Journal of Modern Research in English Language Studies
Print ISSN: 2676-5357
Online ISSN: 2676-5985

Item-Focused Trees Approach in Differential Item Functioning (DIF) Analysis: A Case Study of an EFL Reading Comprehension Test

Masoud Geramipour[1*]

[1]Assistant Professor, Department of Curriculum Studies and Educational Research, Faculty of Psychology and Education, Kharazmi University, Tehran, Iran,
mgramipour@yahoo.com

Abstract

Recent years have witnessed an increasing interest in providing new insights into modern language testing method targeting test bias. In this study, item-focused trees (IFT) approach was applied to identify uniform and non-uniform differential item functioning (DIF) of an English as a foreign language (EFL) reading comprehension test. The multistage cluster sampling method was employed to randomly choose a large sample of 4937 students who took the entrance exam of MA program in English studies. The reading comprehension section of the general English test including 20 items was selected for the IFT analysis. Three categorical and continuous DIF source variables including gender and academic background were concomitantly taken into account for the IFT analysis, which is capable of handling more than one variable with both binary and continuous measurement. Then, in the final stage of IFT analysis within a logistic regression framework, uniform and non-uniform DIF was analyzed using DIF tree package of R. The results showed that 10 items had uniform DIF in which 2 items had 2 joint DIF predictor variables (2 splits) and 8 items had only one split. Additionally, 6 splits and 5 non-uniform DIF items were found in non-uniform DIF analysis in which only 1 item had 2 simultaneous DIF source variables. Furthermore, gender and background knowledge had significant relationships with EFL reading comprehension. This study promises practical implications for addressing gender and background knowledge differences in EFL reading comprehension studies on the one hand, and impacting language testing methodology on the other.

Keywords: Differential Item Functioning (DIF), EFL Reading Comprehension Test, Language Testing, Logistic Regression, Item-focused Trees (IFT)

Received 01 July 2019 Accepted 21 December 2019
Available online 18 February 2020 DOI: 10.30479/jmrels.2019.11061.1379

© xxxx Imam Khomeini International University. All rights reserved.

1.Introduction

Differential item functioning (DIF) has been increasingly considered as a necessary standard in the field of testing in education and psychology (AERA, APA, NCME, 2014). DIF is also widely applied through a variety of methods in the context of EFL testing (e.g., Ahmadi & Bazarvand, 2016; Amirian, Alavi, & Fidalgo, 2014; Aryadoust, 2018; Barati & Ahmadi, 2010; Barati, Ketabi, & Ahmadi, 2006; Birjandi & Amini, 2007; Park, 2008; Ryan & Bachman, 1992). DIF is observed in case of controlling examinee's abilities, where the chance of answering a test item correctly depends on grouping variables (e.g., gender, ethnicity, etc.). Two types of uniform and non-uniform DIF may be observed in the item response data. Uniform DIF occurs if the probability of answering an item correctly is uniformly different in groups across the continuum of students' performance; however, non-uniform DIF between group categories differs across the continuum (Zumbo, 1999).

Methods for investigating DIF are basically implemented on the basis of observed score or latent trait models. Observed score methods like Mantel–Haenszel (Holland & Thayer, 1988), SIBTEST (Shealy & Stout, 1993), or logistic regression (Swaminathan & Rogers, 1990) use sum of raw scores for estimating the examinee's ability. The latent trait models such as confirmatory factor analysis (CFA)-based DIF (Stark, Chernyshenko, & Drasgow, 2006), item response theory (IRT)-based likelihood ratio test (Thissen, Steinberg, & Wainer, 1993), Lord's chi-square (Lord, 1980), or some recently introduced methods such as cognitive diagnostic modeling (CDM)-based DIF (Hou, de la Torre, & Nandakumar, 2014), and Rasch mixture models for DIF detection (Frick, Strobl, & Zeileis, 2015) use latent scores of the examinee's abilities for DIF assessment. There is even a hybrid method of DIF detection named iterative hybrid ordinal logistic regression/item response theory (Choi, Gibbons, & Crane, 2011) which integrates the two aforementioned approaches.

All of the reviewed DIF detection methods can only handle one categorical grouping variable; however, recent DIF methodologies have emerged that can simultaneously investigate the impact of more than one grouping variable. Moreover, continuous variables are also allowed to be analyzed as DIF sources, where no pre-determined subgroups are needed. These are item-focused trees (IFT) methods that may be used in the Rasch model (Tutz & Berger, 2016) or the logistic regression analysis (Berger & Tutz, 2016). However, the Rasch trees method has some limitations compared to the logistic regression-based method. First, the IRT Rasch model only incorporates the difficulty parameter, whereas the Rasch trees method is limited to uniform DIF analysis. Second, it treats the test items at

the global level without detecting the items directly responsible for DIF (Strobl, Kopf, & Zeileis, 2015). To date, only Aryadoust (2018) has applied the Rasch trees model to analyze DIF in an EFL reading comprehension test. Although the model of the IFT, which works under the logistic regression analysis framework, does not have the aforementioned shortcomings, it has not yet been employed to investigate DIF in reading assessment. Moreover, reading comprehension constitutes an integral section of the M.A. entrance exam. More specifically, reading comprehension test items represents a large proportion of the general English part of M.A. entrance exams in English studies. Therefore, it plays a critical role in determining the overall score of the examinees in general, and their final ranking in particular. Additionally, M.A. entrance exams in English studies could guide the future academic lives of the examinees; thus, it is important to reduce the test item bias, especially in reading comprehension section. Accordingly, language test makers need to cooperate with educational measurement experts to concentrate on DIF of test items in order to ensure that the examinees' test scores in reading comprehension section hinge on their overall language proficiency and reading comprehension strategies. Therefore, the present study aimed to probe the current state of the reading comprehension test items through DIF to further facilitate the process of test bias predictability and increase the test fairness potentials.

2. Review of the literature

2.1. The IFT Approach to DIF Investigation under the Logistic Regression Framework

Applying binary logistic regression analysis to detect DIF of binary test item responses (Swaminathan & Rogers, 1990) has a history of about four decades. In doing so, a linear logistic regression equation is set through a test item response as a dependent variable and sum of the test items score, grouping variables (e.g., gender), and their interaction as independent variables. Then, a chi-square statistic is provided to test the effect of the grouping variable and the interaction variable on uniform and non-uniform DIF, respectively (Zumbo, 1999). Alternatively, in a more practical way, there is a model-based approach to investigating DIF through logistic regression analysis in which the likelihood ratio tests are used to detect uniform and non-uniform DIF with their effect sizes (Zumbo & Thomas, 1997).

The binary logistic regression approach could be combined with the recursive partitioning (or item-focused trees) method to study uniform and non-uniform DIF (Strobl, Malley & Tutz, 2009). In this approach, the linear dimension of the logistic regression is replaced by fitting the tree-structured method. Building a tree in this approach means to find a partition (split) in

the DIF source variable successively, where each partition represents a node in the space of the DIF source variable. The number of splits is directly related to the scale of the DIF source variable (Berger & Tutz, 2016). The IFT part of the DIF method is basically rooted in the classification and regression trees method which was originally introduced by Morgan and Sonquist (1963) and further developed by Breiman, Friedman, Olshen, and Stone (1984). A very short introduction to the approach is also found in Hastie, Tibshirani, and Friedman (2009) and the psychometric issues are discussed in Strobl, Malley, and Tutz (2009).

It is possible in the IFT approach to simultaneously investigate the effect of more than one predictor variables (DIF sources) and their interactions, whereas continuous variables could be entered in the DIF analysis with no predetermined intervals. In order to build an IFT model under logistic regression, a closed logistic regression model is fitted to item response data and only the intercept of the equation is partitioned into a set of nodes, proxies, and rectangles depending upon the number and scale of the predictor variables. The terminal node represents the most important variable which induces DIF. The number of rectangles derived from each node also depends on the scale of the DIF predictor variables. Figure 1, which is resulted from the present study, shows an example of IFT diagram with two predictor variables and their interaction.

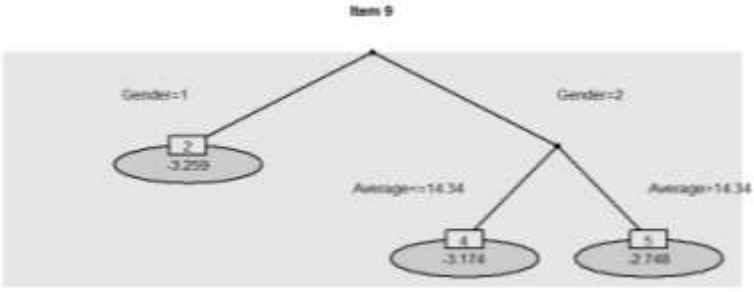

Figure 1. IFT Diagram with Gender and Academic Score Average

As shown in Figure 1, the terminal node of a test item is gender and its intercepts (gammas) for males and females respectively split into two rectangles. Then, category 2 (females) of the terminal node is again divided into two rectangles according to academic average score. Since it is an example of uniform DIF, the magnitude of the intercepts in each node shows that males have generally higher chance to answer item 9 of the test correctly. Females, in this case, have less chance especially when their academic score averages are greater than 14.34. Finally, if no statistically significant split is found in the first terminal node of an item, the fitted tree for an item is a constant and the item is free of DIF.

2.2. DIF Studies on EFL Reading Comprehension Testing

There are numerous DIF studies in the field of EFL reading comprehension testing. In part of a study conducted by Ryan and Bachman (1992), DIF of the vocabulary and reading comprehension section of TOEFL test was investigated across gender and language background. Using the Mantel-Haenszel approach for identifying DIF, they showed that one item in the reading comprehension section of TOEFL favors females. They also revealed that 11 items of the reading comprehension test are against non-native speaking examinees from different language groups. The gender-related result is almost in line with the study of Wainer and Lukhele (1997), but contrary to the result obtained in Carlton and Harris's (1992) study which concluded that the reading comprehension section of TOEFL is differentially against males. Moreover, the language-related DIF results of the study are in contrast with the comments made by Reid (1997) who believes that non-native speakers' performance on the reading section of TOEFL does not differ from native speakers.

Pae (2004) studied gender differences in the reading comprehension subtest of the 1998 Korean national entrance test for universities through IRT likelihood ratio-based approach to DIF investigation. He concluded that there is probably no relationship between item content and DIF type (uniform or non-uniform DIF); however, item content which requires making a logical inference is against females. Additionally, items covering impression, mood, and tone are easier for males. He also showed that all of the items with non-uniform DIF favor females though the items covered different content types. Moreover, using multiple-data analysis, Pae (2012) reported that the item types such as vocabulary and reference in reading comprehension tests mostly exhibit DIF against males.

Barati and Ahmadi (2010) applied an IRT-based approach to detect DIF on the Special English Test of the Iranian National University Entrance Exam (INUEE). They detected gender-related DIF in the reading comprehension subtest of the INUEE. Although subtests such as grammar, language function, cloze test, vocabulary, and word order differentially favor females and males, the researchers reported that the reading section of INUEE favors males and females equally. Amirian, Alavi, and Fidalgo (2014) replicated the same study for the reading comprehension section of the University of Tehran English Proficiency Test (UTEPT). Finally, in another study in the Iranian EFL testing context, Ahmadi and Jalili (2014) found some sources of DIF on a reading comprehension test by applying IRT and logistic regression methods. They reported that different variables such as examinees' text familiarity, gender, topic or text interest, guessing, and the

social variables including location, income, and educational status may induce DIF on a reading comprehension test.

As reviewed above, a discrete variable like gender has usually been supposed as a source of DIF. Other nominal variables including social and background knowledge variables (race, language, academic grade or average score, etc.) may also be considered for DIF analysis. It means that besides gender, the background knowledge of the examinees affects their reading comprehension ability (Zhang & Shanshan, 2011). However, due to the limitation of the previous DIF methodology investigation, continuous variables like age or academic average score with no specific cut-point could not be entered in DIF analysis of reading comprehension tests. Furthermore, the interaction of polytomous and continuous variables in DIF analysis could not be studied due to the same reason.

Thus, to date, no study has been conducted to simultaneously detect gender and background knowledge-related DIF under IFT logistic regression in EFL reading assessment. Therefore, the main purpose of the present study is to do the analysis with three DIF source variables including two polytomous (gender and undergraduate major field of study) and one continuous variable (undergraduate academic average score) by relying on real data of the reading comprehension section of the MA English Test of the Iranian National University Entrance Exam. Hence, the main research questions that guided the study are:

1. Which of the reading comprehension test items have DIF based on the source variables?

2. How do items interact with each other through applying the logistic regression-based IFT methodology?

3. Method

The reading comprehension section of the Iranian National University Entrance Examination was analyzed in this study. The test is composed of 3 reading passages with 20 multiple choice items which are part of a high stakes test held annually to admit the candidates to MA programs in English Language studies. The test is designed for students with a bachelor's degree who aim to pursue education in master degree in state universities. The test has two sections including content knowledge and general English. The general English section is of four sections of structure (10 items), vocabulary (20 items), cloze passage (10 items), and the reading comprehension section (20 items) which was chosen for IFT analysis. The reading comprehension section includes 3 passages with 20 items almost evenly distributed in each passage (see Appendix 1). A sample of 4937 examinees who took the test in 2015 was randomly selected through multistage cluster sampling. In doing

so, one booklet was randomly selected. Then, random cluster samples were proportionately drawn from the population based on the participants' *gender* and *undergraduate field of study*. Table 1 shows the proportion of the three demographic variables of the sample which were used for IFT analysis.

Table1

Descriptive statistics

Gender		Undergraduate field		Average score	
Male	Female	English Studies	Others	Mean	Standard Deviation
25%	75%	78%	22%	15.34	1.65

At last, DIFtree package of R (Berger, 2019) was employed to analyze DIF of the reading comprehension test under the logistic regression-based IFT.

4. Results and Discussion

4.1. Results

The data were entered in DIFtree package of R. Then, uniform and non-uniform DIF was analyzed through logistic regression approach to IFT.

4.1.1. Uniform DIF Analysis

For the purpose of saving time during the analysis in the R system, only uniform DIF was commanded to the R package at the first stage of the IFT analysis. The results of the uniform DIF detection through the IFT are shown in Table 2.

The IFT analysis was done through 1000 permutations and a total number of 12 splits were found in the reading comprehension test items. Then, as shown in Table 2, 10 items were shown to have uniform DIF with different source variables in which 2 items had 2 DIF source variables (2 splits) and 8 items had only one split. No items were detected to have uniform DIF with 3 source variables. As could be judged through the variable and the *gamma* column of Table 2, uniform DIF on items 2, 7, 8, 12, 17 and 20 were only induced by the *undergraduate academic average score* variable. The *gender* variable was the only uniform DIF source for item 18 and *undergraduate field of study* variable induced uniform DIF on item 13 of the reading comprehension test. At last, it was observed that items 9 and 10 had simultaneously two uniform DIF variable sources. *Gender* and *undergraduate academic average score* were related with uniform DIF on item 9 and uniform DIF on item 10 were simultaneously induced by *undergraduate field of study* and *undergraduate academic average score* variables. It is worth noting that the *threshold* column determines the cut-off score for the continuous or dichotomous DIF variable sources and specify the

point at which the scores are split up to make the DIF grouping variable for further uniform or non-uniform DIF analysis through the IFT approach.

Table2

Summary of Uniform DIF Item Statistics

tem	eta	IF	Variables	Number of Splits	Threshold	Gamma
	.27	o	-	-	-	γ_1= - 3.10
	.35	es	Undergraduate Academic Average Score	1	δ_1 =16.54	γ_2= -2.79 γ_3= -2.51
	.35	o	-	-	-	γ_1= - 4.16
	.33	o	-	-	-	γ_1= - 3.80
	.46	o	-	-	-	γ_1= - 3.23
	.29	o	-	-	-	γ_1= -4.18
	.30	es	Undergraduate Academic Average Score	1	δ_1 =17.56	γ_2= -1.78 γ_3= -2.14
	.46	es	Undergraduate Academic Average Score	1	δ_1 =14.83	γ_2= -1.95 γ_3= -1.47
	.38	es	Gender- Undergraduate Academic Average Score	2	δ_1 =1 δ_2 =14.34	γ_2= -3.26 γ_4= -3.17 γ_5= -1.05
0	.32	es	Undergraduate Field of study - Undergraduate Academic Average Score	2	δ_1 =0 δ_2 =13.52	γ_2= -1.41 γ_4= -1.41 γ_5= -1.05
1	.42	o	-	-	-	γ_1= - 3.50
2	.28	es	Undergraduate Academic Average Score	1	δ_1 =14.83	γ_2= -3.06 γ_3= -3.49
3	.35	es	Undergraduate Field of study	1	δ_1 =0	γ_2= -3.38 γ_3= -3.73
4	.30	o	-	-	-	γ_1= -3.43
5	.25	o	-	-	-	γ_1= -2.15
6	.42	o	-	-	-	γ_1= -4.33
7	.27	es	Undergraduate Academic Average Score	1	δ_1 =14.34	γ_2= -2.97 γ_3= -3.46
8	.37	es	Gender	1	δ_1 =1	γ_2= -3.79 γ_3= -3.39
9	.39	o	-	-	-	γ_1= -4.37
0	.27	es	Undergraduate Academic Average Score	1	1 5.77	γ_2= -3.75 γ_3= -4.28

4.1.2. Non-uniform DIF analysis

In the next session of the IFT analysis by the DIFtree package, the non-uniform DIF analysis was commanded to the R package. The results of the non-uniform DIF analysis of the reading comprehension test items through the logistic regression-based IFT are shown in Table 3.

Table3

Summary of Non-uniform DIF Item Statistics

Item	Alpha	DIF	Variables	Number of Splits	Threshold	Gamma
1	α_1=0.27	No	-	-	-	γ_1= -3.10
2	α_1=0.36	No	-	-	-	γ_1= -2.75
3	α_2=0.29 α_3=0.39	Yes	Gender	1	δ_1=1	γ_2= -3.52 γ_3= -4.46
4	α_1=0.33	No	-	-	-	γ_1= -3.80
5	α_1=0.46	No	-	-	-	γ_1= -3.23
6	α_1=0.29	No	-	-	-	γ_1= -4.18
7	α_2=0.33 α_4=0.18 α_5=0.29	Yes	Undergraduate Academic Average Score-Gender	2	δ_1=15.50 δ_2=1	γ_2= -1.94 γ_4= -1.31 γ_5= -1.78
8	α_1=0.46	No	-	-	-	γ_1= -1.70
9	α_2=0.33 α_3=0.41	Yes	Gender	1	δ_1=1	γ_2= -2.91 γ_3= -3.03
10	α_1=0.32	No	-	-	-	γ_1= -1.18
11	α_1=0.42	No	-	-	-	γ_1= -3.50
12	α_1=0.26	No	-	-	-	γ_1= -3.25
13	α_2=0.62 α_3=0.34	Yes	Undergraduate Academic Average Score	1	δ_1=12.70	γ_2= -4.28 γ_3= -3.65
14	α_1=0.30	No	-	-	-	γ_1= -3.43
15	α_1=0.25	No	-	-	-	γ_1= -2.15
16	α_1=0.42	No	-	-	-	γ_1= -4.33
17	α_1=0.26	No	-	-	-	γ_1= -3.24
18	α_1=0.36	No	-	-	-	γ_1= -3.44
19	α_1=0.39	No	-	-	-	γ_1= -4.37
20	α_2=0.18 α_3=0.28	Yes	Undergraduate Field of study	1	δ_1=0	γ_2= -3.17 γ_3= -4.10

The IFT analysis run with 1000 permutations and a total number of 6 splits and 5 non-uniform DIF items were found in the reading comprehension test. According to *alpha* and *variable* columns in Table 3, items 3 and 9 of the reading test had 1 split and possess gender-related non-uniform DIF. Items 13 and 20 also had 1 split and non-uniform DIF across *undergraduate academic average score* and *undergraduate field of study* variables, respectively. Only 1 item (item 7) out of the 6 non-uniform DIF items had 2 variable sources. It had non-uniform DIF against both *undergraduate academic average score* and *gender* variables.

4.2. Discussion

Reading assessment has several objectives in the EFL learning context that include assessing to encourage the students, monitoring learning and providing feedback, diagnosing reading problems, and evaluating proficiency (Nation, 2008). Undoubtedly, fulfilling these goals depends on the reading test validity. Social and consequential aspects of testing have been taken into account extensively for estimating the language test validity (Chapelle, 1999). In doing so, DIF has been proposed as a serious threat to test validity in the last two decades (Gómez-Benito, Sireci, Padilla, Hidalgo, & Benítez, 2018). Accordingly, DIF can investigate the consequences of language testing such as post-hoc reading comprehension test bias to gender.

Striving to spread modern methodologies in language testing (see Geramipour & Shahmirzadi, 2018, 2019), this study also aimed to show an application of item-focused trees (IFT) to study uniform and non-uniform DIF in an EFL reading comprehension test based on the classic logistic regression analysis. The main advantages of the IFT method are to manage more than one DIF source and continuous variables. Accordingly, for the first time, it was decided to apply this methodology in EFL language testing. In doing so, a high stakes EFL reading comprehension test with 20 items was selected for the analysis. Then, three DIF source variables including *gender*, *undergraduate academic average score,* and *undergraduate field of study* were taken into account in the process of IFT analysis. All of the possible interactions among the demographic variables were investigated in one session. Finally, two types of uniform and non-uniform DIFs were detected in the reading comprehension test by at least double DIF source variables.

Looking forward to detecting gender-related DIF in the uniform DIF analysis, item 18 of the reading comprehension test showed only a single effect of gender as seen in Figure 2.

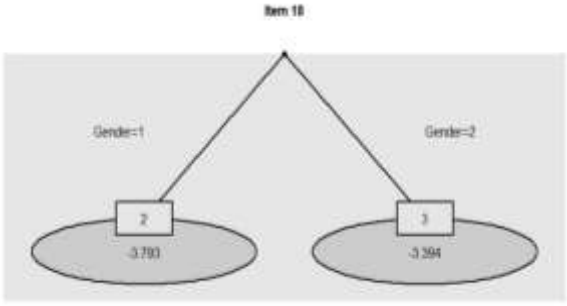

Figure 2. Gender-related Uniform DIF Tree of Item 18 of the Reading Section

Figure 2 shows that male examinees are more likely to answer the item 18 correctly. Considering the key answer to the item reveals that the correct answer to the item had an exercise-related content and may indirectly

remind a kind of sport for men. Then, males may be more attracted to the test option because men are naturally more interested in and busy with sport (Park, 2008). Therefore, ignoring their real ability, they may choose the correct answer more likely compared to females. This result confirms the findings reported by Park (2008) who states that the test contents related to sport favor males.

Looking for a single DIF effect of the *undergraduate academic average score* variable, 6 items (items 2, 7, 8, 12, 17, and 20) were detected to have uniform DIF against it. As the DIF source variable is continuous, the splits found on the average scores show which cut points are biased against the examinees. For instance, Figure 3 shows that students with an undergraduate average score less than or equal to 16.54 have the higher chance to answer the test item correctly.

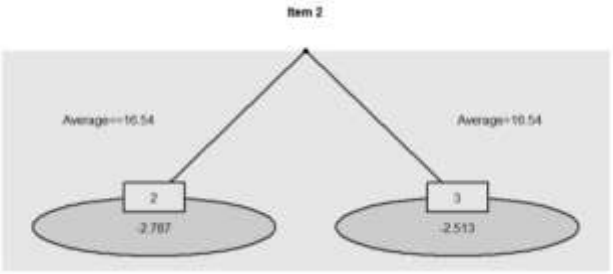

Figure 3. Uniform DIF Tree of Item 2 Based on the Undergraduate Academic Average Score

Academic average score has been considered to be an important dependent variable among college students in some recent studies (Buckingham, 2013; García, López, Icaran, & Burgos, 2014; Tien & Fu, 2008); however, the independent role of the variable as a DIF source was investigated for the first time in the present study. Therefore, the reasons why such a new variable induces DIF in an EFL reading test may call for qualitative investigation.

A single uniform DIF effect of *undergraduate field of study* variable was also detected in item 13 of the reading comprehension test as seen in Figure 4.

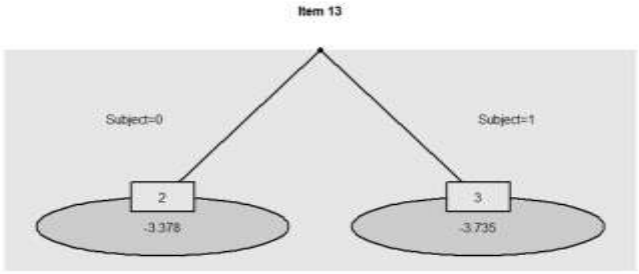

Figure 4. Uniform DIF tree of item 13 based on the single DIF effect of *undergraduate field of study* variable

Most of the students, who took part in the national MA entrance exam of English studies, held a BA in English. The results of the present study indicate that, to some extent, the reading content of such a high stakes test may not be proper for the candidates of other majors.

At last, in the phase of uniform DIF analysis, there are also 2 items which are induced by 2 DIF source variables simultaneously. Item 9 is affected by *gender* and *undergraduate academic average score* variables. *Undergraduate field of study* and *undergraduate academic average score* are responsible for DIF in item 10 of the reading test. Figure 5 shows the DIF tree of item 10 through inducing DIF sources.

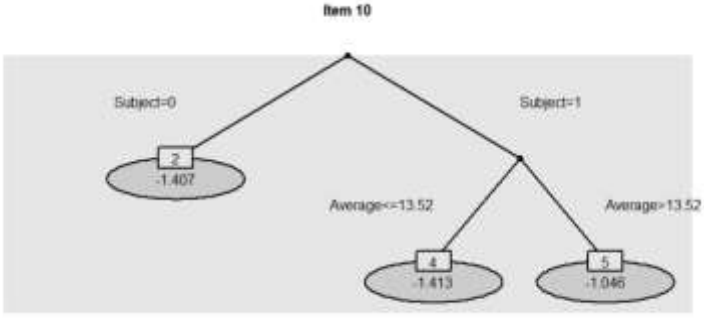

Figure 5. Uniform DIF Tree of Item 10 with 2 DIF Variables

The item shows that the examinees with a related BA degree (subject=1) and an undergraduate academic average score equal or less than 13.52 (average<=13.52) had higher chances to give the correct answer to item 10 of the reading test correctly.

In the non-uniform DIF analysis stage, 2 items (items 3 and 9) were biased toward the *gender* variable. Figure 6 shows the item-focused tree for item 3 of the reading comprehension test.

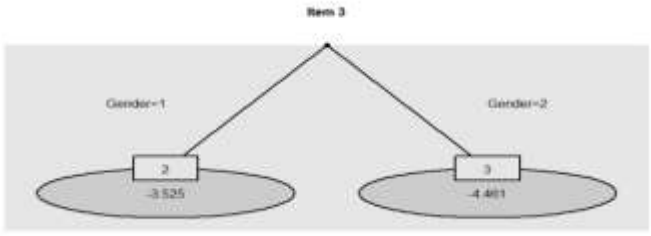

Figure 6. Gender-related Non-uniform DIF Tree of Item 18

Non-uniform DIF, in this case, means that female examinees with higher abilities surpass male counterparts; however, in examinees with average and low ability levels, male candidates perform better in the reading test items. Although most of the research about gender differences in reading had uniformly compared male and female ESL college students (Brantmeier, 2003; Dreyer & Oxford, 1996; Goh & Foong, 1997; Green & Oxford, 1995; Oxford & Nyikos, 1989; Phakiti, 2003; Poole, 2005; Shmais, 2003; Szoke & Sheorey, 2002; Young & Oxford, 1993), none of them were methodologically capable of investigating the non-uniform gender differences among them. Moreover, as shown in Table 2, items 13 and 20 had non-uniform DIF against the *undergraduate academic average score* and *undergraduate field of study* variables, respectively. Figure 7 shows the item focused tree of the item 20 against its DIF source.

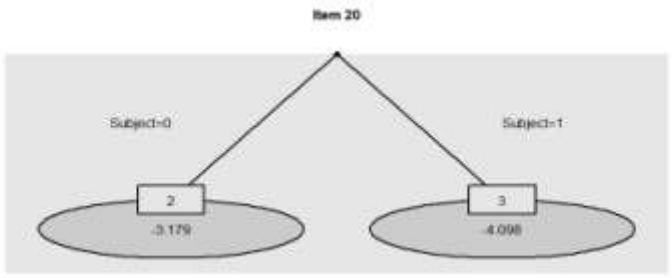

Figure 7. Gender-related Non-uniform Item Focused Tree of Item 20

Non-uniform DIF in item 20 means that only higher ability candidates majoring in a related field in English studies had higher chance to answer the item and the scenario is totally different in lower levels.

At last, as seen in Figure 8, only one item (item 7) had a non-uniform DIF with 2 simultaneous DIF sources including *undergraduate academic average score* and *gender* variables.

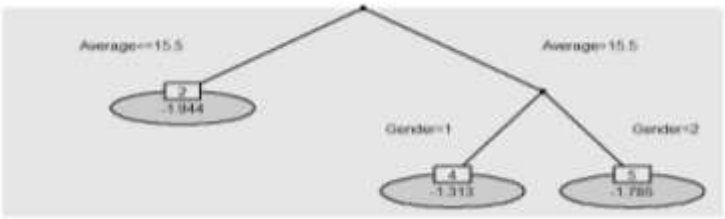

Figure 8. Non-uniform DIF Tree of Item 7 with 2 Simultaneous DIF Source Variables

The same gender-related scenario, like the item 3 and 9 of the reading test, was observed in item 7. However, it only occurs for candidates with an undergraduate average score higher than 15.50. In fact, the candidates' average scores have a screening role for further non-uniform DIF analysis based on the gender variable.

Overall, reader background and reading text characteristics are two important variables, which affect the process of reading comprehension (Woolley, 2011). Considering the former factor, the results of this study showed that gender, academic average score, and the BA students' fields of study are related to their test performance. Although the role of gender is evident with reference to the literature, the other background variables in this study need to be more scrutinized. DIF sources such as academic average scores and the examinees' fields of study are background variables that are directly related to the background knowledge and are consequently associated with EFL reading comprehension.

Finally, in line with the results of this study, the impact of background knowledge on the student's reading comprehension has been extensively addressed in the literature (e.g. Eason, Goldberg, Young, Geist, & Cutting, 2012; Hudson, 1998; Kobayashi & Rinnert, 1996; Sadeghi, 2007; Woolley, 2011). Moreover, it is worth noting that the relationship is bidirectional. More specifically, background knowledge is related to reading ability and English language proficiency conversely affects background knowledge such as academic average score (Maleki & Zangani, 2007).

5. Conclusion and Implications

Reading comprehension section plays a determining role in the examinees' acceptance. Therefore, there is a call for exploring the DIF factors leading to test item bias and taking the statistical measures to reduce the possibility of examinees' distinction based on non-test factors. In the reading comprehension section, the test item developers need to exert more sensitivity due to the biases that may be caused by the examinees' background linguistic-content knowledge and personal orientations than their reading comprehension skills.

After all, for the first time in language testing literature, the present study showed the successful application of logistic regression-based IFT method (Berger & Tutz, 2016) to detect DIF in a high stakes EFL reading test. Future research may target the grammar or cloze test sections to apply the IFT method in EFL testing. Moreover, further research may employ simulation studies mimical to the EFL reading data parameters to investigate the hit rate (true positive rate of DIF detection) of the IFT method. It is hoped this study could make a contribution to advance methodology in language testing on the one hand, and provide hints to promote reading studies on the other.

References

Ahmadi, A., & Bazarvand, D. A. (2016). Gender differential item functioning on a national field-specific test: The case of PhD entrance exam of TEFL in Iran. *Iranian Journal of Language Teaching Research*, *4*(1), 63-82.

Ahmadi, A., & Jalili, T. (2014). A confirmatory study of differential item functioning on EFL reading comprehension. *Applied Research on English Language, 3*(6), 55-68.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, Joint Committee on Standards for Educational, & Psychological Testing (US). (2014). *Standards for educational and psychological testing*. Amer Educational Research Assn.

Amirian, S.M.R, Alavi, S.M., & Fidalgo, A.M. (2014). Analyzing gender differences with an English proficiency test in EFL context. *Iranian Journal of Language Testing*, *4*(2), 187-203.

Aryadoust, V. (2018). Using recursive partitioning Rasch trees to investigate differential item functioning in second language reading tests. *Studies in Educational Evaluation*, *56*, 197-204.

Barati, H., & Ahmadi, A. R. (2010). Gender-based DIF across the subject area: A study of the Iranian National University Entrance Exam. *Journal of Teaching Language Skills, 2*(3), 1-22.

Barati, H., Ketabi, S., & Ahmadi, A. (2006). Differential item functioning in high-stakes tests: The effect of field of study. *Iranian Journal of Applied Linguistics*, *19*(2), 27-42.

Berger, M. (2019). DIFtree: Item Focused Trees for the Identification of Items in Differential Item Functioning. R package version 3.1.4. *https://CRAN.R-project.org/package=DIFtree*

Berger, M., & Tutz, G. (2016). Detection of uniform and nonuniform differential item functioning by item-focused trees. *Journal of Educational and Behavioral Statistics*, *41*(6), 559-592.

Birjandi, P., & Amini, M. (2007). Differential item functioning (test bias) analysis paradigm across manifest and latent examinee groups (on the construct validity of IELTS) [special issue]. *Journal of Human Sciences*, *55*, 1–20.

Brantmeier, C. (2003). Does gender make a difference? Passage content and comprehension in second language reading. *Reading in a Foreign Language*, *15*(1), 1-27.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, J. C. (1984). *Classification and regression trees*. Monterey, CA: Wadsworth.

Buckingham, R. S., & Bush, S. R. (2013). Predictors of academic success for students at the Michigan College of Optometry. *Optometric Education, 38*(3), 92-99.

Eason, S. H., Goldberg, L. F., Young, K. M., Geist, M. C., & Cutting, L. E. (2012). Reader–text interactions: How differential text and question types influence cognitive skills needed for reading comprehension. *Journal of educational psychology*, *104*(3), 515-528.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, *19*, 254-272.

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), 1-30.

Dreyer, C., & Oxford, R. L. (1996). Learning strategies and other predictors of ESL proficiency among Afrikaans speakers in South Africa. *Language Learning Strategies around the World: Cross-Cultural Perspectives*, 61-74.

Frick, H., Strobl, C., & Zeileis, A. (2015). Rasch mixture models for DIF detection: A comparison of old and new score specifications. *Educational and Psychological Measurement*, *75*(2), 208-234.

García, O., López, F., Icaran, E., & Burgos, S. (2014). Relationship between general intelligence, competences and academic achievement among university students. *Personality and Individual Differences, 60,* 48-78.

Geramipour, M., & Shahmirzadi, N. (2018). Application and comparison of multidimensional latent class item response theory on clustering items in comprehension tests. *The Journal of Asia TEFL, 15(2)*, 479-490.

Geramipour, M., & Shahmirzadi, N. (2019). A Gender–Related Differential Item Functioning Study of an English Test. *Journal of Asia TEFL*, *16*(2), 674.

Goh, C., & Foong, K. P. (1997). Chinese ESL students' learning strategies: A look at frequency, proficiency, and gender. *Hong Kong Journal of Applied Linguistics*, *2*(1), 39-53.

Gómez-Benito, J., Sireci, S., Padilla, J. L., Hidalgo, M. D., & Benítez, I. (2018). Differential item functioning: Beyond validity evidence based on internal structure. *Psicothema*, *30*(1), 104-109.

Green, J. M., & Oxford, R. (1995). A closer look at learning strategies, L2 proficiency, and gender. *TESOL Quarterly*, 29(2), 261-297.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning*. New York, NY: Springer-Verlag.

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

Hou, L., de la Torre, J., & Nandakumar, R. (2014). Differential item functioning assessment in cognitive diagnostic modeling: Application of the Wald test to investigate DIF in the DINA model. *Journal of Educational Measurement*, *51*, 98-125.

Hudson, T. (1998). Theoretical perspectives on reading. *Annual Review of Applied Linguistics*, *18*, 43-60.

Kobayashi, H., & Rinnert, C. (1996). Factors affecting composition evaluation in an EFL context: Cultural rhetorical pattern and readers' background. *Language learning*, *46*(3), 397-433.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Maleki, A., & Zangani, E. (2007). A survey on the relationship between English language proficiency and the academic achievement of Iranian EFL students. *Asian EFL Journal, 9(1)*, 86-96.

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association, 58*, 415-435.

Nation, I. S. (2008). *Teaching ESL/EFL reading and writing*. New York: Routledge.

Oxford, R., & Nyikos, M. (1989). Variables affecting choice of language learning strategies by university students. *The Modern Language Journal*, *73*(3), 291-300.

Pae, T. I. (2004). Gender effect on reading comprehension with Korean EFL learners. *System*, *32*(2), 265-281.

Pae, T. (2012). Causes of gender DIF on an EFL language test: A multiple-data analysis over nine years. *Language Testing*, *29*, 533-554.

Park, G.P. (2008). Differential item functioning on an English listening test across gender. *TESOL Quarterly, 42* (1), 115-123.

Phakiti, A. (2003). A closer look at gender and strategy use in L2 reading. *Language Learning*, *53*, 649-702.

Poole, A. (2005). Gender differences in reading strategy use among ESL college students. *Journal of College Reading and Learning*, *36* (1), 7-20.

Reid, J. M. (1997). Which non-native speaker? Differences between international students and US resident (language minority) students. *New directions for teaching and learning*, *1997*(70), 17-27.

Ryan, E, K. & Bachman, F.L. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing*, *9*(1), 12-29.

Sadeghi, K. (2007). The key for successful reader-writer interaction: Factors affecting reading comprehension in L2 revisited. *Asian EFL Journal*, *9*(3), 198-220.

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*(2), 159-194.

Shmais, W. A. (2003). Language learning strategy use in Palestine. *TESL-EJ*, 7(2), 1-17.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests. *Psychological Methods*, *14*, 323-348.

Swaminathan, H., & Rogers, J. (1990). Detecting DIF using logistic regression procedures. *Journal of Educational Measurement*, *27*, 361-370.

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

Strobl, C., Kopf, J., & Zeileis, A. (2015). Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika*, *80*(2), 289-316.

Szoke, E., & Sheorey, R. (2002). A comparative study of the learning strategies of Hungarian and Russian College Students. *Novelty*, *9*(3), 23-36.

Thissen, D. J., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Erlbaum.

Tien, F. F., & Fu, T. T. (2008). The correlates of the digital divide and their impact on college student learning. *Computers and Education*, *50*(1), 421-436.

Tutz, G., & Berger, M. (2016). Item-focussed trees for the identification of items in differential item functioning. *Psychometrika*, *81*(3), 727-750.

Woolley, G. (2011). Reading comprehension: Assisting children with dificulties. Dordrecht:Springer.

Young, D. J., & Oxford, R. L. (1993). Attending to learner reactions to introductory Spanish textbooks. *Hispania*, 76, 593-605.

Zhang, Y., & Shanshan, H. (2011). Background knowledge and reading comprehension. *In International Conference on Computer Technology and Development, 3rd (ICCTD 2011)*. ASME Press.

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, ON: Directorate of Human Resources Research and Evaluation, Department of National Defense*.

Zumbo, B. D., & Thomas, D. R. (1997). A measure of effect size for a model-based approach for studying DIF. *Prince George, Canada: University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science*.

Appendix 1: Reading comprehension test passages and items

PART D: Reading Comprehension

Directions: Read the following three passages and decide which choice (1), (2), (3), or (4) best answers each question. Then mark the correct choice on your answer sheet.

PASSAGE 1:

As the previous two sections have demonstrated, the comparative method relies quite heavily on linguistic evidence to establish genetic relationships among languages.

However, non-linguistic evidence, such as historical information and archeological evidence, can supplement linguistic evidence to help in the classification of languages, especially to help date the origins of proto-languages for which no linguistic evidence exists. The farther back in time one goes, the more sketchy historical information about languages and their speakers becomes. This explains why we know so little about either Proto-Indo-European or Proto-Germanic. In the case of Proto-Indo-European, while the reconstruction of this language has, as Olson (2003:142) comments, provided considerable information concerning how speakers of PIE lived, we currently have no hard evidence about "when and where these people lived." For this reason, we can only guess when this language might have initially been spoken, who spoke it, and how migrations of PIE speakers led to the development of sub-families of PIE (e.g. Proto-Germanic). Dixon (1997:48) states that although the common consensus is that PIE began around 6,000 years ago, he notes that others have provided evidence that the language could have originated up to 10,500 years ago.

We can also only speculate about where PIE was initially spoken. The most widely accepted view of the origins of PIE is the Kurgan Hypothesis, which was originally

proposed by the archeologist Marija Gimbutas (1956). This hypothesis places the original speakers of PIE just north of the Black Sea c. 6,000 years ago. Through a series of migrations, these speakers spread their language all the way to Europe, spawning over time the various sibling languages of PIE, including Proto-Germanic. Archeological and linguistic evidence suggests that original speakers of PIE were warriors who rode horses as they made their way to Europe. An alternative but much less widely accepted hypothesis is Renfrew's (1987) farming-dispersal hypothesis.

41- Which of following is most probably the topic of the paragraph immediately following this passage?

1) Further evidence to solidify the Kurgan Hypothesis
2) More information about how the original PIE speakers lived
3) Another speculation concerning where PIE was initially spoken
4) Scientific ways to use to verify the claims made about the origin of PIE

42- Which of the following does the passage suggest as the primary source of evidence to establish genetic relationships?

1) Historical information and archeological evidence
2) Investigation of languages with no linguistic evidence
3) Reconstruction of an ancient language based on present-day linguistic models
4) A comparative study of different languages in terms of the available linguistic evidence

43- It can be understood from the passage that the view according to which PIE began up to 10,500 years ago is a view which ..... .

1) not all the concerned scientific community members accede to
2) gives the most credence to the tenets of the so-called the Kurgan Hypothesis
3) is the most significant view challenging the common consensus that PIE began around 6,000 years ago
4) is founded on more recent evidence and hence is more reliable than previous related conjectures in this regard

44- According to the passage, Proto-Germanic is ..... .

1) a language that started at the same time as PIE
2) actually an offshoot of what we refer to as PIE
3) a language whose time and place of origin is already well established
4) a sub-family of PIE about whose time and place we are more certain than we are about PIE itself

45- According to the passage, it is NOT true that ..... .

1) PIE may have been taken to Europe by some warriors
2) Dixon was the first to contend that Proto-Germanic is a sibling language of PIE
3) Renfrew's farming-dispersal hypothesis is less widely accepted than the Kurgan Hypothesis
4) the comparative method mentioned in the passage is a method mostly drawing on linguistic evidence as its source of evidence

46- Which of the following best reveals the author's attitude about the controversy regarding the origin of PIE and its sibling languages?

1) Calculated indifference
2) Perfunctory dismissal
3) Disapproval
4) Impartiality

PASSAGE 2:

Intelligence has always been tricky to quantify, not least because it seems to involve most of the brain and so is almost certainly not one "thing". Even so, scores across different kinds of IQ tests have long shown that people who do particularly well–or badly–on one seem to do similarly on all. This can be crunched into a single general intelligence factor, or "g", which correlates pretty well with academic success, income, health and lifespan.

So more intelligence is clearly a good thing, but where does it come from? A large part of the answer seems to be genetics. In 1990, the first twin studies showed that the IQ scores of identical twins raised apart are more similar to each other than those of non-identical twins raised together. Since then a few genes have been linked to IQ, but all of them seem to have a tiny effect and there are probably thousands of genes involved.

That doesn't mean the environment plays no part, at least in childhood. While the brain is developing, everything from diet to education and stimulation plays a huge part in developing the brain structures needed for intelligent thought. Children with a bad diet never fulfill their genetic potential.

But even for educated and well-fed children, the effects of environment wear off over time. By adulthood genes account for 60 to 80 per cent of the variance in intelligence scores, compared with less than 30 per cent in young children. Whether we like it or not, we get more like our close family members the older we get.

So if genes play such a big part, is there anything adults can do to improve IQ? The good news is that one type of intelligence keeps on improving throughout life. Most researchers distinguish between fluid intelligence, which measures the ability to reason, learn and spot patterns, and crystallized intelligence, the sum of all our knowledge so far. Fluid intelligence slows down with age, but crystallized intelligence doesn't. So while we all get a little slower to the party as we get older, we can rest assured that we are still getting cleverer.

47- What is the best title for the passage?

1) Intelligence: Nature and Improvement
2) Intelligence: An Unknown Quantity
3) "G" Factor: A Myth or A Reality?
4) Ways to Enhance Intelligence

48- What does the word "all" in paragraph 1 refer to?

1) people
2) scores
3) tests
4) intelligence components

49- The author mentions the twin studies in paragraph 2 mostly in order to ..... .

1) uncover the complexity of understanding the nature of intelligence
2) support an earlier assertion about the genesis of intelligence
3) cast doubt on the validity of such things as IQ test scores
4) prove that factors other than genetics contribute to intelligence

50- The word "those" in paragraph 2 refers to ..... .

1) twins
2) studies
3) genes
4) IQ scores

51- Which of the following is TRUE about the effect of genetics on intelligence, according to the passage?

1) Genetics exerts its influence on children's intelligence, with the environment playing no role.
2) The level of intelligence is more determined by genes in children than it is in adults.
3) The impact of the environment on the level of intelligence tends to diminish as one ages.
4) The variance in intelligence scores that genes account for in adulthood does not vary from person to person.

52- Which of the following best describes the author's attitude towards the attempts an adult can make to improve their intelligence?

1) Conditional sanguinity
2) Unjustified hope
3) Unguarded optimism
4) Total frustration

53- The passage provides sufficient information to answer which of the following questions?

1) Why does the ability to reason, learn and spot patterns wear off over time?
2) Why does the author state, "We get more like our close family members the older we get,"?
3) What is the correlation coefficient between intelligence and academic success, income, health and lifespan?
4) What are some of the efficacious strategies one can employ in order to ameliorate crystallized intelligence?

PASSAGE 3:

Learning is what your brain does naturally. In fact, it has been doing it every waking minute since about a month before you were born. It is the process by which you acquire and store useful (and useless) information and skills. Can you make it more efficient?

The answer lies in what happens physically as we learn. As it processes information, the brain makes and breaks connections, growing and strengthening the synapses that connect neurons to their neighbors, or shrinking them back. When we are actively learning, the making of new connections outweighs the breaking of old ones. Studies in rats have shown that this rewiring process can happen very quickly–within hours of learning a skill such as reaching through a hole to get a food reward. And in some parts of the brain, notably the hippocampus, the brain grows new brain cells as it learns.

But once a circuit is in place, it needs to be used if it is going to stick. This largely comes down to myelination—the process whereby a circuit that is stimulated enough times grows a coat of fatty membrane. This membrane increases conduction speed, making the circuit work more efficiently.

What, then, is the best way to learn things and retain them? The answer won't come as a huge surprise to anyone who has been to school: focus attention, engage working memory and then, a bit later, actively try to recall it.

Alan Baddeley of the University of York, UK, says it is a good idea to test yourself in this way as it causes your brain to strengthen the new connection. He also suggests consciously trying to link new bits of information to what you already know. That makes the connection more stable in the brain and less likely to waste away through underuse.

The learning process carries on for life, so why is it so much harder to learn when we reach adulthood? The good news is that there seems to be no physiological reason for the slowdown. Instead, it seems to be a lot to do with the fact that we simply spend less

time learning new stuff, and when we do, we don't do it with the same potent mix of enthusiasm and attention as the average child.

Part of the problem seems to be that adults know too much. Research by Gabriele Wulf at the University of Nevada, Las Vegas, has shown that adults tend to learn a physical skill, like hitting a golf ball, by focusing on the details of the movement. Children, however, don't sweat the details, but experiment in getting the ball to go where they want. When Wulf taught adults to learn more like kids, they picked up skills much faster.

This also seems to be true for learning information. As adults we have a vast store of mental shortcuts that allow us to skip over details. But we still have the capacity to learn new things in the same way as children, which suggests that if we could resist the temptation to cut corners, we would probably learn a lot more.

A more tried-and-tested method is to keep active. Ageing leads to the loss of brain tissue, but this may have a lot to do with how little we hare about compared to youngsters. With a little exercise, the brain can spring back to life. In one study, 40 minutes of exercise three times a week for a year increased the size of the hippocampus– which is crucial for learning and memory. It also improved connectivity across the brain, making it easier for new things to stick.

54- The main purpose of the passage is to ……..

1) portray what we can do in order to streamline the act of learning
2) explain the relationship between age and learning strategies
3) compare children and adults with regard to learning
4) delineate the steps involved in the learning process

55- The word "them" in paragraph 2 refers to ……..

1) old connections
2) neurons and their neighbors
3) neurons
4) synapses

56- Which of the following is TRUE about the rewiring process mentioned in the passage?

1) It is a process that solely takes place in the hippocampus.
2) It is triggered when one attempts to learn something new.
3) It is a process that automatically starts, making us learn unconsciously.
4) It is the result of the brain's growing new brain cells as it learns new skills .

57- Which of the following best describes Alan Gabriele Wulf's attitude towards a person's endeavor to begin learning a new skill late in life?

1) Ambivalent but deferential
2) Uncertain but interested
3) Interested and favorable
4) Profound skepticism

58- Which of the following contentions is best supported by the information contained in the passage?

1) Learning is a life-long process that commences right after we are born.
2) Adult's already acquired knowledge almost always facilitates the speed at which they learn new things.
3) Children in schools are unduly required to focus on memory skills, which is a deterrent to active learning.
4) It is interesting to note that physical exercise can give adults a helping hand in the act of learning new skills.

59- Which of the following words is the word "hare" in the last paragraph most related to?

1) Move 2) Think
3) Concentrate 4) Improve

60- Which of the following best represents the main rhetorical function of the passage?

1) Process time order 2) Comparison and contrast
3) Instruction 4) Classification

Bibliographic information of this paper for citing:

Gerampour, M. (2020). Item-focused trees approach in differential item functioning (DIF) analysis: A case study of an EFL reading comprehension test. *Journal of Modern Research in English Language Studies*, *7*(2), 123-147.

Copyright© 2020, Geramipour.