Journal of Pedagogical Research
Volume 7, Issue 5, 2023
https://doi.org/10.33902/JPR.202323599

Research Article

Effects of gender norms on intelligence tests: Evidence from ASIS

Deniz Arslan[1], Ömer Faruk Tamul[2], Murat Doğan Şahin[3] and Ugur Sak[4]

[1]*Anadolu University, Faculty of Education, Eskisehir, Türkiye (ORCID: 0000-0003-2531-7883)*
[2]*Anadolu University, Faculty of Education, Eskisehir, Türkiye (ORCID: 0000-0001-8884-6171)*
[3]*Anadolu University, Faculty of Education, Eskisehir, Türkiye (ORCID: 0000-0002-2174-8443)*
[4]*Western University, Faculty of Education, Canada & Anadolu University, Faculty of Education, Eskisehir, Türkiye (ORCID: 0000-0001-6312-5239)*

An examination of gender-related differential item functioning was conducted on the verbal subtests of the Anadolu-Sak Intelligence Scale. Analyses were conducted using the scale standardization data (N = 4641). A Mantel-Haenszel statistic was used to detect differential item functioning (DIF). A total of 58 verbal analogical reasoning items, 20 verbal short-term memory items, and 70 vocabulary items were analyzed. Initially, items displaying DIF in different age groups were determined, and then experts were consulted to determine whether these items were biased. There were three items with item effects on the Verbal Analogical Reasoning subtest and five items on the Vocabulary subtest. Short-term Memory subtests did not reveal any bias. Several implications regarding cognitive development, gender perceptions, and cultural factors were discussed.

Keywords: ASIS; Verbal items; Differential item functioning; Turkish culture; Gender norms

Article History: Submitted 11 August 2023; Revised 4 November 2023; Published online 13 December 2023

1. Introduction

An intelligence scale should be capable of making the equal, objective, and accurate measurements in terms of culture, gender, and region. Items that make up a scale should not provide an advantage to any subgroups and not include biases against gender, socioeconomic status and regional, and the like. American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [NCME] (2014) recommends that scale scores should represent measured trait similarly across subgroups. When the item response differs between equal ability groups not only by trait but also by group membership, differential item functioning [DIF] occurs (Angoff, 1993). The presence of DIF shows lack of measurement invariance and interpreting the test results and making score comparisons across groups may not be reliable. Also DIF is critically important to the construct validity of the scale (Maller, 2001).

Anadolu-Sak Intelligence Scale (ASIS) is an individually administered intelligence test developed for 4-12 aged children. ASIS was standardized and normed in Türkiye in 2016 (Sak et al., 2016). Number of studies have been conducted to investigate the validity and reliability level of the Anadolu-Sak Intelligence Scale [ASIS] (e.g., Sak et al., 2019; Cirik et al., 2020; Sözel et al., 2018).

Address of Corresponding Author

Deniz Arslan, Anadolu University, Faculty of Education, 26210 Tepebaşı, Eskişehir, Türkiye.

✉ denizarslan@anadolu.edu.tr

How to cite: Arslan, D., Tamul, Ö. M., Şahin, M. D., & Sak, U. (2023). Effects of gender norms on intelligence tests: Evidence from ASIS. *Journal of Pedagogical Research, 7*(5), 374-384. https://doi.org/10.33902/JPR.202323599

ASIS's theoretical validity was confirmed using exploratory and confirmatory factor analysis (Sak et al., 2016). However, its item bias analyses including gender, ethnicity or regional bias have not been tested before. The purpose of the current study was to detecting DIF in the verbal subtests of the ASIS and examining the possible DIF situations in the context of Turkish culture. Because verbal abilities shaped by culture, verbal subtests were examined. The verbal subtests of the ASIS measure crystallized intelligence that includes skills acquired and developed through experiences, such as verbal comprehension, language development, and vocabulary knowledge and the acculturation process (Horn & Blankson, 2012). Since the ASIS was developed in Turkish Culture and its norms were composed of Turkish speaking children, item bias analyses were conducted concerning Turkish Culture.

1.1. Differential Item Functioning (DIF)

DIF is defined as the analysis that reveals systematic differences between the performances of the groups that respond to a test (Osterlind, 1983). DIF can also be expressed as the differentiation of the probability of individuals who are similar in characteristics measured by tests but belong to different groups terms in terms of ethnicity, socioeconomic status, and gender, and the like (Hambleton et al., 1991). In DIF analyses, individuals with the same level of skill are matched, and then are compared. The purpose of matching is to distinguish between bias and real differences between groups in measurements. Item bias indicates that the scale has reliability problems (Kristanjansonn et al., 2005). In addition, bias reduces its construct validity since it shows that another skill area that is incompatible with the structure of the test is also measured except for the skill that is measured (Camilli & Shepard, 1994). The real difference can be explained by experience or knowledge differences that a group has previously acquired about a subject in comparison with the other group. Real differences show that the measurement tool is not defective and indicate differences that originate from various reasons between subgroups. There are many explanations in the literature about the causes of DIF observed in tests. Among the possible sources of DIF, there are (1) differences in the areas of interest, (2) differences in the socioeconomic level, (3) differences in the level of familiarity with concepts, (4) the content of the item, (5) differences in educational status, and (6) differences between the regions of residence (Colom et al., 2004; Doolittle & Clearly, 1987; Kalaycıoglu & Berberoglu, 2010; Li et al., 2004).

DIF analyses are conducted in two steps. Firstly, the relevant item should show DIF as a result of the analyses. In the second step, the reasons for the item showing DIF are discussed, and experts should determine whether the advantage is provided to the relevant group (Camilli & Shepard, 1994). There are many methods to determine DIF: Mantel-Hanszel, Logistic Regression, SIBTEST, Raju Field Measurements, Likelihood Ratio Tests are some of the methods used in DIF analysis (Hambleton et al., 1991).

1.2. DIF in Intelligence Scales

In education and psychology, intelligence scales are used for many purposes, such as diagnosis, educational placement, and ranking, in addition to making decisions on the psychological and cognitive levels of individuals. One of the prerequisites for intelligence scales to be able to make valid and reliable measurements is that the items that make up the scale do not contain bias by subgroups (Nolan et al., 1989). Furthermore, in the creation of the Educational Testing Service [ETS] scale item, it was stated that items that could work for any gender group, difficult words that did not match with the purpose of the test, and items that were unrelated to the structure desired to be measured by the test should be avoided (ETS, 2009). Accordingly, DIF analyses, which may affect the validity of the measurement tool, are frequently performed in intelligence tests and achievement scales, and the causes of bias are examined.

When DIF analyses are examined, the gender variable is expressed as the most commonly used variable. Abad et al. (2004) investigated whether there was an item including gender-related bias in the Advanced Raven Matrices in their study using IRT based DIF methods, conducted with a total of 1970 university students, including 1069 men and 901 women. It was revealed that some

items showed bias in favor of men. The result was shown that the visual-spatial skills of men were more advanced than those of women. There are also studies in which bias analysis is conducted with the standardization samples of scales. Immekus and Maller (2009) conducted a gender-based DIF analysis using Mantel-Haenszel DIF procedure with the standardization sample data of 2000 people of the Kaufman Adolescent and Adult Intelligence Test [KAIT]. While there are items showing bias in favor of girls in subtests in which crystallized intelligence is measured, DIF has not been found in subtests in which fluid intelligence is measured. In his study using the standardization sample of the WISC-III (n = 2200), Maller (2001) investigated whether there was DIF in the items using the IRT Likelihood Ratio detection method. Bias was determined in approximately one-third of the items. The findings obtained by Immekus and Maller (2009) can be evaluated as compatible with the literature, and it can be said that the item biases obtained are caused by the item effect and do not affect the validity of the scale. However, the findings of Maller are noteworthy. The presence of bias in many items may originate from the fact that the item content is formed by words and concepts that can provide an advantage for a gender. Moreover, this result makes it difficult for the total scores to mean the same for women and men. It can be stated that the construct validity of the scale should be questioned. Thus, Immekus and Maller (2009) indicated that the IQ scores obtained from the scale did not have the same meaning for men and women. Therefore, it can be said that arrangements should be made in scoring. There are also DIF analysis studies conducted with the selected sample groups of commonly used intelligence tests. Wechsler et al., (2014) applied four verbal tests to 1191 participants in their studies, in which they investigated gender-based DIF in crystallized intelligence using Rasch model. No significant gender-related difference was obtained in the total scores. It was observed that women were more advantageous in the contents related to daily life, and men were more advantageous in verbal analogies.

There are also DIF analyses conducted using variables in addition to the gender variable. Nolan et al. (1989) investigated whether the K-ABC showed DIF according to the ethnicity and gender variables. While eight items against gifted white students were determined in the study, no item showing DIF was found for gifted black students. Bias was observed in four items for white students with the normal intelligence level and three items for black students. In the context of gender, negative bias was found in two items for gifted men and two items for gifted women. Furthermore, no bias was obtained in terms of gender and ethnicity for gifted or normal students in the total scores.

DIF analyses were also performed in studies conducted with special education groups. Maller (2000) examined whether the items in the four subtests of the UNIT constituted bias between deaf and normal students. No item bias was found in the study using the Mantel-Haenszel method. Murray et al. (2015) investigated items that showed DIF in terms of gender in the items on the Learning Disability Screening Questionnaire [LDSQ]. No significant gender-related DIF was obtained in the items at the end of the analysis.

It was seen that DIF studies are conducted for many intelligence scales and with many subgroups. In these studies, various DIF detecting methods have been used. DIF is examined in two forms, uniform and nonuniform DIF. Uniform DIF occurs when there is no interaction between ability level and group membership. Nonuniform DIF exists when there is interaction between these variables (Rogers & Swaminathan, 1993). ASIS norm data tested in this study do not include ability levels. Because the Mantel-Haenszel method is an effective method for determining uniform DIF (Rogers & Swaminathan, 1993), the MH method was used in the study.

DIF analyses have not been previously conducted for the ASIS. It is important to conduct DIF analyses for the ASIS, which may affect the validity of the scale. Since ASIS was developed in Turkish culture, it is aimed to evaluate the results of DIF analysis in the context of Turkish culture. Verbal subtests measure crystallized ability shaped by culture were used in the analysis.

2. Method

2.1. Participants

Participants included 4641 (2314 girls and 2327 boys) children in the norm study of ASIS. Children's age ranged from 4 to 12. The sample was divided into three age groups based on the ASIS subtests start points. While VAR and VOC subtests have start points at different age groups, VSM has no start point. The number of participants for each subtest and descriptive statistics of subtests are presented in Table 1.

Table 1
Descriptive statistics of VAR, VOC and VSTM subtests

VAR	*4-7 age*	*8-9 age*	*10-12 age*
N	2528	848	1265
Mean	5.34	17.57	26.21
SD	4.60	9.21	11.17
Skewness	1.56	0.63	0.3
Kurtosis	2.96	0.26	−0.53
KR-21	0.82	0.87	0.90
VOC	*4-7 age*	*8-10 age*	*11-12 age*
N	2528	1277	836
Mean	5.46	27.63	40.05
SD	7.33	12.77	14.56
Skewness	1.71	0.35	−0.29
Kurtosis	2.82	0.13	−0.53
KR-21	0.91	0.91	0.93
VSTM	*4-12 age*		
N	4641		
Mean	7.44		
SD	4.11		
Skewness	0.36		
Kurtosis	−0.59		
KR-21	0.76		

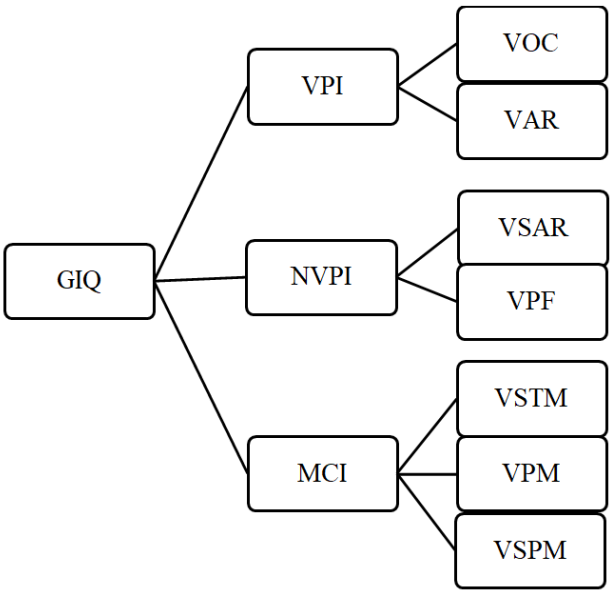
Note. *VAR = Verbal Analogical Reasoning, VOC = Vocabulary, VSTM = Verbal Short-term Memory.

2.2. Data Collection Tools

2.2.1. Anadolu-Sak Intelligence Scale [ASIS]

The ASIS is an individually administered intelligence test. It was developed and standardized in Türkiye (Sak et al., 2016). It consists of seven subtests, 256 items and three factors. ASIS provides three componential and general factor (*g*) score. Componential scores include visual-spatial reasoning, verbal ability and memory. Visual-spatial reasoning is an indicator of fluid intelligence, verbal ability measures crystallized intelligence (Carroll, 2005; Schneider & McGrew, 2012). In the structure of the ASIS General Intelligence Index [GIQ] is on the top. The second layer consist of Verbal Potential Index [VPI], Visual Potential Index [NVPI] and Memory Capacity Index [MCI]. The last layer includes the seven subtests: Visual Sequential Processing Memory [VSPM], Verbal Analogical Reasoning [VAR], Visual Perceptual Flexibility [VPF], Visual Spatial Analogical Reasoning [VSAR], Verbal Short-Term Memory [VSTM], Visual Pattern Memory [VPM] and Vocabulary [VOC]. The GIQ is composed of the seven subtests. The VPI sum of the VAR and VOC subtests. The NVPI is composed of VSAR and VPF subtests. The MCI includes VSPM, VSTM and VPM. VAR subtest measures verbal reasoning, verbal comprehension and crystalized ability. VOC subtest measures language development and vocabulary knowledge. The factor structure of ASIS is presented in Figure 1.

Figure 1
ASIS Factor Structure



Several studies were conducted for the validity and reliability analysis of the ASIS (Sak et al., 2019; Cırık et al., 2020; Sözel et al., 2018). Correlations between ASIS scores and the RIAS and the UNIT intelligence tests scores range from .50 to .82. ASIS scores correlate with students' grades in math, science, language and social studies ranged from .57 to .83 (Sak et al., 2019). Social validity of ASIS was assessed to be very high by test users (Tamul et al., 2020). The discriminant validity of ASIS was examined with clinical groups with autism spectrum disorder, learning disability, attention deficit hyperactivity disorder, giftedness and intellectual disability. All the groups were correctly classified (Cirik et al., 2020; Sözel et al., 2018). In addition, with intelligence-related constructs such as humor production ability (.82) (Arslan et al., 2021), scientific creativity (.55) and math ability (.77) (Köprü & Ayas, 2020) ASIS scores significantly correlate.

2.3. Data Collection and Analysis

Standardization data of the ASIS were used by permission of the Research and Practice Center for High-Ability Education at Anadolu University holding the copyright of the ASIS. The names of the participants were not included in the data set.

The three subtests were examined for DIF. The Mantel-Haenszel method was used to detect DIF in the ASIS subtests across boys and girls. In the first step of the analysis, the Mantel-Haenszel method was used to determine items with DIF. The Mantel-Haenszel method is based on χ^2 statistics. In this method, individuals in the groups determined as focus and reference are balanced according to their test performance (Agresti, 1984; Zieky, 1993). With the Mantel-Haenszel analysis, the DIF ratios of the items are evaluated in three categories:

- A-Level DIF, $|D| < 1$, there is no DIF, or it is at a negligible level,
- B Level DIF, $1 \leq |D| < 1.5$, there is a medium level of DIF,
- C Level DIF, $|D| \geq 1.5$, there is a high level of DIF (Zieky, 1993).

The bias analysis performed in the study was conducted for the first time for the ASIS. The VAR subtest of the ASIS was analyzed for 4-7, 8-9, and 10-12 age groups, and the VOC subtest was analyzed for 4-7, 8-10, and 10-12 age groups. Since the VSTM subtest did not include start point, it was analyzed by examining the responses given by the whole study group to the items. Items demonstrating DIF at levels B and C were determined for each age group, and expert opinion was obtained. A group consisting of seven experts (intelligence test developers, psychologists and linguists) assessed the DIF results.

3. Findings

3.1. DIF Results of the Subtests

The VAR, VOC and VSTM subtests was examined for DIF. The results are presented in Table 2. In VSTM subtest no item was identified for DIF.

Table 2
Results of DIF analysis of VAR and VOC subtests

VAR	*Level*	*Items favored boys*	*Items favored girls*
4-7 age	B	**12**, 15	**2,** 22
8-9 age	B	9, 17, 21, 25	5, **18**, 26
	C	**12**	19
10-12 age	B	17, 25, 27, 32, 38, 42, 47	5, 40, 57
	C	-	**18**
VOC	*Level*	*Items favored boys*	*Items favored girls*
4-7 age	B	-	6
	C	**16**	-
8-10 age	B	**8,** 47	3, 4, 9, **15**, **38**
	C	**16**, **37**	-
11-12 age	B	**16**, 51, 66, 67	**15**, 21, 44, 52
	C	**37**	-

Note. Bold items are biased.

3.2. VAR Subtest Findings

In 4-7 age group, 4 items demonstrated DIF at level B. 2 items were in favor of boys and 2 items were in favor of girls. According to the expert opinions, it was determined that four items measure verbal ability. However, since the toy types with which boys could be more familiar with were mentioned, the 12th item could be accepted to be biased in favor of boys. Since the self-care skills were mentioned, the 2nd item could be accepted to be biased in favor of girls. The 22nd item mentioning sense organs and limbs and the 15th item containing professional knowledge were not accepted as biased. It can be stated that the bias in these items was caused by the item effect, considering differences in the students' knowledge and experience.

In 8-9 age group 9 items displayed DIF. 5 items were in favor of boys, and 4 items were in favor of girls. 7 items were at level B and 2 items were at level C. According to the expert opinions 5th, 9th, 21st, 19th, 17th, 25th and 26th items could not be evaluated as bias. However, it was assumed that the 12th item showing DIF at level C could be biased in favor of boys since it contained the toy types with which boys could be more familiar. And the 18th item showing DIF at level B in girls could be biased in favor of girls since it contained the jewelry, accessory information.

In 10-12 age group 11 items demonstrated DIF. Only one item (18th) was at level C. 7 items were in favor of boys. 4 items were in favor of girls. According to the expert opinions all the items displayed DIF could not be accepted as biased, but the 18th item could be accepted as biased in favor of girls since it contained the jewelry/accessory information. It can be said that the bias in the item was caused by the item effect.

3.3. VOC Subtest Findings

In 4-7 age group 2 items (6th and 16th) displayed DIF at B and C levels. 1 item was in favor of boys and the other item was in favor of girls. Experts stated that item favored girls could not be accepted to be biased. But the item favored boys could be accepted as biased since it contained the car and speed elements. It can be stated that the bias in the item was caused by the item effect due to the more knowledge or experience of boys than girls.

In 8-10 age group 9 items demonstrated DIF at level B and C. 4 items were in favor of boys and 5 items were in favor of girls. According to the expert opinions 8th item containing the strength and

power elements, the 16th item containing the car and speed elements, and the 37th item containing the courage and fearlessness elements could be considered biased in favor of boys. For girls 15th item containing the cleaning/dirtiness elements and the 38th item containing the wet hair concept could be accepted as biased in favor of girls. Experts stated that bias in these items was caused by the item effect.

In 11-12 age group 9 items displayed DIF at B and C levels. 5 items were in favor of boys and 4 items were in favor of girls. According to the expert opinions 16th and 37th items for boys and 15th item for girls could be accepted as biased. Experts stated that bias in these 3 items caused by item effect.

The descriptive values of the items determined to be biased by expert opinions in the VAR and VOC subtests are given in Table 3.

Table 3
Results of DIF for the VAR and VOC subtests

VAR Subtest	*Level*	*Item*	χ^2	p
4-7 age	B	12	15.49	<.001
	B	2	25.76	<.001
8-9 age	B	18	9.57	<.001
	C	12	19.31	<.001
10-12 age	C	18	28.77	<.001
VOC Subtest	*Level*	*Item*	χ^2	p
4-7 age	C	16	40.51	<.001
8-10 age	B	8	16.43	<.001
	B	15	10.11	<.001
	C	16	16.43	<.001
	C	37	29.78	<.001
11-12 age	B	16	4.59	<.001
	B	15	4.41	<.001
	C	37	30.89	<.001

4. Discussion

The purpose of the study was to examine the presence of DIF in the ASIS verbal subtests across boys and girls in the norm sample. Because VAR and VOC subtests have start points by age the analysis were examined by age groups in these subtests. Since VSTM subtest requires all participants to start testing with the first item, the analysis were examined with all groups. According to the expert opinions, while there were 3 biased items in the VAR and 4 items in VOC subtests, no biased item was found in the VSTM subtest.

In the VAR subtest, 12th item accepted as biased. The reason why the item was in favor of boys may be the fact that boys have more knowledge and experience on the means of transportation used in the content of the question. It can be said that the difference in experience is because car toys take an important place in the selection of toys for boys (Bradbard, 1985; Weisgram et al., 2014). A higher number of vehicles, such as cars, planes, trains, trucks, etc., in the game equipment of boys, may have increased their knowledge and experience on means of transportation. The bias in the item is caused by the item effect. While the item was in favor of boys in the 4-7 and 8-9 age groups, it did not display bias in the 10-12 age group. Bias increased in the first stage as the age level increased and disappeared in the final stage. It can be stated that the reason for this situation is the process of development. It can be said that girls complete their knowledge, increase their familiarity levels and close their disadvantages due to the selection of toys with the end of the play period and the transition to adolescence as of the age of 10-12 years.

In the VAR subtest, the 2nd item was in favor of girls showed DIF at level B only in the 4-7 age group. In the item, it is requested to establish an analogy about self-care skills by giving information on combing the hair. The reason why this item was in favor of girls may be the fact

that the item contains concepts with which girls are more familiar. It can be stated that the expression of combing the hair may cause girls to establish analogies easier than boys. Furthermore, the fact that dolls take an important place in the selection of toys for girls (Yagan-Güder, 2014) and self-care skills are realized in these toys may have caused the item to be in favor of girls. Moreover, the fact that the bias in this item was observed only in the 4-7 age group may originate from the fact that boys do not have vocabulary and experience to establish the analogy in question in the 4-7 age group.

The 18th item was in favor of girls at level B in girls in the 8-9 age group and level C in girls in the 10-12 age group. The jewelry knowledge is questioned in the item. It is noteworthy that the item displayed a high level of bias, especially in the 10-12 age group. The perception of gender can be shown as the reason for this situation. Wearing jewelry is considered as one of the behaviors and appearances specific to girls in Turkish culture (Vatandas, 2007). Within the framework of this perception, the use of jewelry is common among girls from a young age. Accordingly, the fact that the item was biased for girls in the 8-9 age group can be explained by the high knowledge and experience of girls, and bias in girls in the 10-12 age group can be explained by the experience of reinforcing the appearance of the girl with jewelry, along with experience and entering adolescence.

In the VSTM subtest, DIF analyses were conducted with all groups. DIF was detected at level B in one item, but this item was not accepted as biased. As a result, no gender bias was observed in any item in the VSTM subtest. Verbal short-term memory is measured by the VSTM subtest. A short story is read to students, and questions about this story are asked. It can be stated that the items forming the VSTM subtest treat boys and girls equally. Furthermore, the fact that there is no information that can be advantageous for any gender in both the story read and the items, the absence of concepts that can be explained by the difference in experience between genders can be regarded as the main reasons for the absence of any DIF in the VSTM subtest.

In the VOC subtest, it was found that two items were in favor of girls and three items were in favor of boys. The 15th item was in favor of girls showed bias at level B in both the 8-10 age group and the 11-12 age group. There are cleaning/dirtiness elements in the item, and the subtest measures the vocabulary. The fact that the item is in favor of girls in the two age groups may be due to the more knowledge and experience of girls (Wood & Eagly, 2015). Gender perceptions and roles can cause this experience. Doing cleaning is regarded as one of the behaviors specific to girls in Turkish society (Vatandas, 2007). Within the framework of this perception and thought, it can be expected that girls have more experience with cleaning than boys. It can be thought that girls with more experience have a wider vocabulary (Borghi et al., 2019). 38th item was in favor of girls displayed bias at level B in the 8-10 age group. The reason why the item was in favor of girls is that the familiarity levels of girls with the concept of wet hair in the item are high and that girls have more experience. Furthermore, since long hair is one of the features found more in girls (Vatandas, 2007) in Turkish society, it can be expected that girls have more experiences lived when their long hair gets wet and the vocabulary they develop is more than that of boys.

While the 8th item was in favor of boys displayed bias at level B in the 8-10 age group, the 37th item showed bias at level C in both the 8-10 age group and the 11-12 age group. There are strength, power, courage, and fearlessness elements in the items. The reason for the high level of bias in these items may be due to gender perception and roles (Zotos & Tsichla, 2014). The perception that boys should be courageous and strong in general (Vatandas, 2007) may have caused boys to experience this perception in their families and acquire more knowledge than girls. The 16th item, which demonstrated bias in favor of boys at level C in the 4-7 age group, at level C in the 8-10 age group, and at level B in the 11-12 age group. It is noteworthy that the item was in favor of boys in all age groups. The car and speed concepts are mentioned in the item. The reason why the item was in favor of boys is that boys have more knowledge and experience in-car and speed subjects. It can be stated that the difference in experience is because boys spend more time with car toys than girls (Onur et al., 1997). This finding is consistent with the research findings in the literature,

stating that men are advantageous in questions involving movement, speed, and automobile concepts (Kalaycıoglu & Kelecioglu, 2011).

Overall the total number of ASIS verbal subtests items displaying DIF was minor. 3 items in the VAR subtest consisting of 58 items; 4 items in the VOC subtest consisting of 70 items displayed DIF. Items displayed DIF may affect individuals' test performance. In ASIS, items are arranged in order of difficulty level. It can be said that since the items showing DIF are not consecutive, they may not cause the test to be terminated due to the stopping rule. However, considering the studies claiming that minor item biases will not produce score differences and decrease of prediction values (Hong & Roznowski, 2001; Ozer-Ozkan, & Acar-Güvendir, 2021; Roznowski & Reith, 1999). It can be said that ASIS measurements are reliable. Also it can be said that items showing DIF should not be used in subsequent editions of ASIS. Cultural and social norms should be taken into account when creating verbal items or subtests.

5. Recommendations

When the items displayed DIF in the verbal subtests of the ASIS and which were considered to be in favor of any gender according to the expert opinions were examined, it was observed that all biases in the items were caused by the item effect. Differences in the students' knowledge and experience about the concepts in the items constituted the basis of the item bias. While this knowledge and experience created gender perception in Turkish culture, children's family experiences, peer interactions, and even the attitudes of grandparents in extended families have been effective in these perceptions (Inci Kuzu, 2015; Yagan-Güder & Alabay, 2016). The social environments of the children shaped their experiences. The fact that the items showing DIF can be explained by gender perception and roles indicates that the DIF in the items is caused by the item effect. In the study, it was observed that boys were more familiar with the concepts of power, courage, car, and speed; girls were more familiar with the concepts of hair, cleaning, self-care skills, jewelry, and accessory, and advantages could be formed in the items in which these concepts were used. In this study, DIF was examined in the context of gender. DIF analyses can also be carried out in the context of the socioeconomic level, the region and educational status variables. The use of concepts, such as power, courage, car, speed, velocity, hair, cleaning, jewelry, etc., can be considered in terms of bias in the creation of verbal items in developing an intelligence test. The findings can be compared with the findings obtained from the ASIS by conducting bias studies in other intelligence scales used in Türkiye.

Author contributions: Deniz Arslan: Conceptualization, Methodology, Investigation, Formal Analysis, Software, Visualization, Writing - original draft. Ömer Faruk Tamul: Conceptualization, Methodology, Investigation, Formal Analysis, Software, Visualization. Murat Doğan Şahin: Conceptualization, Methodology, Data curation, Resources, Supervision, Validation, Writing - review and editing. Ugur Sak: Conceptualization, Funding acquisition, Project administration, Methodology, Data curation, Resources, Supervision, Validation, Writing - review and editing.

Data availability: Data is available upon request at the Center for Research and Practice for High Ability Education at Anadolu University.

Declaration of interest: The authors declare that there is no conflict of interest.

Funding: This study was supported by a grant from Anadolu University (Grant No: 1504E151).

References

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

Agresti, A. (1984). *Analysis of ordinal categorical data*. John Wiley & Sons.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Author.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-24). Lawrence Erlbaum.

Arslan, D., Sak, U., & Atesgöz, N. N. (2021). Are more humorous children more intelligent? A case from Turkish Culture. *Humor, 34*(4), 567-588. https://doi.org/10.1515/humor-2021-0054

Borghi, A. M., Barca, L., Binkofski, F., Castelfranchi, C., Pezzulo, G., & Tummolini, L. (2019). Words as social tools: Language, sociality and inner grounding in abstract concepts. *Physics of Life Reviews*, *29*, 120-153. https://doi.org/10.1016/j.plrev.2018.12.001

Bradbard, M. R. (1985). Sex differences in adults' gifts and children's toy requests at Christmas. *Psychological Reports, 56*(3), 969-970. https://doi.org/10.2466/pr0.1985.56.3.969

Camilli, G. & Shepard L. A., (1994). *Methods for identifying biased test items*. Sage.

Carroll, J. B. (2005). The three-stratum theory of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 69–76). The Guilford Press.

Cırık, M., Sak, U., & Öpengin, E. (2020). An investigation of cognitive profiles of children with attention deficit hyperactivity disorder on Anadolu Sak Intelligence Scale. *Ankara University Faculty of Educational Sciences Journal of Special Education, 21*(4), 663-685. https://doi.org/10.21565/ozelegitimdergisi.570505

Colom, R., Escorial, S. & Rebollo, I. (2004). Sex differences on the Progressive Matrices are influenced by sex differences on spatial ability. *Personality and Individual Differences, 37*(6), 1289-1293. https://doi.org/10.1016/j.paid.2003.12.014

Doolittle, A. E. & Cleary, T. A. (1987). Gender-based differential item performance in mathematic achievement items. *Journal of Educational Measurement, 24*(2), 157-166. https://doi.org/10.1111/j.1745-3984.1987.tb00271.x

Educational Testing Service [ETS]. (2009). *ETS Guidelines for Fairness Review of Assessments*. Author.

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

Hong, S., & Roznowski, M. (2001). An investigation of the influence of internal test bias on regression slope. *Applied Measurement in Education*, *14*(4), 351-368. https://doi.org/10.1207/S15324818AME1404_3

Horn, J. L., & Blankson, A. N. (2012). Foundations for better understanding of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues*. Guilford Press.

Immekus, J. C. & Maller, S. J. (2009). Item parameter invariance of the Kaufman Adolescent and Adult Intelligence Test across male and female samples. *Educational and Psychological Measurement*, *69*(6), 994-1012. https://doi.org/10.1177/0013164409344489

Inci Kuzu, Ç. (2015). The gender prejudice and the toy choice of the children in preschool period and the effect of the parents on this. *Journal of International Social Research, 8*(39), 651- 655.

Kalaycıoglu, D. B. & Berberoglu, G. (2010). Differential item functioning analysis of the science and mathematics items in the University Entrance Examinations in Turkey. *Journal of Psychoeducational Assessment, 29*(5), 467-478. https://doi.org/10.1177/0734282910391623

Kalaycıoglu, D. B. & Kelecioglu, H. (2011). Item bias analysis of the university entrance examination. *Education and Science, 36*(161), 3-13.

Köprü, F., & Ayas, M. B. (2020). An investigation of the criterion validity of Anadolu Sak Intelligence Scale (ASIS): The case of EPTS. *Talent, 10*(2), 110-128. https://doi.org/10.46893/talent.857308

Kristanjansonn E., Aylesworth, R., McDowell, I. & Zumbo, B. D. (2005). A comparison of four methods for detecting differential item functioning in ordered response model. *Educational and Psychological Measurement, 65*(6), 935-953. https://doi.org/10.1177/0013164405275668

Li, Y., Cohen, A. S. & Ibarra, R. A. (2004). Characteristics of mathematics items associated with gender DIF. *International Journal of Testing, 4*(2), 115–136. https://doi.org/10.1207/s15327574ijt0402_2

Maller, S. J. (2000). Item invariance in four subtests of the Universal Nonverbal Intelligence Test (UNIT) across groups of deaf and hearing children. *Journal of Psychoeducational Assessment, 18*(3), 240-254. https://doi.org/10.1177/073428290001800304

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

Murray, A. L., Booth, T. & McKenzie, K. (2015). An analysis of differential item functioning by gender in the Learning Disability Screening Questionnaire (LDSQ). *Research in Developmental Disabilities, 39*(2015), 76-82. https://doi.org/10.1016/j.ridd.2014.12.006

Nolan, R. F., Watlington, D. K. & Willson, V. L. (1989). Gifted and nongifted race and gender effects on item functioning on the Kaufman Assessment Battery for Children. *Journal of Clinical Psychology, 45*(4), 645-650. https://doi.org/10.1002/1097-4679(198907)45:4<645::AID-JCLP2270450422>3.0.CO;2-E

Onur, B., Çelen, N., Çok, F., Artar, M. & Şener Demir, T. (1997). Türkiye'de iki kentte annelerin bakış açısıyla çocukların oyuncak gereksinmesi [Children's need for toys from the perspective of mothers in two cities in Turkey]. *Ankara University Journal of Faculty of Educational Sciences (JFES), 30*(1), 45-74. https://doi.org/10.1501/Egifak_0000000276

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

Özer-Özkan, Y. & Acar-Güvendir, M. (2021). Differential item functioning analysis of a high stake test in terms of statistical regions of Turkey. *Journal of Pedagogical Research, 5*(3), 122-134. https://doi.org/10.33902/JPR.2021371303

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting Differential Item Functioning. *Applied Psychological Measurement*, *17*(2), 105–116. https://doi.org/10.1177/014662169301700201

Roznowski, M., & Reith, J. (1999). Examining the measurement quality of tests containing differentially functioning items: do biased items result in poor measurement? *Educational and Psychological Measurement*, *59*(2), 248-269. https://doi.org/10.1177/00131649921969 83

Sak, U., Bal-Sezerel, B., Ayas, B., Tokmak, F., Özdemir, N. N., Demirel-Gürbüz, Ş., & Öpengin, E. (2016). *Anadolu Sak Intelligence Scale: ASIS practitioner's book.* Eskisehir: Anadolu University UYEP Center.

Sak, U., Bal-Sezerel, B., Dulger, E., Sozel, K., & Ayas, M. B. (2019). Validity of the Anadolu-Sak Intelligence Scale in the identification of gifted students. *Psychological Test and Assessment Modeling, 61*(3), 263-283.

Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 99–144). The Guilford Press.

Sözel, H. K., Öpengin, E., Sak, U., & Karabacak, F. (2018). The discriminant validity of the Anadolu-Sak Intelligence Scale (ASIS) for gifted and other special education groups. *Turkish Journal of Giftedness and Education, 8*(2), 160-180.

Tamul, Ö. F., Sezerel, B. B., Sak, U., & Karabacak, F. (2020). Social validity study of the Anadolu-Sak intelligence Scale (ASIS). *PAU Journal of Education, 49,* 393-412. https://doi.org/10.9779/pauefd.575479

Vatandaş, C. (2007). Toplumsal cinsiyet ve cinsiyet rollerinin algılanışı [Perception of gender and gender roles]. *Istanbul Journal of Sociological Studies*, *35*, 29-56.

Wechsler, S., de Cassia Nakano, T., da Silva Domingues, S. F., Rosa, H. R., da Silva, R. B. F., Silva-Filho, J. H. & Minervino, C. A. D. S. M. (2014). Gender differences on tests of crystallized intelligence. *European Journal of Education and Psychology, 7*(1), 59-72.

Weisgram, E. S., Fulcher, M., & Dinella, L. M. (2014). Pink gives girls permission: Exploring the roles of explicit gender labels and gender-typed colors on preschool children's toy preferences. *Journal of Applied Developmental Psychology, 35*(5), 401-409. https://doi.org/10.1016/j.appdev.2014.06.004

Wood, W., & Eagly, A. H. (2015). Two traditions of research on gender identity. *Sex Roles*, *73*, 461-473. https://doi.org/10.1007/s11199-015-0480-2

Yagan Güder, S. & Alabay, E. (2016). Examination of the toys preferences in children aged 3-6 in the context of gender. *Journal of Kirsehir Education Faculty, 17*(2), 91-111.

Yagan Güder, S. (2014). *Investigating preschool children's perception of gender* (Publication No: 363226) [Doctoral dissertation, Hacettepe University]. Council of Higher Education Thesis Center.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337-347). Erlbaum.

Zotos, Y. C., & Tsichla, E. (2014). Female stereotypes in print advertising: A retrospective analysis. *Procedia-Social and Behavioral Sciences*, *148*(2014), 446-454. https://doi.org/10.1016/j.sbspro.2014.07.064