IZVORNI ZNANSTVENI RAD
UDK:811.112.2'34:81'243:305
DOI: 10.22210/strjez/50-2/3
Primljen: 11. 2. 2021.
Prihvaćen: 21. 5. 2021.

Improving the fairness of ECL listening tests by detecting gender-biased items

Hrisztalina Hrisztova-Gotthardt
hrisztova@inyk.pte.hu
Foreign Language Centre, University of Pécs

Réka Werner
werner.reka@inyk.pte.hu
Foreign Language Centre, University of Pécs

The first main objective of this study is to examine whether ECL listening tests in German at CEFR level B2 are equally fair to both male and female test takers. The second aim is to explore distinctive patterns that can be associated with gender-related differential item functioning (DIF) and potentially result in gender-biased items. For this purpose, two complementary approaches are used:

(1) a multi-faceted Rasch analysis in order to detect items showing DIF in terms of gender, and

(2) item characteristics and content analyses in order to identify potential systematic sources of gender-biased items.

According to the results of the statistical analysis, only 6.5 per cent of the items show differential item functioning, thus verifying that the ECL German listening tests are generally fair toward the two gender groups. The findings of the study did not show any systematic patterns that can be clearly associated with gender-related DIF or can potentially result in gender-biased items.

Keywords: *test fairness, validity, DIF, gender-biased item, ECL German listening tests.*

1. INTRODUCTION

Reliability and validity have been for decades the two main test characteristics language test developers pay the most attention to while designing their tests (cf. Bachman, 1990: 24). Recently, however, the term *test fairness* has been appearing with increasing frequency in papers, studies, and presentations on the topic of language assessment (e.g., Kane, 2010; Kremmel, 2019; Kunnan, 2000; 2004; 2007; 2014; Stoynoff, 2012). Professional guidelines such as the *Code for Fair Testing Practices in Education* (Joint Committee on Testing Practices, 2005: 23), the *ETS International Principles for the Fairness in Assessments* (ETS, 2016: 3–4) and the *ALTE Principles of Good Practice* (ALTE, 2020: 13) also emphasize that test developers should strive to make their tests as fair as possible for candidates of different gender, age, ethnic origin, cultural and language background, and special handicapping conditions and needs. Accordingly, educational and, in particular, language tests should guarantee for all test takers the opportunity to demonstrate their standing on the construct of interest, and should "not advantage or disadvantage some individuals because of characteristics irrelevant to the intended construct" (AERA, APA & NCME 2014: 50). One of the most effective strategies for achieving this goal is to construct bias-free tests.

The ECL[1] language examination system also places great emphasis on test fairness at all stages of test development. The ECL language examinations can be taken at four levels according to the *Common European Framework of Reference for Languages* (CEFR) (Council of Europe, 2001; 2018): A2, B1, B2, and C1 (see https://eclexam.eu). There are various approaches to bias detection carried out by item writers, item reviewers, and psychometricians with the aim of detecting potentially biased items. In this context, special attention is paid to identifying gender-biased items, i.e., items 'favoring' or 'disfavoring' male or female candidates.

Considering the various procedures implemented during the different stages of test development, it may be assumed that ECL tests contain only a limited number of gender-biased items. In order to prove this assumption, the present study aims to examine to what extent ECL listening test items at CEFR level B2 administrated between February 2018 and December 2019 exhibit differential item functioning towards test-taker groups in terms of

[1] ECL is the official abbreviation for the European Consortium for the Certificate of Attainment in Modern Languages. AFU Privates Bildungsinstitut GmbH, the official partner of the consortium in Germany, is responsible for the construction and development of ECL language tests in German.

gender. Additionally, the factors producing these gender-related differences are to be explored. The main objective of the item analysis is to discover potential systematic patterns (i.e., specific item features such as item content, item type, item facility value, and item-fit index) that can potentially lead to gender-biased items.[2] The conclusions drawn in the course of the study can be used for improving the fairness of both ECL tests and high-stakes language tests in general.

2. THEORETICAL BACKGROUND OF THE STUDY AND EMPIRICAL STARTING POINTS

2.1. Test validity and sources of invalidity

As mentioned in Section 1, validity is one of the most important test quality indicators as far as language assessment tests are concerned. In his first validity framework, Messick describes validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (Messick, 1989: 13). A test is considered to be valid if the scores achieved by test takers allow stakeholders and users to make plausible inferences about the specific test takers' ability assessed by the test (see AERA et al., 2014: 11; Bachman, 1990: 25; Wesolowski & Wind, 2019: 438).

The hypothesized ability measured by a test is known as a 'construct'. In language testing, this usually refers to a specific language ability such as listening, reading, writing, or speaking skills (see ALTE, 1998: 139). Experts in the field of psychological, educational, and language testing, however, emphasize that tests do not always assess exclusively the targeted construct and are, therefore, not necessarily valid by default. Experts distinguish between two main sources of invalidity: construct underrepresentation and construct-irrelevant variance. According to the *Standards for Psychological and Educational Testing*, construct underrepresentation "refers to the degree to which a test fails to capture important aspects of the construct" (AERA et al., 2014: 12). On the other hand, construct-irrelevant variance refers to the degree to which test scores are "systematically influenced to some extent by processes that are not part of the construct" (AERA et al., 2014: 12). A process that is extraneous to the test's intended purpose can

[2] For a detailed review of previous research on gender-related DIF and gender-biased items or tests, see Geranpayeh & Kunnan (2007: 191–193) and Pae (2012: 534–537).

be an emotional reaction to the test content, familiarity with the subject matter, or level of interest or motivation, among other factors. Construct-irrelevant variance might, thus, result in unfairness towards individual test takers or groups of test takers, leading to a systematic over- or underestimation of their ability level.

2.2. Test fairness and threats to fairness

Test fairness has gradually become a fundamental issue in educational assessment during the last two decades. Following this positive development, the editors of the *Standards for Psychological and Educational Testing* included a standalone chapter on fairness in testing in the 2014 revision of the volume. According to the definition given by the *Standards*, fairness is closely associated with validity (see also Kane, 2010: 181). A fair test "reflects the same construct(s) for all test takers, and scores from it have the same meaning for all individuals in the intended population; a fair test does not advantage or disadvantage some individuals because of characteristics irrelevant to the intended construct" (AERA et al., 2014: 50). Put differently, if irrelevant test content affects the scores of all test takers to about the same extent, validity is threatened. If irrelevant test content disadvantages some individuals or groups of test takers and, at the same time, advantages other individuals or groups, then fairness as well as validity are threatened (cf. ETS, 2016: 3). Fairness can be threatened by certain latent test-taker characteristics such as age, gender, ethnic and cultural origin, and first language, as well as by various content-irrelevant variables such as prior knowledge, experiences, and level of interest or motivation (see Wesolowski & Wind, 2019: 450).

Language testing experts have devoted considerable attention to fairness in language assessment. Kunnan, for instance, developed a complete test fairness framework (2000; 2004; 2014). Kunnan proposes two general principles and several subprinciples of fairness and justice. Principle 1 and the two subprinciples associated with it are the most relevant for the purposes of this study and are, therefore, cited below:

> Principle 1: *The Principle of Justice.* A test ought to be fair to all test takers; that is, there is a presumption of treating every person with equal respect.
>
> Subprinciple 1: A test ought to have comparable construct validity in terms of its test-score interpretation for all test takers.

> Subprinciple 2: A test ought not to be biased against any test-taker groups, in particular by assessing construct-irrelevant matters (Kunnan, 2004: 33).

2.3. Item bias and potential sources of bias

As stressed by Kunnan (2004: 33), test developers should make efforts to construct fair, unbiased tests. Shepard defines bias as a “systematic error that disadvantages the test performance of one group” (1982: 14). Accordingly, test or item bias occurs when (systematic) differences in test or item performance are not caused by the level of ability being measured but rather by differences in test takers’ individual characteristics that are irrelevant to the targeted construct (see Bachman, 1990: 271; Camilli & Shepard, 1994: 8). For example, an item in a listening test may be easier for men than for women just because of their gender-specific experience or interests, and not due to their listening ability. As a listening comprehension test aims to measure only test takers’ listening ability, the item in question is a biased item that unfairly advantages one group of test takers and, therefore, disadvantages the other group.

In order to avoid or reduce item and test bias, test developers should be aware of the potential sources of bias. Studies in the field of language testing have already identified various sources of bias, including age, gender, cultural and ethnic background, first-language background, background knowledge, and experience related to a particular disciplinary area (cf. Bachman, 1990: 271-279; Elder, 2012: 1; Kunnan, 2007: 110).

2.4. Reducing bias at all stages of test development

As Kunnan (2000: 8) suggests, professionals involved in the process of test development should be aware of the potential sources of bias at all stages of test development and make all possible efforts to prevent the occurrence of bias. For this purpose, various qualitative and quantitative procedures need to be performed. The following section will include a discussion of the fundamental a priori and post hoc procedures for identifying and eliminating item bias that have been strongly recommended by experts in the field of educational and language testing, and that have also been applied by the ECL international examination system.

2.4.1. A priori qualitative procedures

One possible way of ensuring test fairness in terms of test content is to avoid item bias already during test design and development (cf. ALTE, 2011: 80; Elder, 2012: 1; Kremmel, 2019; Kunnan, 2004: 33; Styonoff, 2012: 1). Qualitative procedures aimed at minimizing bias at the writing and reviewing stages include the following:

- Compiling detailed fairness guidelines that include a list of taboo topics (e.g., those that are distressing, gender-specific, or favor candidates with certain professional background knowledge) and detailed instructions on how potential sources of bias can be detected and eliminated (ETS, 2016: 17; Kremmel, 2019).
  Following these recommendations, ECL provides a list of taboo topics and bias checklists to all item writers and reviewers involved in the test development process (Wéber, 2018: 145).
- Organizing regular trainings for item writers and reviewers (ETS, 2016: 17; Kremmel, 2019; Kunnan, 2000: 10). Item writers and reviewers should be regularly trained to examine all aspects of an item or a test for potential bias.
  In this context, ECL organizes annual trainings for item writers and reviewers to review and discuss potential sources of bias in detail.
- Having test material – including texts, rubrics and items – reviewed by experts other than the item writers themselves (ETS, 2016: 17).
  In accordance with this guidance, ECL has a bias sensitive team reviewing all items for potential bias before resolving the tests for piloting.
- Using heterogeneous sets of item writers and reviewers (Elder, 2012: 1; ETS, 2016: 21; Kremmel, 2019; Kunnan 2004: 39). Involving item writers and reviewers from various age, gender, cultural, and ethnic groups could ensure that potential sources of bias are detected at an early stage of test development.
  Following this recommendation, ECL obtains contributions to the tests in German as a foreign language from professionals who represent diverse groups and a variety of perspectives.

2.4.2. Post hoc quantitative procedures

As item writers and reviewers might not always be able to identify potential sources of bias, it remains open to question whether the qualitative procedures described above are sufficient to ensure that their tests are bias-free.

Therefore, additional actions are required. One way of minimizing bias is to examine test items for differential item functioning (cf. AERA et al., 2014: 51; ALTE, 2011: 80; Elder, 2012: 2; ETS, 2016: 20; Ferne & Rupp, 2007: 114; Kunnan, 2007: 109). According to ETS (2016: 20), "DIF occurs when people in different groups perform in substantially different ways on a test item, even though the people have been matched in terms of their relevant knowledge and skill as measured by the test." In this case, there should be a latent variable (e.g., gender, age, ethnic origin, etc.) in addition to the construct being measured (e.g., listening comprehension) that influences the item response (cf. Steinberg &Thissen, 2013: 349).

Since early 2018, ECL has performed DIF analysis on live test results from listening and reading tests in German as a foreign language.

2.4.3. Post hoc qualitative procedures

DIF analyses, however, only help in detecting potentially biased items. The statistical detection of DIF does not always automatically mean that the item in question is a biased item. Items displaying DIF need to be reviewed for potential bias. In the course of this qualitative analysis, expert panels explore the question of which feature of the items may have resulted in DIF (Ferne & Rupp, 2007: 141).

In search of explanations for DIF and potential sources of bias, ECL also performs qualitative analysis on items exhibiting DIF. All items that have been clearly identified by experts as biased items are neutralized and post hoc score correction is applied before the test results are released. In this manner, the fairness (and validity) of the ECL tests is improved.

3. AIMS OF THE PRESENT STUDY

As explained in Section 2.4., ECL pays special attention to reducing the number of gender-biased items by performing various qualitative and quantitative procedures at different stages of test development. Accordingly, the following hypothesis can be raised:

Due to the variety of a priori and post hoc procedures applied by the ECL examination system with the aim of reducing item bias, it can be assumed that the ECL listening tests in German at CEFR level B2 show only a limited number of gender-biased items.

Testing this hypothesis is the first major objective of the present study. For this purpose, the following research question needs to be answered:

RQ1: Do ECL listening test items exhibit DIF towards test-taker groups in terms of gender? If so, to what extent?

Determining the exact extent of items displaying gender-related DIF is only one of two major objectives of this study. The factors that possibly produce gender-related differences should also be empirically examined. This leads us to the second aim of the present study: exploring potential systematic patterns (e.g., specific item features such as item type, item facility, item content, and item-fit index) that can be clearly associated with gender-related DIF and identified as potential sources of gender bias. In order to accomplish the second major objective of the study, two additional research questions need to be addressed:

RQ2: Are there any systematic patterns of observed score differences between the two groups that constantly favor one group over the other and lead to inappropriate interpretations of test results?

RQ3: Do the items flagged for DIF effectively advantage or disadvantage one of the gender groups being examined, thus indicating the presence of gender bias?

4. METHODOLOGY

In line with its two main objectives, the present study comprises two parts: (1) gender-related DIF analysis and (2) analysis of the potential causes of gender-related DIF and sources of gender bias. As currently over 95 per cent of the live test population taking the ECL exams in German consists of candidates having the same first language (Hungarian) and age range (between 17 and 25 years), in the quantitative post-test analyses only the variable of gender is considered.

4.1. Instrument

ECL listening tests in German at CEFR level B2 were considered for this study. The tests being examined were administrated between February 2018 and December 2019. As ECL examinations in German are administered five times a year, ten sets of listening tests were examined for gender-related DIF. The DIF analysis was based on Verhelst's conditional statement that "[a]n item shows no DIF if in the (conceptual) population of boys with an arbitrary but fixed level of proficiency and the (conceptual) population of girls with the same level of proficiency, the p-values of the item are identical" (2004: 11).

Investigating DIF on a long-term basis is still considered a desideratum in the area of language assessment (cf. Pae, 2012: 534). Accordingly, the present study aims to examine the presence of gender-related DIF not just at a single point in time (i.e., a single administration term) but across multiple data collection points.

The ECL listening comprehension tests assess test takers' listening skill through two different task types: a multiple-choice task with three options and an open-ended question task requiring short answers. Each task contains 10 items, which makes a total of 20 items for each test. All items are scored dichotomously, i.e., each response to an item is scored as either correct or incorrect.

4.2. Data

For the purposes of the present study, the following data was collected and analyzed:

- item-level responses (correct or incorrect responses) for each item and test taker, and
- test takers' gender (male or female).

The test takers whose exam results were analyzed in the course of this study were previously informed that their background and performance information may be used anonymously for the purposes of statistics, research, and quality assurance (see *ECL Exam Regulations* and *Regulations for Protecting Data from ECL Language Examinations*).

4.3. Analytical approach

Two complementary approaches were used in this study: (1) statistical (DIF) analysis and (2) item characteristics and content analysis.

4.3.1. Statistical (DIF) analysis

To date, different DIF detection methods have been developed with the purpose of identifying items that function differently for two groups of test takers who have been statistically matched based on their ability (cf. Pae, 2012: 533). The statistical approach applied in this study is based on the *Many-Facet Rasch Management* model (hereinafter MFRM) developed by John Linacre (1994). One of the most important advantages of MFRM is that the model provides "a fine-grained analysis of multiple variables potentially having an impact on test or assessment outcomes" (Eckes, 2009: 3). Hence, the MFRM model incorporates more variables than the two considered by the basic Rasch model (cf. Rasch, 1980), namely items and test takers. As one of the major objectives of this study is to examine to what extent the variable of gender has an impact on test results, MFRM proved to be a suitable statistical approach for investigating this research question.

The analysis was performed with the help of the MFRM-based software *Facets* (Green, 2013: Chapter 14). *Facets* was designed to construct measures from complex data involving combinations of different variables (or facets). The facets examined in the course of the present study were the following: test takers' ability, item facility, and test takers' gender.

Regarding the variable 'gender', the following two test takers' groups were defined: male and female examinees. The groups of test takers for whom items should be investigated for DIF are commonly referred as "the reference group and the focal group, where the suspicion is that the focal group(s) might be unfairly disadvantaged on items due to DIF" (Ferne & Rupp, 2007: 115). The objective of this study is to test the hypothesis that the ECL listening tests in German at CEFR level B2 show only a limited number of gender-biased items and that examinees with a similar level of listening ability have a nearly equal chance to answer the items correctly, regardless of their gender. Accordingly, in this particular case it is not relevant which group is designated as reference and which as focal (cf. Eckes, 2011: 363).

4.3.2. Additional analyses for detecting systematic patterns and potential bias

Experts in psychological and educational testing agree that statistical evidence of DIF in a test does not automatically suggest the presence of bias in the items (cf. AERA et al., 2014: 51; Camilli & Shepard, 1994: 7-8; Wesolowski & Wind, 2019: 450). As Eckes points out, "[w]hen an item shows DIF, this is no more than statistical information that something unexpected has happened, something that needs to be explained" (2011: 363–364). DIF only provides a starting point for further investigations that should help detect potential biased items. As part of this study, several additional analyses were carried out with the goal of exploring both content-related and not-content-related causes of DIF and potential sources of gender bias. In the search for systematic patterns that may result in gender bias, the following four relationships were analyzed:

- items exhibiting DIF – item facility value,
- items exhibiting DIF – item fit-index,
- items exhibiting DIF – item type, and
- items exhibiting DIF – item content.

For receptive skills (i.e., reading and listening), the ECL examination system performs Classical Test Theory (CTT) analysis on each data set where item facility values are calculated. The *p*-values are expected to be between 0.3 and 0.7 (Fulcher, 2010: 182). In the course of the analysis, it was examined whether the facility value of the items displaying DIF showed any systematic deviations from the predefined range.

As mentioned in Section 4.2.1., the statistical software *Facets* performs Rasch-based analysis on the data. The original Rasch model is based on probability theory, which assumes that the probability of a person answering an item correctly is a function of the person's ability and the item's difficulty (Green, 2013: 151; Henning, 1987: 107–108). Accordingly, when Rasch analysis is performed, item difficulty and person ability logit figures are estimated along with item and person fit statistics. Item fit-indices (i.e., infit MnSq[3] and outfit MnSq values) are expected to be between 0.5 and 1.5 (Green, 2013: 169; Linacre & Wright, 1994: 370). MnSq values that fall outside this range indicate that the particular item or test taker does not perform as expected and does not fit the model. Such items or persons are called misfits. Misfitting items can be, among other reasons, an indicator of a validity issue. In this context, the MnSq values (or item fit-indices) of the items flagged for gender-related DIF were also taken into account in the course of the analysis. The question of whether there is a systematic relationship between items showing DIF and misfitting items was investigated.

An additional analysis was performed with the goal of determining whether one of the two item formats applied in ECL German listening tests (e.g., multiple-choice items and open-ended questions requiring short answers) shows DIF more frequently when compared to the other item type.

There can be, however, a construct-irrelevant factor such as a "secondary dimension that is not part of the test construct" (Eckes, 2011: 364), but which influences the performance of a particular examinee group. Therefore, items flagged for DIF should be carefully reviewed by content experts. The goal of the content analysis is to find a possible explanation as to why DIF occurred and whether the items displaying DIF effectively favor or disfavor one of the two test-taker groups.

In the course of this study, content analysis was performed on the listening items showing DIF. Nine content experts representing different ages, genders, native languages, and cultural backgrounds were asked to review the items

[3] MnSq stands for mean square value.

and to decide if there was any evidence suggesting that the items favored either the male or female examinees. A questionnaire originally created by Geranpayeh and Kunnan (2007) was adapted for the purposes of the content analysis. The experts were asked to rate each item on a scale of 1 to 5 for each of the two examinee groups. The point values were defined as follows:

- 1 – an item that strongly advantages the particular gender group
- 2 – an item that slightly advantages the particular gender group
- 3 – a neutral items in terms of gender
- 4 – an item that slightly disadvantages the particular gender group
- 5 – an item that strongly disadvantages the particular gender group

In addition, the experts were asked to write a short comment explaining why an item could favor one of the gender groups. It was assumed that if there was evidence that any of the items exhibiting DIF clearly advantaged or disadvantaged one of the two test-taker groups being examined, then such an item could be biased in terms of gender.

4.3.3. Wilcoxon Signed Rank Test

Kunnan suggests that “evidence from mean score differences between relevant subgroups should be examined and if such differences are found, an investigation should be undertaken to determine that such differences are not attributable to a source of construct underrepresentation or construct-irrelevance variance” (2007: 110).

In order to test whether there was a significant difference in the performance of male and female test takers, a Wilcoxon Signed Rank Test was performed on the 10 ECL Listening tests under investigation. The Wilcoxon Signed Rank Test is a nonparametric statistical test that compares two paired sets or groups. The goal of the test is to determine whether the two groups are different from one another to a statistically significant degree.

5. RESULTS AND DISCUSSION

The following section describes and interprets the results of the quantitative and qualitative analyses carried out in the course of the study.

5.1. Results of the DIF analysis

As mentioned above, altogether 200 listening items from 10 tests were examined for gender-related DIF. The results of the statistical analysis, performed with the MFRM-based software *Facets,* showed differential item functioning for 13 items (**see Appendix**), which corresponds to 6.5 percent of the total number of items. DIF items were found in 6 of the 10 listening tests examined.

Table 1 illustrates the 13 items exhibiting DIF and the respective values.

Table 1. Items displaying differential item functioning in terms of gender

Item number	**Exam period**	**Number of candidates**	**Original item number**	***Target measr Male***	***Target measr Female***	***Target contrast male-female***	***Prob (p)***[a]
1	April 2018	462 Male: 170 Female: 292	12	-0.91	0.05	-0.96	0.0003
2			14	-0.52	-1.19	0.66	0.0108
3	June 2018	394 Male: 149 Female: 245	4	0.43	-0.11	0.54	0.0275
4			6	0.06	0.64	-0.59	0.0178
5	December 2018	280 Male: 108 Female: 172	1	-0.43	0.38	-0.81	0.0073
6			13	0.89	1.67	-0.78	0.0084
7	April 2019	528 Male: 191 Female: 337	12	1.02	0.43	0.59	0.0061
8			13	0.73	0.31	0.42	0.0443
9			16	0.73	1.28	-0.55	0.0129
10	June 2019	435 Male: 167 Female: 268	14	-0.38	-1.05	0.67	0.0081
11	October 2019	250 Male: 92 Female: 158	6	-0.80	-1.49	0.69	0.0389
12			11	0.14	0.80	-0.66	0.0377
13			15	-0.35	0.32	-0.67	0.0330

[a] In the field of applied linguistics and language testing, in general, a probability level of 0.05 is used. (Green, 2013: 90)

For each of the 13 items displaying DIF, the following values were calculated:

- *Target Measr* (male and female) shows the difficulty level of each item in logits. The lower the logit value, the easier the item, and the higher the logit value, the more difficult the item (cf. Green, 2013: 168).

Target Contrast indicates the difference between the logits of the two groups under investigation. In this particular case, a negative value indicates that the item proved to be more difficult for women, whereas a positive value indicates that men had more difficulty completing the item. Accordingly, 7 items (1, 4, 5, 6, 9, 12, and 13) were more difficult for the female examinees and 6 items (2, 3, 7, 8, 10, and 11) for the male examinees. Except for December 2018, when both items flagged for DIF favored male candidates, in all examination periods under investigation there was an approximately equal number of DIF items favoring male and female test takers, respectively. Therefore, an advantage for a group on some items could be neutralized by a disadvantage for the same group on other items (cf. Elder, 2012: 3).

- A t-test was performed to calculate the probability of the two test-taker groups responding differently to an item. If the *p*-value is less than or equal to 0.05 ($p \leq 0.05$), the result is considered to be statistically significant. Based on the *p*-values listed in Table 1, it can be concluded that the male and the female examinees performed significantly differently on the 13 items in question.

This finding answers the first research question of the present study: there are, indeed, several ECL listening test items that exhibit DIF towards test-taker groups in terms of gender. In total, 6.5 percent of all items under investigation were classified as having gender-related DIF.

5.2. Results of the additional analyses

The various item characteristics analyses described below aimed at finding systematic patterns that can be clearly associated with gender-related DIF. The goal of the content analysis, on the other hand, was to investigate if any of the 13 items showing DIF effectively advantage or disadvantage one of the two test-taker groups, thus indicating the presence of bias.

5.2.1. Items showing DIF – item facility value

As described in Section 4.3.2, item facility values (*p*) are expected to be between 0.3 and 0.7. Based on the results of the CTT analyses (see Table 2 below), it can be concluded that the 13 items displaying gender-related DIF do not show systematic deviation from the suggested range.

Table 2. Items displaying DIF and their facility value (p-value) and discrimination index

Item number	***p*-value**	***D*[b]**
1	0.60	0.62
2	0.71	0.37
3	0.61	0.59
4	0.55	0.68
5	0.63	0.56
6	0.40	0.60
7	0.38	0.55
8	0.42	0.47

Item number	*p*-value	*D*[b]
9	0.31	0.47
10	0.76	0.54
11	0.70	0.34
12	0.42	0.73
13	0.51	0.68

[b] *D* stands for discrimination index. The discrimination index is a measure of how well an item is able to distinguish between high-performing and low-performing examinees. In all 13 cases the discrimination index is within the recommended range of *D* ≥ 0.30 (Crocker & Algina, 2016: 315).

Except for item 2 ($p = 0.71$) and item 10 ($p = 0.76$), all *p*-values are within the suggested range. Thus, there is no explicit relationship between item difficulty and DIF.

5.2.2. Items showing DIF – item fit-index

Considering the results of the Rasch-based analysis, it can be concluded that there is no obvious correlation between the misfit items and the items exhibiting DIF. There are only two items for which the outfit index falls outside the range: item 2 (*Outfit MnSq* = 1.68) and item 11 (*Outfit MnSq* = 1.65) (see Table 3). In the course of the post hoc analysis carried out during the respective examination periods, both items were neutralized and post hoc score correction was applied before the test results were released.

Table 3. Item displaying DIF and their item fit-index

Item number	*Outfit MnSq*	*Infit MnSq*
1	0.83	0.91
2	1.68	1.10
3	1.01	1.03
4	0.88	0.92
5	1.17	1.01
6	1.02	1.07
7	1.03	1.05
8	1.14	1.11
9	0.96	0.99
10	0.73	0.86
11	1.65	1.30
12	0.80	0.91
13	0.86	0.93

5.2.3. Item showing DIF – item type

As described in Section 4.3.2, two item types can be found in the ECL German listening tests, namely multiple-choice items and open-ended questions requiring short answers. Four of the 13 items flagged for DIF (3, 4, 5, and 11) belong to the multiple-choice task type. The remaining 9 items are open-ended questions requiring short answers. Based on these results, it can be assumed that the items with open-ended questions are more likely to display gender-related DIF. This finding could serve as an important clue for item writers, who should pay greater attention to potential advantages or disadvantages for one of the gender groups when creating items of this type.

There is, however, no empirical evidence that either item type systematically advantages or disadvantages one of the two test-taker groups under investigation. In the case of multiple-choice items, two items (4 and 5) were more difficult for female candidates, and two items (3 and 11) were more challenging for male candidates. As for the other item type, a similar pattern emerges: five items (1, 6, 9, 12, and 13) were more difficult for female candidates, and four items (2, 7, 8, and 10) were more challenging for male candidates.

5.2.4. Content analysis

As already described in Section 4.3.2., nine content experts were asked to examine the 13 items displaying DIF and to rate them on a 1–5 scale based on their potential to result in gender bias. Table 4 illustrates the average ratings of the content experts for the 13 items (rounded up to one decimal point).

Table 4: Average content expert ratings by gender

Item	Item content	Average rating for male examinees	Average ratings for female examinees
1	Different cigarette parts and their function	3	3
2	Cigarette manufacturing process	3	3.2
3	Ordering organic food on the internet	3.4	2.6
4	Content of the organic food box	3.2	2.8
5	Possible reasons for moving to Berlin	3	3
6	Places for getting Christmas trees	2.8	3.2
7	Organizing an exhibition on the topic of fast fashion	3.4	2.6

Item	Item content	Average rating for male examinees	Average ratings for female examinees
8	Growing areas of cotton	3	3
9	The original color of cotton products	2.9	3
10	Another name for environmentally conscious students	3.1	2.9
11	Cultural events and modern communication media	3	2.8
12	The hunting partner of the US president	2.8	3.2
13	The first creation of Margarethe Steiff, the "mother" of the Teddy bear	3.3	2.6

In the case of items 1, 2, 5, 8, 9, 10, and 11, the average ratings suggest that the items were judged to be either gender neutral (3.0) or to slightly advantage (2.8 – 2.9) or disadvantage (3.1 – 3.2) one gender group, while there is neither an advantage nor a disadvantage for the other gender group (3.0). Due to space constraints, these items are not discussed in detail here. Only the items for which the average score showed greater deviations for both test taker groups are examined more closely below.

The *Target contrast* values of item 3 (0.54) and item 7 (0.59), estimated by *Facets*, indicate that these items proved to be more difficult for male examinees.

For item 3, the average ratings of the content analysis assume a slight advantage (2.6) for female test takers and a slight disadvantage (3.4) for male test takers. The content experts justified their opinion using the following argument: the topic of the item is shopping for groceries and organic food, in particular. As women buy groceries more often and they are more interested in organic food than men, female candidates could have a slight advantage when answering this question. The experts' opinion thus verifies the results of the DIF analysis. It can be therefore concluded that the subject matter of the item could be the reason for the relatively higher performance of female examinees on this item. The item may be therefore biased in terms of gender.

For item 7, the average ratings of the content analysis assume a slight advantage (2.6) for female test takers and a slight disadvantage (3.4) for male test takers. As the topic of the task is 'fast fashion' and women seem generally to be more interested in fashion than men, several experts assumed that female test takers were listening more carefully to the recording than male test takers. The item itself, however, was judged as rather neutral and unbiased in terms of gender.

The *Target contrast* values of item 4 (-0.54), 6 (-0.78), 12 (-0.66), and 13 (-0.67), as estimated by *Facets*, indicate that these items proved to be more difficult for female examinees.

For item 4, the average ratings of the content experts suggest exactly the opposite: the experts assume a slight advantage (2.8) for female test takers and a slight disadvantage (3.2) for male test takers. There were, however, only two experts who expressed a vague assumption that the text topic (i.e., ordering organic food on the internet) could give a small advantage to female test takers. There was no clear evidence for the presence of gender bias in item 4.

For item 6, the average ratings of the content analysis assume a slight advantage (2.8) for male test takers and a slight disadvantage (3.2) for female test takers. According to three content experts, usually men are responsible for getting a Christmas tree, which makes them more experienced on the topic. This could be the reason for the better performance of male examinees on this item and could indicate the presence of a gender bias.

For item 12, the average ratings of the content analysis assume a slight advantage (2.8) for male test takers and a slight disadvantage (3.2) for female test takers. Three content experts believed that the "hunting" topic of the task and the "business partners" answer to item 12 are themes generally more relevant in men's everyday lives. This could be the reason for the better performance of male examinees on this item and could indicate a gender bias.

For item 13, the average ratings of the content experts suggest exactly the opposite: the experts assumed a slight advantage (2.6) for female test takers and a slight disadvantage (3.3) for male test takers. The experts claimed that the vocabulary used in the paragraph to which the item refers ("sew", "felt", "pincushion", etc.) could, to a lesser degree, favor female examinees. The experts pointed out, however, that the correct answer to item 13 ("clothes") should not cause difficulties for male test takers. Accordingly, there is no reason to believe that the item was biased toward any gender group.

5.3. Wilcoxon Signed Rank Test

The Wilcoxon Signed Rank Test was applied in order to compare the overall performance of the two test-taker groups on the six listening tests containing DIF items and to determine whether there was a significant difference between the performances of male and female examinees. Although the analysis is based on the item-level results of the two groups, the estimated *p*-value refers to the performance on the whole test. As the *p*-value is higher than 0.05 for all listening tests (see Table 5), it can be concluded that there is no significant difference in the overall results of the two groups. In other words, there is no evidence for an advantage or disadvantage on the ECL language tests towards either of the test-taker groups.

Table 5. Wilcoxon test: Asymptotic Significance (p-value)

Exam period	Number of examinees	Number of male examinees	Number of female examinees	*p*-value
April 2018	462	170	292	0.082
June 2018	394	149	245	0.837
December 2018	280	108	172	0.911
April 2019	528	191	337	0.784
June 2019	435	167	268	0.797
October 2019	250	92	158	0.857

6. CONCLUSION AND OUTLOOK

In this study, 10 ECL German listening tests at CEFR level B2 were examined for potential gender-biased items. Two complementary approaches were used: First, the 200 listening items were examined for DIF, and second, the items exhibiting DIF were subjected to content and non-content analysis.

The DIF analysis showed differential item functioning for 13 items on 6 tests, which corresponds to only 6.5 percent of the total number of items. This finding verifies the hypothesis of the study – namely, that, due to the variety of a priori and post hoc procedures applied by the examination system with the aim of reducing item bias, ECL tests would show only a limited number of potentially gender-biased items.

In the course of the item characteristics analyses, no clear pattern emerged in terms of why particular items were identified as having DIF. The only evidence of a potential interaction effect was observed in the *item exhibiting DIF – item type* relationship: the majority of the items showing DIF belong to the task type of open-ended questions requiring short answers. Nevertheless, the empirical evidence was not sufficient to conclude that this particular item type systematically advantages or disadvantages a particular gender group.

The main results of the item content analysis showed that, although statistical procedures detected differential item functioning in a few items, expert judges were not able to clearly identify sources of gender bias in the majority of the items. Only three items were judged as potentially gender-biased items by the content experts due to their subject matter or vocabulary. But in these few cases, the differences in the ratings were negligible.

In summary, it can be stated that the findings of the study did not show systematic patterns that can be clearly associated with gender-related DIF or that definitively compromise the fairness of the ECL German listening tests under investigation. Though it has been observed that certain test topics and item types could potentially advantage or disadvantage a particular test-taker group and result in gender-biased items, more systematic analysis needs to be conducted before a definite recommendation is offered.

Future studies in the field of language testing can build on the manifold analytical approach developed for the purposes of this survey, applying a more systematic approach to exploring potential distinctive patterns that can be associated with gender-, age- or L1-related DIF, and examining the presence of DIF across multiple data collection points.

As mentioned in Section 4, the current ECL test population is extremely homogeneous in terms of first language and age. For this reason, currently only the variable of gender can be considered for the purposes of DIF-analysis. In 2020, however, the ECL language tests in German successfully completed an audit and were awarded a Q-Mark by the Association of Language Testers in Europe. As the test population gradually becomes more heterogeneous, more complex analyses can be performed in the near future.

Moreover, in addition to the quantitative and qualitative methods presented in this study, another powerful statistical tool, known as profile analysis, can be used for detecting systematic deviations from the predictions of a measurement model for binary items (Verhelst, 2012). With the help of the program *Profile-G*, launched by Verhelst (2018), systematic deviations at the group level can be identified.

The conclusions drawn in the course of such studies may help test developers rule out bias with certainty and ensure test fairness.

Acknowledgements

The authors would like to thank the content experts for reviewing the 13 items flagged for DIF and providing insightful comments. The authors would like to express their gratitude to Gábor Szabó for his helpful advice

regarding the statistical methods applied in this study. A special thanks goes to Peter Sabath for proofreading the first version of the manuscript.

REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA) & National Council on Measurement in Education (NCME) (2014) *Standards for Educational and Psychological Testing*. Washington DC: American Educational Research Association.

Association of Language Testers in Europe (ALTE) (1998) *Multilingual Glossary of Language Testing Terms*. Cambridge: Cambridge University Press.

Association of Language Testers in Europe (ALTE) (2011) *Manual for Language Test. Development and Examining. For Use with the CEFR*. Strasbourg: Council of Europe.

Association of Language Testers in Europe (ALTE) (2020) *ALTE Principles of Good Practice*. Cambridge: ALTE.

Bachman, L. F. (1990) *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Camilli, G. & Shepard, L. (1994) *Methods for identifying biased test items*. Thousand Oaks, CA: SAGE Publications.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)*. Cambridge: Cambridge University Press.

Council of Europe (2018) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) Companion Volume with New Descriptors*. Strasbourg: Strasbourg Cedex.

Crocker, L. & Algina, J. (2006) *Introduction to Classical and Modern Test Theory*. Mason, OH: Cengage Learning.

Eckes, T. (2009) Many-facet Rasch Measurement. In Takala, S. (Ed.) *Reference Supplement to the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (Section H). Strasbourg: Council of Europe/Language Policy Division, 52 p.

Eckes, T. (2011) A Study of Differencial Item Functioning in the TestDaF Reading and Listening Section. In Galaczi, E. D. & Weir C. J. (Eds.) *Exploring Language Frameworks: Proceedings of the ALTE Kraków Conference, July 2011* (*Studies in Language Testing*). Cambridge: Cambridge University Press, 362-388.

ECL Exam Regulations. https://eclexam.eu/wp-content/uploads/Exam_Regulations_2020.pdf (accessed 28 January 2021)

ECL Language Examination System: https://eclexam.eu (accessed 28 January 2021)

Educational Testing Service (ETS) (2016) *ETS International Principles for the Fairness of Assessments. A Manual for Developing Locally Appropriate Fairness Guidelines for Various Countries*. https://www.ets.org/s/about/pdf/fairness_review_international.pdf (accessed 28 January 2021)

Elder, C. (2012) Bias in Language Assessment. In Chapelle, C. A. (Ed.) *The Encyclopedia of Applied Linguistics* (Online Edition). pp. 7. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781405198431.wbeal1198 (accessed 25 September 2020)

Facets = https://www.winsteps.com/facets.htm (accessed 28 January 2021)

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. doi: 10.1080/15434300701375923

Fulcher, G. (2010) *Practical Language Testing*. London: Hodder Education. doi: 10.4324/980203767399

Geranpayeh, A. & Kunnan, A. J. (2007) Differential Item Functioning in Terms of Age in the Certificate in Advanced English Examination. *Language Assessment Quarterly* 4 (2), 190-222. doi: 10.1080/15434300701375758

Green, R. (2013) *Statistical Analyses for Language Testers*. London: Palgrave Macmillan. doi: 10.1057/9781137018298

Henning, G. (1987) *A Guide to Language Testing: Development, Evaluation and Research*. Cambridge, MA: Newbury House.

Joint Committee on Testing Practices (2005) Code of Fair Testing Practices in Education (revised). *Educational Measurement: Issues and Practice* 24, 23-29. doi: 10.1111/j.1745-3992.2005.00004.x

Kane, M. (2010) Validity and Fairness. *Language Testing* 27 177-182. doi: 10.1177/02655 32209349467

Kremmel, B. (2019) *Avoiding Bias in Language Test Development*. (Paper presented at the "ATLE 54th Meeting and Conference", Ljubljana, 6-8 November 2019).

Kunnan, A. J. (2000) Fairness and Justice for All. In Kunnan, A. J. (Ed.). *Fairness and Validation in Language Assessment*. Cambridge: Cambridge University Press, 1-13.

Kunnan, A. J. (2004) Test Fairness. In Milanovic, M. & Weir, C. J. (Eds.) *European Language Testing in a Global Context*. Cambridge: Cambridge University Press, 27-48.

Kunnan, A. J. (2007) Test Fairness, Test Bias, and DIF. *Language Assessment Quarterly* 4 (2), 109-112. doi: 10.1080/15434300701375865

Kunnan, A. J. (2014) Fairness and Justice in Language Assessment. In Kunnan, A. J. (Ed.) *The Companion to Language Assessment*. Hoboken, New Jersey: John Wiley & Sons, 1-17. doi: 10.1002/9781118411360.wbcla144

Linacre, J. M. (1994) *Many-Facet Rasch Measurement*. Chicago: Mesa Press.

Linacre, J. M. & Wright, B. D. (1994) Reasonable Mean-square Fit Values. *Rasch Measurement Transactions* 8 (3), 370. https://rasch.org/rmt/rmt83b.htm (accessed 28 January 2021)

Messick, S. (1989) Validity. In Linn, R. L. (Ed.) *Educational Measurement*. Washington, DC: American Council on Education and National Council on Measurement in Education, 13–103.

Pae, Tae-Il (2012) Causes of Gender DIF on an EFL Language Test: A Multiple Data Analysis Over Nine Years. *Language Testing* 29 (4), 533–554. doi: 10.1177/0265532211434027

Rasch, G. (1980) *Probabilistic Models for Some Intelligence and Attainment Tests*. Chicago: University of Chicago Press. (Original work published 1960). doi: 10.1016/s0019-9958(61)80061-2

Regulations for Protecting Data from ECL Language Examinations. https://eclexam.eu/wpcontent/uploads/Regulations_for_Protecting_Data_University_of_Pecs.pdf (accessed 28 January 2021)

Shepard, L. A. (1982) Definition of Bias. In Berk, R. A. (Ed.): *Handbook of Methods for Detecting Bias*. Baltimore: John Hopkins University, 9-30.

Steinberg, L. & Thissen, D. (2013) Item Response Theory. In Comer, J. S. & Kendall, P. C. (Eds.) *The Oxford Handbook of Research Strategies for Clinical Psychology*. Oxford: Oxford University Press, 336-373. doi: 10.1093/oxfordhb/9780199793549.013.0018

Stoynoff, S. (2012) Fairness in Language Assessment. In Chapelle, C. A. (Ed.) *The Encyclopedia of Applied Linguistics* (Online Edition). https://onlinelibrary.wiley.com/doi/abs/10.1002/9781405198431.wbeal0409. (accessed 25 September 2020)

Verhelst, N. D. (2004) *Reference Supplement to the Preliminary Pilot version of the Manual for Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Section C: Classical Test Theory*. Strasbourg: Language Policy Division.

Verhelst, N. D. (2012) Profile Analysis: A Closer Look at the PISA 2000 Reading Data. *Scandinavian Journal of Educational Research* 56 (3), 315–32. https://doi.org/10.1080/00313831.2011.583937. (accessed 24 September 2021)

Verhelst, N. D. (2018) *Profile-G*. http://www.ealta.eu.org/documents/resources/ProfileG-1-2-3PL.zip. (accessed 24 September 2021)

Wéber, K. (2018) Topic Variety and Topic Constraints in ECL Language Examinations. In Zelenická, E. (Ed.) *Moderné jazyky v súčasnej Európe*. Nitra. Filozofická fakulta, Univerzita Konštantína Filozofa, 143–150.

Wesolowski, B. & Wind, S. (2019) Validity, Reliability, and Fairness in Music Testing. In Brophy, T. S. (Ed.) *The Oxford Handbook of Assessment Policy and Practice in Music Education.* Vol 1. Oxford: Oxford University Press, 436-460. doi: oxfordhb/9780190248093.013.20

APPENDIX

Items exhibiting DIF

Item 1

Examination period: April 2018
Original item number: 12

Welcher Teil der Zigarette besteht aus einem Zelluloseextrakt?
der Filter

[…] In der Regel werden Zigaretten mit Filter hergestellt. Dieser besteht aus einem Zelluloseextrakt. Er hält bestimmte Partikel zurück und dient auch der Verfeinerung des Geschmacks. […]

Item 2

Examination period: April 2018
Original item number: 14

Wozu benutzen Tabakkonzerne Zusatzstoffe?
Verbesserung des Geschmacks / (bessere) Konservierung / (bessere) Verbrennung

Dem natürlichen Tabak wird neben der Herstellung eine Vielzahl von Stoffen zugesetzt. [...] Die Tabakkonzerne heben die Zusatzstoffe als Verbesserung des Geschmacks der Zigarette, einer besseren Konservierung und Verbrennung hervor.

Item 3

Examination period: June 2018
Original item number: 4

Die Kunden ...

A/ können bald auch im Internet bestellen.
B/ dürfen die Lieferzeit nicht selbst bestimmen.
C/ <u>erhalten ihre Bestellung direkt in der Wohnung.</u>

[...]
A:
Wie sieht das denn praktisch aus, wenn ich so eine Biokiste kaufen will?
B:
Das ist so einfach, wie wir es von anderen Produkten kennen. Die Kunden klicken sich durch die Anbieter, wählen die Lieferzeit und bekommen die Biokiste an die Haustür geliefert.
[...]

Item 4

Examination period: June 2018
Original item number: 6

Der Inhalt einer Biokiste ...

A/ steht bereits bei der Bestellung im Internet fest.
B/ <u>kann den Kunden **überraschen**.</u>
C/ bleibt Woche für Woche gleich.

[...] Außerdem funktioniert die Biokiste nicht wie ein Einkauf sondern wie ein Abonnement. Die Kiste kommt zum Beispiel einmal pro Woche oder pro Monat. Der größte Unterschied ist aber, dass der Kunde bei einer Biokiste im Normalfall nicht genau weiß, was tatsächlich geliefert wird. [...]

Item 5

Examination period: December 2018
Original item number: 1

Heiko Janssen ist nach Berlin gezogen, weil ...

A/ er die Einwohner Berlins für besonders kultiviert hielt.
B/ er in seiner Heimatstadt keine Arbeit als Taxifahrer fand.
C/ er das große kulturelle Angebot der Stadt attraktiv fand.

[...] Wenn man Heiko Janssen fragt, was ihn aus seiner kleinen Heimatstadt mit ihren knapp 50.000 Einwohnern in die ferne Großstadt gezogen hat, dann verweist er auf die kulturelle Vielfalt Berlins. Damit meint Janssen vor allem das Nachtleben, die Musikszene, die Kinos. [...]

Item 6

Examination period: December 2018
Original item number: 13

Woher kann man sich einen echten Baum holen?
aus dem (naheliegenden) Wald

[...] früher oder später sieht das Plastikbäumchen nicht mehr so schön aus und wandert auf den Müll. Dann sollte man lieber in den naheliegenden Wald gehen und sich einen grünen Baum aussuchen. [...]

Item 7

Examination period: April 2019
Original item number: 12

Wer hat sich an der Organisation der Ausstellung beteiligt? (2)
eine Lehrerin und ihre Schüler

[...] Mit diesem Thema beschäftigt sich die Ausstellung „Fast Fashion. Die Schattenseiten der Mode“ in Dresden. [...] Eine Lehrerin aus Dresden hat mit ihren Schülern einen Teil der Ausstellung gestaltet. [...]

Item 8

Examination period: April 2019
Original item number: 13

Wo wird Baumwolle in erster Linie angebaut?
Indien

Viele T-Shirts bestehen aus Baumwolle. Die wird vor allem in Indien, aber zum Teil auch in China, in den USA und in Afrika angebaut. [...]

Item 9

Examination period: April 2019
Original item number: 16

Was für eine Farbe haben Stoffe aus Baumwolle eigentlich?
(die Farbe von einer) Eierschale

[…] Die Fäden werden dann mit Hilfe von Maschinen zu Stoffen gewebt oder gestrickt. Dieser Stoff muss dann weiter verarbeitet werden, denn von Natur aus hat der Baumwollstoff die Farbe von einer Eierschale. […]

Item 10

Examination period: April 2019
Original item number: 14

Wie nennt man die Schüler, die sich besonders für Umweltschutz einsetzen?
Energie-Experten

[…] Außerdem ist es wichtig, dass die Schülerinnen und Schüler eine besondere „Umwelt-Bildung“ bekommen. Dafür wählt jede Klasse einen oder zwei Schüler, die besonderes Interesse am Umweltschutz zeigen und mit den Lehrern über wichtige Umwelt-Fragen sprechen. Diese Kinder sind die so genannten „Energie-Experten“.[…]

Item 11

Examination period: October 2019
Original item number: 6

In den modernen Kommunikationsmedien …
A/ wird das Theatertreffen live übertragen.
B/ können Zuschauer ihre Eindrücke miteinander teilen.
C/ ist Kunst ein selten diskutiertes Thema.

[…] Gerade das Treffen in Berlin zeigt aber auch, wie eng das Theater und die verschiedenen modernen Kommunikationsmedien inzwischen miteinander verknüpft sind. Das Theatertreffen bekommt inzwischen eine Vielzahl von Reaktionen und Rückmeldungen durch live-Chats und Blogs. Da haben Zuschauer die Möglichkeit sich untereinander auszutauschen. […]

Item 12

Examination period: October 2019
Original item number: 11

Mit wem ging der amerikanische Präsident gerne auf die Jagd?
mit seinen Geschäftspartnern

[...] Der Präsident war ein begeisterter Jäger und freute sich, wenn seine Geschäftspartner für ihn einen Jagdausflug organisierten und ihn dabei begleiteten. [...]

Item 13

Examination period: October 2019
Original item number: 15

Was produzierte Margarethe Steiff in ihrem Laden ursprünglich?
Kleidungsstücke (aus Filz)

[...] Anfangs hatte Frau Steiff ein kleines Geschäft, wo sie Kleidungsstücke aus Filz herstellte und verkaufte. Sie dachte sich aber, dass man aus dem Stoff eigentlich auch andere Sachen herstellen könnte. So nähte Frau Steiff aus Stoffresten einen kleinen Elefanten, der eigentlich als Nadelkissen gedacht war. [...]

Poboljšavanje nepristranosti ECL testova slušanja detekcijom čestica koje diskriminiraju prema spolu

Hrisztalina Hrisztova-Gotthardt
hrisztova@inyk.pte.hu
Foreign Language Centre, University of Pécs

Réka Werner
werner.reka@inyk.pte.hu
Foreign Language Centre, University of Pécs

Prvi je cilj ovoga istraživanja ispitati jesu li ECL testovi slušanja na njemačkom jeziku razine B2 prema ZEROJ-u jednako nepristrani prema muškim i ženskim osobama koje rješavaju test. Drugi je cilj istražiti posebne obrasce koji se mogu povezati s razlikovnošću čestica povezanih sa spolom. U tu svrhu primijenjena su dva komplementarna pristupa:
1) Raschova analiza kako bi se pronašle čestice koje pokazuju funkcioniranje diferenciranih stavki prema spolu
2) analiza karakteristika čestica i sadržaja kako bi se identificirali potencijalni sistematični izvori čestica koje diskriminiraju prema spolu.

Prema statističkoj analizi samo 6,5 % čestica pokazuje razlikovno funkcioniranje i time potvrđuje da je ECL slušni test njemačkoga općenito nepristran prema ijednom spolu. Rezultati studije nisu pokazali obrazac koji se prema razlikovnom funkcioniranju čestica može jasno povezati s određenim spolom ili rezultira česticama koje su pristrane prema određenom spolu.

Ključne riječi: *ECL, nepristranost u testiranju, ,razlikovnost* čestica*, slušni test njemačkoga, validnost.*