Journal of Teaching Language Skills (JTLS)
38(4), Winter 2020, pp. 1-42- ISSN: 2008-8191
DOI: 10.22099/jtls.2020.36702.2794

Using Multiple-Variable Matching to Identify EFL Ecological Sources of Differential Item Functioning

Touraj Jalili [*] **Hossein Barati** [**]

Ahmad Moein Zadeh [***]

Abstract

Context is a vague notion with numerous building blocks making language test scores inferences quite convoluted. This study has made use of a model of item responding that has striven to theorize the contextual infrastructure of differential item functioning (DIF) research and help specify the sources of DIF. Two steps were taken in this research: first, to identify DIF by gender grouping via logistic regression modeling, an inventory of mostly cited DIF sources was prepared, based on which a list of demographic items was appended to the TOEFL reading paper only to be administered to the intermediate Iranian undergraduates; second, using multiple-variable matching regression (Wu & Ercikan, 2006), a built-in sequence was followed to let every potential DIF source be considered as a covariate, over and above the conditioning variable, and specify whether a particular ecological variable could reduce DIF value/status. Then, all significant variables were analyzed together to show the final DIF predictors. The same procedures, i.e., individual/collective analyses, were employed after the purification of the test. The results indicated three ecological predictors affecting DIF before and after purification: income, administration convenience, and SES. The ultimate predictors helped create an EFL configuration of the ecological model of item responding.

Keywords: DIF sources, Ecological model, Multiple-variable matching, Context, Reading

Received: 26/05/2020 Accepted: 21/07/2020

[*] Ph.D. Candidate, University of Isfahan- Email: touraj.jalili@yahoo.com
[**] Assoicate Professor, University of Isfahan - Email: h.barati@gmail.com, Corresponding author
[***] Assoicate Professor, University of Isfahan - Email: moein@fgn.ui.ac.ir

Language assessment is an enterprise in which relevant and adequate information should be collected to find a way of circumventing (serious or whatever) detrimental consequences of the decisions. To this end, validity arguments must help abate the amount of construct-irrelevant variance leading to a disparate performance on the part of test-takers. This disparity, in reading assessments, helps specify and screen out a host of irrelevant, yet influential, variables, including *inter alia* individual differences (ID) and text variables (Alderson, 2000). The former is related to the ways readers read and affect the process of reading; whereas, the latter pertains to the linguistic and contextual characteristics of the text. Studies on the role of ID factors beyond linguistics knowledge (Brantmeier, 2001), passage content and gender (Brantmeier, 2003), gender and test methods (Brantmeier, 2007), interest (Pae, 2012), etc. as key variables in reading comprehension abound (Newman, Groom, Handelman, & Pennebaker, 2008).

All influential factors, including textual, personal, psychological, affective, methodological, demographic, educational, socioeconomic, cultural, and contextual/situational factors (Zumbo & Gelin, 2005), can be reformulated in language assessment in terms of 'ecology' (Zumbo et al., 2015). The pragmatic and social aspects of the ecology, in language testing, are investigated under the rubric of *bias analysis* to help get more insight into whether, how, and why social groups (e.g., men vs. women) make different interpretations and uses of language. The potential differences are statistically analyzed via *differential item functioning (DIF)* tests as a piece of evidence in the explanation of variations, hence the validity of test scores' inferences (Zumbo, 2009). DIF researchers, thus, need to delve into item property, ID, and contextual predictors of test-takers' differential performance.

Differential Item Functioning

DIF is a statistical technique that is applied to uncover the differential item response patterns between groups of test-takers and thereby helps detect

potentially biased items (Zumbo & Gelin, 2005). DIF is a necessary but insufficient condition for bias (Roever, 2005). Thus, to explain the bias, it needs to be backed up by ample ecological pieces of evidence. Bias studies in tests began at the end of the 1960s and developed exponentially in educational, social, and legal debates. In the 1980s legal cases such as the *Golden Rule* settlement led to the development of methods for identifying DIF (Gómez-Benito, et al., 2018). Thus, to develop the building blocks of the DIF domain, a number of defining terms were coined (e.g., item impact, reference group, uniform DIF, DIF cancellation, etc.) (see Sireci & Rios, 2013).

The above defining terms, together with DIF identification, designated the frontiers of the first generation of DIF research (Zumbo, 2007). In the second generation, new statistical software packages were developed (see Hidalgo and Gómez-Benito (2010, p. 37); McNamara and Roever (2006, p. 93); and Zumbo et al. (2015, p. 140) for different taxonomies of DIF detection methods). Moreover, employing the multidimensional approach to DIF detection (Abbott, 2007), researchers chose to focus *only* on the psychological, cognitive, or unexplained item-specific sources of DIF. However, pursuing hidden subgroups informed and predicted by 'testing situation' (Zumbo & Gelin, 2005; Zumbo, 2007; Zumbo et al., 2015) could transfer DIF praxis and theorizing into the third generation of DIF.

Third Generation DIF Theorizing

An explanatory model that has attempted to elucidate the potential psychometric, linguistic, psychological, and contextual sources of differential performance in language tests is *an ecological model of item responding* (Zumbo et al., 2015). Zumbo, Liu, Wu, Shear, Olvera Astivia, and Ark (2015), building on Zumbo's (2007) five general uses of DIF, introduced a novel ecological model of item responding to cluster the received predictors of variations in test scores. The model indicates a relationship between organisms and situational variables and presumes that the previous research

on DIF sources have not done justice to the field to justify variations as a relatively small number of factors have been investigated mainly within the first two layers of the model (see Figure 1, below) but disregarded the equally important factors within the remaining layers. As such, sufficient pragmatic explanations for the validity of observed variations are yet to develop. Lamenting that more fertile ecological bundles could have been unthreaded in the literature, Zumbo et al. (2015, p. 139) proposed and sketched the graphic representation of a general ecological model of item responding, with a particular focus on language testing. This model, which is one of the ecological models of item responding, is depicted in Figure 1.

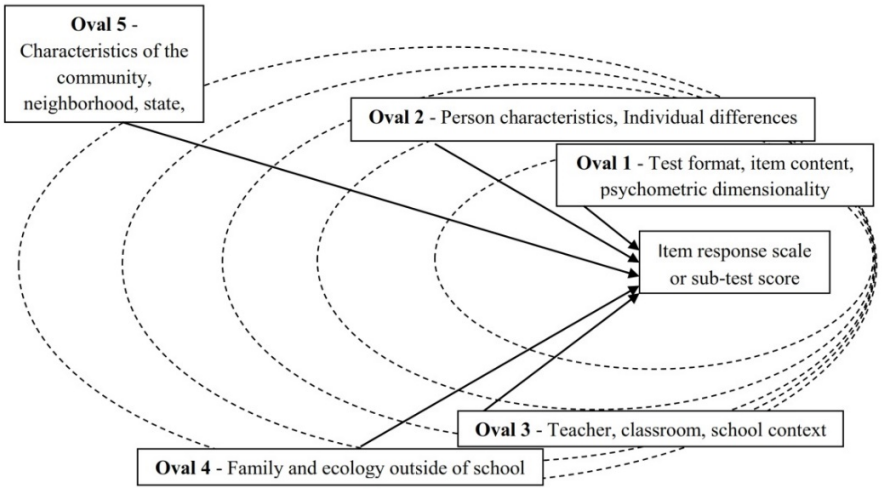


Figure 1. An ecological model of item responding (Zumbo et al., 2015, p. 139)

Presumably, the five layers or concentric ovals of the ecological model were conceptualized in ascending order. The authors took *ecology* as a superordinate term encompassing all relevant predictors of DIF within second and third generations of DIF research and acknowledged that these are not the only five layers of explanatory concepts or necessarily measured or manifest

variables as more layers (latent variables/nested layers) (p. 140) can be considered resulting in mediated/moderated DIF (Zumbo & Gelin, 2005). It is more than clear that *ecological models* with their contextual components would vary with different purposes, participants, tests (topics), and settings.

Table 1.
Levels of Explanatory Variables in the Ecological Model (Zumbo et al., 2015, p. 145)

Oval 1 – Item Content
• Meta-cognition: student reports of the usefulness of the strategy "summarizing" of a long and rather difficult two-page text • Meta-cognition: student reports of the usefulness of the strategies such as concentration, quickly read, discuss with others for understanding and memorizing the text
Oval 2 –*Person characteristics, individual differences*
• Gender/Sex • Like read – fiction • Like read – non-fiction books • Joy/like reading
Oval 3 –*Teacher, classroom, school context*
• At school – Group work • Time – Language lessons • Time – Other language lessons • Teachers stimulation of reading engagement • Teacher student relations • At school – homework
Oval 4 –*Family and Ecology outside of School*
• Index of economic, social and cultural status • Amount of time spent reading for enjoyment • Highest parental education in years • Wealth • Highest educational level of parents • Home educational resources • Online reading

As Figure 1, above, demonstrates, Zumbo et al. (2015) hypothesized that the five concentric ecological ovals along with their sub-categories may,

presumably, impact test-takers' performance in language assessments. The authors (p. 145) also reported on an example study in which they enumerated, through ancillary supplemental data, some ecological factors within each oval and used them as the observed predictors of unobserved latent groups. The factors are shown in Table 1, above. It is not clear, however, why the 'student-reports-of-the-usefulness-of-strategy' variable was put in oval one, despite the fact that it is an ID factor (Cohen & Macaro, 2007, p. 22).

The ecological model can inform the contextual analysis of DIF occurrence. As regards to the model, four points were highlighted by Zumbo et al. (2015, p. 140). First, the model was informed by the ecological systems theory. Second, test-takers (i.e., social present and history) and cognitive processes (i.e., item responding) are inseparable entities. Third, the model articulates what is meant by "context" in Zumbo's (2009) view of validity as a contextualized and pragmatic explanation. Fourth, the model serves as a foundation for the psychometric methodology of DIF analysis. Following is a list of purposes followed in this research.

a) to accumulate the literature-backed DIF sources to illustrate the contextual makeup of DIF theorizing and praxis,
b) to apply the ecological model, in an EFL language testing in general and L2 reading in particular, to clarify whether contextual/cultural factors could predict gender group membership,
c) to discuss the entangled nature of cognitive, cultural, and social variables leading to an argument about the contextual view of validity, and
d) to embark on the ecological model to serve a novel methodology of DIF identification introduced by Wu and Ercikan (2006). To this end, a built-in sequence of entering variables into logistic regression (LR) analysis, with ecological sources/predictors of DIF, was followed in the study.

In LR methodology, comparison groups are conditioned on an ability estimate akin to covariate treatment in a regression analysis (Zumbo, 2007b). In effect, if, as Wu and Ercikan (2006) cogently argued for the cultural source of DIF, we consider each ecological factor as a covariate in a regression analysis, the conventional LR methodology will reveal whether or not the presence of the covariate reduces the status of DIF. Figure 2 represents the path diagram of LR analysis with multiple covariates.

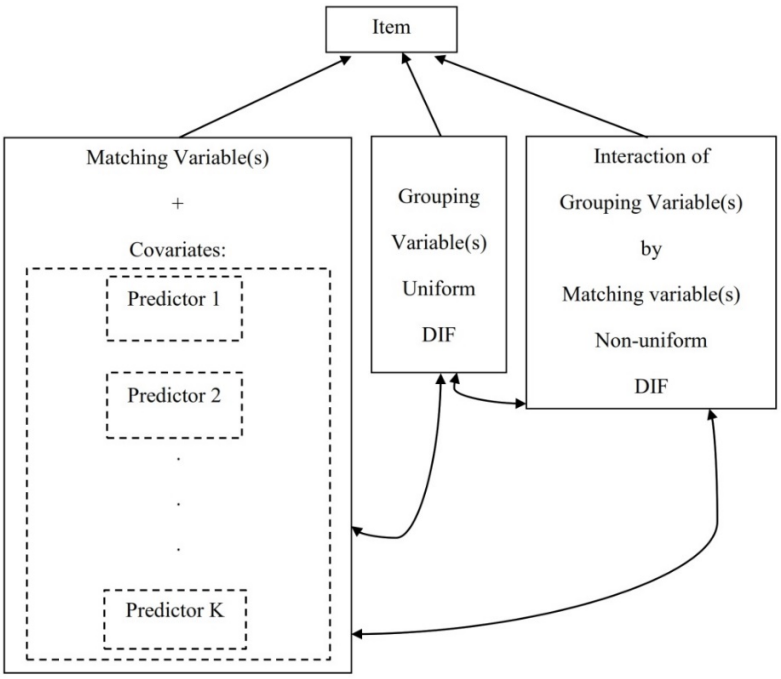


Figure 2. Multiple-variable matching of DIF predictors in logistic regression analysis

In this diagram, all variables are manifest and included in the squares, single-headed arrows are paths or coefficients, curved double-headed arrows represent correlations, and dashed variables denote the sources of DIF effects or predictors of group membership.

The model in Figure 2 shows a built-in sequence of entering variables into the regression model: first, the matching variable(s), then the grouping, and finally, the interaction term(s). We modified Zumbo's (1999) and Zumbo et al.'s (2015) diagrammatic representation of LR analysis to include Wu and Ercikan's (2006) conceptualization of DIF sources as covariates in the matching variable of the analysis. Thus, functioning in the Third Generation of DIF, the path diagram could help detect the ecological predictors of DIF. This statistical LR model includes two analytical steps (see the methodology section, below). Thus, the ecological model (Figure 1) and LR statistical model (Figure 2) were used, in this inquiry,

a. to identify DIF by gender grouping in an EFL context of reading assessment, and
b. to discuss ecological sources of gender-related DIF.

The literature of DIF studies seems to have been oblivious to the crucial role in test scores' variations of the ecological variables/models (Zumbo et al., 2015). This is the point we discuss in the following section.

The Literature of DIF Sources

DIF researchers have long attempted to investigate the matter of measurement invariance across different groupings, such as gender (Aryadoust et al., 2011; Cheong & Kamata, 2013; Cho et al., 2012; Cohen & Bolt, 2005; Lee & Geisinger, 2014; Mendes-Barnett & Ercikan, 2006; Oshima et a., 1998; Pae, 2012; Ryan & Bachman, 1992; Stricker & Emmerich, 1999; Suh & Talley, 2015; Takala & Kaftandjeva, 2000), and language assessments, such as reading tests (Abbott, 2007; Banks, 2012; Chen & Henning, 1985; Elder et al., 2003; Koo et al., 2014; Lee & Geisinger, 2014; Oshima et a., 1998; Pae, 2004b, 2012; Ryan & Bachman, 1992; Sasaki, 1991; Stricker & Emmerich, 1999; Taylor & Lee, 2012; Uiterwijk & Vallen, 2005), in different ESL/EFL contexts (Ferne & Rupp, 2007). Despite all attempts to provide

accurate accounts of DIF sources, previous studies have infrequently investigated ample contextually-relevant sources (Zumbo, 2007).

For example, Allalouf and Abramzon (2008) found DIF, across language groups, due to *L1 transfer*, *cognates*, *critical period hypothesis (CPH),* and *wide-context items.* Abbott (2007) identified DIF related to top-down and bottom-up *strategies*. Aryadoust et al.'s (2011) used gender as the grouping factor and found DIF related to *gender, guessing/distractor,* the difficulty of *linguistic elements,* and *item stem length.* Similar to Aryadoust et al. (2011) and Abbott (2007), Taylor and Lee's (2012) study fell into Zumbo's (2007) second DIF generation and analyzed DIF based on gender. They hypothesized DIF associated with *attitudes toward and willingness* to respond to test items, *interpretation* ways, *females' verbalism,* and *wide-context items.* Wu and Ercikan (2006) found *culture* as an explanatory source of DIF across two language groups. Koo et al. (2014), utilizing gender, ethnicity, and ELL status as the grouping variables, found no significant effect for gender, but *prior educational experiences in L1* and *ethnicity* resulted in DIF. Suh and Talley (2015) compared detection methods to identify gender differential distractor functioning (DDF); however, no DIF was reported. Ercikan, et al. (2014) argued DIF based on the diversity of linguistic minority. They found no DIF related to the curriculum; however, they identified the *linguistic load of test items, familiarity with specific vocabulary,* and *sentence structure complexity* as possible sources of DIF. Ercikan (2002), on the other hand, provided interpretation for DIF due to *test adaptation, curricular differences, instruction methods, cultural differences,* and *limitations in definitions of topics*. Cho, Lee, and Kingston (2012) investigated the relationships between DIF, item types, and students' accommodation status and content knowledge. The sources of DIF were speculated to be *developmental, instrumental,* or *random error.* The study did not find any relationship between *students' accommodation status* and academic ability. Item type was not related to DIF direction, and nuisance dimensions (gender, ethnicity, and disability status)

had a minimal association with the observed DIF. Elosua and Lopez-Jauregui (2007) reported on DIF sources based on test adaptation. *Grammatical* and *semantic* factors affected DIF but *translation problems* and *cultural factors,* could not, significantly, affect the test adaptation. Finally, Oliveri et al. (2014) concentrated on linguistic heterogeneity DIF occurring with English language learners (ELLs). The findings revealed that focal group heterogeneity led to reduced false positive rate but increased false-negative DIF. Failures to correctly detect DIF and the presence of false negatives were due to *grouping factors* or the use of *traditional methods*. Table 2, below, represents a list of highly cited specific/general DIF sources in the literature.

Table 2.

Research-Based DIF Sources and the Studies

Sources of DIF	Researchers
1. Age	Cohen & Bolt, 2005
2. Race, Ethnicity	Koo et al., 2014; Oliveri et al., 2014
3. Opportunity to Learn	Takala & Kaftandjieva, 2000
4. Familiarity (with Item Type)	Ercikan, 2002; Kunnan, 1990; Oliveri et al., 2014; Pae, 2012; Stricker & Emmerich, 1999
5. Higher propensity to take (college) courses	McNamara & Roever, 2006
6. Interest in Subject Matter	McNamara & Roever, 2006; Pae, 2012; Stricker & Emmerich, 1999
7. Gender	Aryadoust et al., 2011; Cheong & Kamata, 2013; Cohen & Bolt, 2005; O'Neill & McPeek, 1993; Pae, 2004b; Ryan & Bachman, 1992; Takala & Kaftandjieva, 2000
8. Native Speaker (NS) Status	Elder et al., 2003; McNamara & Roever, 2006
9. Language	Abbott, 2007; Ercikan et al., 2014; Finch et al., 2016; Harding, 2011; Kim, 2001; Le, 2009; Uiterwijk & Vallen, 2005
10. Academics	Pae, 2004b; Kunnan, 1990
11. Emotional reactions to items	Pae, 2012; Stricker & Emmerich, 1999
12. Test-wiseness	Wu & Ercikan, 2006
13. Attitudes toward test	Wu & Ercikan, 2006

Sources of DIF	Researchers
14. L1 Transfer	Allalouf & Abramzon, 2008; Koo et al., 2014; Kunnan, 1990
15. Critical Period Hypothesis (CPH)	Allalouf & Abramzon, 2008
16. Wide-Context Items	Allalouf & Abramzon, 2008; Taylor et al. 2012
17. Top-Down & Bottom-Up/ Strategies	Abbott, 2007
18. Guessing/Item stem length	Aryadoust et al., 2011
19. Difficulty of Linguistic Elements	Aryadoust et al., 2011; Bolt & Thurlow, 2007; Cho et al., 2012; Cohen & Bolt, 2005; Ercikan, 2002; Helwig et al. 1999; Oliveri et al., 2014; Roth et al., 2013; Santelices & Wilson 2012
20. Cognitive Level	Mendes-Barnet & Ercikan, 2006; Pae, 2012
21. Attitudes/Willingness	Taylor & Lee, 2012
22. How to Interpret	Taylor & Lee, 2012
23. Females' Verbalism	Taylor & Lee, 2012
24. Vocabulary Knowledge	Chen & Henning, 1985; Ercikan et al., 2014; Jang & Roussos, 2009; Oliveri et al., 2014; Roth et al., 2013
25. Short-term memory	Aryadoust et al., 2011
26. Sentence structure complexity	Ercikan et al., 2014
27. Rating Rubrics	Kim, 2001
28. Person * Rater	Lynch et al., 1998
29. Person * Items	Lynch et al., 1998
30. Rater * Learner's background variable (e.g., accent)	Elder et al., 2003; Lynch et al., 1998
31. Rater * Item	Lynch & McNamara, 1998
32. Facets' interaction: Person * Rater * Item	Lynch & McNamara, 1998
33. Test Adaptation/Translation	Allalouf et al. 1999; Elosua & Lopez-Jauregui, 2007; Ercikan, 2002; Jang & Roussos, 2009; Wu & Ercikan, 2006
34. Curricular Differences	Ercikan, 2002; Wu & Ercikan, 2006
35. (Unintended) Cultural Effects	Allalouf et al. 1999; Banks, 2012; Elosua & Lopez-Jauregui, 2007; Ercikan, 2002; Oliveri et al., 2014; Uiterwijk & Vallen, 2005; Wu & Ercikan, 2006
36. Instruction methods	Ercikan, 2002; Oliveri, et al., 2014
37. Limitations in definitions of topics	Ercikan, 2002
38. Developmental	Cho et al., 2012
39. Instrumental	Cho et al., 2012
40. Random error	Cho et al., 2012

Sources of DIF	Researchers
41. Accommodation status	Cheong & Kamata, 2013; Cho et al., 2012
42. Grammatical differences across languages	Elosua & Lopez-Jauregui, 2007
43. Semantic differences across languages	Elosua & Lopez-Jauregui, 2007
44. Heterogeneity	Oliveri et al., 2014
45. Scale Indeterminacy problem	Cheong & Kamata, 2013
46. (Traditional) DIF methods	Oliveri et al., 2014; Sasaki, 1991
47. DDF (differential distractor functioning)	Banks, 2012; Jang & Roussos, 2009; Suh & Talley, 2015; Tsaousis, et al., 2018
48. Dialect	Harding, 2011; Oliveri et al., 2014
49. Multidimensional item impact	Mazor et al., 1995
50. Schooling/Environment	Cheong, 2006
51. Socioeconomic Status (SES)	Banks, 2012; Oliveri et al., 2014; Oshima et al., 1998; Shermis et al. 2017
52. Home educational resources	Finch et al., 2016
53. Contextual variables	Lee & Geisinger, 2014; Oshima et al., 1998

The above list is not meant to be exhaustive as more sources of DIF can be included. The purpose is to acquaint the reader with the main thrusts of DIF findings in so far as they are associated with the second or third generations of DIF research.

As the results of the above studies with particular designs indicate, many potential ecological predictors of variation are left unanalyzed. Thus, their explanatory power is not rigorous enough. Ferne and Rupp (2007) reviewed research on DIF between 1990 and 2005 and found either *a priori* expert judgments for item coding/review or *post hoc* judgments for item analysis. In the following years the situation for DIF explanation was no better (Zumbo et al., 2015). Even in the context of Iran, the justifications proved fragile and feeble. For instance, Barati et al. (2006) speculated DIF occurrence due to the benefited group's more exposure to courses involving the skills of logic, inferencing, and holistic view. Focusing on gender DIF, Ahmadi and Darabi (2016) hypothesized that the female-friendly DIF might refer to the "widespread belief of female superiority in language learning" (p. 75).

Overall, only a few studies made use of all the ecological layers in Figure 1 to provide a rather comprehensive account of DIF effects. Two such studies are Zumbo et al. (2015) in Canada and Ahmadi and Jalili (2014) in Iran. Therefore, following this ecological line of inquiry, the present research aimed at

1. applying the literature-supported DIF sources and the ecological model in quest for preparing relevant EFL gender-related sources of variations in reading test performance, and
2. elucidating the extent to which the ecological variables could reduce DIF status, hence predicting variations in item responding.

Method

Sample of Participants

A total of 866 intermediate Iranian EFL university undergraduates from private/public universities and language institutes were invited to sit the reading test. Similar to many DIF studies, a larger number of test-takers could have been assessed had the data from large-scale (inter)national assessments been utilized but, then, the study could suffer from a lack of access to demographic/ecological information about test-takers (see the appendix). The study focused on potential DIF by gender grouping. There were, overall, 459 (53%) males and 407 (47%) female students.

Instrument

Two instruments were utilized in this study: a reading test and a questionnaire. To assess the EFL test-takers' reading ability, the reading paper of the TOEFL iBT test (2010) was selected. The reading test included 36 binary items addressing the following 9 reading subskills:

- Vocabulary,
- Pronoun Reference,
- Terms,

- Exception,
- Purpose,
- Cause,
- Authors' Opinion,
- Sentence Insertion,
- Paraphrase.

The reading test was divided into two subtests: the main reading test (24 items) and the screening test (12 items). The subtests were similar in terms of the reading subskills. Moreover, given the use of short tests in DIF studies, e.g., Finch et al.'s (2016) reading test including only 13 items, the employed test in this research does not seem to be a very short test. Drawing an analogy between the 12-item screening test and the ETS scores for the reading module of TOEFL tests, we considered those who scored 7 or higher as the intermediate/upper-intermediate group whose performance on the main reading test was, subsequently, analyzed for DIF. The results of a pilot study, with 30 upper-intermediate undergraduates, revealed, according to facility and discrimination indices, that among the items in the test, item 2 (*Reference*) and 6 (*Purpose*) were the easiest items. In contrast, item 9 (*Purpose*) and 13 (*Cause*) were the most difficult ones. Using coefficient alpha, the reliability of the pilot study of scores on the 36-item test was estimated as 0.897 rounded to .90.

The second instrument in this research was a questionnaire, appended to the reading test paper, comprising an array of ecological items (see the appendix). To develop the questionnaire, relevant DIF sources from Table 2 as well as suggested predictors from Table 1 were reformulated as a set of questions. The questionnaire, with .80 (.798) reliability index, had two piloting phases.

In the initial piloting (Dörnyei, 2003, p. 66), an item pool was prepared including DIF sources and the opinions of three specialists (applied linguists) and three non-specialists (language teachers) in the field. The latter group,

unaware of the field jargon, agreed with all the questions and added some extra items, such as the effect on the test performance of the testing room's *architecture, temperature/weather, health condition,* etc. Later, such items were decided to lie under the heading *contextual variables' effect* (Lee & Geisinger, 2014; Oshima et al., 1998). The specialists, on the other hand, included the effect on the test performance of test-takers' *vocabulary/grammar knowledge* and *reading ability* (Ercikan et al., 2014; Oliveri et al., 2014; and Roth et al., 2014).

Thus, for the vocabulary and grammar knowledge in the questionnaire, the test takers' scores in the same/previous semester or the evaluative judgment of their teachers/lecturers were registered. However, for reading ability, because a large number of respondents had either no reading score or scores based on a very short test, the examinees' scores in the 12-item screening test were taken into consideration. For reading strategy, the test takers were asked to show their proclivity for one of two strategies, i.e., whether they passively decoded sequential graphic-phonemic-syntactic-semantic systems (bottom-up) or activated relevant schemata and mapped incoming information onto them (top-down) (Alderson, 2000, p. 17). The significant effect on variations of this variable could corroborate the role of strategy in reading assessment (Abbott, 2007).

Moreover, some other items (either backed up in the literature or recommended by the specialists), including *prior education, opportunity to learn, familiarity with vocabulary/MC items, political decisions, university curricula, personality type, anxiety level, family relations,* and *occupation* were discarded, in the final phase, because of missing responses and/or lack of variation in the scores (Dörnyei, 2003).

Data Collection

The primary data included:

a) the gender groups' performance on the reading test items with four options, and
b) the respondents' answers to the items of the questionnaire (see the appendix).

The items of both the reading test and the questionnaire were of binary type. The reading items were scored as correct/incorrect, whereas the questionnaire items were collected as either/or preferences. For the questionnaire, a Likert scale was not used because a number of items could not have more than two options and, thus, to have a consistent format throughout the questionnaire and consistent comparisons across the ecological variables as covariates, all items were written in the binary type.

Data Analysis

Stage 1: DIF Detection

As discussed above, the current research adopted logistic regression (LR) (Shimizu & Zumbo, 2005) in the way used by Wu and Ercikan (2006). The items of the reading test were all binary items; however, all the items were analyzed through multinomial regression because, as Wu and Ercikan acknowledged, binary items are, in essence, a special case of ordinal items. The conventional LR analysis was run to find DIF items by gender grouping. Thus, a built-in sequence was utilized in the LR model.

Stage (1): Matching on 'Total' score only
Matching Model: only 'Total score' entered the model as a covariate first.
$Logit = b_0 + b_1 * \text{'Total'}$
(1 degree of freedom)
Full model: 'Total' (as a covariate), 'Gender', and their interaction
$Logit = b_0 + b_1 * \text{'Total'} + b_2 * \text{'Gender'} + b_3 * \text{'Gender * Total'}$
(3 degrees of freedom)

The difference in Chi-square values between the full and matching models with 2 degrees of freedom was flagged as gender-related DIF. The stage-one analysis was carried out only once for every reading item. To report the effect size, two criteria were used (see Table 3, below).

Table 3.
DIF Types Labeling (Hidalgo & López-Pina, 2004; and Jodoin & Gierl, 2001)

DIF Types	Hidalgo & López-Pina (2004)	Jodoin & Gierl (2001)
Negligible	$\Delta R^2 < 0.13$	Nagelkerke R^2 difference < 0.035
Moderate	$0.13 \leq \Delta R^2 \leq 0.26$	0.035 < Nagelkerke R^2 difference < 0.070
Large	$\Delta R^2 > 0.26$	0.070 < Nagelkerke R^2 difference

As you see in the table, the classifications bear little resemblance to each other. It seems that Hidalgo and López-Pina's (2004) labeling would be an underestimation of DIF size, whereas that of Jodoin and Gierl (2001) would be an overestimation of LR effect size, that is, using the latter most DIF items turn out to be of large DIF size.

Stage 2: Analysis of DIF Sources

Following Wu and Ercikan's (2006, p. 293) lead, all 32 ecological variables (see the appendix) were analyzed, individually, as covariates, over and above the total score, in the matching variable (see Figure 2). Unlike stage-one analysis that was conducted only once for every item, a stage-two analysis was done 32 times for every reading item. That is, in case an item displayed DIF based on gender at stage one, we employed all ecological variables one at a time as covariates and checked whether DIF value or status was reduced. If the presence of a particular ecological covariate (e.g., income) helped reduce DIF status, that very variable would be referred to as 'DIF

predictor.' The built-in sequence at stage two for *income,* one of the ecological variables, is as follows:

Stage (2): Extra Matching on 'Income' and interaction terms in addition to 'Total'
Matching Model: two matching variables, 'Total', and 'Income', and their interaction term as covariates first
$Logit = b_0 + b_1 * 'Total' + b_2 * 'Income' + b_3 * 'Total * Income'$
(3 degree of freedom)
Full Model: 'Gender' as a grouping variable and the interaction terms entered the model.
$Logit = b_0 + b_1 * 'Total' + b2 * 'Income' + b_3 * 'Total * Income' + b_4 * 'Gender' + b_5 * 'Gender * Total' + b_6 * 'Gender * Income' + b_7 * 'Total * Income * Gender'$
(7 degrees of freedom)

At Stage Two, the difference in Chi-square values was tested with 4 degrees of freedom for each and every item. Provided that the magnitude of 'gender-related DIF' at Stage one would decrease at Stage two, 'income' was considered as a source of DIF.

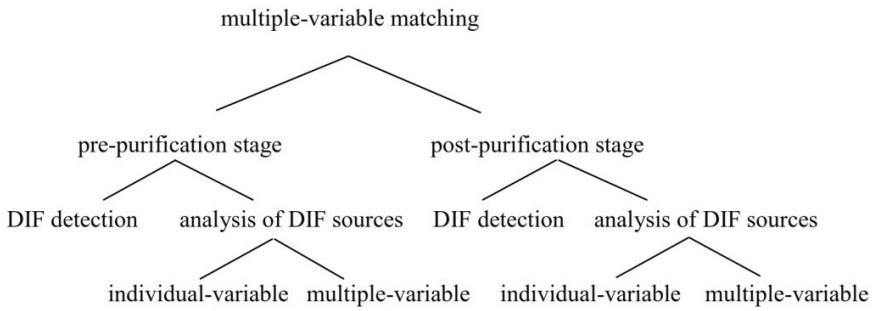
Having investigated the effect on DIF reduction of individual ecological variables, we analyzed all significant ecological covariates together in a regression analysis to see whether or not they could reduce DIF status at the presence of each other. This analysis could have been run in the primary stage-two analysis; however, the presence of all 32 covariates besides 'total score,' as a matching variable, would have jeopardized the statistical rigor of the test, resulting in inflated Type I or II errors. As such, only significant covariates, from individual stage-two analysis, were selected for collective regression analysis.

The same analyses (i.e., individual and collective analyses) were run after the purification stage (i.e., after DIF items were deleted and made the

matching criterion free of DIF contamination). The purification procedure was employed to remove sizeable DIF after the first detection run to repeat the process without those contaminated items (McNamara & Roever, 2006, p. 109). The reason for this final analysis was to check and compare DIF status reduction with the DIF-embedded and DIF-reduced matching criteria.

Overall, the analysis was run at two stages: pre- and post-purification stages. Each stage was divided into two sub-stages: DIF detection and analysis of DIF sources. The analysis of DIF sources, itself, fell into two stages: individual-variable matching (in which an individual ecological variable was added over-and-above the total score as the matching variables) and multiple-variable matching (in which all significant variables at the individual level were analyzed in a multiple regression analysis. All these binary steps are represented in Figure 3, below.

Figure 3. Multiple-variable matching stages



Results

DIF Detection

Having analyzed all 24 items at stage one, we found 9 items (nearly one third or 33% of the test) displaying DIF by gender groups. The males and females functioned significantly differently in the following items: items 1, 3, 9, 12, 15, 16, 17, 22, and 23. Out of the nine reading subskills (see the

methodology section), only the exception and paraphrase item types did not turn out to exhibit DIF across gender groups (see Table 4, below). Employing Jodoin and Gierl's (2001) taxonomy of DIF effect size, we found that all items were of large type. In contrast, using Hidalgo and López-Pina's (2004) classification, we detected only one large DIF (item 9), six moderate DIF (items 12, 15, 16, 17, 22, and 23), and two negligible effects (items 1 and 3). The results of this section, together with DIF sources at stage one and stage two, are indicated in Table 4.

Analysis of DIF Sources

An array of 32 ecological variables, as potential DIF predictors (see the appendix), was added, as covariates one at a time, to the LR analysis to assess gender-related DIF status reduction. Unlike Wu and Ercikan's (2006) conservative .01 ≤ α level for step two, we employed α < .05 at stage one and both .05 ≤ α and .01 ≤ α significance levels at stage two. Note that the uniform/non-uniform DIF, as well as direction (i.e., gender group advantage) of DIF, would not be accounted for because the focus of the study was on determining DIF reduction across the two stages. The results of stage one, item type, effect size, DIF type, and stage two (DIF reduction) are presented in Table 4.

Table 4 indicates that five items (items 12, 15, 16, 17, and 22) did not exhibit significant DIF by gender at stage two, meaning that the presence of ecological covariates helped reduce their DIF status. In contrast, the remaining items stayed uninfluenced by the ecological covariates. The five items were all moderate (Hidalgo & López-Pina, 2004) or large (Jodoin & Gierl, 2001) DIF type. They included two *vocabulary* items, one *sentence insertion,* one *pronoun reference,* and one *cause* item. At .05 ≤ α significance level, in item 12 two ecological variables (*rubric difficulty* and *home resources*), in item 15 one variable (*interest* in the passage content), in item 16 two variables (*SES* (socioeconomic status) and *grammar* knowledge), and in item 17 two

covariates (*reading* score and interest in *non-fiction* books) made the stage-one significant gender variable function as non-significant at stage two.

The final collective regression analysis of all significant covariates (both at .05 ≤ α and .01 ≤ α levels) clarified that in item 12 *rubric difficulty, home resources*, *income,* and *administration convenience*, in item 15 *interest* and *language class time,* in item 16 *SES* and *item difficulty,* and in item 17 only *L1* reduced DIF status at stage two.

Table 4.

The Results of DIF Detection (Stage 1) and DIF Sources by Multiple-Variable Matching (Stage 2)

Stage 1					Stage 2	
DIF items and Item Types	Significant gender (α < .05)/effect size	Effect size criteria		Reduced at stage 2?	Sources of non-significant gender (.05 ≤ α)	Sources of non-significant gender (.01 ≤ α)
		Hidalgo & López-Pina (2004)	Jodoin & Gierl (2001)			
1 (Terms)	.000/.110	N	L	No		
3 (Vocabulary)	.000/.112	N	L	No		
9 (purpose)	.000/.269	L	L	No		
12* (Sentence Insertion)	.011/.155	M	L	Yes	rubric difficulty (.083), home resources (.050)	income (.012), administration convenience (.019), teacher's stimulation (.021), MC interest (.027), out-of-school events (.042)
15* (Vocabulary)	.000/.142	M	L	Yes	interest (.066)	L class time (.010)
16* (Vocabulary)	.018/.182	M	L	Yes	SES (.055), grammar knowledge (.051)	propensity to study (.017), item difficulty (.024)

17* (Pronoun Reference)	.029/.165	M	L	Yes	reading ability (.077), enjoy non-fiction (.055)	ethnicity (.029), L1 (.038), parents' education (.048), enjoy fiction (.038)
22* (Cause)	.005/.138	M	L	Yes		MC interest (.011)
23 (Author's Opinion)	.000/.199	M	L	No		

Note. * denotes non-DIF items at stage 2, L represents large effect size, M stands for moderate effect, N signifies negligible effect size, and the decimals in the parentheses indicate non-significant values of gender.

Having identified all 9 DIF items as sizeable/large effect (Jodoin & Gierl, 2001) at stage one, we removed them and operated a two-stage analysis after purification (Hidalgo & Gómez-Benito, 2010, p. 40) to analyze the degree to which the remaining items could function as anchor items. The results of the post-purification analysis indicated that two items (items 4 and 11) displayed significant DIF by gender at stage one while indicating non-significant gender at stage two. Table 5 represents the results of stage-one and stage-two analyses of purified items.

In item 4, no covariate reduced DIF at $.05 \leq \alpha$ level; however, at $.01 \leq \alpha$ level nine ecological variables turned the significant gender at stage one into non-significant at stage two; hence, reduced DIF status. In item 11, on the other hand, seven covariates reduced DIF at $.05 \leq \alpha$ level, and nine ecological variables made gender non-significant at $.01 \leq \alpha$ level (see Table 5).

Table 5.

The Results of DIF Detection (Stage 1) and DIF Sources by Multiple-Variable Matching (Stage 2) After Purification

Stage 1				Stage 2		
DIF items and Item Types	Significant gender (α < .05)/effect size	Effect size criteria		Reduced at stage 2?	Sources of non-significant gender (.05 ≤ α)	Sources of non-significant gender (.01 ≤ α)
		Hidalgo & López-Pina (2004)	Jodoin & Gierl (2001)			
4* (Purpose)	.022/.269	L	L	Yes		field (.047), income (.019), administration convenience (.036), distractor (.034), grammar knowledge (.017),MC interest (.014), enjoy reading (.042), enjoy fiction (.027), home resources (.029)
11* (Author's Opinion)	.033/.130	M	L	Yes	field (.051), distractor (.055), grammar knowledge (.058), reading ability (.072), reading strategy (.202), enjoy reading (.075), online reading (.105)	SES (.017), content familiarity (.046), income (.030), group work (.026), teacher's manner (.027), physical setting (.017), enjoy fiction (.046), teacher's stimulation (.030), home resources (.049)

Note. * denotes non-DIF items at stage 2, L represents large effect size, M stands for moderate effect, and the decimals in the parentheses indicate non-significant values of gender.

The final collective regression analysis of all significant covariates (both at $.05 \leq \alpha$ and $.01 \leq \alpha$ levels) clarified that not all ecological predictors in Table 5 could reduce gender-based DIF at the presence of each other. Thus, in item 4 *income*, *administration convenience, distractor, grammar knowledge,* and enjoy *fiction* books were the DIF sources. In item 11, however, *distractor, reading ability, enjoy reading, SES, content familiarity, income, teacher's manner*, and *teacher's stimulation of reading* were introduced as DIF reduction sources. Income and distractor appeared in both items.

Table 6 shows a comparison between (nearly ten) DIF sources found together at pre- and post-purification stages. The item types at the stages were completely different; however, three common DIF sources were found at both stages: income, administration convenience, and SES. These three DIF sources are highlighted in the table.

Table 6.

Item Types and DIF Sources before/after the Purification

Stage	Item Types	DIF Sources
Pre-Purification	Cause, Pronoun Reference, Sentence Insertion, and Vocabulary	*administration convenience,* home resources, *income,* interest, item difficulty, L1, language class time, MC interest, rubric difficulty, and *SES*
Post-Purification	Author's Opinion and Purpose	*administration convenience,* content familiarity, distractor, enjoy fiction, enjoy reading, grammar knowledge, *income,* reading ability, *SES,* teacher's manner, and teacher's stimulation of reading

The 18 DIF sources in Table 6 were subjected to factor analysis (principal components analysis). The KMO (.522) and Bartlett's test (.000) supported the factorability of the correlation matrix. However, the 18 items were not strongly distributed in the five components meant to be similar to the Zumbo et al.'s (2015) five ecological layers. As such only 46.5% of the variance was

explained by five components. Ideally, the components with 3 or more items loaded are retained, therefore, in this research, that could factorize 12 items, a three-factor solution was obtained as only three components were loaded by five, three, and four items.

Discussion

Due to the influence of cross-cultural variables, it is highly likely that the reader variable (gender) and particularly the interaction between text type and ecological variables by gender do not apply to certain cultures. Thus, the findings of gender DIF studies (Cohen & Bolt, 2005; Lee & Geisinger, 2014; Mendes-Barnett & Ercikan, 2006; Pae, 2012; Stricker & Emmerich, 1999; Taylor & Lee, 2012) in different contexts/cultures may not be corroborated in a new context. Therefore, we introduced our EFL framework of the ecological model because, as Wu and Ercikan (2006, p. 298) acknowledged, a source variable (in a particular setting) is context- and purpose-dependent and is not a fixed characteristic of an item. To have an EFL configuration of the ecological model (Zumbo et al., 2015), the current study applied a multiple-variable matching LR analysis (Wu & Ercikan, 2006) to determine the proportion of DIF status reduction, hence gender-related ecological DIF predictors. The reasons for the size of the detected items varied. The difference in effect sizes of the DIF items might pertain, besides contextual variables, to the item types and their level of cognitive processing (Mendes-Barnet & Ercikan, 2006; Pae, 2012). For instance, item 1 assessed knowledge subskill and turned out to show negligible (Hidalgo & López-Pina, 2004) size. However, item 9 assessed learners' analytical reasoning and, as such, showed a large effect size. This item, in the pilot study, was detected as the most difficult item. The next step, in the analysis, was to compare the results after removing the large DIF items.

Sireci and Rios (2013, p. 183) indicate that purification, presumably, might (not) improve DIF detection in the LR method. Gómez-Benito et al.

(2018) argue that the purification procedure is an aspect of the consequential evidence of construct validation; thus, researchers should ask "does eliminating DIF items lead to construct underrepresentation?" (p.107). As such, the decision (not) to remove sizeable DIF items might end in sizeable consequences. Therefore, researchers should analyze DIF both with and without purification to compare the results to see which results are most interpretable.

Vocabulary turned out to be the prominent reading subskills that differentiated the performance of the males and females in this research. In the pilot study, a number of examinees articulated they had difficulty understanding a few *vocabulary* items in the reading passages or test items. Even though vocabulary items, most often, are candidates for exhibiting DIF (Ercikan et al., 2014; Jang & Roussos, 2009; Oliveri et al., 2014), they are the sine qua non of every kind of reading comprehension, yet "the desirability of using dictionaries during reading tests" to clarify if vocabulary is construct relevant needs to be accounted for (Alderson, 2000, p. 99). One can argue that the emergence (stage I) and reduction (stage II) of the DIF items both at the pre- and post-purification procedures might pertain to linguistic competence (e.g., vocabulary), illocutionary competence (e.g., sentence insertion), strategic competence (e.g., pronoun reference and author's opinion) (Cohen & Macaro, 2007), and cognitive ability (e.g., cause and purpose) (Mendes-Barnett & Ercikan, 2006). However, such hypothetical reasoning requires in-depth qualitative investigation, corpus analysis, and expert judgments that were beyond the scope of the study.

In this research, the two-stage analysis before and after purification represents a model-building approach to DIF-source analysis in which ecological variables entered the equation one at a time. The emergent DIF predictors helped develop an EFL grounded ecological model of variation demonstrating what ecological factors affected the EFL gender groups' performance on the reading test items. The makeup of the ecological model

(Zumbo et al., 2015) helps boost our understanding by locating the gender-related DIF sources in the layers of the model. Figure 4 can represent one of

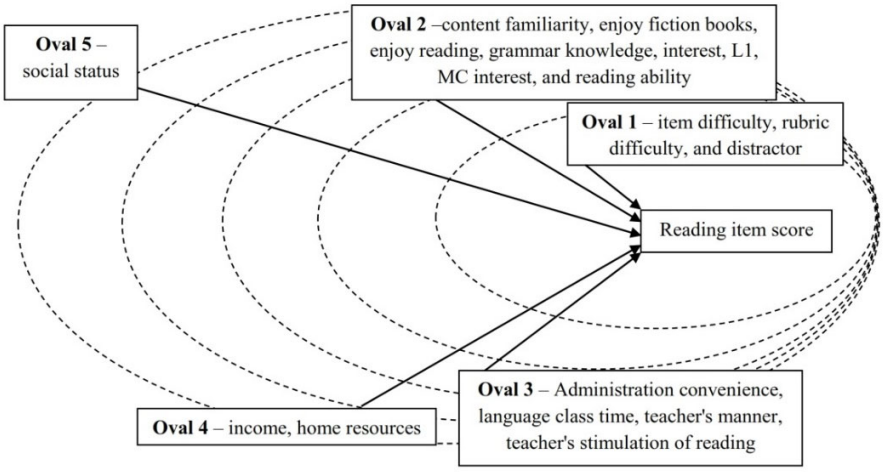


Figure 4. An EFL Configuration of the ecological model of item responding

the EFL frameworks/configurations of the ecological model of item responding in a reading assessment.

As Figure 4 demonstrates, the five concentric ecological ovals, along with their sub-categories, turned out to impact the EFL male and female test-takers' performance in the reading assessments. A comparison between the findings of this research and those of other studies in the literature (see Table 2) can indicate the strength of our grounded model in Figure 4. The question why only these explanatory variables, and not others, should remedy the effect of gender on the difficulty of any given item was partly answered by the comparison across DIF findings. However, qualitative analysis and judgments by content experts were beyond the scope of the study. Besides, had linguistic investigation been done, the sources of gender DIF would have solely been considered as certain item properties; however, the emphasis in the third generation of DIF research is on 'testing situation.' Furthermore, *ad hoc*

explanations for the functioning of the covariates were not contrived as there was no way of knowing, subsequently, how valid such explanations could be. Apparently, the statistical interpretation could explain the differential functioning of covariates. The covariates that decrease gender-based DIF substantially can add new information to the conditioning variable, and as such, "the improvement in matching is greater, or, stated another way, more of the latent ability space is accounted for." (Mazor et al., 1995, p. 139). That is why, they could change gender DIF into non-gender-DIF at the second stage, whereas the information provided by the other variables turned out to be redundant and could not increase the matching power of the conditioning criterion. Overall, 18 ecological components of the model functioned as DIF predictors in this research in the following descending order:

Oval 2 (ID variables):

1. content familiarity (Ahmadi & Jalili, 2014; Ercikan, 2002; Kunnan, 1990; Oliveri et al., 2014; Pae, 2012; Stricker & Emmerich, 1999),
2. enjoy fiction books (Zumbo et al., 2015),
3. enjoy reading (Alderson, 2000; Zumbo et al., 2015),
4. grammar knowledge (Elosua et al., 2007),
5. interest (Ahmadi & Jalili, 2014; McNamara et al., 2006; Pae, 2012; Stricker & Emmerich, 1999),
6. L1 (Abbott, 2007; Ercikan et al., 2014; Finch et al., 2016; Kim, 2001; Le, 2009; Uiterwijk et al., 2005),
7. MC (multiple-choice) interest (Alderson, 2000; Allalouf et al., 2008; Cho et al., 2012; Taylor et al. 2012),
8. reading ability (Alderson, 2000; Pae, 2004b; Zumbo et al., 2015),

Oval 3 (school context):

9. Administration convenience (Cheong et al., 2013; Cho et al., 2012),
10. Language class time (Cheong, 2006; Zumbo & Gelin, 2005),
11. teacher's manner (Cheong, 2006; Zumbo et al., 2015),

12. teacher's stimulation of reading (Lee & Geisinger, 2014; Zumbo et al., 2015),

Oval 1 (item property variables):

13. item difficulty (Aryadoust et al., 2011; Bolt et al., 2007; Cho et al., 2012; Cohen et al., 2005; Ercikan, 2002; Helwig et al. 1999; Oliveri et al., 2014; Roth et al., 2013; Santelices and Wilson 2012),
14. rubric difficulty (Ercikan, 2002; Roth et al., 2013),
15. distractor (Banks, 2012; Jang & Roussos, 2009; Suh et al., 2015; Tsaousis et al., 2018),

Oval 4 (family context):

16. income (Ahmadi & Jalili, 2014; Zumbo, 2007a; Zumbo & Gelin, 2005),
17. home resources (Finch et al., 2016), and

Oval 5 (community context):

18. social status (Banks, 2012; Oliveri et al., 2014; Oshima et al., 1998; Shermis et al., 2017).

Despite the categorization of the above 18 variables into five ecological ovals/layers (Zumbo et al., 2015), a factor analysis test revealed that the variables were clustered in *three components.* The three-factor solution includes:

1. interest (ID variable), teacher's stimulation of reading (school context), enjoy fiction books (ID), reading ability (ID), and enjoy reading (ID),
2. income (family context), and home resources (family context), teacher's manner (school context), and
3. MC interest (ID variable), content familiarity (ID), L1 (ID), and social status (community context).

Thus, it seems that more *ID* variables than contextual items were loaded on the first and third components, whereas the second component was loaded by only *contextua*l variables. The interpretation of the three components was

proportionately consistent with Zumbo et al.'s *item property, ID,* and *contextual* layers of the ecological model.

Zumbo et al. (2015, p. 140) argued that "conventional first and second generation DIF practices have focused on the first oval with some modest attempts at the second oval as sources for an explanation for DIF." The pervasive application of ID factors in the DIF literature might relate to Dörnyei and Skehan's (2003, p. 589) contention that, "individual differences in second language learning … have generated the most consistent predictors of second language learning success." Besides, as Zumbo et al. (ibid, p. 139) acknowledged, ID factors should be properly treated as social constructions that need to be explained by contextual or situational/ecological variables. In this research, in a similar vein, approximately half of the detected sources belonged to the ID layer. This might indicate that the present research can identify with both second and third generations of DIF research. Interestingly, all ecological layers, in the study, were filled by the DIF predictors; thus, we may contend that the mission in the third generation was proportionately carried out.

However, the most crucial requirement of the third generation, i.e., addressing the "why" question of validity, was not satisfactorily fulfilled in the study because the psychometric modeling in the third generation "should explore and allow for latent class and mixture models" (Zumbo, 2009, p. 76). In fact, the mission in the last generation is incomplete unless mixture modeling for multilevel data sets is utilized (Cohen & Bolt, 2005; Vermunt, 2008). Besides, as McNamara and Roever (2006, p. 92) put it, verbal protocols on (gender) groups' thought processes would further our understanding of DIF, but it is almost never done. Sireci and Rios (2013, p. 172) state that for an item to represent bias, both DIF findings and qualitative explanations need to be accounted for. But both quantitative and qualitative pieces of evidence can hardly dispense with theory. The Zumbo et al.'s (2015) ecological model can function as a stand-in for the theoretical rationale for item responding in

ESL/EFL contexts. This is in keeping with interpretive/use validity arguments to support actions based on test scores. The theoretical models can inform and be informed by various contextual manifest/hidden layers. The development of global/local ecological models, however, is an onerous task in practice because

1. enumerating the contributory variables in the cognitive and strategic processes of item responding engenders an unwieldy inventory which needs to be explained, categorized, clustered, minimized, correlated, and ironed out in different settings;
2. researchers need to prioritize and conceptualize the confounding influence of each and every ecological variable;
3. the layers of the model are, presumably, in isolation and ample plausible evidence is required to justify the inclusiveness and relevance of the layers. Furthermore, one can argue, as Widdowson (2001, p. 17) contended for communicative competence models, that the disparate layers of the ecological model might be a static set of components, and as such cannot account for the dynamic interrelationships engaged in the prediction of variations across learners. Thus, it seems that conceptualizing a myriad of contextual parameters will aggravate the shortcomings of measurement and validation, but does not rectify it. Perhaps, as Widdowson (ibid, p. 19) put it, "there must be some potential in the [item] itself that is contextually realized." That is, context is immanent in the test item as an intrinsic valency.
4. even the enumeration of associated layers and inclusive clusters is unidirectional, in that only plausible contributory variables are included; we cannot make any inference about other irrelevant variables that need to be excluded from the model; however, this multidirectional task is quite direful as opening the Pandora's Box might end in formidable consequences;

5. the model needs to be related, advocated, or even differentiated from a theoretical model of test method facets to elucidate the inclusive/exclusive nature of both models and the disparity of direct and indirect tests; and finally,
6. if performance variables are added to the ecological model (Zumbo et al., 2015) the demarcation between administration/methodology and ability gets blurred and might not be explained, e.g. if a female test-taker might get a female-friendly item (O'Neill & McPeek, 1993) correct more often than a male student, does this mean she is more able than the male test taker? Or if she cannot answer it correctly, does it mean that the item is not female-friendly or she does not have enough competence to answer the item in question?

Despite a number of drawbacks in the research, including a rather small sample, lack of mixture modeling and qualitative supporting evidence, and application of a single statistical test, the study had a few beneficial effects, including the application of an ecological model and a questionnaire which are rarely used in the literature, the use of multi-variable matching of DIF predictors, and a collection of literature-backed DIF sources. Analyzing DIF items with reference to the sources approved by the literature helped develop a grounded model of item responding for the EFL reading assessment. Thus, a few implications might be conceived of as: the administration of demographic questionnaires in high-stakes assessments, the judicious application of ecological and statistical models, the employment of mixture modeling and mixed-methods approach, the use of literature-based sources as (counter-) evidence for the findings, and putting DIF explanations to an empirical test by intentionally inducing DIF and having the comparison groups answer the items (Ahmadi & Jalili, 2014).

However, the most important implication of the study is the application of the ecological variables as covariates to inform DIF to provide relations-to-

other-variables validity evidence. This piece of evidence investigates "if the relationships between item/test responses and additional variable or covariates … follow the same patterns for identifiable groups of the intended population" (Gómez-Benito et al., 2018, p. 107). Gómez-Benito et al. (2018), further, explicated the exponential exploitation of DIF explanations in all sources of evidence to support explicit purposes, and expounded on the lack of DIF as "assumptions" that must be examined to expand interpretive/use arguments.

References

Abbott, M. (2007). A confirmatory approach to differential item functioning on an ESL reading assessment. *Language Testing, 24*(1), 7-36.

Ahmadi, A. & Darabi Bazvand, 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*(2), 55-68.

Alderson, J. C. (2000). *Assessing reading.* Cambridge: Cambridge University Press.

Allalouf, A. & Abramzon, A. (2008). Constructing better second language assessments based on differential item functioning analysis. *Language Assessment Quarterly, 5*(2), 120-141.

Allalouf, A. Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36*(3), 185-198.

Aryadoust, V., Goh, C. C. M., & Kim, L. O. (2011). An investigation of differential item functioning in the MELAB listening test. *Language Assessment Quarterly, 8,* 361-385.

Banks, K. (2012). Are inferential reading items more susceptible to cultural bias than literal reading items? *Applied Measurement in Education, 25,* 220-245.

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

Bolt, S. & Thurlow, M. (2007). Item-level effects of the read-aloud accommodation for students with reading disabilities. *Assessment for effective Intervention, 33,* 15-28.

Brantmeier, C. (2001). Second language reading research on passage content and gender: Challenges for the intermediate-level curriculum. *Foreign Language Annals, 34*(4), 325-333.

Brantmeier, C. (2003). Beyond linguistics knowledge: Individual differences in second language reading. *Foreign Language Annals, 36*(1), 33-43.

Brantmeier, C. (2007). Adult second language reading in the USA: The effects of readers' gender and test methods. *Forum on public policy, 14,* 1-34.

Chen, Z. & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing, 2*(2), 155-163.

Cheong, Y. F. (2006). Analysis of school context effects on differential item functioning using hierarchical generalized linear models. *International Journal of Testing, 6*(1), 57-79.

Cheong, Y. F., & Kamata, A. (2013). Centering, scale indeterminacy, and differential item functioning detection in hierarchical generalized linear and generalized linear mixed models. *Applied Measurement in Education, 26*, 233-252.

Cho, H-J., Lee, J., & Kingston, N. (2012). Examining the effectiveness of test accommodation using DIF and a mixture IRT model. *Applied Measurement in Education, 25,* 281-304.

Cohen, A. S. & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement, 42*(2), 133-148.

Cohen, A. & Macaro, E. (Eds.). (2007). *Language learner strategies: Thirty years of research and practice.* Oxford, UK: Oxford University Press.

Dörnyei, Z. (2003). *Questionnaires in second language research.* Lawrence Erlbaum Associates, Inc.

Dörnyei, Z. & Skehan, P. (2003). Individual differences in L2 learning. In C. Doughty & M. Long (Eds.), *The handbook of second language acquisition* (pp. 589-630). Malden, MA: Blackwell Publishing.

Elder, C., McNamara, T. F., & Congdon, P. (2003). Understanding Rasch measurement: Rasch techniques for detecting bias in performance assessment: An example comparing the performance of native and non-native speakers on a test of academic English. *Journal of Applied Measurement, 4,* 181-197.

Elosua, P. & Lopez-Jauregui, A. (2007). Potential sources of differential item functioning in the adaptation of tests. *International Journal of Testing, 7(1),* 39-52.

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage assessments. *International Journal of Testing, 2*(3&4), 199-215.

Ercikan, K., Roth, W., Simon, M., Sandilands, D., & Lyons-Thomas, J. (2014). Inconsistencies in DIF detection for sub-groups in heterogeneous language groups. *Applied Measurement in Education, 27,* 273-285.

Ferne, T. & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly, 4*(2), 113-148.

Finch, W. H., Hernández Finch, M. E., & French, B. F. (2016). Recursive partitioning to identify potential causes of differential item functioning in cross-national data. *International Journal of Testing, 16,* 21-53.

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

Harding, L. (2011). Accent, listening assessment and the potential for a shared-L1 advantage: A DIF perspective. *Language Testing, 29*(2), 163-180.

Helwig, R., Rozek-Tedesco, M. A., Tindal, G., Heath, B., & Almond, P. J. (1999). Reading as an access to mathematics problem solving on multiple-choice tests for six-grade students. *Journal of Educational Research, 93,* 113-125.

Hidalgo, M. D. & Gómez-Benito, J. (2010). Differential item functioning. In P. Peterson, E. Baker, & B. McGaw, (Eds.), *International encyclopedia of education, 4,* (pp. 36-44). Oxford: Elsevier.

Hidalgo, M. D. & López-Pina, J. A. (2004). DIF detection and effect size: A comparison between logistic regression and Mantle-Haenszel variation. *Educational and Psychological Measurement, 64,* 903-915.

Jang, E. E. & Roussos, L. (2009). Integrative analytic approach to detecting and interpreting L2 vocabulary DIF. *International Journal of Testing, 9*(3), 238-259.

Jodoin, M. G., & Gierl, M. J. (2001). Type-one error and power rates using an effect size measure with the logistic regression for DIF detection. *Applied Measurement in Education, 14,* 329-349.

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing, 18*(1), 89-114.

Koo, J., Becker, B. J., & Kim, Y-S. (2014). Examining differential item functioning trends for English language learners in a reading test: A meta-analytical approach. *Language Testing, 31*(1), 89-109.

Kunnan, A. J. (1990). DIF in native language and gender groups in an ESL placement test. *TESOL Quarterly, 24,* 741-746.

Le, L. T. (2009). Investigating gender differential item functioning across countries and test languages for PISA science items. *International Journal of Testing, 9*(2), 122-133.

Lee, H., & Geisinger, K. F. (2014). The effect of propensity scores on DIF analysis: Inference on the potential cause of DIF. *International Journal of Testing, 14,* 313-338.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and Many-Facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*(2), 158-180.

Mazor, K. M., Kanjee, A., & Clause, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement, 32*(2), 131-144.

McNamara, T. & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA & Oxford: Blackwell.

Mendes-Barnett, S., & Ercikan, K. (2006). Examining sources of gender DIF in mathematics assessments using a confirmatory multidimensional model approach. *Applied Measurement in Education, 19*(4), 289-304.

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 1400 text samples. *Discourse Processes, 45,* 211-236.

Oliveri, M. E., Ercikan, K., Zumbo, B. D. (2014). Effects of population heterogeneity on accuracy of DIF detection. Applied Measurement in Education, 27(4), 286-300.

O'Neill, K. A. & McPeek, W. M. (1993). Item and test characteristics that are associated with differential item functioning. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning,* 255-276. Hillsdale, NJ: Lawrence Erlbaum Associates.

Oshima, T. C., Raju, N. S., Flowers, C. P., & Slinde, J. A. (1998). Differential bundle functioning using the DFIT framework: Procedures for identifying possible sources of differential functioning. *Applied Measurement in Education, 11*(4), 353-369.

Pae, T. I. (2004b). Gender effect on reading comprehension with Korean EFL learners. *System, 32*(3), 265-281.

Pae, T. I. (2012). Causes of gender DIF on an EFL language test: A multiple-data analysis over nine years. *Language Testing, 29*(4), 533-554.

Roever, C. (2005). That's not fair: Fairness, bias, and differential item functioning in language testing. *SLS Brownbag*, 1-14.

Roth, W. M., Oliveri, M. E., Sandilands, D. D., & Lyons-Thomas, J. (2013). Investigating linguistic sources of differential item functioning using expert think-aloud protocols in science achievement tests. *International Journal of Science Education, 35*(4), 546-576.

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

Santelices, M. V. & Wilson, M. (2012). On the relationship between differential item functioning and item difficulty: An issue of methods? Item response theory approach to differential item functioning. *Educational and Psychological Measurement, 72*(1), 5-36.

Sasaki, M. (1991). A comparison of two methods for detecting differential item functioning in an ESL placement test. *Language Testing, 8*(2), 95-111.

Shermis, M. D., Mao, L., Mulholland, M., & Kieftenbeld, V. (2017). Use of automated scoring features to generate hypotheses regarding language-based DIF. *International Journal of Testing, 17*(4), 351-371.

Shimizu, Y., & Zumbo, B. D. (2005). Logistic regression for differential item functioning: A primer. *Japan Language Testing Association Journal, 7,* 110-124.

Sireci, S. G., & Rios, J. (2013). Decisions that make a difference in detecting differential item functioning. *Educational Research and Evaluation, 19*(2-3), 170-187.

Stricker, L. J., & Emmerich, W. (1999). Possible Determinants of differential item functioning: Familiarity, interest, and emotional reaction. *Journal of Educational Measurement, 36*(4), 347-366.

Suh, Y., & Talley, A. E. (2015). An empirical comparison of DDF detection methods for understanding the causes of DIF in multiple-choice items. *Applied Measurement in Education, 28*, 48-67.

Takala, S. & Kaftandjieva, F. (2000). Test fairness: A DIF analysis of an L2 vocabulary test. *Language Testing, 17*(3), 323-340.

Taylor, C. S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education, 25*, 246-280.

Tsaousis, I., Sideridis, G., & Al-Saawi, F. (2018). Differential distractor functioning as a method for explaining DIF: The case for a national admissions test in Saudi Arabia. *International Journal of Testing, 18*(1), 1-26.

Uiterwijk, H. & Vallen, T. (2005). Linguistic sources of item bias for second generation immigrants in Dutch tests. *Language Testing, 22*(2), 211-234.

Vermunt, J. K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical Methods in Medical Research, 17,* 33-51.

Widdowson, H. (2001). Communicative language testing: The art of the possible. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara, & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Essays in honour of Alan Davis* (pp. 12-21). Cambridge: Cambridge University Press.

Wu, A. D. & Ercikan, K. (2006). Using multiple-variable matching to identify cultural sources of differential item functioning. *International Journal of Testing, 6*(3), 287-300.

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

Zumbo, B. D. (2007). Three generations of DIF analysis: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223-233.

Zumbo, B. D. (2009). Validity as contextualized and pragmatic explanation, and its implications for validation practice. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 65-82). Charlotte, NC: IAP-Information Age Publishing, Inc.

Zumbo, B. D. & Gelin, M. N. (2005). A matter of test bias in educational policy research: bringing the context into picture by investigating sociological/community moderated (or mediated) test and item bias. *Journal of Educational Research and Policy Studies, 5*(1), 1-23.

Zumbo, B. D., Liu, Y., Wu, A. D., Shear, B. R., Olvera Astivia, O. L., & Ark, T. K. (2015). A methodology for Zumbo's third generation DIF analyses and the ecology of item responding. *Language Assessment Quarterly, 12,* 136-151.

Appendix

Questionnaire: Ecological DIF predictors and their frequencies in the respondents' answers

1.	Your (family's) socioeconomic status	High (73%)	Average/Low (27%)
2.	Ethnicity	Fars (72.7%)	non-Fars (27.3%)
3.	Your propensity to take (college) courses	High (73.9%)	Average/Low (26.1%)
4.	Your Interest in the subject matter	High (45.8%)	Average/Low (54.2%)
5.	L1	Persian (71%)	non-Persian (29%)
6.	Your familiarity with content	High (38.7%)	Low (61.3%)
7.	Academics	English (48.2%)	non-English (51.8%)
8.	The difficulty level of the test	High (68.9%)	Low (31.1%)
9.	Your accommodation (convenience) status	Satisfactory (64.3%)	Unsatisfactory (35.7%)
10.	The effect of plausible distracters on your choice	High (65.6%)	Low (34.4%)
11.	Vocabulary knowledge	High (32.2%)	Average/Low (67.8%)
12.	Grammar knowledge	High (39.4%)	Average/Low (60.6%)
13.	Reading score	High (35.3%)	Average/Low (64.7%)
14.	Reading strategy	Bottom-up (52.2%)	Top-down (47.8%)
15.	The emotional reaction to item/test	Yes (13.3%)	No (86.7%)
16.	Test rubric/instruction difficulty	Yes (27.6%)	No (72.4%)
17.	Motivation	High (29.6%)	Low (70.4%)
18.	Your Interest in multiple-choice items	High (71.1%)	Low (28.9%)
19.	The effect on your performance of contextual variables (e.g., classroom size, temperature, etc.)	High (57.6%)	Low (42.4%)
20.	The effects of unanticipated/out-of-school events on your performance	High (59.2%)	Low (40.8%)
21.	Joy/like reading	High (71.4%)	Low (28.6%)
22.	Like read – Fiction	High (49.5%)	Low (50.5%)
23.	Like read – non-fiction	High (73.2%)	Low (26.8%)
24.	Time – Language lessons	Ample (76.8%)	Insufficient (23.2%)
25.	At school – Group work	Yes (50.5%)	No (49.5%)

26.	Teachers stimulation of reading engagement	High (78.3%)	Low (21.7%)
27.	Teacher-student relations	High (77.8%)	Low (22.2%)
28.	Parental education	Academic (59.7%)	Non-academic (40.3%)
29.	Wealth/family's income	High (44.2%)	Average/Low (55.8%)
30.	Home educational resources	High (69.7%)	Low (30.3%)
31.	Online reading	High (50.7%)	Low (49.3%)
32.	Neighborhood	Uptown/Up (50%)	Downtown/Down (50%)