GENDER DIFFERENTIAL ITEM FUNCTIONING IN SLOVAK VERSION OF INTELLIGENCE STRUCTURE TEST 2000 - REVISED

Michal KOHÚT[1], Peter HALAMA[1], Vladimír DOČKAL[2], Peter ŽITNÝ[1]

[1]University of Trnava
Trnava, Slovak Republic
E-mail: michalkohut.tn@gmail.com

[2]Research Institute for Child Psychology and Pathopsychology
Bratislava, Slovak Republic

Abstract: The study focused on the gender differential item functioning in Slovak version of the Intelligence Structure Test 2000 - Revised (Amthauer et al., 2011). The sample included 744 middle and high school students with mean age of 16.94 years. The non-parametric method SIBTEST for identification of items with differential functioning was used in order to detect uniform and non-uniform DIF. The analysis showed that the I-S-T 2000 R includes several items with DIF favoring either males or females, but in most subtests, with no or small effect on differences between genders. Substantial but nonsignificant effect of DIF items on subtest score was found for Verbal Analogy, which contained six items with DIF all favoring females. These items included verbal content related to areas more common for females such as diet or food. The results suggest that specific content of verbal intelligence items can be a potential source of gender bias.

Key words: intelligence testing, differential item functioning, gender, I-S-T 2000 R

The question of test validity is of a great importance in intelligence testing. Possible favoritism towards one of the groups tested for intelligence disturbs test validity and can be a source of serious violation of fair testing (AERA, APA & NCME, 1999). Comparability of tests results across different groups is an inevitable condition for meaningful test use and if not guaranteed, test bias is present. Test bias is defined as systematic error in the estimation of a value. A biased test is one that systematically overestimates or underestimates the value of the measured variable. Presence of test bias is always negative, as biased tests systematically underrepresent target groups' true aptitudes or abilities (Reynolds & Suzuki, 2012). In the past decades, the differential item functioning (DIF) approach has been developed to address the issue of test bias at the item level (Osterlind & Everson, 2009). DIF is present when examinees from different groups have different probabilities or likelihoods of success on an item, after they have been matched on the ability of interest (Clauser & Mazor, 1998). DIF is not a pure difference between groups in item response, rather, it is a difference in response between members of different groups with the same level of measured ability. If items with differential func-

Aknowledgement
The study was supported by Grant Agency VEGA No. 1/0234/15.

DOI: 10.21909/sp.2016.03.720

tioning are included in the test, overall score of the test can favor one or another group and it can produce socially, legally or politically significant bias affecting the different groups (McDonald, 2013). To date, several statistical methods for detecting DIF have been applied, such as the Mantel-Haenszel procedure (Holland & Thayer, 1988; Fidalgo & Madeira, 2008), logistic regression (French & Miller, 1996; Zumbo, 1999) or item response theory approach (Jelínek, Květon, & Vobořil, 2011). There are two types of DIF: uniform and non-uniform. Uniform DIF means that the item gives advantage to one group across all levels of ability. On the other hand, item with non-uniform DIF changes the direction of advantage at different levels of the ability continuum (Osterlind & Everson, 2009). Differential item functioning analysis has been used in many studies related to intelligence testing. Maller (2000) analyzed DIF in deaf and hearing children and Simos et al. (2011) in different ethnic groups in Greece. Several studies also confirmed that one of the sources of DIF can be found in language skills and language origin (Martiniello, 2009; Roomaney & Koch, 2013; Schaap, 2011). All these studies suggested that the DIF approach is a good tool for the identification of bias at item and test level.

Difference between male and female performance in intelligence tests is the subject of frequent scientific discussion, sometimes with contradictory findings and inconsistent conclusions. Colom et al. (2000) found in their research with more than 10 000 adult subjects that there are only negligible sex differences in general intelligence. The same conclusions were drawn in their latter study using the Spanish version of WAIS-III (Colom et al., 2002). On the other hand, Jackson and Rushton (2006) concluded that in a sample of 102 516 subjects (17-18 years old adolescents), there is a clear evidence of small but non-trivial differences in general intelligence favoring males. This difference had a point-biserial effect of 0.12 (equivalent to 3.63 IQ points) and was confirmed across all socio-economic levels and ethnic groups. Male outperformance was confirmed also in a metaanalysis of studies using Raven's Progressive Matrices (RPM), which are usually considered a measure of *g* intelligence (Irwing & Lynn, 2005). More agreement is on the existence of gender differences in partial intelligence abilities, such as male outperformance in spatial ability (Linn & Petersen, 1985; Voyer et al., 2005) or female outperformance in verbal ability (Hyde & Linn, 1988; Weiss et al., 2006). However, even in these areas contradictory opinions and findings and also can be found. E.g., for verbal abilities, Lynn (2005) assert that examination of literature leads to the conclusion that in adults, males have slightly higher verbal abilities than females. Lemos et al. (2013) also found that boys outperformed girls in all subtests including verbal reasoning. These differences are explained by *g* intelligence.

One of the possible explanations for inconsistent findings in this area is the characteristics of the tests used in research. Blinkhorn (2005) suggested, that if sex differences in intelligence are to be found, detailed study of the internal workings of the test tends to show why. He emphasized the importance of gender-fair tests and explained how biased test can negatively influence the results of meta-analysis. His ideas are strongly supported by results of studies focused on gender differential item functioning in intelligence tests, even in measuring the *g*-factor. Results of many studies confirmed that several intelligence tests contain

items which exhibit gender DIF (e.g., Immekus & Maller, 2009; Brown & Rodgers, 2009; Simos et al., 2011). Some of these studies focused on the sources of gender DIF in item content. Abad et al. (2004) studied sex DIF in Raven's Advanced Progressive Matrices, which is supposed to measure *g* intelligence. They found that test contains several items, which are biased against female performance, especially those where spatial performance is especially pronounced. Deleting these items reduced sex difference in overall score of the test. However, later analysis of Chiesi et al. (2012) showed no gender DIF items in RPM and they concluded that there were no gender-related advantages or disadvantages in this test. Maller (2000) studied gender DIF in American standardization sample of Wechsler Intelligence Scale for Children WISC-III. She found that almost one-third of the items show some kind of DIF either uniform or non-uniform. Concerning the content, items from the Information subtest, which were more difficult for girls, contained content pertaining to science, whereas items containing knowledge about months, seasons, or a famous girl in history were all easier for girls. Items on the Similarities subtest that involved numbers or measurements were easier for boys, whereas an item containing words to describe emotions was easier for girls. An item from the Vocabulary subtest difficult for girls contained content related to sports, teams, or competition. Most of the items from the Picture Completion subtest showed uniform DIF related to picture content. Items with a picture of a male or a female were more difficult for girls or boys, respectively. Tuzt and Berger (in press) applied DIF analysis to verbal subtest Sentence Completion of the Intelligence Structure Test 2000 - Revised. Three of the 20 items showed gender DIF. Items related to social relations was easier for females, while items related to nature and technics were easier for males. All these results suggest that item content can be a significant source of gender bias in intelligence testing in both verbal and non-verbal (sub)tests.

In this study, the focus is on the Intelligence Structure Test 2000 - Revised (Amthauer et al., 2001). IST in different versions is one of the most used complex intelligence tests in some European countries (Evers et al., 2012). However, in spite of its frequent usage, there is a lack of independent psychometric evaluation in research literature. There is also an absence of empirical evidence related to the functioning of I-S-T 2000 R items across gender groups. To address this issue, the purpose of this study was to investigate DIF in the I-S-T 2000 R items across boys and girls in a Slovak adolescent standardization sample. We aimed to identify whether such items are present in the test, and if yes, whether they influence overall score of the subtests. We also aimed to analyze the content of the DIF items in order to identify possible sources of DIF.

Method

Sample

The sample used in this study comes from Slovak standardization sample of I-S-T 2000 R. It included 744 Slovak middle (N = 68) and high school (N = 676) students, 424 (57%) males and 320 (43%) females. Mean age of participants was 16.94 years (SD = 1.36), from 13 years to 22 years. Mean age for males was 17.01 years (SD = 1.37) and 16.85 (SD = 1.33) for females. Participants

came from different parts of Slovakia: west (22.3%), east (26.1%), north (25.8%) and south (25.8%). Majority of participants (34.4%) lived in cities with population between 10 000 and 49 999. Second most frequent were participant from cities with population between 2000-9999 (26.2%), then the towns with population between 500-1999 citizens (18.5%), cities with more than 50000 citizens (15.7%) and villages with population under 499 citizens (5.1%).

Measure

The Intelligence Structure Test - Revised (I-S-T 2000 R form A; Amthauer, Brocke, Liepmann, & Beauducel, 2001) is an intelligence test battery measuring three intelligence areas, each with three subtests. Verbal Intelligence is assessed through Sentence Completion, Verbal Analogies and Similarities subtests. Sentence Completion (SC) contains sentences with a missing word which are completed by one of five options. In Verbal Analogies (VA) subtest, the task is to detect a relation between two words and find a word with similar relationship to another word. Similarities subtest (VS) presents groups of six words and a person has to find two words with common collective term. The Numerical Intelligence is assessed through subtests: Numerical Calculations, Number Series and Numerical Signs. Numerical Calculations (CA) contains arithmetical tasks with real numbers. Number Series (NS) presents series of numbers formed according to a specific rule with tasks to choose the next number in series. In Numerical Signs (SI) subtest, a person has to choose correct mathematical operators to an equation. The third area of intelligence, the Figural Intelligence is assessed through Figure Selection, Cubes and Matrices sub-tests. Figure Selection (FS) items present geometrical shapes together with some pieces resulting from cutting up one of the shapes with tasks to identify the whole shape which can be constructed from individual pieces. Items in Cubes subtest (CU) require the identification of a rotated cube among different options presented with only 3 faces visible. The last subtest, Matrices (MA), contains items with a set of figures arranged according to a particular rule and the task is to choose the figure from options provided that conforms to this rule.

Each of these 9 subtests consists of 20 items, that is 180 items for the test. Every item has a choice of 5 options, one key option and four distractor options. The I-S-T 2000 R is frequently used in psychological assessment in several European countries, e.g. Germany or Czech Republic (Evers et al., 2012). It is also used in psychological research for measurement of general intelligence (e.g., Dislich et al., 2012; Steinmayr & Spinath, 2015).

Analysis

For a data preparation and analysis Statistical Package for Social Sciences (SPSS) was used. Differential item functioning was examined with Simultaneous Item Bias Test (SIBTEST). Item and test characteristic curves were plotted in IRTPRO 2.1 under 2PL model and finalized in MS Excel.

SIBTEST is a non-parametric statistical method for DIF analysis based on the Shealy and Stout's (1993) multidimensional item response theory model. Examinees are matched on latent ability (scale score, or other user-selected variable), which is also called *primary* dimension. DIF may occur when two groups of examinees (e.g., males and females)

with the same latent ability, or primary dimension, differ in secondary dimension (Roussos & Stout, 1996a). SIBTETS can detect uniform as well as nonuniform DIF (Li & Stout, 1996). Additionally, SIBTEST can analyze different functioning of more items as one item set for accessing differential bundle functioning (DBF). DBF testifies if effects of several differentially functioning items favoring focal or reference group cancel out or favor one group of examinees (Douglas, Roussos, & Stout, 1996).

Results

I-S-T 2000 R Scales Gender Differences

Gender differences were examined with a Student's t-test for independent samples. The results are shown in Table 1. Females scored significantly higher in 5 of 10 scales, namely Sentence Completion (1), Verbal Analogies (2), Similarities (3), Numerical Calculations (4) and Matrices (9). Males scored significantly higher in the scale Numerical Signs (6). Largest gender difference was in the scale Similarities (d = -.56). Generally, gender differences were from small to medium size.

Differential Item Functioning

Table 2 presents quantity and percentage of items suspicious of DIF identified by SIBTEST with DBF computed. Scales Numerical Calculations (4) and Cubes (8) show no differentially functioning items within the SIBTEST analysis. The highest number of DIF items was identified in the scale Verbal Analogies (2). In case of the first scale Sentence Completion, the DBF was non-significant because the effect of items favoring males and females cancelled out.

List of all items suspicious of DIF, identified by the SIBTEST is shown in the Table 3. Totally, 18 items were identified. Roussos and

Table 1 *Gender differences for I-S-T 2000 R subtest scores*

Subtest	Males		Females		*d*	*t*
	M	SD	M	SD		
SC	9.12	3.52	9.91	3.05	0.24	3.27**
VA	9.03	3.43	10.28	3.15	0.38	5.18***
VS	9.52	4.60	11.92	4.04	0.55	7.56***
CA	10.51	5.03	12.27	4.50	0.37	5.03***
NS	10.09	5.78	10.29	5.23	0.04	0.49
SI	11.21	4.71	10.56	4.07	0.15	2.01*
FS	9.66	4.20	9.47	3.79	0.05	0.63
CU	10.13	4.22	10.22	3.70	0.02	0.34
MA	8.90	3.24	10.17	2.90	0.41	5.62***

Note. SC - Sentence Completion, VA - Verbal Analogies, VS - Similarities, CA - Numerical Calculations, NS - Number Series, SI - Numerical Signs, FS - Figure Selection, CU - Cubes, MA – Matrices

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2 *Number and percentage of items showing uniform DIF identified by the SIBTEST with DBF*

Subtest	Differential item functioning (β)						DBF (β/N)
	N	%	Favor males		Favor females		
			N	M	N	M	
SC	3	15	2	0.17	1	-0.14	*n.s.*
VA	6	30	0	-	6	-0.11	-0.66
VS	2	10	0	-	2	-0.14	-0.28
CA	0	0	-	-	-	-	-
NS	1	5	1	0.08	0	-	-
SI	1	5	0	-	1	-0.10	-
FS	2	10	0	-	2	-0.16	-0.32
CU	0	0	-	-	-	-	-
MA	1	5	0	-	1	-0.10	-

Note. SC - Sentence Completion, VA - Verbal Analogies, VS - Similarities, CA – Numerical Calculations, NS - Number Series, SI - Numerical Signs, FS - Figure Selection, CU - Cubes, MA - Matrices

Number of items in each scale is 20. DIF and DBF with $p < 0.01$.

Stout (1996b) proposed a DIF classification system for SIBTEST results. An item shows negligible or small level of uniform DIF if absolute value of β is smaller than 0.059; medium if absolute β is between 0.059 and 0.088; medium to large if β is equal or higher than 0.088; and β significantly differ from 0. By this classification, all but one of the uniform DIF is medium to large size.

In Figure 1, we present test characteristic curves for males and females of all subtests plotted under the 2PL model. Almost in all cases, characteristic curves overlap, suggesting there is no significant difference between genders in the expected test score. A difference can be seen in the second subtest Verbal Analogies where 6 differentially functioning items favoring females were identified. In spite of this, the expected score for females is higher by approximately only 1 point. Figures suggest some gender differences also for the Cubes subtest (especially at the higher level of the score), but this difference is not caused by DIF because no DIF items were detected for this subtest. To better understand the effect that DIF items may have on the total scores of subtests we accessed the gender differences again, but only with the items that do not show DIF. Table 4 displays Cohen's *d* coefficients calculated for original and "purified" subtests together with its 95% confidence intervals. As seen, confidence intervals are overlapping for all subtests, so no significant changes in Cohen's *d* occurred after deleting the DIF items. However, the subtest Verbal Analogies did show partial decrease of difference amount (from 0.38 to 0.19). Although statistically non-significant, this decrease suggests that gender difference in this subtest

Table 3 *List of items showing DIF identified by SIBTEST with values of Beta estimates*

Subtest	Item	SIBTEST	
		Beta estimate	Standard error
SC	7	-0.13	0.03
SC	11	0.11	0.04
SC	15	0.13	0.03
VA	23	-0.10	0.03
VA	27	-0.12	0.03
VA	28	-0.10	0.04
VA	32	-0.10	0.03
VA	36	-0.15	0.03
VA	38	-0.12	0.03
VS	44	-0.13	0.04
VS	46	-0.14	0.03
CA	65	-0.09*[1]	0.03
NS	86	0.08	0.03
SI	115	-0.10	0.03
FS	136	-0.18	0.04
FS	139	-0.13	0.04
MA	170	-0.10	0.03
MA	172	0.11*[2]	0.03

Note. SC - Sentence Completion, SC - Verbal Analogies, VS - Similarities, CA - Numerical Calculations, NS - Number Series, SI - Numerical Signs, FS - Figure Selection, CU - Cubes, MA – Matrices

$p < 0.01$

* β value for Crossing SIBTEST indicating nonuniform DIF.

1 β for males = 0.03; β for females = 0.6

2 β for males = 0.06; β for females = 0.5

may be partially caused by the effect of differentially functioning items that favor females.

In the next step, we focused on DIF items especially in the subtest Verbal Analogies. To solve the item in this subtest, examinees have to understand the relationship or pattern between a word pair and then apply that pattern to choose the correct word to complete the second pair. In Figure 2 we present item characteristic curves of 6 items in which differential functioning was detected. In all cases DIF was uniform and significant. Nonuniform crossing DIF in items 32 and 36 suggested in figures did not show a significance. We looked at the content of the items with

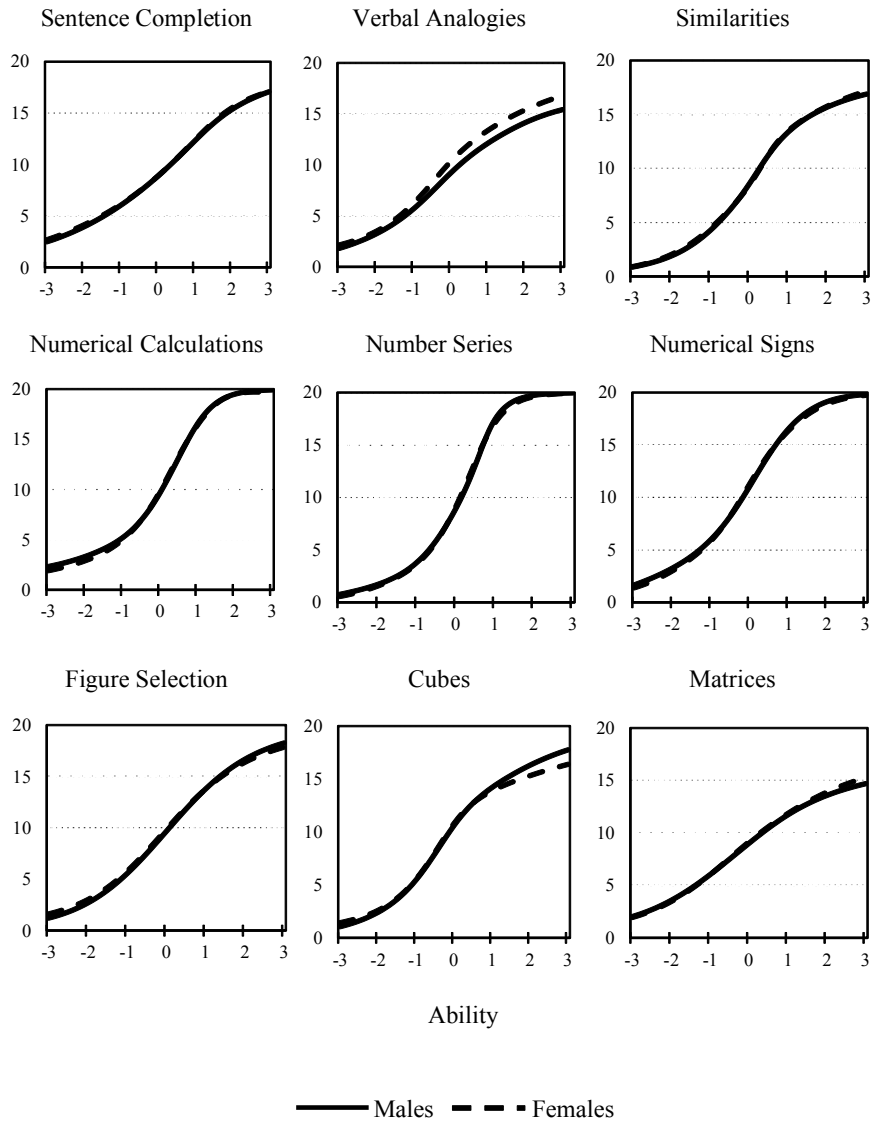

Figure 1 Characteristic curves of I-S-T 2000 R subtests

Table 4 *Cohen's d for gender differences in I-S-T 2000 R subtests with and without DIF items*

Subtest	*d* and 95% CI for difference in subtest with all items	*d* with 95% CI for differences in subtest without DIF items
SC	0.24 [0.38, 0.09]	0.26 [0.41, 0.12]
VA	0.38 [0.52, 0.23]	0.18 [0.33, 0.04]
VS	0.55 [0.70, 0.40]	0.49 [0.64, 0.35]
CA	0.37 [0.51, 0.22]	-
NS	0.04 [0.38, 0.09]	0.05 [0.19, 0.10]
SI	0.15 [0.00, 0.29]	0.16 [0.02, 0.31]
FS	0.05 [0.10, 0.19]	0.12 [0.03, 0.26]
CU	0.02 [0.17, 0.12]	-
MA	0.41 [0.56, 0.26]	0.37 [0.51, 0.22]

Note. SC - Sentence Completion, SC - Verbal Analogies, VS - Similarities, CA - Numerical Calculations, NS - Number Series, SI - Numerical Signs, FS - Figure Selection, CU - Cubes, MA - Matrices

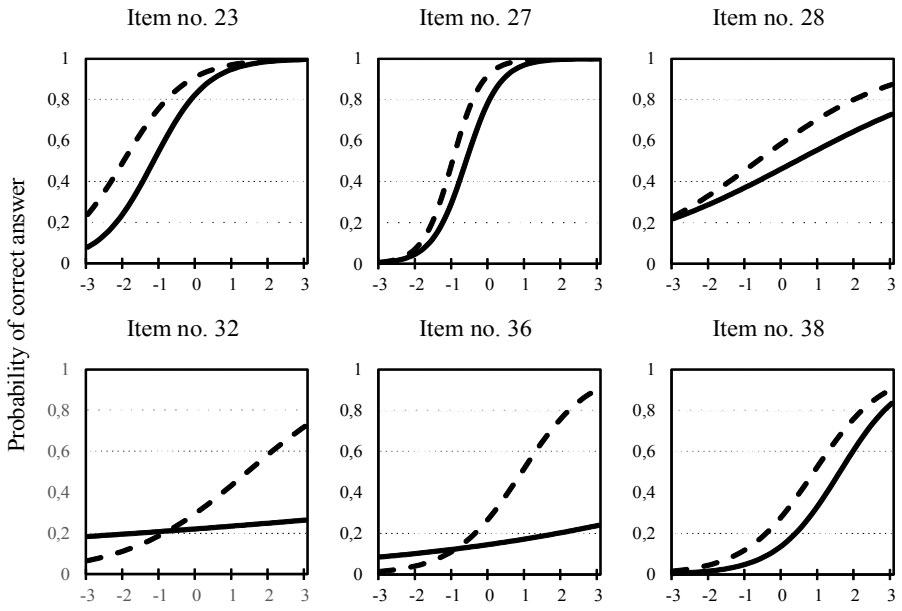

Figure 2 Differential functioning of items in Verbal Analogies subtest

DIF. Item 36 has the highest beta estimate value in this sub-test and its content deals with topic of diet and health. Other items' content focused on kitchen stuff (27), emotions and feelings (38), civil matters (28), pictures and reading (23) and adverbs of time and frequency (32). Concerning items from other verbal subtests, those favoring males are dealing with nature (11) or mechanics (15), those favoring females are dealing with feelings (7), jewels (46) or buildings (44).

Discussion

Analysis of gender differences in I-S-T 2000 R subtests confirmed that six subtests showed significant differences between males and females. In five of these subtests, females scored higher than males, which does not correspond with previous findings suggesting higher general intelligence in males (Jackson & Rushton, 2006; Irwing & Lynn, 2005). The only subtest with significant outperformance of males was Numerical Signs, but the effect size of this difference was rather small (Cohen *d* 0.15). On the other hand, all subtests related to verbal intelligence showed significant differences with higher score for females, even at the middle level of the effect size for Similarities (Cohen *d* 0.55). Our results confirmed previous findings related to female outperformance in verbal ability (Hyde & Linn, 1988; Weiss et al., 2006).

In subsequent analysis, we focused on the DIF analysis to identify whether these differences are caused by items displaying possible gender bias. 18 items showed significant level of differential item functioning, majority of them (14) favoring females. More than a half of these items come from verbal subtests, which suggests that items with verbal contents are potentially more affected by gender bias (Steinmayr et al., 2015). Special attention should be devoted to Verbal Analogy subtest, which shows substantial, although not significant effect of DIF items on the overall score. As a decrease of 0.2 in Cohen's *d* coefficient after deleting DIF items corresponds to a decrease of about 3 points in subtest IQ score, we consider it not negligible, because empirical research showed that even such small difference can have some effect on specific relevant variables (see Deary et al., 2004 for example). In this subtest, the task is to detect the relation between two words and find a word with similar relationship to another word (Amthauer et al., 2001). The content analysis of the items showing DIF from this subtest but also from other verbal subtests, suggests that differences in male and female preferences could be a source of differential functioning. Steinmayr et al. (2015), when discussing gender DIF in German knowledge tests, considers Ackerman's (1996) PPIK theory of adult intellect, which emphasizes that adult intellect is predicated on four components: intelligence-as-process (fluid abilities), personality traits, interests, and intelligence-as-knowledge. Personal interests are an important factor of intelligence development because higher interest in particular domain could lead to better achievement in tests related to this domain. Personal interests and preferences of males and females, which are shaped by specific gender roles as defined in particular micro and macro environment (Lindsey, 2015), can lead to higher capability to solve problems and tasks related to the area of preference. Inclusion of items related to an interest area of males or females in intelligence test can be a source of differential functioning and possible gender bias. As

suggested by the results, most of the I-S-T 2000 R subtests do not show substantial inclusion of such items with the exception of Verbal Analogy. This subtest includes 6 items with DIF favoring females and some of them have clear relations to female preferences such as diet and health (item 36) or kitchen stuff (item 27). Items with relation to general gender preferences are included in other subtests, e.g. items favoring males from the Sentence Completion subtest with clear tendency to male preferences as mechanics (item 15) or nature (item 11), but these items are less frequent and their effect on the subtest score is negligible.

In the light of our results, we can conclude that I-S-T 2000 R is not an intelligence test strongly affected by gender differential item functioning in adolescent sample. No subtest score seemed to be influenced by the presence of items with differential item functioning toward one or another gender, with the exception of Verbal Analogy, which shows nonsignificant but substantial effect. Also, the overall number of items with detected DIF is rather small when compared with previous studies focusing e.g., on WISC-III (e.g., Maller, 2000), in which one-third of the items showed some kind of DIF. However, future revisions of I-S-T 2000 R should take into account that specific content of verbal items and its relation to male or female preferences is a potential source of DIF presence and, subsequently, potential source of gender bias.

The limitation of our study comes from specific sample used in the research. We used Slovak standardization sample of I-S-T 2000 R, which includes only adolescents and no other samples are available in this moment. Adolescent sample is specific due to the fact that gender roles or preferences are under development during the whole life and transition to adult age can bring some changes in this area (Steensmae et al., 2013). Adult gender roles and preferences can be different from adolescent ones and this can be manifested also in a test situation. Based on this fact, we can assume that DIF analysis with an adult sample could lead to some differences in results related to differential item functioning. Further research should focus on the different age samples, especially adults in different developmental stages, to reveal whether the results in our study could be generalized to the entire target population of the test or whether these results are developmentally specific.

Conclusions

Slovak version of the Amthauer's Intelligence Structure Test 2000 - R shows significant gender differences in a sample of adolescents, since in several subtests females scored higher than males. Analysis of different item functioning revealed that this effect is only partially caused by items with gender bias as only one subtest score (Verbal Analogy) appears to be substantially affected by DIF items. However, DIF analysis of items from verbal subtests showed that if an item has gender relevant content, it can be biased toward one or the other gender and potentially, it can affect the overall score of the subtest. In general, I-S-T 2000 R, although showing gender differences in several subtests, does not seem to be strongly affected by gender bias coming from items displaying DIF. Further research on different age sample should clarify whether these results are developmentally stable or whether they are changing through developmental stages.

Received March 2, 2016

References

Abad, F. J., Colom, R., Rebollo, I., & Escorial, S. (2004). Sex differential item functioning in the Raven's Advanced Progressive Matrices: Evidence for bias. *Personality and Individual Differences*, *36*(6), 1459-1470.

Ackerman, P. L. (1996). A theory of adult intellectual development: Process, personality, interests, and knowledge. *Intelligence, 22*, 227-257.

AERA, APA & NCME (1999). *Standards for educational and psychological tests*. Washington, DC: American Psychological Association, American Educational Research Association, National Council on Measurement in Education.

Amthauer, R., Brocke, B., Liepmann, D., & Beauducel, A. (2001). *Intelligenz-Struktur-Test 2000 R*. Göttingen: Hogrefe.

Blinkhorn, S. (2005). Intelligence: A gender bender. *Nature*, *438*(7064), 31-32.

Brown, T., & Rodger, S. (2009). An evaluation of the validity of the Test of Visual Perceptual Skills-Revised (TVPS-R) using the Rasch Measurement Model. *The British Journal of Occupational Therapy*, *72*(2), 65-78.

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

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31-44.

Colom, R., Juan-Espinosa, M., Abad, F., & García, L. F. (2000). Negligible sex differences in general intelligence. *Intelligence*, *28*(1), 57-68.

Colom, R., García, L. F., Juan-Espinosa, M., & Abad, F. J. (2002). Null sex differences in general intelligence: Evidence from the WAIS-III. *The Spanish Journal of Psychology*, *5*(01), 29-35.

Deary, I. J., Whiteman, M. C., Starr, J. M., Whalley, L. J., & Fox, H. C. (2004). The impact of childhood intelligence on later life: Following up the Scottish mental surveys of 1932 and 1947. *Journal of Personality and Social Psychology*, *86*(1), 130-147.

Dislich, F. X., Imhoff, R., Banse, R., Altstötter-Gleich, C., Zinkernagel, A., & Schmitt, M. (2012). Discrepancies between implicit and explicit selfconcepts of intelligence predict performance on tests of intelligence. *European Journal of Personality*, *26*(3), 212-220.

Douglas, J. A., Roussous, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33*, 465-484.

Evers, A., Muñiz, J., Bartram, D., Boben, D., Egeland, J., Fernández-Hermida, J. R., ... & Iliescu, D. (2012). Testing practices in the 21st century: Developments and European psychologists' opinions. *European Psychologist*, *17*(4), 300-319.

Fidalgo, A., & Madeira, J. (2008). Generalized Mantel-Haenszel methods for differential item functioning detection. *Educational and Psychological Measurement*, 68, 940-958.

French, A. W., & Miller, T. R. (1996). Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement 33,* 315-332.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, *104*(1), 53.

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*, *35,* 623-642.

Irwing, P., & Lynn, R. (2005). Sex differences in means and variability on the progressive matrices in university students: A meta analysis. *British Journal of Psychology*, *96*(4), 505-524.

Jackson, D. N., & Rushton, J. P. (2006). Males have greater g: Sex differences in general mental ability from 100,000 17- to 18-year-olds on the Scholastic Assessment Test. *Intelligence*, *34*(5), 479-486.

Jelínek, M., Květon, P., & Vobořil, D. (2011). *Testování v psychologii. Teorie odpovědi na položku a počítačové adaptívni testování*. Praha: Grada Publishing.

Lemos, G. C., Abad, F. J., Almeida, L. S., & Colom, R. (2013). Sex differences on g and non-g intellectual performance reveal potential sources of STEM discrepancies. *Intelligence*, *41*(1), 11-18.

Li, H., & Stout, W. (1996). A new procedure for detection of crossing DIF. *Psychometrika, 61*(4), 647-677.

Lindsey, L. L. (2015). *Gender roles: A sociological perspective*. London,. New York: Routledge.

Linn, M., & Petersen, A. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development, 56*, 1479-1498.

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*, 240-254.

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.

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational Assessment*, *14*, 160-179.

McDonald, R. P. (2013). *Test theory: A unified treatment*. New York: Routledge.

Osterlind, S. J., & Everson, H. T. (2009) *Differential item functioning*. New York: Sage Publications.

Reynolds, C. R., & Suzuki, L. A. (2012) Bias in psychological assessment: An empirical review and recommendations. In I. B. Weiner, J. R. Graham, & J. A. Naglieri (Eds.), *Handbook of Psychology*, *Volume 10, Assessment Psychology, 2nd Edition* (pp. 82-113). New York: Wiley.

Roomaney, R., & Koch, E. (2013). An item and construct bias analysis of two language versions of a Verbal Analogies Scale. *South African Journal of Psychology*, *43*, 314-326.

Roussos, L., & Stout, W. (1996a). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355-371.

Roussos, L., & Stout, W. (1996b). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33,* 215-230.

Schaap, P. (2011). The differential item functioning and structural equivalence of a nonverbal cognitive ability test for five language groups. *SA Journal of Industrial Psychology*, *37*, 1-16.

Shealy, R., & Stout, W. (1993). A model based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Simos, P. G., Sideridis, G. D., Protopapas, A., & Mouzaki, A. (2011). Psychometric evaluation of a receptive vocabulary test for Greek elementary students. *Assessment for Effective Intervention*, *37*(1), 34-49.

Steensma, T. D., Kreukels, B. P., de Vries, A. L., & Cohen-Kettenis, P. T. (2013). Gender identity development in adolescence. *Hormones and Behavior*, *64*(2), 288-297.

Steinmayr, R., Bergold, S., Margraf-Stiksrud, J., & Freund, P. A. (2015). Gender differences on general knowledge tests: Are they due to Differential Item Functioning? *Intelligence*, *50*, 164-174.

Steinmayr, R., & Spinath, B. (2015). Intelligence as a potential moderator in the internal/external frame of reference model. An exploratory analysis. *Journal for Educational Research Online/Journal für Bildungsforschung Online*, *7*(1), 198-218.

Tutz, G., & Berger, M. (in press). Item focused trees for the identification of items in Differential Item Functioning. *Psychometrika*.

Voyer, D., Voyer, S., & Bryden, M. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin, 117*, 250-270.

Weiss, E. M., Ragland, J. D., Brensinger, C. M., Bilker, W. B., Deisenhammer, E. A., & Delazer, M. (2006). Sex differences in clustering and switching in verbal fluency tasks. *Journal of the International Neuropsychological Society*, *12*(04), 502-509.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa: Directorate of Human Resources Research and Evaluation, Department of National Defense.