Heliyon 9 (2023) e16548

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Investigating rater–student interaction, gender bias, and major bias in the assessment of research seminar presentation

Fitria Arifiyanti [a,*], Soeharto Soeharto [a], Stephen Amukune [a], Son Van Nguyen [a,b], Khalil Aburezeq [a], Achmad Hidayatullah [a], Eri Sarimanah [c]

[a] *Doctoral School of Education, University of Szeged, 30-34, Petőfi S. sgt., H-6722, Szeged, Hungary*
[b] *Thuyloi University, No. 175, Tay Son Street, Dong Da District, Hanoi, Viet Nam*
[c] *Faculty of Language and Literature Education, Pakuan University, Indonesia*

ARTICLE INFO

Keywords:
Gender bias
Major bias
Rater–Student interaction
Many-facet Rasch measurement
Assessment

ABSTRACT

The study aims to investigate the severity of raters toward examinees' performances and to analyze biases subject to genders and academic majors based on the rater–examinee interactions. Data were collected across a 13-week research seminar course in English. Six raters were selected to rate 33 examinees using 18-item criteria. The many-facet Rasch measurement was utilized to analyze rater–examinee interactions and bias factors. The results confirmed that the instrument is reliable and valid. This study depicted the interaction between raters and other facets using a variable map, where the raters exhibited different levels of severity/leniency in scoring students' performances in oral presentations at the research seminar course. The result based on the Rasch analysis also confirmed that gender and academic majors contaminated rater assessment. Bias interaction between raters and student gender was detected, and Rater 6 displayed bias based on gender due to the tendency to give higher scores to female than male participants with a target contrast of 2.05 logits. Bias interaction between rater academic majors and student academic majors was also identified among raters with linguistics and psychology majors.

1. Introduction

Presentation skills are crucial to the 21st century, especially among academics, because they must deliver ideas, reports, research, and projects in seminars, webinars, defenses, and conferences. In this manner, their work can reach more audiences, and they can become more influential, particularly in their fields or industries. Presentation skills can be fostered among students in higher education who are required to present their work in different subjects. These students are typically assessed by leaderboards, academic councils, lecturers, and peers. The results from the raters' assessments can impact students' achievements (Engelhard & Wind, 2018). Therefore, the quality of the assessment of rates should be ensured such that the final result of the examinees will not deviate from actual performance. However, scholars have proven that raters' assessment is influenced by various factors, such as preferences and backgrounds. Demographic factors, including gender (e.g., Refs. [1–3]), and majors (e.g., Refs. [4,5]) are also proposed to influence the decisions of raters regarding the presentation of examinees. Another point of concern is that presentations are difficult to assess; thus, developing a rubric is necessary to achieve valid judgments relevant to tasks [6,7].

[*] Corresponding author.
E-mail address: fitria.arifiyanti@edu.u-szeged.hu (F. Arifiyanti).

https://doi.org/10.1016/j.heliyon.2023.e16548
Received 3 March 2022; Received in revised form 17 May 2023; Accepted 18 May 2023
Available online 23 May 2023

1.1. Gender and academic major in assessment

In higher education, rater assessment is beneficial for ensuring the quality of the evaluation of student performance in oral presentations, especially when possible sources of bias can be detected and mitigated [8]. One of the factors influencing assessment is gender. Scholars demonstrated that gender exerts an impact on decisions and social settings because the judgments of women are based on the ethics of refraining from creating discomfort or harm for others, according to psychosocial studies on gender [1,9]. This tendency is a likely result of the propensity to estimate the capabilities of males. However, males tend to judge females; in other words, men believe that they are doing women a favor by making sensible decisions, which can lead to overestimating female talents [2,8]. The gender of the target individual substantially influences the perceptions of evaluators about the specific presentations of individuals [10,11]. Along with this thought, female presenters obtained the lowest ratings on all described dimensions in early studies on oral presentations conducted by Wilson and Bayard [12]. However, Aryadoust [1] revealed the opposite effect of gender in which male evaluators provided women with high scores while females scored males highly. Correspondingly, Swim [13] proposed that gender has a small but significant influence on the social judgment of raters regarding the work performance of individuals, which is consistent with the findings of Shore [14]. Alternatively, Bauer and Baltes [8] found that gender-based stereotypes led to a negative perception of female interviewers. However, a free-recall intervention reduced the negative bias, demonstrating education's positive role in eliminating gender-related bias. Wu and Tan [3] added that females consider males more competent, which may explain why females give males higher marks, whereas opposite-gender evaluations were inflated (overrated). Moreover, gender exhibits a significant but weak effect on the social judgment of raters [2,12,14].

Another factor is academic majors. According to the sociology of education research, majoring in the same field may help students form ties and result in in-group allocation. Furthermore, individuals who work with in-group members were more pro-social and accommodating than those who work with out-group members [15,16]. Scholars provided evidence and exacerbated that apprehension can considerably impact students' performance in oral presentations [1,10]. For this reason, presenters prefer same-major evaluators to dissimilar-major evaluators. This sense of belonging can also be considered prejudice against out-group members who are viewed as competitors, which leads to the undervaluation of the abilities of out-group members [17]. In the same vein, the empathy of peer raters for presenters may be another source of bias in performance rating. When presenting to an audience, empathic peer raters enable themselves to discover and acknowledge the experiences of peer presenters that reflect their feelings [5]. In other words, people empathize more with people who are similar to them. In the academic setting, empathic raters will identify more strongly with presenters from similar backgrounds/majors and establish a more vital link between their performance and that of presenters is reasonable [18,19].

1.2. Bias in performance evaluation

Despite the increase in awareness among academicians in measuring students' success from multiple aspects, such as performance in an oral presentation, bias in evaluation related to students' performance during an assessment is undeniable. The presence of bias could reduce the integrity of the assessment outcomes and the reliability of the rater assessment. Aryadoust [1] stated that the concept of bias is comparable to that of differentiated person functioning (DIF), in which students perceived competency interacts with their characteristics. As such, bias must not be by chance but substantial (large magnitude) and logically justified during the evaluation process.

Scholars have long examined bias in performance evaluation. For example, in the economic domain, Kraiger and Ford [20] investigated the effect of race on rating and found that supervisors tend to produce higher ratings for subordinates of the same race as theirs. Hoyt [21] examined rater bias in psychology research and found that bias in the superior rating is driven by job-specific elements in response to an overall effectiveness item. However, factors not observed in the overall effectiveness item influence peer- and self-ratings bias. For example, a superior rating was given due to the greater experience and responsibility in evaluating the job. Stonewall et al. [22] argued that bias is grounded on biased attitudes that implicitly and explicitly shape the manner in which individuals perceive the environment. In the classroom, the manifestation of bias evaluation may form negative assumptions about others, body language, words, and interactions. In the context of multicultural classes where students originate from various backgrounds, preformed bias in evaluation frequently occurs. The differences in background, cultures, and ethnicities lead to different styles, such as pronunciation and presentation in English. Students from an English-domain background have the possibility to present better than students from other domains such as mathematics and science, which implies bias in the evaluation of performance if a rater tends to consider the language aspect only. Bias terms such as negative and positive in opposite directions tend to cancel each other out [23]. Therefore, if a rater overestimates the student's performance using the current ratings, whereas another rater or group underestimates the student by the same amount, bias terms will cancel each other out.

1.3. Presentation skills

Oral presentations are multimodal communications where students transform the content of the written texts into visuals or slides that contain pictures, text, and, occasionally, sound. During oral presentations, the text is transformed into verbal or nonverbal communication [24]. These forms of oral presentations build the personality and academic growth of students. Additionally, students' anxiety influences the quality of oral presentations, which may occur at the beginning and end of presentations [25]. However, anxiety can be reduced if the learner is given appropriate instructions prior to the presentation or is continuously exposed to the feared stimuli familiar with oral presentations [1,25]. Another technique to reduce anxiety during presentations is facilitating the presenter's meditation, relaxing, and developing positive thoughts associated with the oral presentations [25].

Oral presentation skills are grouped into three categories: verbal communication, non-verbal communication, and content and organization [1,23]. Effective language improves verbal communication skills, and success lies in the communication skills of the presenter during oral presentations [1]. However, the quality of oral presentations can be influenced by effective speech delivery rate, enunciation, lexical and grammatical competency, and presentation techniques [26,27]. Moreover, the tone of speech is vital for ensuring that the oral presentation is lively and that clarity is maintained [28]. Apart from the tone, possessing the appropriate vocabulary and grammatical skills is critical to ensure that the presenter will select the right words and use them effectively in passing the content [29]. In terms of content and organization skills, a significant feature of academic presentations is the organization of content using discourse markers or phrases [30]. Discourse markers are “a class of lexical expressions drawn primarily from the syntactic classes of conjunctions, adverbs, and prepositional phrases” [31], p. 931. Aryadoust [32] pointed out that discourse markers identify the beginning, continuation, and end of an oral presentation. Conversely, the failure of listeners to distinguish discourse markers can hamper their comprehension of oral presentations [33]. Another factor that impacts the quality of oral presentation is the ability of the presenter to negotiate, discuss, and explain facts to the level of the audience to reduce misinterpretation [34]. Therefore, possessing a repertoire of content and organization skills enables the speaker to tailor the content to the level of the audience to ensure comprehension [35]. Moreover, many factors contribute to nonverbal communication, including time management, body language [36], confidence [25], appearance [37], audio or visual aids such as pictures and slides [38], and eye contact with participants [28,36]. Additionally, the presenter should offer cues between verbal and nonverbal communication, such as friendly facial expressions, to build rapport compared to staring blankly at the listeners [36]. According to Mayer [39], the cognitive theory of multimedia learning and combining text, verbal input, and pictures, such as PowerPoint slides, enhance the effectiveness of message delivery and understanding.

1.4. The many-facet Rasch measurement

The many-facet Rasch measurement (MFRM) is a classical method based on item response theory and an extension of the Rasch model [40]. MFRM intends to detect and measure the interaction of student ability and other factors, such as examiners, students themselves, station, gender, and majors, among others, which may influence student scores [41]. The relevant factors that exert an impact on scores are typically called facets. For example, variations in student scores may emerge due to united bias from examiners. This bias, independently or synergistically, threatens the validity of the assessment outcomes [32,41]. Therefore, although any measurement may have many facets, the students should maintain the objectivity of measurement. However, examiner facets such as bias, behaviors, characteristics, and variability are common errors [41,42]. Unlike the classical test theory, MFRM overcomes sample bias and enables the estimation of the probability of a student with a certain ability to exhibit a particular outcome related to a set of facets. For example, student M of Gender X from station Y was rated by examiner L on domain W. As a diagnostic tool, MFRM can help provide information about the functioning of the rating scale, the behavior of each examiner, and individual student performance [40].

Regarding issues in the introduction and study literature, this study intends to evaluate the presence of bias, assess presentation performances based on raters, and examine the influence of gender and majors on the assessment of students’ presentations using a series of seminars where Doctoral students present their research.

1.5. Research questions

This study aims to assess the interaction between raters and examinees and to investigate bias due to gender and academic majors in research on seminar presentation using the MFRM. The following research questions were formulated:

1. Do raters, students, and the scoring criteria achieve fit validity and reliability criteria based on the MFRM?
2. Does the scale category function appropriately based on the facet model?
3. How is the raters’ severity toward examinee performance?
4. Can gender bias be observed from the rater assessment of the students?
5. Do the academic majors of raters interact with the academic majors of students?

2. Methods

2.1. Participants

The university granted ethics approval with reference number: 22/2021, Date: 26 November 2021, for the research before the research project implementation in the seminar presentation. Participants participated voluntarily by signed the research consent before data collection using the convenience sampling. This study employed a cross-sectional research design using quantitative methods. Data were collected during a 13-week research seminar course in English. Six raters were selected from doctoral students of the previous year with various majors with at least two years of teaching experience, two females and four males. The selected raters are also having teaching experiences in higher education. The minimum requirement to perform MFRM is two raters whereby six raters in models is acceptable [41,43]. The participants (examinees) were 33 students in their first and second years in the doctoral school of education using convenience sampling.

2.2. Instruments

Student presentations were assessed using an evaluation form formulated by Aryadoust [1] (see Appendix A). The evaluation form consisted of 18 items and three scoring rubric categories, namely, 1 = fair/needs improvement, 2 = good, and 3 = excellent/outstanding. The evaluation form focuses on identifying three presentation skills, namely, content and organization skills (four items), non-verbal communication skills (six items), and verbal communication skills (eight items). This study used an online assessment instead of a paper-based one. Therefore, as examinees conduct their research seminar presentation, raters can fill out the online form using smartphones or laptops through Google Chrome, Mozilla Firefox, and other standard Internet browsers.

2.3. Procedures

To converge the perceptions of rates in evaluating examinee performances, the researchers conducted a training session for raters to use the instruments and understand the criteria outlined in the evaluation form. The training comprised rating sessions, presentation practices, and an introduction to the online evaluation form. At the initial stage of the training session, the raters were introduced to the indicators of successful presentation performances, including content and organization skills, non-verbal communication skills, and verbal communication skills. In addition, the raters watched several sample videos on YouTube of famous speakers in academic and public presentations. These sample videos were analyzed by contrasting the presentation performance made by an inexperienced speaker. Finally, the raters were introduced to the use of the scales of the evaluation form based on different subscales, followed by rating session practices, where one of the raters gave an unreal presentation. The training session was conducted for six meetings, with 2 h for each meeting for two weeks.

After the training sessions, the raters were required to rate student presentations in a research seminar course conducted once per week in the doctoral school of education for 13 weeks. The study maintained the anonymity and confidentiality of the identity information during the data analysis. The students gave presentations on various research topics related to their dissertation projects. The students were given 15 min to present their research projects in English. Raters gave an evaluation for about two or three students per week. The researchers requested that the raters fill up the student and rater backgrounds, including gender, academic majors, and age, using the online evaluation form. At the end of the program, 33 students were evaluated successfully by six raters. Finally, this study collected 6 raters × 33 ratee/students × 18 items = 3564 data points with no missing data. This dataset was coded into FACETS software to perform MFRM analysis (see Appendix A).

The MFRM analysis generated the logit scale (logits) for all facets, including raters, examinees, item criteria, and scale categories. However, we only used the related facets to answer the related research questions. There was no particular range score for logits ranging from positive to negative infinity. Raters should fit the Rasch model based on the infit MNSQ, outfit MNSQ, and ZSTD. Scale category functioning analysis was used to identify how well the rater can understand the scale category in the rating instrument. The DIF analysis was utilized to detect bias interaction between gender and academic majors on related facets.

2.4. Data analysis

Data consisting of ratings using the online evaluation form were analyzed using MFRM via FACETS, Version 3.83.5 [43]. First, raw scores from the rating evaluation were imported to Microsoft Excel; afterward, the researchers formulated a specific coding (Appendix A) for performing the multi-rater analysis using FACETS. The original standard version of the mathematical model was modified to cover the interaction between gender and academic majors. The following equation represents the modified model for the MFRM analysis:

$$log\frac{\mathrm{P}_{\mathrm{nmgijmgk}}}{\mathrm{P}_{\mathrm{nmgijmgk}-1}} = B_n - D_i - C_j - M_j - G_j - M_n - G_n - E_h \tag{1}$$

where:

$\mathrm{P}_{\mathrm{nmgijmgk}}$ denotes the probability that examinee n with gender g majoring in m will be awarded a rating of k on scoring criterion i (individual scoring criterion i) by rater j with gender g majoring in m;

$\mathrm{P}_{\mathrm{nmgijmgk}-1}$ stands for the probability that examinee n with gender g majoring in m will be awarded a rating of k − 1 on scoring criterion i by rater j with gender g majoring in m;

B_n = ability (proficiency) of examinee n;

D_i = difficulty level of scoring criterion i;

M_j = academic major of rater j;

G_j = gender of rater j;

C_j = severity of rater j (teacher and peer);

M_n = academic major of examinee n;

G_n = gender of student presenter n; and

E_h = difficulty level of the threshold from categories k − 1 to k of the scale unique to scoring criterion i [41,43].

The probability of student scores based on rater assessment using a certain criterion is a function of multiple facets. This study considered seven facets: student, rater, item criteria, student gender, student major, rater gender, and rater major. The severity and leniency of the rater assessment would be calculated based on every facet. The log-odd unit (logit) would represent all the facets, an interval scale. Fair average measures (FAMs) are reported based on the MFRM analysis [41,43]. Strata and reliability would be estimated for all facets. Moreover, reliability is used to estimate the ratio of true variance to observed variance, which indicates the internal consistency of each facet. The strata indicate the wider spread of facet elements, higher separation values, and wider spread of the elements within a certain facet Wu & Tan, [3,43]. Interrater agreement was reported as percentages to evaluate rater agreement in the evaluation process [32,41].

FACETS also calculates the chi-square test for statistical significance. The chi-square test functions to differentiate the distinct level of student performance, item difficulty, and rater severity. A significant *p*-value indicates the level has an equal measure instead of different (e.g., $p < .05$ or 0.01). In assessing the validity criteria; fit statistics are evaluated based on the infit mean square (MNSQ), outfit MNSQ, infit z-standardized (ZSTD), and outfit ZSTD. For infit MNSQ and outfit MNSQ, the acceptable values range from 0.5 to 1.5. However, values from 1.6 are acceptable, given that the element has a positive correlation [44,45]. The infit and outfit ZSTD values of the facet element are acceptable if they range from −2 to +2. However, the appropriate sample size for obtaining the reliable value of ZSTD is less than 250. Therefore, the ZSTD value larger than a sample size of 250 can be ignored [45,46].

The rating scale function of the evaluation form is evaluated based on the Rasch–Andrich thresholds, where the average measures of the scoring categories should improve monotonically from the lower to the higher criteria [41,43,45]. The bias interactions between certain facets are also evaluated using model bias facets (Appendix A). Bias analysis is useful for verifying the validity of ratings based on the interaction among background variables, such as gender and academic majors, whereas the differential item functioning (DIF) graph represents rater bias interaction. DIF analysis can be confirmed using DIF contrast and significant probability ($p < .01$) [47,48].

3. Result

In this study, the participants were between 29 and 40 years old and registered in an international class at a Hungarian university. Most international students completed their Master's and Bachelor's degrees in English. The participants are majoring in linguistics ($n = 4$), mathematics ($n = 6$), science education ($n = 7$), and psychology ($n = 16$). The sample was composed of 14 females (42%) and 19 (58%) males, which achieves gender balance in the dataset. Lastly, the data was analysed using FACETS software to rum MFRM Model.

3.1. *Fit validity and reliability (RQ1)*

First, the 18-item criteria were analyzed based on standard deviation, mean, kurtosis, and skewness to determine the normal distribution of the dataset. According to Table 1, item criteria 11 was the most difficult one, whereby the raters tended to give a low score to rate (FAM = 2.31; Measure (logits) = 0.53, SE = 0.16), whereas item criteria 14 was the easiest one whereby the raters tended to give a high score to rate (FAM = 2.59; Measure (logits) = − 0.51, SE = 0.17). The item criteria indicated a normal distribution through kurtosis and skewness values ranging from −1.188 (scoring criterion 11) to −0.313 (scoring criterion 2) for kurtosis and from −0.696 (scoring criterion 14) to −0.06 (scoring criterion 2) for skewness. The acceptable values for achieving a normal distribution

Table 1
The summary statistics of MFRM analysis.

Psychometrics	Students	Criteria	Raters
Number	33	18	6
Measure			
Mean	0.00	0.00	1.79
SD	0.92	0.31	0.77
FAM			
Mean	2.43	2.45	2.45
SD	0.23	0.08	0.20
SE	0.02	0.00	0.02
Outfit MNSQ			
Mean	0.99	1.01	1.05
SD	0.29	0.17	0.18
Infit MNSQ			
Mean	0.98	1.00	1.06
SD	0.26	0.17	0.19
Outfit ZSTD			
Mean	−0.2	−0.6	0.1
SD	1.9	1.6	2.4
Infit ZSTD			
Mean	−0.2	0.0	0.4
SD	1.8	1.5	2.8
Strata	5.64	2.49	10.19
Reliability	0.94	0.72	0.98
Chi-square/df	566,5/32*	64,3/17*	382/5*
Inter-rater reliability			
Observed Exact Agreements			56.7%
Expected %			61.6%

*$p < .01$, SD = Standard Deviation, SE = Standard error, FAM = Fair average measure.

range from −2 to +2 [49]. The mean measure of students and criteria was 0.00 logit, whereas the standard deviation for students pointed to a wide dispersion of measures with 0.95 logits. Last but not least, the standard deviation of criteria indicated the absence of a wide dispersion of measure across the logit scale at 0.31 logit. The mean of the FAM values showed that all the facets in the model influenced one another, which ranged from 2.43 to 2.45 logits, with a low dispersion based on the standard deviation. The result of interrater reliability revealed a high level of agreement among the raters (>50%), which indicated the effectiveness of the training session for raters in ensuring the same perspective for the scoring evaluation form [41,43].

For fit validity, the mean of infit and outfit MNSQ for all facets ranged from 0.98 logits to 1.06 logits, which indicated that fit validity was achieved. The acceptable range is from 0.5 to 1.5, where the ideal values for fit criteria are close to 1.00 logits [44,45]. The mean of infit and outfit ZSTD of all facets also achieved fit validity, where the threshold ranged from −0.2 logit to 0.4 logits, and the acceptance criteria should range from −2 logits to 2 logits [45,50]. The strata of all facets revealed that all facets in the model ranged from easy/leniency to difficult/severity. The strata values could be imperative if striking 2 logits in which the larger the separation index, the superior the quality of the facets [45,50,51].

The acceptable reliability criteria ranged from 0.68 to 1.00 [45,50]. Fisher (2007) stated that values more than 0.67 represent acceptable reliability. All main facets exhibit acceptable values for reliability, which consist of students (0.94), criteria (0.72), and raters (0.98). The results of the chi-square test for all the facets displayed significant *p*-values. In other words, each facet in the model had a different level of measure (unequal). Table 1 represents the Rasch parameters based on the MFRM analysis for three major facets. Facets, such as gender and academic major, would be used to demonstrate bias interaction.

To ensure fit validity for the item criteria level, the parameters for the item criteria were evaluated based on the MFRM analysis (Table 2). The infit and outfit MNSQ values fell under acceptable values, which ranged from 0.73 logits to 1.44 logits. However, four items, namely, item criteria 4, 7, 11, and 12, deviated from the acceptable values of the infit and outfit ZSTD, <−2 or >+2. We retained these items in the analysis because they exhibited a positive point measure correlation (PTMAC), which contributed to the evaluation of student measures at various levels based on the recommendations of previous studies [45], Andrich & Marais, [52,53]. All the item criteria possessed low standard errors (0.16 and 0.17), which indicated high precision in item estimation. The small differences between FAM and the observed average score specified that.

In the model, the influence of other aspects exerted a substantial but minimal effect on estimating item difficulty. Fig. 1 represents the model-expected item characteristic curve (ICC) plot. The ICC plot demonstrated that the expected measure relative to item difficulty matched those of the probability model, where all empirical item estimates followed the expected values in the model. This result proved that all item criteria achieved excellent fit validity based on the ICC plot evaluation.

3.2. *Rating scale category and threshold (RQ2)*

Fig. 2 illustrates the category probability curve based on the model. It revealed that each rating scale from Category 1 (Fair) to Category 3 (Excellent) reached a clear peak without an overlapping line, where the expected and observe lines were close to each other. This result proved that the rater could distinguish each category in the instrument evaluation or that the scale categories possess good functionality [45,54].

The statistics for the rating scale category in Table 3 describe the three-rating scale from fair to excellent, which reveals a monotonic improvement based on the average measures [55,56]. In addition, the Rasch–Andrich thresholds pointed to a monotonic improvement from low to 1.92 logits, which showed that the raters easily understood the rating scale categories [50]. Outfit MNSQ values indicated that all scale categories have good fit validity (1.00; [43,45]. Moreover, the percentages of the rating scale category were 10% (306), 51% (1490), and 39% (1138) for Categories 1 (Fair), 2 (Good), and 3 (Excellent), respectively. The percentages of the

Table 2
The *item parameter criteria.*

item Criteria	Total	Observed Average	FAM	Measure (logits)	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	OUTFIT ZSTD	PTMAC
1	371	2.28	2.44	0.04	0.16	0.92	−0.7	1	0	0.59
2	364	2.23	2.39	0.22	0.16	0.88	−1.1	0.88	−1.1	0.61
3	371	2.28	2.44	0.04	0.16	0.94	−0.5	0.91	−0.7	0.67
4	369	2.26	2.43	0.09	0.16	0.74	−2.7	0.76	−2.2	0.68
5	387	2.37	2.56	−0.4	0.17	0.92	−0.7	0.88	−0.9	0.66
6	361	2.21	2.37	0.3	0.16	0.87	−1.2	0.91	−0.7	0.66
7	372	2.28	2.45	0.01	0.16	0.73	−2.8	0.73	−2.6	0.7
8	387	2.37	2.56	−0.4	0.17	1.03	0.2	1.07	0.5	0.64
9	357	2.19	2.34	0.4	0.16	0.93	−0.7	0.91	−0.8	0.75
10	360	2.21	2.36	0.33	0.16	0.91	−0.8	0.88	−1.1	0.74
11	352	2.16	2.31	0.53	0.16	1.25	2.3	1.23	2	0.69
12	370	2.27	2.43	0.06	0.16	1.44	3.8	1.44	3.5	0.68
13	382	2.34	2.52	−0.26	0.17	1.14	1.3	1.07	0.6	0.65
14	391	2.4	2.59	−0.51	0.17	1.1	0.9	1.15	1.1	0.69
15	389	2.39	2.57	−0.46	0.17	1.06	0.5	1.05	0.3	0.7
16	384	2.36	2.54	−0.32	0.17	1.13	1.2	1.21	1.6	0.62
17	371	2.28	2.44	0.04	0.16	1.01	0.1	1.09	0.8	0.51
18	362	2.22	2.38	0.28	0.16	0.96	−0.3	0.98	−0.1	0.66

SE = Standard Error, FAM = Fair average measure, PTMAC = Point measure correlation.

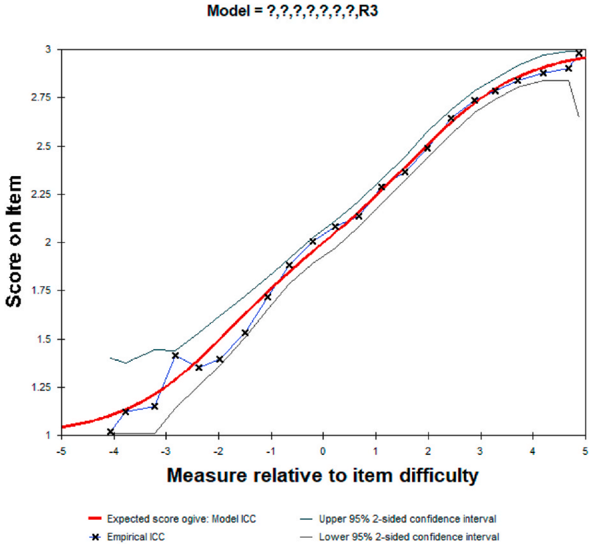


Fig. 1. The ICC plot for item criteria.

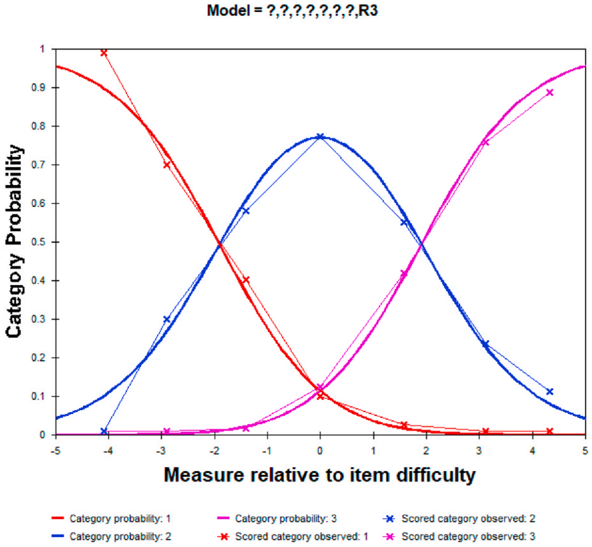


Fig. 2. The category probability curve based on observed scores and expected scores.

Table 3
Rating scale category statistics.

Rating scale category	Total	Percent	Average measure	Expected measure	Outfit MNSQ	Rasch-Andrich Thresholds
1(Fair)	306	10%	−1.22	−1.24	1.00	low
2(Good)	1490	51%	0.66	0.67	1.00	−1.92
3(Excellent)	1138	39%	2.49	2.48	1.00	1.92

scale categories demonstrated an improvement in the scale category from 1(Fair) to 3(Excellent), which reflected high levels of ability in evaluating student presentations related to the research projects.

3.3. Rater measurement and variable map (RQ3)

The variable map of the seven facets was generated using MFRM analysis. Fig. 3 illustrates the variable map for student gender, student majors, rater, rater gender, rater majors, and item criteria. The first column denotes the ruler scale on the left side, which ranges from −2 to +3. The scale units are logits with an anchor at zero point. The second column represents students with ability estimates from low to high, where students 11 and 29 exhibit the highest and lowest levels of ability, respectively. Each student was rated by 6 raters using the 18-item criteria. The fourth column indicates various rater severity/leniency, where raters 6 and 3 were the most severe and lenient, respectively. The dummy facets were included in the model to depict student gender, rater gender, student academic majors, and rater academic majors in the variable map. The measures for student gender were close to zero, whereas the measures for rater gender were located between 1 logit and −1 logit, which indicates that rater gender might influence rater assessment. The male raters had 0.61 logits, and female raters had −0.61 logits, which confirms that female raters were more severe than male raters. Furthermore, the rater's academic major designated a more widespread severity than the student's academic major, which suggests that the academic majors of the raters in the model influenced the rater's severity in evaluating student performance. The last column denotes the scale criteria with a relatively wide range, from 1 to 3. Table 3 explains the functionality of the scale criteria.

Table 4 presents further rater measurement information. All the MNSQ values achieve the fit validity criteria with positive PTMAC. Rater measures range from 0.62 logits to 2.61 logits.

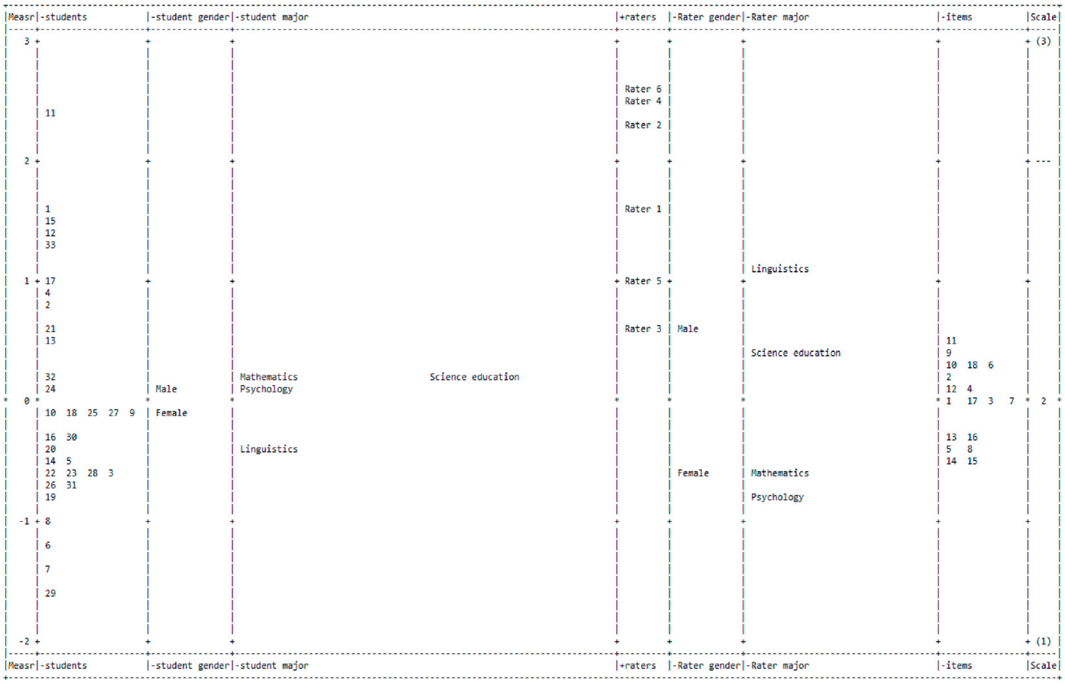
3.4. Gender bias in rater assessment (RQ4)

Table 5 illustrates how raters gave ratings based on gender. The table demonstrates that only Rater 6 displayed significant differences with the female (4.06 logits) and male (2.01) measures, which resulted in a target contrast (DIF Size) of 2.05 logits with a significant probability ($p < .01$). Rater 6 tended to be more severe in rating the male than female participants compared with the other raters. The fixed chi-square test for bias interaction between the rater and student gender was significant ($\chi^2(12) = 50.4$; $p < .001$), which suggests that the magnitude of the observed measurement variation was substantive and not coincidental. Although the rater and gender interaction were insignificant, this result was important. However, if a significant bias exists in a certain rater measure, the final results need to be canceled to achieve fairness in evaluating student performance.

DIF analysis was also performed to verify any rater bias assessment based on student gender. DIF analysis indicated rater responses based on subgroups (student gender) for each rater in the rating evaluation [32,41]. Based on the absolute measure, the DIF graph also depicted rater bias based on student gender, whereas Rater 6 exhibited significant range differences in student ratings based on gender. Fig. 4 illustrates the bias interaction between the rater and student gender.

3.5. Interaction between rater academic major and student academic major (RQ5)

Table 6 points out three statistically significant bias interactions between rater academic majors and student academic majors ($p <$ .01). Interestingly, the rater with the linguistics major exhibited a positive bias size (1.08 logits) to students majoring in linguistics (t (71) = 3.4, $p < .01$). In other words, the rater with the linguistics major tended to give a higher score to students majoring in linguistics



```
+-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
|Measr|-students           |-student gender|-student major                                                         |+raters  |-Rater gender|-Rater major                        |-items          |Scale|
|-----+--------------------+---------------+-----------------------------------------------------------------------+---------+-------------+-----------------------------------+----------------+-----|
|   3 +                    +               +                                                                       +         +             +                                   +                + (3) |
|     |                    |               |                                                                       |         |             |                                   |                |     |
|     |                    |               |                                                                       |         |             |                                   |                |     |
|     |                    |               |                                                                       |         |             |                                   |                |     |
|     |                    |               |                                                                       | Rater 6 |             |                                   |                |     |
|     |                    |               |                                                                       | Rater 4 |             |                                   |                |     |
|     | 11                 |               |                                                                       |         |             |                                   |                |     |
|     |                    |               |                                                                       | Rater 2 |             |                                   |                |     |
|     |                    |               |                                                                       |         |             |                                   |                |     |
|     |                    |               |                                                                       |         |             |                                   |                |     |
|   2 +                    +               +                                                                       +         +             +                                   +                + --- |
|     |                    |               |                                                                       |         |             |                                   |                |     |
|     |                    |               |                                                                       |         |             |                                   |                |     |
|     |                    |               |                                                                       |         |             |                                   |                |     |
|     | 1                  |               |                                                                       | Rater 1 |             |                                   |                |     |
|     | 15                 |               |                                                                       |         |             |                                   |                |     |
|     | 12                 |               |                                                                       |         |             |                                   |                |     |
|     | 33                 |               |                                                                       |         |             |                                   |                |     |
|     |                    |               |                                                                       |         |             |                                   |                |     |
|     |                    |               |                                                                       |         |             | Linguistics                       |                |     |
|   1 + 17                 +               +                                                                       + Rater 5 +             +                                   +                +     |
|     | 4                  |               |                                                                       |         |             |                                   |                |     |
|     | 2                  |               |                                                                       |         |             |                                   |                |     |
|     |                    |               |                                                                       |         |             |                                   |                |     |
|     | 21                 |               |                                                                       | Rater 3 | Male        |                                   |                |     |
|     | 13                 |               |                                                                       |         |             |                                   | 11             |     |
|     |                    |               |                                                                       |         |             | Science education                 | 9              |     |
|     |                    |               |                                                                       |         |             |                                   | 10  18  6      |     |
|     | 32                 |               | Mathematics                        Science education                  |         |             |                                   | 2              |     |
|     | 24                 | Male          | Psychology                                                            |         |             |                                   | 12  4          |     |
*   0 *                    *               *                                                                       *         *             *                                   * 1   17  3   7  *  2  *
|     | 10  18  25  27  9  | Female        |                                                                       |         |             |                                   |                |     |
|     |                    |               |                                                                       |         |             |                                   |                |     |
|     | 16  30             |               |                                                                       |         |             |                                   | 13  16         |     |
|     | 20                 |               | Linguistics                                                           |         |             |                                   | 5   8          |     |
|     | 14  5              |               |                                                                       |         |             |                                   | 14  15         |     |
|     | 22  23  28  3      |               |                                                                       |         | Female      | Mathematics                       |                |     |
|     | 26  31             |               |                                                                       |         |             |                                   |                |     |
|     | 19                 |               |                                                                       |         |             | Psychology                        |                |     |
|     |                    |               |                                                                       |         |             |                                   |                |     |
|  -1 + 8                  +               +                                                                       +         +             +                                   +                +     |
|     |                    |               |                                                                       |         |             |                                   |                |     |
|     | 6                  |               |                                                                       |         |             |                                   |                |     |
|     |                    |               |                                                                       |         |             |                                   |                |     |
|     | 7                  |               |                                                                       |         |             |                                   |                |     |
|     |                    |               |                                                                       |         |             |                                   |                |     |
|     | 29                 |               |                                                                       |         |             |                                   |                |     |
|     |                    |               |                                                                       |         |             |                                   |                |     |
|     |                    |               |                                                                       |         |             |                                   |                |     |
|     |                    |               |                                                                       |         |             |                                   |                |     |
|  -2 +                    +               +                                                                       +         +             +                                   +                + (1) |
|-----+--------------------+---------------+-----------------------------------------------------------------------+---------+-------------+-----------------------------------+----------------+-----|
|Measr|-students           |-student gender|-student major                                                         |+raters  |-Rater gender|-Rater major                        |-items          |Scale|
+-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------+
```

Fig. 3. Variable map the location of seven facets in the model. Student gender, student major, rater gender, and rater major were anchored to zero.

Table 4
Rater measurement.

Rater	Total	Observed Average	FAM	Measure (logits)	SE	Infit MNSQ	Infit ZSTD	Outfit MNSQ	Infit ZSTD	PTMAC
Rater 1	1270	2.14	2.4	1.62	0.08	0.91	−1.7	0.91	−1.6	0.48
Rater 2	758	2.63	2.59	2.31	0.13	1.37	4.3	1.33	3.3	0.32
Rater 3	1020	1.72	2.15	0.62	0.08	0.86	−2.5	0.86	−2.5	0.53
Rater 4	1562	2.63	2.65	2.54	0.09	1.12	2.2	1.14	2	0.44
Rater 5	1273	2.28	2.24	1.01	0.08	0.87	−2.4	0.87	−2.5	0.48
Rater 6	817	2.67	2.66	2.61	0.13	1.22	2.9	1.2	2.1	0.43

SE = Standard Error, FAM = Fair average measure, PTMAC = Point measure correlation.

No substantial differences were noted between the observed average score and FAM, indicating a high rater assessment precision. Table 1 indicates that the rater strata and reliability were 10.19 and 0.98, respectively. The chi-square statistic also exerted a significant effect ($\chi^2 = 382.6$; df = 5; $p <$ .001), demonstrating high heterogeneity among raters. This result revealed that raters displayed various degrees of severity, which influenced student performance ratings.

Table 5
Bias interaction between rater and student gender.

Rater	Target measure (logits)	Student gender	Target measure (logits)	Student gender	Target contrast	t	df	p
Rater 1	1.6	Female	1.64	Male	−0.04	−0.22	573	0.8297
Rater 2	2	Female	2.51	Male	−0.51	−1.9	262	0.0579
Rater 3	0.72	Female	0.54	Male	0.18	1.11	574	0.2683
Rater 4	2.33	Female	2.73	Male	−0.39	−2.13	588	0.0337
Rater 5	0.84	Female	1.16	Male	−0.32	−1.88	552	0.0601
Rater 6	4.06	Female	2.01	Male	2.05	5.93	185	0

p = probability(significant if p < .01), t = t-statistic, df = degree of freedom.

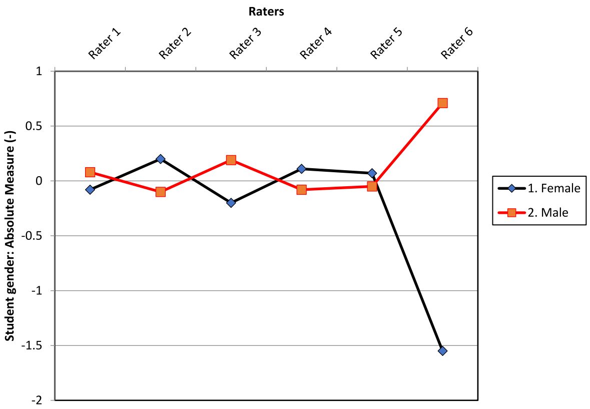


Fig. 4. DIF graph for bias interaction between rater and student gender.

compared with students with other academic majors. This bias case displayed the most extensive bias among other significant bias interactions. The rater with the linguistics major also displayed a negative bias interaction in assessing students majoring in psychology (bias size [−0.32 logits]; ($t(287) = -2.6$, $p < .01$). This result indicates that the rater with the linguistics major tended to be more severe in giving low scores to students majoring in psychology compared with students with other academic majors. The last bias interaction identified was between the raters with a psychology major, who tended to give low scores to students majoring in linguistics with a significant negative bias size (−0.67 logits; ($t(125) = -3.3$, $p < .01$). The fixed chi-square test for rater academic major and student academic major also pointed to significant results ($t(16) = 40$, $p < .01$), which indicates that bias interaction between academic majors displayed a substantive variation. As such, the bias of peer academic majors should be canceled out to achieve fairness in performance evaluation. The DIF graph in Fig. 5 illustrates the bias interaction between the rater academic major and the student academic major.

4. Discussion

Rater assessment helps evaluate student performance in presentations, especially at higher education institutions where biased sources can be identified and excluded from the final evaluation [8,42]. Raters must hold the same perception in providing assessments and scores based on the rubrics on the evaluation form to achieve an effective assessment, [1]. However, raters tend to employ biased

Table 6
Bias interaction between rater academic major and student academic major.

Student major	Measure (logits)	Rater major	Measure (logits)	Bias Size	t	df	Prob.
Mathematics	0.16	Mathematics	−0.61	1.09	1.45	17	0.1665
Linguistics	−0.45	Linguistics	1.08	0.81	3.4	71	0.0011
Mathematics	0.16	Psychology	−0.82	0.22	1.22	161	0.2258
Science education	0.15	Linguistics	1.08	0.19	1.06	125	0.29
Psychology	0.14	Psychology	−0.82	0.13	1.15	395	0.2491
Psychology	0.14	Science education	0.36	0.08	0.97	557	0.3318
Mathematics	0.16	Linguistics	1.08	0.07	0.37	107	0.7105
Linguistics	−0.45	Mathematics	−0.61	0.07	0.24	71	0.813
Science education	0.15	Science education	0.36	0.02	0.14	233	0.8867
Psychology	0.14	Mathematics	−0.61	0	−0	89	0.9887
Linguistics	−0.45	Science education	0.36	−0.02	−0.1	143	0.8987
Science education	0.15	Psychology	−0.82	−0.07	−0.5	215	0.6243
Science education	0.15	Mathematics	−0.61	−0.14	−0.7	107	0.4746
Mathematics	0.16	Science education	0.36	−0.22	−1.6	215	0.1135
Psychology	0.14	Linguistics	1.08	−0.32	−2.6	287	0.0094
Linguistics	−0.45	Psychology	−0.82	−0.67	−3.3	125	0.0011

p = probability(significant if p < .01), t = t-statistic, df = degree of freedom.

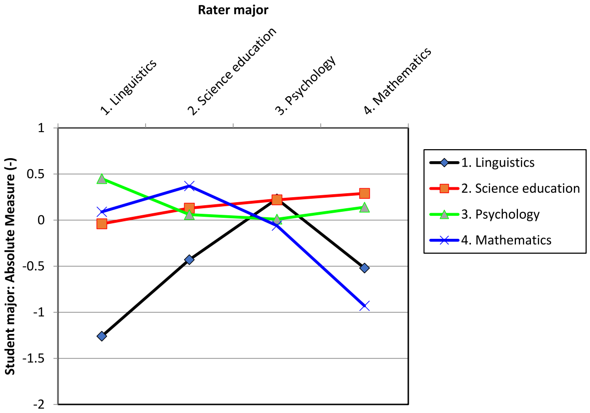


Fig. 5. DIF graph for bias interaction between rater academic major and student academic major.

sources based on the presenter's background, such as gender and academic major [57]. Therefore, applying an objective measurement using MFRM was required to identify rater and student interaction to ensure validity and reliability and to measure severity and ability levels in the evaluation process.

The results confirmed that the adapted instrument for assessing student performance in oral presentation achieved the validity and reliability criteria based on the Rasch parameters. The ICC plot and the category probability curve also confirmed that all items and scale categories in the evaluation form functioned well. Moreover, the study revealed that raters in the model assessment possess varying levels of severity. This finding corroborated with Aryadoust [1], who revealed that rater assessment could lead to different measure estimates in assessing oral presentations. In addition [58], found that raters exhibited a widespread separation in assessing student performance regarding communication skills.

Investigating the cause of bias in the results of an evaluation is a challenging issue because raters tend to be inconsistent in evaluation concerning theoretical concepts and presenter backgrounds [7,23,27]. These issues typically occur during peer assessment, where reliability and bias factors are ignored. Previous research revealed that raters were highly dependent on factors such as academic majors and gender in conducting peer assessments [16,59,60]. Consequently, an appropriate measurement is required to achieve a fair and reliable measurement.

This study proved that MFRM analysis could help researchers identify bias interaction in gender and academic majors. Rater 6 displayed bias based on gender due to the tendency to give higher scores to female than male participants with a target contrast of 2.05 logits (DIF Size). Therefore, if an evaluation result without bias is desired, Rater 6 can be excluded from the assessment model. This finding is in line with [16], who found that examinee gender can influence rater assessment during evaluation. Furthermore, Aryadoust [1] found that gender plays an essential role in influencing rater judgment in the context of peer assessment. The current study also confirmed three statistically significant biases based on rater academic majors and student academic majors bias interactions. Interestingly, these interactions resulted from rater and student interaction with linguistics and psychology majors, as previously mentioned in the results for academic major interaction. Academic majors as a biased source were also noted across the assessment contexts when assessing items and students majoring in science education [42,61]. [1] also found that interaction between different academic majors could influence the quality of rater assessment in evaluating examinee performance.

5. Limitations and recommendations for future study

The study has four main limitations. First, the sample size (n = 33) was relatively small compared with the whole population of doctoral students in Hungary, and we only have six raters whereby only two females raters selected. Therefore, we recommended the future research have comparable number of rater between males and females. We formulated a dataset combination with six raters, 33 examinees and 18 item criteria, and this study collected 3564 data points. However, we still recommend that other researchers recruit more students or examinees to explain and generalize the result in a comprehensive manner. Second, this study did not assess student knowledge about the research seminar course but only the presentation. As such, students' insights may influence their presentation performances. Thus, it is advised that an initial assessment should be conducted using a test or questionnaire to measure the cognitive aspects of the students in certain subjects. Third, most students were non-native speakers; thus, their mother tongue was not English. To address this aspect, further research could endeavor to compare bias in presentation performance based on language. Fourth, to avoid gender bias and academic major bias, other researchers should implement rater training sessions with bias interaction and objective evaluation material to minimize the bias interaction across raters and examinees. Finally, the study hopes that the results will inspire other researchers to conduct facet research to enrich interactions and biased sources related to rater assessment.

6. Conclusions

To summarise, this study confirmed that the instrument was reliable and valid based on Rasch parameters in the result section. The interaction between raters and other facets was depicted using a variable map where the raters expressed different levels of severity in scoring the presentations related to the research seminar course. Moreover, the study found that gender and academic majors contaminated rater assessment. Bias interaction between rater academic majors and student academic majors was also identified among raters with linguistics and psychology majors. Finally, this study observed that bias based on background variables contaminates rater assessment in higher education for low-stake assessments. Through the study, a sample of fair assessments can be made, and the identified biased rater can be excluded to obtain fair and reliable assessment results.

Credit author statement

Fitria Arifiyanti: Wrote the paper; Conceived and designed the experiments; Analyzed and interpreted the data; Performed the experiment; Contributed reagents, materials, analysis tools or data.

Soeharto Soeharto; Eri Sarimanah: Wrote the paper; Performed the experiment; Analyzed and interpreted the data.

Stephen Amukune; Son Van Nguyen; Khalil Aburezeq; Achmad Hidayatullah: Wrote the paper; Performed the experiments.

Data availability statement

Data will be made available on request.

Additional information

No additional information is available for this paper.

Funding statement

This work was supported by the University of Szeged Open Access Fund (grant number: **5663**).

Declaration of competing interest

The authors declare no conflict of interest.

Appendix A

Data commands.

The specification of Facets input file is as follows:
Title = FACETS oral presentation in research seminar course.
Facets = 7; Raters + Students + Student gender + student major + rater gender + rater major + item category
inter-rater = 1; compute inter-rater agreements (Inter-rater = rater facet number).
Positive = 1; the first facet has positive ability.
Noncentered = 1; measure from the center of the measures for each facet.
Left = yes;
Models =

```
?,?,?,?,?,?,?,R3; put the model statement for your facets and elements here.
?B,?B,?,?,?,?,?,R3; bias interaction between raters and students.
?B,?,?B,?,?,?,?,R3; bias interaction between raters and students’ gender.
?,?,?,?B,?,?B,?,R3; bias interaction between raters’ major and students’ major.
*
Labels =
*
1, raters.
1 = Rater 1.
2 = Rater 2.
3 = Rater 3.
4 = Rater 4.
5 = Rater 5.
6 = Rater 6.
*
2, students.
1–33.
*
3, student gender, A.
1 = Female.
2 = Male.
*
4, student major, A.
1 = Linguistics.
2 = Science education.
4 = Psychology.
5 = Mathematics.
*
5, Rater gender, A.
1 = Female.
2 = Male.
*
6, Rater major, A.
1 = Linguistics.
2 = Science education.
4 = Psychology.
5 = Mathematics.
*
7, items.
1–18.
*
data =
1,1,1,2,2,2,1-18a,1,1,2,1,1,2,1,2,1,1,1,1,1,1,1,2,2,2.
2,1,1,2,2,5,1-18a,2,1,2,3,1,3,1,2,3,2,1,2,1,3,1,2,3,2.
…
6,33,2,5,2,4,1-18a,2,3,3,3,3,2,3,3,2,2,3,2,2,2,3,2,3,3.
3,33,2,5,2,1,1-18a,2,2,1,2,1,1,1,1,1,1,1,2,1,1,1,1,1.
```

Evaluation form.

Q1 Interesting opening to capture the attention.
Q2 Clear context for the controversy leading to thesis/stance.
Q3 Well-developed presentation of information for explanation or substantiation.
Q4 Convincing main and supporting ideas; persuasive and sufficient to support stance.
Q5 Sound analysis in a manner suitable for an audience of educated lay-persons.
Q6 Suitable transitional devices to link sections of presentation coherently.
Q7 Suitably strong conclusion that advances thesis and points the way forward.
Q8 Answered audience’s questions politely, clearly, knowledgeably & effectively.
Q9 Various sentence patterns & well-formed sentences, correct grammar.
Q10 Appropriate vocabulary & style.
Q11 Clear & correct pronunciation - stressed syllables & words properly.
Q12 Appropriate volume/pitch/tone/speed.
Q13 Enthusiastic and confident– e.g., through words, facial expressions, smiles, voice, tone, etc.

Q14 Confident posture – e.g., faced the audience straight & tall - not slouching, not stiff, not pacing every other minute, etc.

Q15 Meaningful and fitting gestures – e.g., those that are natural and those that complement the verbal language.

Q16 Frequent eye contact with the audience.

Q17 Good time management.

Q18 Effective audio/visual aids – e.g., relevant PowerPoint slides, and graphs, tables, pictures, objects; were attractively and professionally designed and crafted.

References

[1] V. Aryadoust, Gender and academic major bias in peer assessment of oral presentations, Lang. Assess. Q. 13 (1) (2016) 1–24, https://doi.org/10.1080/15434303.2015.1133626.

[2] A. Milanowicz, B. Bokus, Gender and moral judgments: the role of who is speaking to whom, J. Gend. Stud. 22 (4) (2013) 423–443.

[3] S.M. Wu, S. Tan, Managing rater effects through the use of FACETS analysis: the case of a university placement test, High Educ. Res. Dev. 35 (2) (2016) 380–394, https://doi.org/10.1080/07294360.2015.1087381.

[4] B. Soeharto, E. Csapó, F. Sarimanah, I. Dewi, T. Sabri, A review of students' common misconceptions in science and their diagnostic assessment tools, Jurnal Pendidikan IPA Indonesia 8 (2) (2019) 247–266, https://doi.org/10.15294/jpii.v8i2.18649.

[5] F. De Waal, The Age of Empathy, Crown, New York: NY, 2009.

[6] M. Maryati, Z.K. Prasetyo, I. Wilujeng, B. Sumintono, Measuring teachers' pedagogical content knowledge using many-facet rasch model, Jurnal Cakrawala Pendidikan 38 (3) (2019) 452–464, https://doi.org/10.21831/cp.v38i3.26598.

[7] H. Kim, Effects of Rating Criteria Order on the Halo Effect in L2 Writing Assessment: a Many-Facet Rasch Measurement Analysis, vol. 10, 2020, p. 1, https://doi.org/10.1186/s40468-020-00115-0. Language Testing in Asia.

[8] C.C. Bauer, B.B. Baltes, Reducing the effects of gender stereotypes on performance evaluations, Sex. Roles 47 (9) (2002) 465–476.

[9] B. Simpson, Sex, fear, and greed: a social dilemma analysis of gender and cooperation, Soc. Forces 82 (1) (2003) 35–52.

[10] M. Abu Taha, K. Abu Rezeq, Oral communication apprehension among English senior majors at Al-Quds Open University in Palestine, Int. J. Res. Engl. Educ. 3 (1) (2018) 44–58.

[11] H.K. Davison, M.J. Burke, Sex discrimination in simulated employment contexts: a meta-analytic investigation, J. Vocat. Behav. 56 (2) (2000) 225–248.

[12] J. Wilson, D. Bayard, Accent, gender, and the elderly listener: evaluations of NZE and other English accents by rest home residents, Te Reo 35 (1) (1992) 19–56.

[13] J. Swim, In search of gender bias in evaluations and trait inferences: the role of diagnosticity and gender stereotypicality of behavioral information, Sex. Roles 29 (3) (1993) 213–237.

[14] T.H. Shore, Subtle gender bias in the assessment of managerial potential, Sex. Roles 27 (9–10) (1992) 499–515.

[15] Y. Chen, S.X. Li, Group identity and social preferences, Am. Econ. Rev. 99 (1) (2009) 431–457.

[16] A.M. Langan, et al., Peer assessment of oral presentations: effects of student gender, university affiliation and participation in the development of assessment criteria, Assess Eval. High Educ. 30 (1) (2005) 21–34.

[17] V. Aryadoust, P. Mehran, M. Alizadeh, Validating a computer-assisted language learning attitude instrument used in Iranian EFL context: an evidence-based approach, Comput. Assist. Lang. Learn. 29 (3) (2016) 561–595, https://doi.org/10.1080/09588221.2014.1000931.

[18] J. Rifkin, The Empathic Civilization: the Race to Global Consciousness in a World in Crisis, Penguin, 2009.

[19] M.L. Hoffman, Empathy and Moral Development: Implications for Caring and Justice, Cambridge University Press, Cambridge, UK, 2000.

[20] K. Kraiger, J.K. Ford, A meta-analysis of ratee race effects in performance ratings, J. Appl. Psychol. 70 (1) (1985) 56.

[21] W.T. Hoyt, Rater bias in psychological research: when is it a problem and what can we do about it? Psychol. Methods 5 (1) (2000) 64.

[22] J. Stonewall, M. Dorneich, C. Dorius, J. Rongerude, A Review of Bias in Peer Assessment', in *2018 CoNECD-The Collaborative Network for Engineering and Computing Diversity Conference*, 2018.

[23] V. Aryadoust, Understanding the role of likeability in the peer assessments of university students' oral presentation skills: a latent variable approach, Lang. Assess. Q. 14 (4) (2017) 398–419, https://doi.org/10.1080/15434303.2017.1393820.

[24] C. Sundrarajun, R. Kiely, The oral presentation as a context for learning and assessment, Innovat. Lang. Learn. Teach. 4 (2) (2010) 101–117.

[25] R.R. Behnke, C.R. Sawyer, Public speaking anxiety as a function of sensitization and habituation processes, Commun. Educ. 53 (2) (2004) 164–1173.

[26] J. Kormos, M. Dénes, Exploring measures and perceptions of fluency in the speech of second language learners, System 32 (2) (2004) 145–164.

[27] W. Cheng, M. Warren, Peer assessment of language proficiency, Lang. Test. 22 (1) (2005) 93–121.

[28] R. Hincks, Measures and perceptions of liveliness in student oral presentation speech: a proposal for an automatic feedback mechanism, System 33 (4) (2005) 575–591.

[29] S. Luoma, Assessing Speaking, Ernst Klett Sprachen, 2004.

[30] K. Gillett, B. O'Neill, J.G. Bloomfield, Factors influencing the development of end-of-life communication skills: a focus group study of nursing and medical students, Nurse Educ. Today 36 (2016) 395–400.

[31] B. Fraser, What are discourse markers? J. Pragmat. 31 (7) (1999) 931–952.

[32] V. Aryadoust, Self- and peer assessments of oral presentations by first-year university students, Educ. Assess. 20 (3) (2015) 199–225, https://doi.org/10.1080/10627197.2015.1061989.

[33] J. Flowerdew, S. Tauroza, The effect of discourse markers on second language lecture comprehension, Stud. Sec. Lang. Acquis. 17 (4) (1995) 435–458.

[34] I. Bejar, D. Douglas, J. Jamieson, S. Nissan, J. Turner, TOEFL 2000 Listening Framework, Educational Testing Service, Princeton, NJ, 2000.

[35] A.B.M. Tsui, J. Fullilove, Bottom-up or top-down processing as a discriminator of L2 listening performance, Appl. Linguist. 19 (4) (1998) 432–451.

[36] D. Vogel, M. Meyer, S. Harendza, Verbal and non-verbal communication skills including empathy during history taking of undergraduate medical students, BMC Med. Educ. 18 (1) (2018) 1–7.

[37] D. Maloney, G. Freeman, D.Y. Wohn, Talking without a voice" understanding non-verbal communication in social virtual reality, Proc. ACM Hum.-Comput. Inter. 4 (2020) 1–25. CSCW2.

[38] K.A. Soretire, B.O. Patrick, Effect of game-based interactive PowerPoint on Studentsâ€™ learning outcomes in civic education in zaria education zone of kaduna state Nigeria, ATBU J. Sci. Technol. Educ. 9 (1) (2021) 12–21.

[39] R.E. Mayer, Applying the Science of Learning, Pearson/Allyn & Bacon Boston, MA, 2011.

[40] M. Tavakol, G. Pinner, Using the Many-Facet Rasch Model to analyze and evaluate the quality of objective structured clinical examination: a non-experimental cross-sectional design, BMJ Open 9 (9) (2019), e029208.

[41] T. Eckes, Introduction to Many-Facet Rasch Measurement, Peter Lang Verla, Berlin, Germany, 2011.

[42] W.J. Boone, J.S. Townsend, J.R. Staver, Utilizing multifaceted rasch measurement through FACETS to evaluate science education data sets composed of judges, respondents, and rating scale items: an exemplar utilizing the elementary science teaching analysis matrix instrument, Sci. Educ. 100 (2) (2016) 221–238, https://doi.org/10.1002/sce.21210.

[43] J.M. Linacre, A User's Guide to FACETS Rasch-Model Computer Programs Program, July. IEEE, 2021, https://doi.org/10.1109/COMPSAC.2011.4.

[44] D. Andrich, Advances in social measurement: a Rasch measurement theory, in: Perceived Health and Adaptation in Chronic Disease, Routledge, 2018, pp. 66–91.

[45] W.J. Boone, J.R. Staver, M.S. Yale, Rasch Analysis in the Human Sciences, Springer, New York, NY, 2014.

[46] N.H. Azizan, Z. Mahmud, A. Rambli, Rasch rating scale item estimates using maximum likelihood approach: effects of sample size on the accuracy and bias of the estimates, Int. J. Adv. Sci. Technol. 29 (4) (2020) 2526–2531.

[47] W.J. Boone, J.R. Staver, M.S. Yale, Rasch Analysis in the Human Sciences, Springer, 2013.

[48] R. Zwick, D.T. Thayer, C. Lewis, An empirical Bayes approach to Mantel-Haenszel DIF analysis, J. Educ. Meas. 36 (1) (1999), https://doi.org/10.1111/j.1745-3984.1999.tb00543.x.

[49] A. Field, Discovering Statistics Using IBM SPSS Statistics, sage, 2013.

[50] W.P.J. Fisher, Rating scale instrument quality criteria, Rasch Meas. Trans. 21 (1) (2007) 1095.

[51] M. Planinic, W.J. Boone, A. Susac, L. Ivanjek, Rasch analysis in physics education research: why measurement matters, Phys. Rev. Phys. Educ. Res. 15 (2) (2019), 20111, https://doi.org/10.1103/PhysRevPhysEducRes.15.020111.

[52] D. Andrich, I. Marais, A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences, Springer Nature Singapore Pte Ltd, Singapore, 2019.

[53] S. Soeharto, B. Csapó, Evaluating item difficulty patterns for assessing student misconceptions in science across physics, chemistry, and biology concepts, Heliyon 7 (2021) 11, https://doi.org/10.1016/j.heliyon.2021.e08352.

[54] M.S. Khine, Rasch Measurement (2020), https://doi.org/10.1007/978-981-15-1800-3.

[55] B. Wright, J.M. Linacre, Reasonable mean-square fit values, Rasch Meas. Trans. 8 (3) (1994) 370.

[56] M. Planinic, W.J. Boone, A. Susac, L. Ivanjek, Rasch analysis in physics education research: why measurement matters, Phys. Rev. Phys. Educ. Res. 15 (2) (2019) 1–14, https://doi.org/10.1103/PhysRevPhysEducRes.15.020111.

[57] V. Aryadoust, Self- and peer assessments of oral presentations by first-year university students, Educ. Assess. 20 (3) (2015) 199–225, https://doi.org/10.1080/10627197.2015.1061989.

[58] M.F. Bin Mohd Noh, M.E.E. Bin Mohd Matore, Rating performance among raters of different experience through multi-facet rasch measurement (MFRM) model, J. Meas. Eval. Educ. Psychol. 11 (2) (2020) 147–162, https://doi.org/10.21031/epod.662964.

[59] A. Darmana, A. Sutiani, H.A. Nasution, N.S.A. Sylvia, N. Aminah, T. Utami, Analysis of multi rater with facets on instruments HOTS of solution chemistry based on tawheed, J. Phys. Conf. 1819 (1) (2021) 1–7, https://doi.org/10.1088/1742-6596/1819/1/012038.

[60] T.H. He, W.J. Gou, Y.C. Chien, I.S.J. Chen, S.M. Chang, Multi-faceted Rasch measurement and bias patterns in EFL writing performance assessment, Psychol. Rep. 112 (2) (2013) 469–485, https://doi.org/10.2466/03.11.PR0.112.2.469-485.

[61] S. Yamtinah, et al., Examining the content validity of android-based augmented reality media for chemical bonding using rasch model, Jurnal Penelitian Pendidikan IPA 7 (2021) 320–325, https://doi.org/10.29303/jppipa.v7ispecialissue.1094. SpecialIssue.