking skills (Baron, 1985; Dewey, 1930; Ennis, 1991; Facione & Facione, 1992; McPeck, 1991; Paul, 1990; Perkins et al., 1993). Furthermore, it is claimed that motivation to think critically contributes to the use of critical thinking skills (Garcia & Pintrich, 1992; Ingle, 2007; Valenzuala et al., 2011). As reported in the literature motivation related beliefs and behaviors of both males and females are influenced by cultural stereotyping of gender roles (Meece et al., 2006). Studies on feelings of success and motivation have also revealed that males attribute their successes to their abilities; on the other hand, females attribute not their successes, but their failures to their abilities (Bar Tal, 1978, Crandall et al., 1965; Frieze, 1975). There are also views reported in the literature that, in areas culturally associated with gender, females are more disposed to experience learned helplessness when compared to males (Eccles et al., 1983, Farmer & Vispoel, 1990). On the other hand, a number of research findings also indicate that these gender related differences are not behavioral but only emerge in causal relationships (Eccles et al., 1983, Kloosterman, 1990; Parsons et al., 1984). Hence, it is important to examine affective factors related to cognitive skills in order to understand these constructs and their cultural associations.

With respect to characteristics regarding critical thinking, such as sustaining a discussion on a topic or refuting certain views, it is stated that females display a more accommodationist approach than males do. However, females are reported to display more behaviors than those of males in critically evaluating their own class performance (Feingold, 1994; Ruble et al., 1993). While some studies on critical thinking report gender differences (King et al., 1990; Serin et al., 2010), some others report no gender differences (Ersözlü & Arslan, 2009; McLean & Miller, 2010). French et al. (2012) claim that before such evaluations regarding these kinds of differences are made, it is important to examine measures for any indications of DIF.

While Ernst and Monroe (2004) stated that education has a positive impact on developing critical thinking, Tsui (2000) investigated how campus culture develops critical thinking and

highlighted an increase in students' critical thinking skills and also dispositions in universities that support freedom of thinking and are run with a democratic understanding, whereas the condition in high school education where students are more passive and not made to engage actively in the learning process have a negative impact on the development of critical thinking. Taking such information into consideration together with other research findings and expert opinions, it can be said that the source of DIF in terms of the level of education variable could be language and expression.

Accordingly, in the present study, it has been considered that such a difference could emerge in tests measuring beliefs and perceptions related to cognitive skills such as critical thinking; thus, whether there were such gender and level of education related differences in the critical thinking, motivation test was investigated by means of DIF.

1.2. Differential Item Functioning

Differential item functioning emerges when individuals are at the same ability level but in different groups that have different probabilities of providing responses to items (Gierl et al., 1999). The concept of ability is defined in the Turkish Language Assosiation Updated Turkish Dictionary (n.d.) as an individual's attribute, capability, talent, or capacity to understand or to do. Based on this definition, it can be deduced that ability is more to do with the process of performing cognitive or psychomotor skills.

It may not be appropriate to use the concept of ability when defining DIF since when measuring affective features, the responses are based on individuals' self-reports, and there is no right or wrong behavior or response. Hence, as the scale in the present study measures an affective feature, the definition of DIF is operationalized as the differentiation in the response patterns given to some items by individuals at the same affective level but in different groups. Moreover, in the discussion on the findings obtained from DIF analyses, the concept critical thinking disposition level is used instead of ability level.

The presence of DIF in an item is believed to be a threat to construct validity (Jensen, 1980; Steinberg & Thissen, 2006). Thus, when DIF is found to be present in an item, it is recommended that the source should be investigated. This can be done by receiving expert opinions on the content of items with DIF in terms of, for example, conceptual or cultural features (Ateşok Deveci, 2008; Karakaya & Kutlu, 2012; Yıldırım & Büyüköztürk, 2018).

When studies on DIF and item bias in the related literature are examined, it is observed that DIF is mostly researched in tests measuring cognitive characteristics (e.g., French et al., 2014; Kurnaz & Kelecioğlu, 2008; Kurnaz Adıbatmaz & Yıldız, 2020; Maller, 2001; Stump et al., 2005; Yıldırım & Büyüköztürk, 2018), in national and international measurement tools (e.g., Altıntaş & Kutlu, 2019; Kalaycıoğlu & Kelecioğlu, 2011; Karakaya & Kutlu, 2012), and in studies on the development or adaptation of measurement tools (e.g., do Nascimento et al., 2021; Nielsen & Dammeyer, 2019). In recent years, the number of studies investigating DIF or item bias in measurement tools measuring affective characteristics (Gök et al., 2014; Garcia et al., 2021; Köse, 2015; Lau et al., 2020; Şengül Avşar & Emons, 2021; Usta, 2020) is becoming increasingly prevalent. It is believed that the present study will contribute to the literature in terms of DIF identification and the investigation of its sources based on data obtained from the administration of the measurement tool measuring an affective characteristic, namely critical thinking motivation.

In the Critical Thinking Motivation Scale used in the present study, the scores obtained from the items are evaluated with a mark ranging from 1 to 6: high scores indicate high critical thinking motivation levels. When DIF is found to be present in the items of the measurement tool, it is concluded that individuals at the same critical thinking motivation level but in different groups have a varying probability of providing responses to items. When this is the case, it is

important that the items be examined for any expression or content that may be causing DIF. The findings of the present study can be instructive for researchers in two ways: first, if there are words and expressions that have an informative effect during the development or adaptation stage of a measurement tool, a finding can be generated on the discussion of how these can be eliminated; second, findings can be generated on whether results obtained from measurement tools create a difference stemming from items across the groups in terms of male and female scores or by level of education. In the measurement of cognitive or affective features, comparisons by gender and level of education are highly common; hence, the present study was designed to take into consideration the variables of gender and level of education.

In the present study, the responses to the following research questions were sought:

1. Do the items in the Critical Thinking Motivation Scale include DIF based on gender and level of education?
2. If there are items with DIF in the Critical Thinking Motivation Scale, how can the source of DIF in these items be accounted for?

2. METHOD

2.1. Study Group

In the present study, data were collected from 1050 individuals residing in various provinces in Türkiye and examined for univariate and multivariate outliers, while some part of the data were removed from the dataset in order to meet the fundamental statistical assumptions.

In total, data from 817 individuals were utilized in the DIF analysis. The age mean of the study group was 22.02±2.8 years. Of the participants, 47.5% were female, while 52.5% were male. The study group characteristics are presented in Table 1.

Table 1. *Study group characteristics.*

	Variable	Number	Percentage
Gender	Female	429	47.5
	Male	388	52.5
Province	İstanbul	92	11.3
	Ankara	92	11.3
	Karabük	80	9.8
	Konya	77	9.4
	Kastamonu	45	5.5
	Ağrı	39	4.8
	Mersin	36	4.4
	Afyon	32	3.9
	Bursa	46	5.6
	Çankırı	48	5.9
	Gaziantep	42	5.1
	Hatay	40	4.9
	Samsun	31	3.8
	Sakarya	44	5.4
	Other	73	8.9
Level of education	High school graduate	109	13.3
	University student	547	67.0
	University graduate	161	19.7

The data were collected from individuals residing in different provinces, namely İstanbul (11.3%), Ankara (11.3%), Konya (9.4%), and Karabük (9.8%). The collection of data from individuals living in different provinces is believed to increase the generalizability of the findings. In consideration of the measurement tool features, it was decided that the participants needed to be at least a high school graduate, which was set as a criterion in data collection. The study group was comprised of individuals who were high school graduates (n= 109, 13.3%), university students (n= 547, 67.0%), and university graduates (n=161, 19.7%).

2.2. Data Collection Tools

In the present study, the Critical Thinking Motivation Scale (Valenzuala Nieto & Saiz, 2011) adapted to the Turkish culture by Dönmez and Kaya (2016) was utilized. The Scale consisted of five subfactors, namely expectancy, attainment, intrinsic/interest value, utility, and cost and 19 items and the highest and lowest scores that could be obtained from the Scale were 114 and 19, respectively. The participants were expected to mark one of the six degrees of agreement in the Likert scale that they found most appropriate: (1 = "Strongly disagree", 6 = "Strongly agree"), while the Scale did not have any items that required inverse marking.

The items in the Scale aimed to measure the participants' expectations regarding critical and conscientious thinking (expectation) and the meaning they attributed to such thinking (value). The higher the total score obtained from the Scale was, the higher the participant's critical thinking disposition (that is critical thinking expectation and value) was interpreted to be; conversely, the lower the total score of the participant was, the lower the participant's critical thinking disposition (i.e. critical thinking expectation and value) was interpreted to be.

The scale was administered to 312 university students during its adaptation to the Turkish culture. The data collected from these participants were analyzed and the analysis results showed that all 19 items were categorized into five factors with eigenvalues values higher than 1 and they accounted for 67.91% of the total variance. The χ^2/df fit index value of the confirmatory factor analysis was 1.53. The NFI, CFI, and RMSEA were found to be 0.85, 0.94, and 0.58, respectively. Cronbach alpha coefficient of the scale was calculated between .73 and .85 for sub-dimensions and total score. These findings suggest that the research is valid at an acceptable level.

2.3. Data Collection and Analysis

The study was reported to be ethically appropriate in terms of ethical principles by the Karabük University Social and Human Sciences Research Ethics Committee (Decision number: E-78977401-050.02.04-49379). The items in the data collection tool and the questions found essential in the personal information form were used to develop an electronic Google form. This online form was sent to the participants, who voluntarily participated in the study.

The data were collected with the assistance of Karabük University students volunteering to contribute to the study. These students were asked to send the data collection form via Google forms to university students or high school graduates they knew. The collection of data via Google forms prevented the loss of data in the data set. Data were collected from 1050 individuals living in different provinces in Türkiye. However, during the stage of testing the fundamental assumptions, 233 data were removed from the data set after checking for the univariate and multivariate outliers.

It is recommended in the literature on scale adaptation that data obtained from the scale adapted should be checked for reliability in all the studies in which the scale is used. Hence, to check the reliability of the data obtained in the present study, the Cronbach alpha coefficient was calculated, and the reliability was found to be 0.88. The internal consistency of the sub-dimensions ranged between .76 and .80.

Prior to DIF analyses, unidimensionality and the normal distribution of the data were examined. To examine whether the data obtained from the measurement tool met the normality assumption, the skewness and kurtosis values were used. In the distribution, the skewness and kurtosis values were found to be 0.252 and -0.571, respectively; the standard error of skewness was calculated to be 0.086 and the standard error of kurtosis was 0.171. These values indicate that the distribution met the normality assumption (Büyüköztürk, 2021).

To examine the unidimensional outlier values in the distribution, *Z* standard scores were calculated for each item. All the items had *Z* standard scores ranging between 1.019 and -4.93. The unidimensional outliers in the distribution were eliminated, and after each outlier value was removed, the *Z* standard scores were recalculated for all the items and for all the participants. In the final data, the *Z* standard scores were found to range between 1.44 and -3.95. When the sample size is large, *Z* standard score that is ±3 is an expected condition. When this is the case, it is more appropriate to interpret the *Z* standard scores together with the mean, standard deviation, and the lowest and highest values (Tabachnick & Fidell, 2007).

On the other hand, multidimensional outliers were compared with the Mahalonobis distances (α=.001) and the critical chi-square value for K-1 degrees of freedom for all the items in the test of all the participants. The Mahalonobis distances ranged between the values of 1.11 and 113.6. At this stage, the data that showed deviation higher than the critical chi-square value was removed from the data set; subsequently, the *Z* score distributions and the Mahalonobis distances were reexamined. In the final data (N=817), the Mahalonobis distances were found to range between 42.2 and 1.72. The critical chi-square value for 18 degrees of freedom was 42.31. As there was no critical chi-square value exceeding the Mahalonobis distance, it could be concluded that there were no multidimensional outlier values in the data distribution (Mertler & Vannatta, 2005). On the other hand, kurtosis and skewness values were also calculated for all the items in order to check the multidimensional normality assumption, and these values were found to be between -1.2 and 0.89. It can therefore be said that each item is normally distributed separately and together.

After the removal of the unidimensional and multidimensional outlier values, which is essential for the administration of parametric tests that have multivariate data, other assumptions were tested. Exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were performed with the data in order to examine unidimensionality in the distribution. The scatter plot obtained and the factors, the eigenvalues of the factors, and their contribution to the total variance are presented in Figure 1 and Table 2, respectively.

Figure 1. *Scatter plot.*

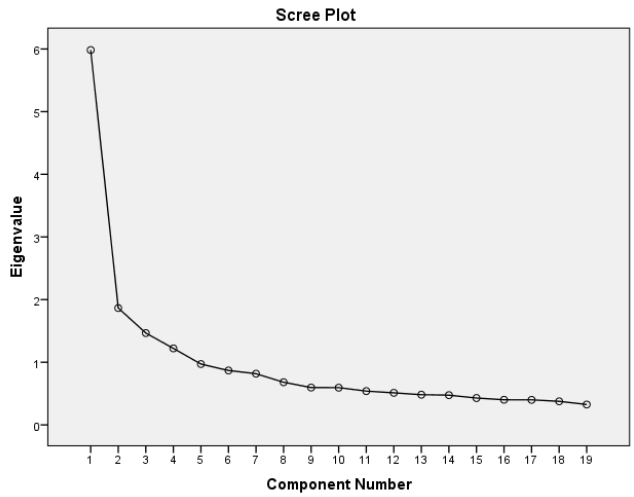


Table 2. *Factor eigenvalues and their contribution to the total variance based on the EFA results.*

Factor	Eigenvalue	Explained total variance
1	5.983	31.490
2	1.865	9.817
3	1.466	7.713
4	1.221	6.426
5	.971	5.111

It can be observed in the scatter plot in Figure 1 that while there is an abrupt fall after the slope of the first factor, the slope for the second and third factors has formed a plateau. The EFA results in Table 2 show that the difference between the eigenvalue of the first factor and the eigenvalue of the second value is higher than the differences between the eigenvalues of the other factors; the contribution of the first factor to the total variance is higher than the contribution of the other factors, which indicates that the unidimensionality assumption is met (Hambleton & Swaminathan, 1989). Meeting the unidimensionality assumption provides evidence for having met the local independence assumption (Hambleton et al., 1991).

The validity of the Critical Thinking Motivation Scale was checked with CFA for the sample group in this study and it was concluded that the measurement tool produced valid results (χ^2=590,415; χ^2/df=4,2; CFI=.92; GFI=.93; AGFI=.91; RMR=.045; NFI=.89; RMSEA=.06). Based on the evidence obtained as a result of the assumption checking analyses, it was concluded that the fundamental assumptions were met. The DIF analyses of the data collected were performed via the '*lordif*' function in the *R* '*lordif*' package (Choi et al., 2011). The lordif package was used because when DIF is identified in items that are scored across multiple categories and when there are more than two group variables, one of these variables is included in the model as a set of puppet variables. During the analysis process, the Generalized Partial Credit Model from the Item Response Theory was used (Muraki, 1992). In the Generalized Partial Credit Model, when DIF analysis is performed in the items to which weighted scoring is applied, the discriminatory parameters are also included in the model.

A form was developed to obtain expert opinions regarding the sources of DIF found to exist in some of the items; opinions were obtained from five measurement and evaluation experts, a child development expert who had worked on critical thinking, and a sociologist who had studied social classes and sexism. In the expert opinion form, an explanation of DIF was provided, the items with DIF and which groups these items favored were stated, and their opinions were asked about what the sources of the DIF could be. The results were interpreted and discussed based on these expert opinions.

3. RESULTS

The present study initially investigated whether there were items with DIF in the Critical Thinking Motivation Scale based on gender and level of education and then examined the sources of the items having DIF based on experts' opinions. Hence, this section is presented under two subtitles, which report results obtained from the analysis of DIF and results obtained from expert opinions regarding the sources of DIF.

3.1. The DIF Analysis Results

Whether or not the items displayed DIF based on gender and level of education was examined in the study and the results obtained are presented in Table 3.

Table 3. *DIF results based on the variables of gender and level of education.*

Item Number	Variable			
	Gender		Level of Education	
	Uniform (p)	Non-uniform (p)	Uniform (c^2)	Non-uniform (c^2)
1	0.065	0.422	0.017	**0.004***
2	0.934	0.962	**0.005***	0.651
3	0.011	0.105	**0.003***	0.575
4	0.325	0.863	**0.001***	0.130
5	**0.005***	0.021	0.692	0.447
6	0.034	0.357	0.313	0.515
7	**0.006***	0.820	0.646	0.851
8	0.199	0.379	0.472	0.748
9	0.746	0.234	**0.006***	0.025
10	0.237	0.311	**0.001***	0.557
11	0.932	0.745	0.057	0.682
12	0.502	0.844	**0.002***	0.104
13	0.308	0.986	**0.008***	0.980
14	0.071	0.562	**0.008***	0.308
15	**0.006***	0.929	0.790	0.669
16	0.021	0.288	0.090	0.773
17	0.084	0.150	0.651	0.064
18	0.197	0.038	0.612	0.415
19	0.887	0.720	0.618	0.661

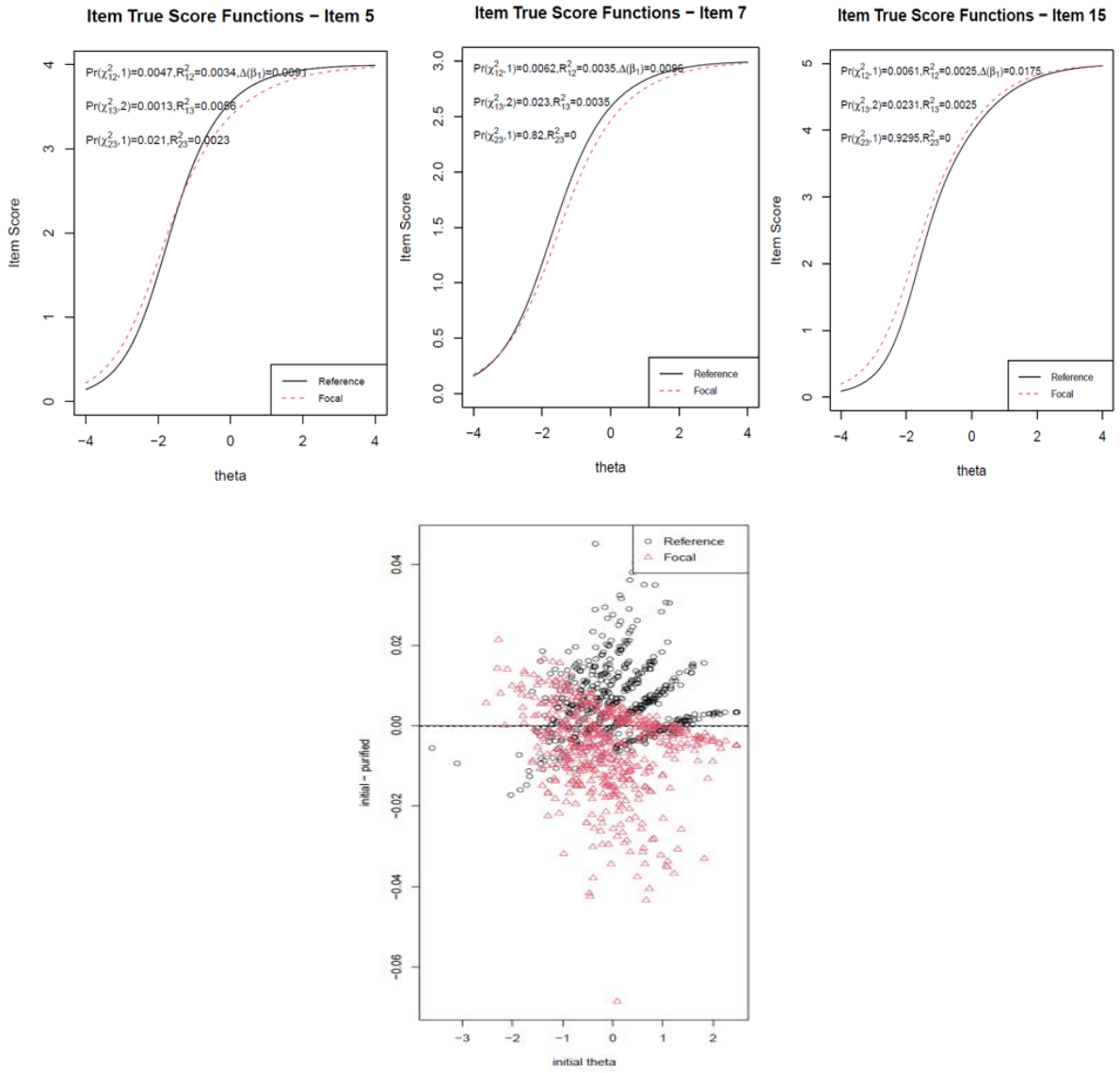
* Identification of DIF at .01 significance level

In the present study, the analyses based on the gender variable yielded three items with DIF, namely Items 5, 7, and 15. On the other hand, the analyses based on level of education yielded a total of nine items with DIF, namely Items 1, 2, 3, 4, 9, 10, 12, 13, and 14.

The scatter plot of the difference between the test characteristic curves of the items identified with DIF based on the gender variable and the predicted levels of critical thinking dispositions after the items with DIF were removed from the test is displayed in Figure 2.

When the test characteristic curves of the items with DIF are examined in Figure 2, it can be revealed that the items displaying DIF based on gender were in favor of female participants with low levels of critical thinking dispositions. In the scatter plot depicting the differences among the predicted critical thinking dispositions levels after items with DIF were removed from the test, the values on the y axis represent the difference between the predicted critical thinking disposition levels obtained from the entire scale and the critical thinking disposition levels after the items with DIF were removed from the test. It can be stated that individuals with a positive value on the vertical axis were influenced negatively from the items with DIF, while those with a negative value were positively influenced by the items with DIF. Accordingly, it can be said that items with DIF generally functioned in favor of female participants.

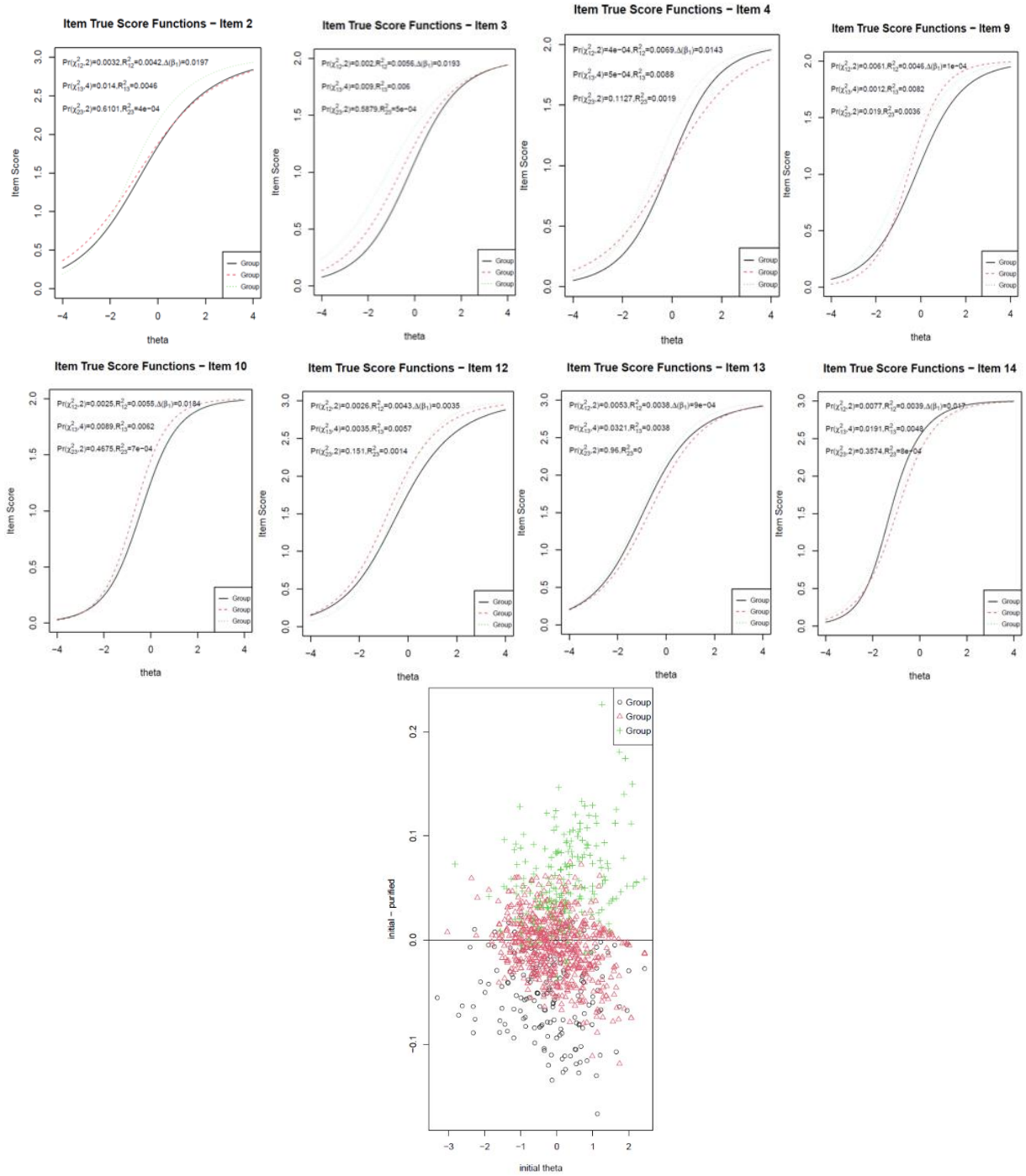
Figure 2. *The scatter plot of the difference between the test characteristic curves of the items identified with DIF based on the gender variable and the predicted levels of critical thinking dispositions after the items with DIF were removed from the test.*



The scatter plot of the difference between the test characteristic curves of the items identified with DIF based on the level of education variable and the predicted levels of critical thinking dispositions after the items with DIF were removed from the test is displayed in Figure 3. In the graphs, Group 1 represents individuals who are high school graduates and work in a job; Group 2 represents the university students; and Group 3 represents university graduates who have an occupation.

When the test characteristic curves of items identified as having DIF are examined in Figure 3, it can be observed that Items 1, 2, 3, 4, and 13 function in favor of university graduates, Items 9, 10 and 12 function in favor of university students, and Item 14 functions in favor of high school graduates. One other finding that was obtained was that results varied in items at low and high critical thinking disposition levels. It was generally observed that in items with DIF, the difference between university graduates and high school graduates was larger. Since the identification of the sources of DIF in the related items may provide important information to researchers who develop or adapt scales, the content of the items with DIF examined by the experts and the results obtained are provided in Figure 3.

Figure 3. *The scatterplot of the difference between the test characteristic curves of the items identified with DIF based on the level of education variable and the predicted levels of critical thinking dispositions after the items with DIF were removed from the test.*



3.2. Results on Expert Opinions on DIF Resources

The items identified as having DIF were examined in terms of item bias based on expert opinion. The experts were asked whether the items with DIF based on gender/level of education constituted a source of bias. The items with DIF by gender and level of education are presented in Table 4.

It was revealed that expert opinions had two foci as regards the source of DIF in items with DIF based on the gender factor. The first was that the ways of expression in some items in the measurement tool (e.g., reasoning correctly) could lead to DIF. Second, in Items 5, 7 and 12,

expressions such as "…learning is important", "For me it is important to use my intellectual skills", …I like to think" could have increased women's inclination to provide "the response expected by the environment".

Table 4. *Items identified to have DIF.*

Variable	Item number	Sub- factor	Item with DIF	Group
Gender	5	Attainment	For me it is important to learn how to reason correctly.	In favor of women
	7	Attainment	For me it is important to use my intellectual skills.	
	15	Utility	I like to think critically.	
Level of education	1	Expectancy	Concerning reasoning correctly, I am better than most of my peers.	In favor of university graduates
	2	Expectancy	I am capable of understanding everything related to thinking in a rigorous way.	
	3	Expectancy	I am able to learn how to think in a rigorous way.	
	4	Expectancy	I am able to learn how to reason correctly better than most of my peers.	
	13	Utility	I like to reason properly before deciding about something.	
	9	Intrinsic value/ interest	Thinking critically will help me to become a good professional.	In favor of university students
	10	Intrinsic value/ interest	Thinking critically will be useful for my future.	
	12	Intrinsic value/ interest	Thinking critically is useful for other subjects and courses.	
	14	Utility	I like to learn things that will improve my way of thinking.	In favor of high school graduates

In items with DIF based on the subfactor of level of education, it was revealed that expert opinions regarding sources of DIF had three foci. The first opinion was that the expressions of some of the items in the measurement tool (e.g., reasoning correctly, being a good professional, how to think in a rigorous way) could be the source of DIF. The second was that being a university student or being a university graduate could increase individuals' motivation to think critically. The third opinion was the probability of high school graduates' refraining from answering items with high scores when the content was based on such expressions as being better or being a professional. Such findings are addressed in the discussion section in detail with samples from expert opinions.

According to expert opinions, the formation of DIF in three items (Items 5, 7, and 15) based on gender can be attributed to the fact that women with low critical thinking dispositions have a high tendency to meet societal expectations. Below are direct quotations from experts' views regarding this issue:

> *"that women need to develop correct reasoning skills to free themselves from the secondary position they are in when compared to men is a social reality. That women who do not learn to reason correctly will be eliminated from the system faster than men has been*

engrained into women's mind as a cultural code. Conversely, the errors that men make in society or their incorrect reasonings are tolerated more when compared to those of women."

Expert A

"*Regarding this topic, the metaphor of "leaking pipe" explains this topic in more detail. According to this approach, as a result of the challenges women face, they are eliminated within the process. Women who do not want to be eliminated must learn to think more accurately. For women, critical thinking is an important step to move out of the patriarchal system they are a part of. It is by this means that they can question the system and can struggle to raise themselves to the position they 'desire/deserve'*."

Expert D

Moreover, based on these findings, it can be highlighted that the adaptation of a scale to a new culture does not merely consist of psychometric calculations, and thus examining the cultural background of the measurement tool being adapted is important. In terms of level of education, the experts were of the common opinion that being a university student, or a university graduate could increase their motivation to think critically. Direct quotations from experts' opinions on the topic are provided as follows:

"…it reveals that not only education, but the university environment is also influential in the development of critical thinking. By creating a learning environment where students are encouraged to participate in discussions and debates on social and political topics, it appears that a campus culture with social and political awareness is conducive to development of critical thinking skills. The factor underlying the fact that university graduates evaluate the item with a high score when compared to high school graduates at the same ability level is not only about level of education but also the learning environment and the campus culture, which should not be disregarded."

Expert A

"Thus, it could be that university graduates felt a higher need for thinking skills and the need to think. It is known that when compared to other people, those with a high need to think are more realistic in terms of their self-predictions. And when I look at the items here it seems that people were asked to make predictions about their own performance regarding critical thinking. University graduates could be more conscious about this as well."

Expert B

"Reasoning correctly. "It doesn't look appropriate to the Turkish language structure to me. "Does it mean evaluating events accurately? How will inaccurate reasoning occur? These could stem from the unclarity of the expressions, from the university graduates' getting a different meaning from the item.

Expert C

4. DISCUSSION and CONCLUSION

The accuracy of the evaluation of the results obtained from the administration of a measurement tool depends on the aim of the measurement tool in subject and in its technical adequacy (Glover & Albers, 2007). When a measurement tool developed in one culture is adapted to another culture, the linguistic and cultural differences between the respondents can substantially threaten the validity and the psychometric properties of the measurement tool (Hambleton et al., 2004). When measurement tools are adapted, in some circumstances, words or phrases used in the developed and adapted tools do not convey the same meaning either linguistically or culturally. When such a condition is present, the equivalence of the original and the adapted form is distorted, and the validity of the adapted tool becomes questionable (Poortinga, 1989).

The items and item content of adapted scales are expected to accurately reflect the differences between subgroups in the target culture. Examination of changing item function or item bias is a common way to investigate such differences. If the response behavior for an item varies between two individuals from the same culture who have the same level of the measured feature, and if this creates variance against or in favor of one of the groups, then this can cause wrong decisions to be made in between-group comparisons. DIF can provide crucial information to test developers or adaptors to identify such conditions and to investigate the sources of DIF.

In a study by Gallos (1995) it is reported that there is a significant relationship between critical thinking and gender, and that the reason underlying this is a learning environment that is in favor of males; it is also stated that females have more doubts than males have about their abilities/talents and intellectual competences; when females encounter failure, they more often impose the causes of failure upon themselves, while males do so on external conditions; and it was revealed that females are less likely than men to initiate small learning groups and to participate in these; however, when they are encouraged to do so, they are as successful as males.

Taking into consideration experts' opinions as well as the findings reported in the study by Gallos (1995), the reason why the items with DIF in the present study that are in favor of females at the same level in terms of the feature measured could be related to cultural codes and gender based cultural experiences. The non-existence of DIF in the other levels of the measured feature – that DIF only existed in low levels – could be attributed to the fact that women at low levels regarding the measured feature could have changed the meaning they derived from the items or caused a social acceptance error.

In items measuring affective features, the individual reads the items, attributes meaning to them, and then selects the item found most appropriate. As in maximum success tests, there is no response that is the most accurate nor an expected response. The responses are based on what the respondent finds appropriate. Hence, when interpreting the item, the individual is expected to remain completely independent of social norms or social doctrines; however, this may not be an easy task for test implementers or evaluators in real life. In this case, when writing items, many elements, such as social doctrines, collective subconscious, and culture need to be taken into consideration, and the feature measured through items should be freed of these contexts. To illustrate, in the fifth item (For me, reasoning correctly is important), a female respondent who has a low level of the measured feature can be disposed to select 'strongly agree' in an item to meet societal expectations; that is because she accepts the society's expectations of her to provide the correct response. If individuals in different groups (e.g., men and women) who have the same level of the measured trait understand the item or the meaning they attribute to the item changes, it can be said that the item does not represent the construct to the same degree in these groups (Davidov et al., 2014; Millsap, 2012).

Schwartz and Meyer (2010) state that all research areas are influenced by cultural practices (e.g., language, traditions), cultural values (e.g., individual versus group), and cultural identity (e.g., allegiance to a particular group). At the outset, it is important to examine how the motivation for critical thinking differs in the cultural context between men and women, as well as from those with higher to those with lower levels of education. In this respect, it is important to examine the psychological and sociological contexts of test development or adaptation processes and to examine what the meanings attributed to the language used in the items mean for individuals in different groups. The development or adaptation of a measurement tool is an effort to find the best meaning to represent the measured construct.

Kholberg (1973) stated that the majority of female participants displayed a moral tendency to be a 'good child' in terms of the responses given to conflict entailing questions in the moral

development theory; Gilligan (1979) attributed this to cultural doctrines. In the phase of 'being a good child' in Kholberg's moral development theory, the individual tends to display behaviors accepted to be appropriate by the society in order to get others' approval. It would not be wrong to state that the stories that entail conflicts in Kholberg's theory requires critical thinking and critical evaluation. Hence, the results from Kholberg (1973) and Gilligan (1979) support the findings obtained from the present study.

When Tedesco was writing about her book titled *Women's Ways of Knowing* in 1991, she stated that women believed that language was not dependable, that they experienced difficulties in expressing their self-identity and preferred to remain silent, that women who possessed learned knowledge did not believe they could provide the correct responses, and that they would echo others' voices rather than express their own; she stated that apart from those whom they decide to be the same in terms of background, conditions or views, they were generally reluctant to share their inner world with others. Considering this, it can be stated that in items measuring females' affective features, there is an important cultural process, and that this cultural process should be carefully examined when creating items in a test or scale.

When Jensen (1980) explained the relationship between culture, language, and test bias, s/he explained culture sterility of a test as 'distance from culture' and stated that when a measure tool is translated into another language, it will have a different content and the meaning attributed to the items may vary. Considering that the groups responding to the items are from different subcultures in terms of gender, level of education etc., it may be important in terms of the construct validity of measurement tools to be reconstructed so that items with DIF convey the same meaning to all the subgroups.

Hambelton and Rogers (1995) stated that to prevent items in a test from creating bias in favor of/against a prevalent culture or subcultures, the following questions need to be answered:

(1) Does the item include words that express different meanings to different sub sociocultural groups or words that are unfamiliar to those subgroups?
(2) Does the item include words that are difficult to understand?
(3) Does the item include words that are peculiar to a certain region or words that are not used frequently across the country?

When this information and the expert opinions in the present study are examined in combination, it can be concluded that there may be content that causes DIF in the language and expressions of the items. It is possible to state that an examination of the items with DIF revealed that university graduates, when compared to the other education level participants but with the same level of the feature being measured, had more often marked the options that yielded higher scores in items such as "…*I am better than most of my peers*", "…*I find myself proficient*", and "*I like to reason before I decide about something*." As for the university students, they more often marked the higher end of the Likert scale in items when compared to the other participants with the same level of the feature being measured in items such as "…*it will help me become a good professional*", and "…*it will be helpful for my future*." On the other hand, high school graduates, when compared to the other participants with the same level of features being measured, seemed to mark the 'strongly agree' option more often in the item that read '*I like to learn things that will improve my way of thinking'*. The respondents' item response behaviors seem to be related to how they perceive themselves based on their level of education and what they expect from themselves based on their social status. This could indicate that when the content of items is interpreted, individuals create meaning based on their social status and what is expected of them; this can create a difference in the scores of individuals in different groups but with the same level of critical thinking disposition.

Lau et al. (2023) administered a scale measuring gelotophobia, gelotophilia, and catagelasticism to university students in Taiwan and Canada. The Canadian English version was adapted from the German version. English version was then adapted into Taiwanese Chinese. While there were no items with DIF in the data obtained from the Canada sample, five items with DIF were found in the Taiwan sample. Only one of these items had a significant level. Then, in the data collected from Canada, DIF was calculated for the subgroups defined as Chinese living in Canada and answering the English form. In the data obtained from the English form, it was determined that there was no DIF for this subgroup and the reason for the DIF in the item was explained by the meaning changes in the words during the translation process. These results obtained from the study of Lau et al. (2023) confirm the argument of this study. In the adapted tests, it can be said that the translation processes and the meanings of the items affect the power to represent the construct.

Osterlind (1983) and Jensen (1980) highlighted that DIF in items or item bias can be caused by external factors such as culture and environment. Accordingly, based on the results of the present study, it can be valid to say that there may be external bias causing DIF, but the language and expressions in the measurement tool also increase the probability of DIF in the related items.

Considering the results of the present study, it can be said that validity evidence based solely on translation processes and psychometric computations of the measurement tools adapted to the Turkish culture may not be sufficient. In data obtained from the administration of developed or adapted measurement tools, investigating DIF can also yield significant evidence regarding the validity of a scale. Furthermore, it can be argued that it is important to examine the nature of the impact of how items are understood in the sub cultural groups by receiving opinions of experts in such areas as sociology and psychology.

One of the limitations of this research is that most of the data collected in this study were from university students. It is not known in which direction increasing the number of high school students and high school graduates would change the results. Since the research is based on individuals' self-report, it is assumed that the participants answered the items sincerely and accurately and that their reading comprehension skills were at a similar level. The evidence of reliability and validity in the study confirms these assumptions.

Researchers can examine DIF in tools measuring different affective features. In achievement tests and tests measuring affective features, respondent behaviors will show variation based on the structure of the feature being measured. Hence, in tools measuring affective features, investigating DIF can lead to different results. In addition, two different tools measuring critical thinking and critical thinking motivation can be administered to the same group, and the scores obtained from the achievement test can be used as an external criterion. In this way, findings based on the relationship between real performance and the affective feature related to the performance can be obtained.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Karabük University, 28/07/2021, 2021/07-05.

Authorship Contribution Statement

Fatma Betül Kurnaz: Investigation, Resources, Visualization, Software, Formal Analysis, and Writing-original draft. **Hüseyin Yıldız**: Methodology, Software, Formal Analysis.

Orcid

Fatma Betül KURNAZ https://orcid.org/0000-0002-7042-2159
Hüseyin YILDIZ https://orcid.org/0000-0003-2387-263X

REFERENCES

Altıntaş, Ö., & Kutlu, Ö. (2019). Investigating differential item functioning of Ankara University examination for foreign students by recursive partitioning analysis in the Rasch model. *International Journal of Assessment Tools in Education*, *6*(4), 602–616. https://doi.org/10.21449/ijate.554212

Ateşok Deveci, N. (2008). *Examination of Inter-university Board foreign language test in the frame of item bias*. [Doctoral dissertation, Ankara University]. https://tez.yok.gov.tr

Athman Ernst, J., & Monroe, M. (2004). The effects of environment-based education on students' critical thinking skills and disposition toward critical thinking. *Environmental Education Research, 10*(4), 507-522. https://doi.org/10.1080/1350462042000291038

Bar-Tal, D. (1978). Attributional analysis of achievement-related behavior. *Review of Educational Research, 48*(2), 259–271. https://doi.org/10.3102/00346543048002259

Baron, J. (1985). *Rationality and intelligence*. Cambridge University Press.

Büyüköztürk, Ş. (2021). *Sosyal bilimler için veri analizi el kitabı* [Data analysis handbook for social sciences]. PegemA.

Byrne, B.M., Oakland, T., Leong, F.T.L., Van De Vijver, F.J.R., Hambleton, R.K., Cheung, F. M., & Bartram, D. (2009). A critical analysis of cross-cultural research and testing practices: Implications for improved education and training in psychology. *Training and Education in Professional Psychology, 3*, 94-105. https://doi.org/10.1037/a0014516

Choi, S.W., Gibbons, L.E., & Crane, P.K. (2011). Lordif: An R package for detecting differential item functioning using iterative hybrid ordinal logistic regression/item response theory and Monte Carlo simulations. *Journal of Statistical Software, 39*(8), 1. https://doi.org/10.18637/jss.v039.i08

Cole, D.A., Maxwell, S.E., Avery, R., & Salas, E. (1993). Multivariate group comparisons of variable systems: MANOVA and structural equation modeling. *Psychological Bulletin, 114*, 174-184. https://doi.org/10.1037/0033-2909.114.1.174

Crandall, V.C., Katkovsky, W., & Crandall, V.J. (1965). Children's belief in their own control of reinforcement in intellectual-academic achievement situations. *Child Development, 36*, 91–109. https://doi.org/10.2307/1126783

Crane, P.K., Gibbons, L.E., Jolley, L., & Van Belle, G. (2006). Differential item functioning analysis with ordinal logistic regression techniques: DIF detect and difwithpar. *Medical Care, 44*(11), 115-123. https://www.jstor.org/stable/41219511

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Thomson Learning.

Davidov, E., Meuleman, B., Cieciuch, J., Schmidt, P., & Billiet, J. (2014). Measurement equivalence in cross-national research. *Annual Review of Sociology, 40*, 55-75. https://doi.org/10.1146/annurev-soc-071913-043137

Dewey, J. (1930). *Human nature and conduct*. The Modern Library.

do Nascimento, C.D., Peter, W.F., Ribeiro, I.M., Moreira, B.S., Lima, V.P., Kirkwood, R.N., & Bastone, A.C. (2021). Cross-cultural validity of the animated activity questionnaire for patients with hip and knee osteoarthritis: A comparison between the Netherlands and Brazil, *Brazilian Journal of Physical Therapy*, *25*(6):767-774. http://doi.org/10.1016/j.bjpt.2021.06.002

Dönmez, B., & Kaya, F. (2016). Eleştirel Düşünme Motivasyonu Ölçeği'nin Türkçe'ye uyarlanması [Turkish adaptation study of Critical Thinking Motivational Scale]. *HAYEF*

Journal of Education, 13-2(25), 159-173. https://dergipark.org.tr/tr/pub/iuhayefd/issue/24491/259590

Eccles, J.S., Adler, T.F., Futterman, R., Goff, S.B., Kaczala, C.M., & Meece, J.L. (1983). Expectancies, values and academic behaviors. In J.T. Spence (Ed.), *Achievement and Achievement Motives* (pp. 75–146). San Francisco Freeman.

Ennis, R.H. (1991). Critical thinking: A streamlined conception. *Teaching Philosophy, 14*, 5-25. http://doi.org/10.1057/9781137378057_2

Ennis, R.H. (1993). Critical thinking assessment. *Theory into Practice, 32*, 179-186. https://doi.org/10.1080/00405849309543594

Ennis, R.H. (1996). Critical thinking dispositions: Their nature and assessability. *Informal Logic, 18*(2), 165-182. https://doi.org/10.22329/il.v18i2.2378

Ersözlü, A. N., & Arslan, M. (2009). The effect of developing reflective thinking on metacognitional awareness at primary education level in Turkey. *Reflective Practice, 10*, 683–695. https://doi.org/10.1080/14623940903290752

Facione, P.A., & Facione, N.C. (1992). *The California critical thinking dispositions inventory*. California Academic Press.

Farmer, H.S., & Vispoel, W.P. (1990). Attributions of female and male adolescents for real-life failure experiences. *Journal of Experimental Education, 58*(2), 127–140. https://doi.org/10.1080/00220973.1990.10806529

Feingold, A. (1994). Gender differences in personality: A meta analysis. *Psychological Bulletin, 116*, 429-456. https://doi.org/10.1037/0033-2909.116.3.429

Ferne, T., & Rupp, A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly, 4*, 113-148. https://doi.org/10.1080/15434300701375923

French, B.F., Hand, B., Nam, J., Yen, H.J., & Vazquez, J.A.V. (2014). Detection of differential item functioning in the Cornell Critical Thinking Test across Korean and North American students. *Psychological Test and Assessment Modeling, 56*(3), 275.

French, B.F., Hand, B., Therrien, W.J., & Vazquez, J.A.V. (2012). Detection of sex differential item functioning in the Cornell Critical Thinking Test. *Europen Journal of Psychological Assessment, 28*(3), 201-207. http://doi.org/10.1027/1015-5759/a000127

Frieze, I.H. (1975). Women's expectations for and causal attributions of success and failure. In T. Mednick, S. Tangi, & L.W. Hoffman (Eds.), *Women and achievement. Social and motivational analysis.* (pp. 158–171). John Wiley and Sons.

Galic, Z., Scherer, K.T., & Leberton, J.M. (2014). Examining the measurement equivalence of the conditional reasoning test for aggression across U.S. and Croatian samples. *Psychological Test and Assessment Modeling, 56*, 195-216. http://darhiv.ffzg.unizg.hr/id/eprint/5547

Gallos, J.V. (1995). Gender and silence. *Collage Teaching, 43*(3), 101-105. http://doi.org/10.1080/87567555.1995.9925525

Garcia, J.M., Gallagher, M.W., O'Bryant, S.E., & Medina, L.D. (2021). Differential item functioning of the Beck Anxiety Inventory in a rural, multi-ethnic cohort. *Journal of Affective Disorders, 293*, 36-42. http://doi.org/10.1016/j.jad.2021.06.005

Garcia, T., & Pintrich, P.R. (1992). The effect of PBL curriculum on students' motivation and self-regulation. Paper presented at The Biennial Conference of The European Association for Research on Learning and Instruction, Italy.

Gierl, M., Khaliq, S.N., & Boughton, K. (1999). Gender differential item functioning in mathematics and science: Prevalence and policy implications. In Improving Large-Scale Assessment in Education Symposium at the Annual Meeting of the Canadian Society for the Study of Education, Canada.

Gilligan, C. (1979). Woman's place in man's life cycle. *Harvard Educational Review, 49*, 431-446. https://doi.org/10.17763/haer.49.4.h13657354113g463

Glevey, K.E. (2006). Promoting thinking skills in education. *London Review of Education, 4*(3), 291-302. http://doi.org/10.1080/14748460601044005

Glover, T.A., & Albers, C.A. (2007). Consideraitons for evaluating universal screening assessments. *Journal of School Psychology, 45*(2), 117-135. http://doi.org/10.1016/j.jsp.2006.05.005

Gök, B., Kabasakal, K.A., & Kelecioğlu, H. (2014). PISA 2009 öğrenci anketi tutum maddelerinin kültüre göre değişen madde fonksiyonu açısından incelenmesi [Analysis of attitude items in PISA 2009 student questionnaire in terms of differential item functioning based on culture]. *Journal of Measurement and Evaluation in Education and Psychology, 5*(1), 72-87. https://doi.org/10.21031/epod.64124

Halpern, D.F. (1998). Teaching critical thinking for transfer across domains. *American Psychologist, 53*, 449-455. https://doi.org/10.1037/0003-066X.53.4.449

Hambelton, R., & Rogers, J. (1995). Item bias review. *Practical Assessment, Research, and Evaluation, 4*(6), https://doi.org/10.7275/jymp-md73

Hambleton, R.K., & Rovinelli, R.J. (1986). Assessing the dimensionality of a set of test items. *Applied Psychological Measurement, 10*(3), 287-302. https://doi.org/10.1177/0146621686010003 07

Hambleton, R.K., & Swaminathan, H. (1989). *Item response theory: Principles and applications*. Kluwer Nijhoff.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Sage.

Ingle, C.O. (2007). *Predictors of critical thinking ability among college students*. [Doctoral dissertation]. Available from ProQuest Dissertations and Theses Database. (UMI No. 3263681).

Jensen, A.R. (1980). *Bias in mental testing*. Free Press.

Kalaycioğlu, D.B., & Kelecioğlu, H. (2011). Öğrenci Seçme Sınavı'nın madde yanlılığı açısından incelenmesi [Item Bias Analysis of the University Entrance Examination]. *Education and Science, 36*(161), 3-11.

Karakaya, İ., & Kutlu, Ö. (2012). Seviye Belirleme Sınavındaki Türkçe alt testlerinin madde yanlılığının incelenmesi [An Investigation of Item Bias in Turkish Sub Tests in Level Determination Exam]. *Education and Science, 37*(165). http://egitimvebilim.ted.org.tr/index.php/EB/article/view/1342

King, P.M., Wood, P.K., & Mines, R.A. (1990). Critical thinking among college and graduate students. *The Review of Higher Education, 13*, 167-186. http://doi.org/10.1353/rhe.1990.0026

Kholberg, L. (1973). Continuities and discontinuities in childhood and adult moral development revisited. In P.B. Baltes & L.R. Goutlet (Eds.), *Lifespan developmental psychology: Research and theory* (pp. 179-204). Academic Press.

Kloosterman, P. (2001). Attributions, performance following failure, and motivation in mathematics. In E. Fennema & G.C. Leder (Eds.), *Mathematics and gender*. Teachers College Press.

Köse, İ.A. (2015). PISA 2009 öğrenci anketi alt ölçeklerinde (Q32-Q33) bulunan maddelerin değişen madde fonksiyonu açısından incelenmesi [Investigation of items in PISA 2009 student questionnaire subscales (Q32-Q33) in terms of differential item functioning]. *Kastamonu Education Journal, 23*(1), 227-240. https://dergipark.org.tr/en/pub/kefdergi/issue/22600/241461

Kurnaz, F.B., & Kelecioğlu, H. (2008). Investigation of Peabody Picture Vocabulary Test from the point of item bias. *World Applied Sciences Journal, 3*(2), 231-239.

https://www.academia.edu/7282678/Investigation_of_Peabody_Picture_Vocabulary_Test_from_the_point_of_item_bias_peabody_picture_vocabulary_test

Kurnaz Adıbatmaz, F.B., & Yıldız, H. (2020). The Effects of distractors to differential item functioning in Peabody Picture Vocabulary Test. *Journal of Theoretical Educational Science, 13*(3), 530-547. https://dergipark.org.tr/tr/pub/akukeg/issue/54987/621581

Lau, C., Chiesi, F., Saklofske, D.H., Yan, G., & Li, C. (2020). How essential is the essential resilience scale? Differential item functioning of Chinese and English versions and criterion validity. *Personality and Individual Differences, 155*, 109666. http://doi.org/10.1016/j.paid.2019.109666

Lau, C., Swindall, T., Chiesi, F., Quilty, L.C., Chen, H.C., Chan, Y.C., ... & Torres-Marín, J. (2023). Cultural differences in how people deal with ridicule and laughter: Differential item functioning between the Taiwanese Chinese and Canadian English versions of the PhoPhiKat-45. *European Journal of Investigation in Health, Psychology and Education, 13*(2), 238-258. https://doi.org/10.3390/ejihpe13020019

Maller, S.J. (2001). Differential item functioning in the WISC-III: Item parameters for boys and girls in the national standardization sample. *Educational and Psychological Measurement, 61*(5), 793-817. https://doi.org/10.1177/00131640121971527

McLean, C.P., & Miller, N.A. (2010). Changes in critical thinking skills following a course on science and pseudoscience: A quasi-experimental study. *Teaching of Psychology, 37*, 85-90. https://doi.org/10.1080/00986281003626714

Mcpeck, J.E. (1990). Teaching critical thinking. Routledge.

Meece, J.L., Glienke, B.B., & Burg, S. (2006). Gender and motivation. *Journal of School Psychology, 44*(5), 351-373. https://doi.org/10.1016/j.jsp.2006.04.004

Mertler, C.A., Vannatta, R.A., & LaVenia, K.N. (2005). *Advanced and multivariate statistical methods: Practical application and interpretation*. Pyrczak. https://doi.org/10.4324/9781003047223

Millsap, R.E. (2012). *Statistical approaches to measurement invariance*. Routledge.

Muraki, E. (1992). *A generalized partial credıt model: Applicatıon of an em algorithm*. ETS Research Report Series. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Nielsen, T., & Dammeyer, J. (2019). Measuring higher education students' perceived stress: An IRT-based construct validity study of the PSS-10. *Studies in Educational Evaluation, 63*, 17-25. http://doi.org/10.1016/j.stueduc.2019.06.007

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

Parsons, J., Adler, T.F., & Kaczala, C.M. (1984). Socialization of achievement attitudes and beliefs: Parental influences. *Child Development, 53*, 322-339. https://doi.org/10.2307/1128973

Paul, R.W. (1990). *Critical thinking: What every person needs to survive in a rapidly changing world*. Center for Critical Thinking and Moral Critique, Sonoma State University.

Perkins, D.N., Jay, E., & Tishman, S. (1993). Beyond abilities: A dispositional theory of thinking. *Merrill-Palmer Quarterly, 39*(1), 1-21. https://www.jstor.org/stable/23087298

Poortinga, Y.H. (1989). Equivalence of cross-cultural data: An overview of basic issues. *International Journal of Psychology, 24*(6), 737-756. https://doi.org/10.1080/00207598908247842

Ruble, D., Greulich, F., Pomerantz, E.M., & Gochberg, B. (1993). The role of gender-related processes in the development of sex differences in self-evaluation and depression. *Journal of Affective Disorders, 29*(1), 97-128. https://doi.org/10.1016/0165-0327(93)90027-H

Serin, Q., Serin, N.B., Saracaloğlu, A.S., & Ceylan, A. (2010). The examination of critical thinking styles of university students (TRNC Sample). *Procedia Social and Behavioral Sciences, 9*, 864–868. https://doi.org/10.1016/j.sbspro.2010.12.250

Steinberg, L., & Thissen, D. (2006). Using effect sizes for research reporting: examples using item response theory to analyze differential item functioning. *Psychological Methods, 11*, 402-415. https://doi.org/10.1037/1082-989X.11.4.402

Stump, T.E., Monahan, P., & Mchorney, C.A. (2005). Differential item functioning in the short portable mental status questionnaire. *Research on Aging, 27*(3), 355-384. https://doi.org/10.1177/0164027504273784

Schwartz, S., & Meyer, I.H. (2010). Mental health disparities research: The impact of within and between group analyses on tests of social stress hypotheses. *Social Science & Medicine, 70*(8), 1111-1118. https://doi.org/10.1016/j.socscimed.2009.11.032

Şengül Avşar, A., & Emons, W.H.M. (2021). A cross-cultural comparison of non-cognitive outputs towards science between Turkish and Dutch students taking into account detected person misfit. *Studies in Educational Evaluation, 70*(101053). http://doi.org/10.1016/j.stueduc.2021.101053

Tabachnick, B.G., & Fidell, L.S. (2007). *Using multivariate statistics*. Pearson.

Tedesco, J. (1991). Women's ways of knowing/women's ways of composing. *Rhetoric Review, 9*(2), 246-256. http://doi.org/10.1080/07350199109388931

Tsui, L. (2000). Effects of campus culture on students' critical thinking. *The Review of Higher Education, 23*(4), 421-441. http://doi.org/10.1353/rhe.2000.0020

Turkish Language Assosiation (n.d.). Ability. In Updated Turkish Dictionary. Retrieved February 28, 2021. https://www.tdk.gov.tr/

Usta, H.G. (2020). Sınav kaygı ölçeği maddelerinin çeşitli yöntemlere göre değişen madde fonksiyonlarının incelenmesi [Analysis of Test Anxiety Scale items in terms of differential item functioning by different methods]. *Cumhuriyet International Journal of Education, 9*(4), 1225-1242. https://doi.org/10.30703/cije.703337

Valenzuela, J., Nieto, A.M., & Saiz, C. (2011). Critical thinking motivational scale: A contribution to the study of relationship between critical thinking and motivation. *Journal of Research in Educational Psychology, 9*(2), 823-848. http://repositorio.ual.es/bitstream/handle/10835/819/Art_24_588.pdf?sequence=1

Yıldırım, H., & Büyüköztürk, Ş. (2018). Using the Delphi Technique and focus-group interviews to determine item bias on the Mathematics Section of the Level Determination Exam for 2012. *Educational Sciences: Theory & Practice, 18*(2), 447-470. http://doi.org/10.12738/estp.2018.2.0317